

AUGMENTATION OF FINANCIAL DATASETS AND EVALUATING
FINANCIAL TEXT GENERATED BY A.I.

by

Stacey Dianne Taylor

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
May 2024

Dalhousie University is located in Mi'kma'ki, the
ancestral and unceded territory of the Mi'kmaq.
We are all Treaty people.

© Copyright by Stacey Dianne Taylor, 2024

Table of Contents

LIST OF TABLES	ix
LIST OF FIGURES	xi
ABSTRACT	xiii
LIST OF ABBREVIATIONS USED	xiv
GLOSSARY	xvi
ACKNOWLEDGEMENTS	xix
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 General Background	3
1.2.1 Financial Statement Data and Augmentation Challenges	3
1.3 Contributions	7
1.4 Publications	8
1.5 Outline	10
Chapter 2 Literature Review	12
2.1 The First Generative Models	13
2.1.1 EM Algorithm	13
2.1.2 The Boltzmann Machine	13
2.1.3 Markov Random Field	14
2.2 Contemporary Generative Models	14
2.2.1 Variational Autoencoders	15

2.2.2	Generative Adversarial Networks	16
2.2.3	ChatGPT	21
2.2.4	SMOTE	21
2.3	Consideration of State of the Art and Corporate Work	23
Chapter 3	Augmentation of Tabular Financial Statement Data	25
3.1	Chapter Introduction	25
3.2	Project 1: Predicting the Distress of Large US Financial Intermediaries using CNNs	27
3.2.1	Introduction	27
3.2.2	Related Work	28
3.2.3	Methodology	30
3.2.3.1	Initial Selection of Financial Intermediaries	30
3.2.3.2	Sentiment Analysis	32
3.2.3.3	LIBOR Rates	33
3.2.3.4	Equity Ratio	33
3.2.3.5	Critical Accounting Keywords	34
3.2.3.6	Data Augmentation	34
3.2.3.7	Dataset Summary	35
3.2.3.8	Scaling	35
3.2.3.9	Proposed Method	37
3.2.4	Experiments	41
3.2.5	Results and Discussion	41
3.2.5.1	Post Hoc Experimentation	45
3.2.6	Conclusion, Future Work, and Limitations	46
3.3	Project 2: Tabular Augmentation Strategies	49
3.3.1	Introduction	49
3.3.1.1	Related Work	50
3.3.2	Methodology	51

3.3.2.1	Dataset	51
3.3.2.2	Data Labelling	53
3.3.3	Algorithms, Models, Experiments, Results, and Discussion	54
3.3.3.1	Parameters for Synthetic Data Validity	54
3.3.3.2	Initial Testing	55
3.3.3.3	Features	56
3.3.3.4	Generative Adversarial Networks	57
3.3.3.5	Variational AutoEncoders	60
3.3.3.6	ChatGPT	61
3.3.3.7	Synthetic Minority Oversampling Technique	65
3.3.4	Evaluation Using a Downstream Task	69
3.3.5	Conclusion, Limitations, and Future Work	72
Chapter 4	Augmentation of Textual Financial Data	75
4.1	Chapter Introduction	75
4.2	Project 3: Fine-Tuning GPT-3 for Financial Summary Data	76
4.2.1	Introduction	76
4.2.2	Methodology	76
4.2.2.1	Dataset	76
4.2.2.2	Fine-Tune Training Model Selection	78
4.2.2.3	Cosine Similarity	79
4.2.2.4	Prompt Sample Selection	79
4.2.2.5	Sentiment Analysis	79
4.2.2.6	Flesch Reading Ease Score	80
4.2.2.7	Statistical Significance Tests	82
4.2.3	Experiments	83
4.2.3.1	Base Model Testing	83
4.2.3.2	Fine-Tune Training	83
4.2.4	Results and Discussion	83
4.2.5	Conclusion and Future Work	89

4.2.6	Limitations and Ethical Considerations	90
4.3	Project 4: Using ChatGPT to Generate Financial Disclosure Notes	90
4.3.1	Introduction	90
4.3.2	Background	91
4.3.3	Methodology	92
4.3.3.1	Dataset	92
4.3.3.2	Cosine Similarity	94
4.3.3.3	Jaccard Similarity	94
4.3.3.4	Burrows' Delta	95
4.3.3.5	Flesch Reading Ease Score	96
4.3.4	Experiments	96
4.3.5	Results and Discussion	97
4.3.5.1	Cosine Similarity and Jaccard Similarity	97
4.3.5.2	Burrows' Delta	98
4.3.5.3	Flesch Reading Ease	100
4.3.5.4	Side-By-Side Comparison of Text	100
4.3.6	Conclusion, Future Work, and Limitations	105
4.4	Project 5: ChatGPT for Financial Literacy	106
4.4.1	Introduction	106
4.4.2	Methodology	107
4.4.2.1	User Definitions	107
4.4.2.2	Prompts	109
4.4.2.3	Dataset	110
4.4.3	Experiments and Hypotheses	112
4.4.4	Results and Discussion	113
4.4.4.1	GPT-3.5	113
4.4.4.2	GPT-4	117
4.4.5	Conclusion and Future Work	119

4.5	Project 6: Using Raw Accounting Data To Augment Financial Text Using ChatGPT	121
4.5.1	Introduction	121
4.5.2	Related Work	122
4.5.3	Methodology	123
4.5.3.1	Dataset	123
4.5.3.2	Semantic Similarity Using Word Embeddings	124
4.5.3.3	Readability	125
4.5.4	Experiments	125
4.5.5	Results and Discussion	125
4.5.6	Conclusion, Future Work, and Limitations	128
Chapter 5	Foreshadowing the Need for Company Author Attribution	130
5.1	Project 7: Company Authorship Attribution and Verification	130
5.1.1	Introduction	130
5.1.2	Problem Statement	134
5.1.3	Methodology	134
5.1.3.1	Data Collection and Processing	134
5.1.4	Experiments	136
5.1.4.1	Experiment 1 - Latent Dirichlet Allocation	136
5.1.4.2	Experiment 2 - Burrows' Delta	139
5.1.4.3	Experiment 3 - Common N-gram Distance	140
5.1.5	Results and Discussion	140
5.1.5.1	Experiment 1: Latent Dirichlet Allocation	140
5.1.5.2	Experiment 2: Burrows' Delta	145
5.1.5.3	Experiment 3: CNG Distance	146
Chapter 6	Conclusion and Future Work	150
6.1	Achieved Objectives	150
6.2	Future Work	155

6.3 Coda	156
References	157
Appendix A Financial Statement Overview	184
Appendix B Big and Small Data	186
Appendix C US Financial System	189
C.1 Intermediation	189
C.2 Going Concern and Concentration Risk	190
C.3 Financial Distress	190
Appendix D Tabular Dataset Company Listing	192
Appendix E Financial Literacy Results	193
E.1 Cosine Similarity Results	193
E.2 Readability Results	194
Appendix F Selection of Peers	201
Appendix G ChatGPT Transcripts for Synthetic Generation	202
Appendix H Coherence Scores	205
Appendix I Features Removed for Project 7 - Experiment 1	208
I.1 Textual features removed prior to Latent Dirichlet Allocation	208
I.2 Topic features for Latent Dirichlet Allocation evaluation	211

Appendix J Topic Scores 213

LIST OF TABLES

3.1	Company Listing [309]	32
3.2	Dataset Summary Information.	36
3.3	Comparative Convolutional Neural Network Results	43
3.4	Tabular Dataset Descriptive Statistics	52
3.5	Classification Results For Each Generation Algorithm	74
4.1	Fine-Tuning Prompt and Completion Pairs	77
4.2	Cosine Similarity Results	84
4.3	Wilcoxon Signed Rank test for Cosine	86
4.4	Wilcoxon Signed Rank test for Readability	86
4.5	Sentiment Breakdown And Variance	87
4.6	Disclosure References by Bank and Year	93
4.7	Results for Cosine Similarity.	97
4.8	Results for Jaccard Similarity	99
4.9	The Results of Burrows' Delta	99
4.10	Results for Flesch Reading Ease	101
4.11	Financial Literary Dataset Summary Statistics	111
4.12	Fine-Tuning Prompt and Completion Pairs	125
4.13	Results for Semantic Similarity Using GPT-3.5 and GPT-4	127
4.14	Change in Sentiment Scores	128
5.1	Bank Listing	135
5.2	Jurisdictional accuracy — top features removed	142
5.3	Jurisdictional accuracy — bottom features removed	142
5.4	Company accuracy — top features removed	143
5.5	Company accuracy — bottom features removed	144

5.6	Classification accuracy by jurisdiction	145
5.7	Canada — word n-grams	147
5.8	Europe — word n-grams	147
5.9	US — word n-grams	148
5.10	Canada — character n-grams	148
5.11	Europe — character n-grams	149
5.12	US — character n-grams	149
D.1	Tabular Dataset — Companies and Records	192
E.1	Cosine Similarity Results for GPT-3.5	195
E.2	Cosine Similarity Results for GPT-3.5 (Continued)	196
E.4	Cosine Similarity Results for GPT-4	197
E.5	Cosine Similarity Results for GPT-4 (Continued)	198
H.1	Coherence Scores - Known	206
H.2	Coherence Scores - Unknown	207

LIST OF FIGURES

1.1	Financial statement relationships	3
1.2	An overlay of Note 4 (Securities)	4
2.1	Basic Autoencoder Architecture	15
2.2	Variational Autoencoder Architecture	16
2.3	GAN Architecture	17
2.4	Example of creating synthetic data records with SMOTE	22
3.1	Convolutional Neural Network for this research	39
3.2	Model loss for the unaugmented dataset	42
3.3	Model loss for the fully augmented dataset	43
3.4	Range of random perturbations	46
3.5	Model losses for three different random perturbation (10%)	47
3.6	Model losses for three different random perturbation (20%)	47
3.7	Model losses for three different random perturbations (30%)	47
3.8	Example of synthetic financial statements	56
3.9	Example of accounting data to be passed to the generator.	57
3.10	TabGAN Sufficiency Issue	59
3.11	CTGAN 1000 Synthetic Records	60
3.12	Variational AutoEncoder Synthetic Records	61
3.13	Prompt provided to ChatGPT	62
3.14	Sample of the full tabular synthetic generation by ChatGPT	62
3.15	Visual Inspection of synthetic JPM data	64
3.16	Distributions of Net Interest Income	65
3.17	Peer augmentation process	68

4.1	Flesch Reading Ease Scoring	81
4.2	Side-by-Side comparison for Citigroup	85
4.3	Sentiment Shifts Between Models For Wells Fargo	88
4.4	GPT-3’s output of questions	89
4.5	GPT-3’s output in Thai	89
4.6	Comparison of Note Disclosure 1	103
4.7	Comparison of Note Disclosure 2	104
4.8	Illustration of semantic similarity	124
4.9	Ground truth and ChatGPT text for Bank of New York Mellon	126
4.10	ChatGPT’s generated text for Q1	126
5.1	A comparison of RBC’s 2012 and 2013 annual report	132
5.2	Partial sample of the DBT matrix,	139
G.1	Prompt given to ChatGPT	202
G.2	Response 1 for ChatGPT	203
G.3	Response 2 for ChatGPT	204
J.1	Topic scores — Canada	214
J.2	Topic scores — Europe	215
J.3	Topic scores — US	216
J.4	Topic scores — Canada “Unnamed”	217
J.5	Topic scores — Europe “Unnamed”	218
J.6	Topic scores — US “Unnamed”	219

ABSTRACT

Information is fundamental to decision-making. Yet, data is very sparse for the financial domain, even though, in this era of big data, it seems abundant. The work presented in this thesis addresses that scarcity over seven projects which investigate and examine creating synthetic financial data, both quantitative and textual. In the first two projects, we examine methods to generate synthetic financial statement data as well as the effects of synthetic data on a downstream classification task. The next four projects evaluate how well ChatGPT generates textual financial data for the notes to the financial statements, selected parts of financial reports, as well as how it adapts its responses based on the identified knowledge of its end users, ranging from a non-financial user to a financially sophisticated user.

The authorship attribution project is of the utmost importance particularly since company authorship attribution has not been studied yet, to the best of our knowledge. We have author profiles and a good understanding for identified authors such as William Shakespeare, Mary Shelley, or George Washington, but we do not yet have that depth of understanding and identifiability for corporate communication. This attribution task is a non-trivial problem given that lengthy corporate communication is often collaboratively written by many authors, many (or all) of which are never identified, with contributions by non-writing authors as well who vet and review the text or sign off on the text, for example. This plethora of unidentified authors means that we have to treat the text as a single “figurehead” author, with the understanding that many (likely) unidentified authors (writing and not) have contributed to the work. In our experiments, the Common N-Gram Distance algorithm provided the best and most consistent results, achieving between 95% and 100% accuracy for character n-grams and 100% accuracy for word n-grams. Tools like ChatGPT can be exploited and used to commit fraud. Given the potential for significant effect and harm on the capital markets, tools that can quickly detect fraudulent corporate communication will be needed. Our research contributes to that effort.

LIST OF ABBREVIATIONS USED

AA Authorship Attribution. 7, 8

AR Annual Report. 1, 2, 4, 5, 7, 32, 92, 93, 98, 102, 130

AV Authorship Verification. 7, 8

BERT Bidirectional Encoder Representations from Transformers. 32, 33

BS Balance Sheet. 3, 4, 5, 52, 53, 56, 57, 67, 73, 184, 185

CNN Convolutional Neural Network. iii, 11, 27, 28, 37, 38, 69

DL Deep Learning. 25

EDGAR Electronic Data Gathering, Analysis, and Retrieval system. 30, 31

FML Financial Machine Learning. 12

FS financial statements. xvi, 1, 2, 3, 4, 5, 6, 7, 8, 12, 14, 25, 26, 30, 50, 51, 52, 55, 56, 57, 58, 60, 67, 70, 71, 73, 121, 128, 152

GAN Generative Adversarial Network. 16, 17, 19, 20

IS income statement. 3, 4, 5, 52, 56, 57, 67, 73, 184

LLM Large Language Model. 8

LR Linear Regression. 69

ML Machine Learning. 4, 7, 12, 25, 26

NB Naïve Bayes. 69

RF Random Forest. 69

SCF Statement of Cash Flows. [184](#), [185](#)

SEC U.S. Securities and Exchange Commission. [30](#), [31](#), [35](#), [36](#), [51](#), [52](#)

SMOTE Synthetic Minority Over-sampling Technique. [22](#)

SVM Support Vector Machines. [69](#)

GLOSSARY

10-K A 10-k is a regulatory filing required by public companies which provide comprehensive details on the company’s operations, financial health, risks [327]. The 10-K also includes the audit opinion as well as the [FS](#) and [Notes](#). [30](#), [33](#), [34](#)

augment The act or result of using tools to add more records to increase the size of the dataset. [2](#), [11](#), [26](#), [35](#), [50](#), [51](#), [66](#), [72](#), [90](#), [106](#), [121](#), [150](#), [151](#)

BLEU bilingual evaluation understudy, better known as BLEU, is a metric that evaluates the quality of translated text based reference translations, with a score ranging from 0 to 1 [241]. [122](#)

Concentration Risk Concentration Risk is a risk that must be disclosed by the company if it poses a severe danger to the company. This type of risk can manifest in many ways, but it most notably manifests itself in concentrations of the sources or types of credit that a bank extends”. [27](#), [34](#), [35](#), [36](#), [45](#), [47](#), [138](#)

embeddings Embeddings are vector representations of text. Embeddings can be calculated for either words, sentences, paragraphs, and documents. [27](#)

Equity Ratio The Equity Ratio is calculated as $EquityRatio = TotalEquity/TotalAssets$. This ratio is important as it provides information on how it funds its assets. If the ratio is less than 0.5, it means that the company is funded with more debt than equity. [33](#)

Financial distress We define financial distress using two conditions: when a company is not able to meet its financial obligations and when [Going Concern](#) has been flagged by the auditors. [27](#)

GAAP Generally Accepted Accounting Principles are the rules that govern corporate accounting to ensure consistency and transparency [141]. [5](#), [91](#)

Going Concern Going Concern is the ability for a company to meet its financial obligations and continue business for the foreseeable future.. [xvi](#), [27](#), [34](#), [35](#), [36](#), [40](#), [45](#), [46](#), [52](#), [138](#)

IFRS The International Financial Reporting Standards are accounting standards that must be followed by public companies in most countries for the preparation of their interim and annual financial statements. At this time, US public companies must continue to use US GAAP rather than IFRS. [5](#)

Intermediaries Together, banks and shadow banks are known as “financial intermediaries”. [28](#)

LIBOR The London Interbank Offered Rate is the interest rate at which institutions will lend to each other [143]. [27](#), [33](#), [36](#), [40](#), [46](#), [48](#)

MD&A The Management Discussion & Analysis section of the annual report that provides information on the company’s performance and how the Management expects the company to perform in the future. [29](#), [30](#), [31](#), [32](#), [36](#), [40](#)

Non-GAAP measures Industry accepted financial measures that are outside of the accounting rules. [31](#)

Notes Note Disclosures. Also called the “Notes” to the financial statements or just “notes”. The notes accompany the financial statements and are presented directly after the financial statements. Items in the financial statements that have accompanying notes are numbered so that the reader can then consult the appropriate number in the notes to get the additional information. [xvi](#), [2](#), [3](#), [4](#), [5](#), [7](#), [30](#), [121](#), [122](#), [128](#), [152](#)

Notes A critical part to the financial statements are the *note disclosures*, often referred to simply as “Notes”. These disclosures provide critical additional information about items recognized in the financial statements [50, 167], as well as those that are not [148].. [1](#)

QDAP The Quantitative Discourse Analysis Package (available in R) is a collection of text analysis dictionaries. [32](#)

ROUGE Recall-Oriented Understudy for Gisting Evaluation, better known as ROUGE, is a metric that evaluates summarized text based on “ideal” human summaries [187]. 122

ticker A ticker (stock symbol) is a unique abbreviation label provided to each company on the stock exchange [3]. For example, Apple’s ticker symbol is AAPL. 55

ACKNOWLEDGEMENTS

Unbeknownst to me, life had set things in motion which would give me the opportunity to do a PhD — something I had always wanted to do. Even though I had an MBA, I still needed a master degree with a thesis. There began my PhD journey seven and a half years ago. I returned to do a Master of Electronic Commerce at Dalhousie University and it completely revolutionized my life and has taken me in new directions, challenged me to “up my game”, and has brought me into work and research projects that I never thought I would ever be doing. A lot of people have helped me directly and indirectly in achieving my goal, and so I would like to recognize them here for what they have done for me as a way to express my immense gratitude.

Dr. Vlado Kešelj has been my supervisor for almost seven years and has guided me through my Master of Electronic Commerce and PhD. Without his support, insight, and him taking a chance on a non-traditional computer science aspirant, I would not be where I am today. To those who follow me — traditional computer science student or not — you will not find a better supervisor (in my humble opinion). Thank you Dr. Kešelj for everything you have done for me.

As a very green e-Commerce student, I took a data mining class with Dr. Jacek Wołkowicz. One day after class, he asked me what my future plans were after I finished the Master of e-Commerce. I am sure he thought nothing of the question at the time. He listened to my plans to go on and do an accounting PhD and what I wanted to do for my research. When I finished my explanation, he suggested I check out this thing called “Natural Language Processing”. I had no idea what it was or that it even existed. After looking into it, at his suggestion, I made the decision to change my PhD aspirations and go in a new direction because he was right — what I needed was Natural Language Processing for me to do the research that I wanted to do. Thank you, Dr. Wołkowicz — I hope you know what an impact you had on me as my instructor and now colleague.

I am grateful for the guidance and support that my internal committee — Dr. Vladimir Lucic, Dr. Evangelos Milios, and Dr. Malcolm Heywood — has given me over the years. I would also like to thank Dr. Howard Hamilton for agreeing to be my external examiner and for his feedback and comments on my dissertation.

It has been an honour to be part of the Dalhousie Natural Language Processing (DNLP) group over the years. Special thanks to Dr. Kešelj for organizing it and bringing us together. To my fellow students who have sat through my dry-run presentations and given me feedback on how to improve, thank you. I also thank Dr. Yannick Marchand and Dr. Nauzer Kalyaniwalla for their support and feedback over the years.

I would like to sincerely thank Dr. Sylvain Charlebois who has been a great mentor and provided me with counsel, opportunities, and support. It has been an honour to work with you and be part of the Agri-Foods Analytics Lab. I would also like to thank other members of the Agri-Foods Analytics Lab – Janet Music and Janèle Vézeau — for their collaboration and friendship throughout my journey. I look forward to continuing our partnership!

The Dalhousie Writing Centre has also been a tremendous help to me over the years, beginning with “Writing Week” where I met Kala Hirtle, Janice McDonald-Eddington, and Adam Auch. I am so grateful for their insights into my writing, helping me prepare for presentations, and giving me support on how to overcome challenges as a PhD student and candidate.

A special thanks goes out to Helena Martel, the Faculty of Graduate Studies’ Senior Thesis Clerk who is a tour de force. Helena — we would all be lost without you. Thank you for everything you do.

Dr. Linda Macdonald — you have been an absolutely amazing colleague, mentor, and friend to me throughout the years. I am so grateful for the serendipitous way that we came together and for everything you have done for me, opportunities you have given me, and knowing that you were always in my corner.

There are a number of very special people who have, in their own way, helped me in my PhD journey: Dr. David Roach, Dr. Suzanne Le-May Sheffield, the late Dr. Anne Marie Ryan, James Fleming, Anthony Jones, Heather Hector, Michael Potwin, Dr. Shannon Lin, and Antonia Prlic. Thank you for your friendship, guidance, and support.

To my amazing friends Barnaby Kerekes, Jennifer Baird, Dijana Kosmajac, Colin Conrad, Jenny Farbstein, Anna-Marie Manley, Samantha Taylor, Bryce Cross, Karen MacRury-Sweet, Wayne Sweet, Allison Sweet, and Rebecca Oulton — you are all amazing and I would be lost without you. I love you all so much.

Finally, thank you to my family for their support. I wish that my mother, Gail, who passed away in 2012, could have been here to see me complete such a momentous journey. I also thank my fur babies for all their love — Shelby, Alley, Honey Pie (HP), and Cleo. You were always there when I needed you. I love you so much. I especially thank my sister, Kylie, my brother-in-law Dave, and Ashley and Colton, my niece and nephew, for their unwavering support. It means the world to me to know that through the ups and downs you have been in my corner and always will be.

Chapter 1

Introduction

1.1 Motivation

Information is fundamental to decision-making. Information is created through the transformation of data so that we, as humans, can understand it and make informed decisions with it. Investors need information to make investments, companies need information on how to move forward, creditors need information on who to lend to, and regulators need information to ensure that companies are in compliance. Yet, data is very sparse for the financial domain, even though, in this era of big data, it seems to be abundant.

Where we get the data to transform into information depends on the decisions that we are trying to make. Public companies are very secretive and protect their information (usually referred to as “firm-level” data), as this internal information comprises things like transactions, products in development, and planned strategic moves. As a result, this information provides insight into what makes a company competitive [94]. Companies do not usually share this level of data as it can erode their competitive advantages (past, current, and future), give away proprietary information and provide micro-level information on transactions that are aggregated into larger numbers on their financial statements (FS).

That said, public companies are required to provide quarterly and annual disclosures on performance which include the FS, the Notes to the FS (Notes), and the Annual Report (AR). These requirements mean that, at minimum, there are four reports (three quarterly and one annual) made available to the public each year. Therefore, over a 20-year period, for example, 80 data records (reports) are available, provided that the company has remained in business, not missed any filing deadlines, has not been acquired, and has not converted to a private company. At first glance, 80 records seems like a lot until we compare to other types of publicly available financial data such as securities data. Returning to our example of 80 records that took 20 years to acquire, researchers can scrape 80 records of securities data in 80 minutes (using the smallest data collection interval of 1 minute) from sites like

Yahoo Finance.

Therefore, we need to find ways to **augment** (increase the size of) the information that public companies provide to help inform decision-making. This need to supplement the data leads to a two-part connected challenge: if we **augment** the **FS**, that also means that we need to **augment** the **Notes** (through generation) and the financial reports that include the **FS** and **Notes**. Otherwise, we limit the decision-making to only the quantitative data. As well, these two documents are presented as part of the Annual Report (**AR**). Therefore, it follows that the **AR** would also need to be augmented. To date, the only work that has been done on financial augmentation is for the securities market¹. At the time of this research, we can find no evidence of any work being done on the augmentation of the **FS**, the **Notes**, or the **AR**.

Augmentation of these elements, however, creates technological challenges. The main **FS** are all linked together (briefly discussed in Section 1.2.1) and must balance (i.e., tie together) through specific line item amounts shared between statements. It is this balancing that creates a constraining element which adds to the challenge. If the **FS** do not balance, then the data is largely unusable. Therefore, the overarching motivation of this thesis is to address the scarcity of accessible public company data. This is done by investigating current methods for quantitative tabular augmentation and proposing a new method specific to the **FS**, as well as evaluating ChatGPT as a method for augmenting the **Notes** and the financial reports. We also provide a foreshadowing² on the impacts of augmentation and the need to expand tools and detection methods as a result.

¹The “securities market” is often colloquially referred to as the “stock market”. However, this implies that the only product available are stocks, which is incorrect. Therefore, we use the term “securities market” for accuracy, but also to recognize that other products are available besides stocks such as bonds, warrants, and options, for example.

²We use the term “foreshadowing” here purposefully as we believe it conveys the gravity of a warning more than other terms such as anticipation or prediction.

1.2 General Background

1.2.1 Financial Statement Data and Augmentation Challenges

There are four main FS: the Income Statement (IS), the Balance Sheet (BS)³, the Statement of Retained Earnings, and the Statement of Cash Flows [169, 185]. Each financial statement is related to the other FS, and in order for all of statements to balance, the relational inputs have to be correct. Figure 1.1 provides a basic diagram of the important relationships and each statement is briefly discussed in Appendix A.

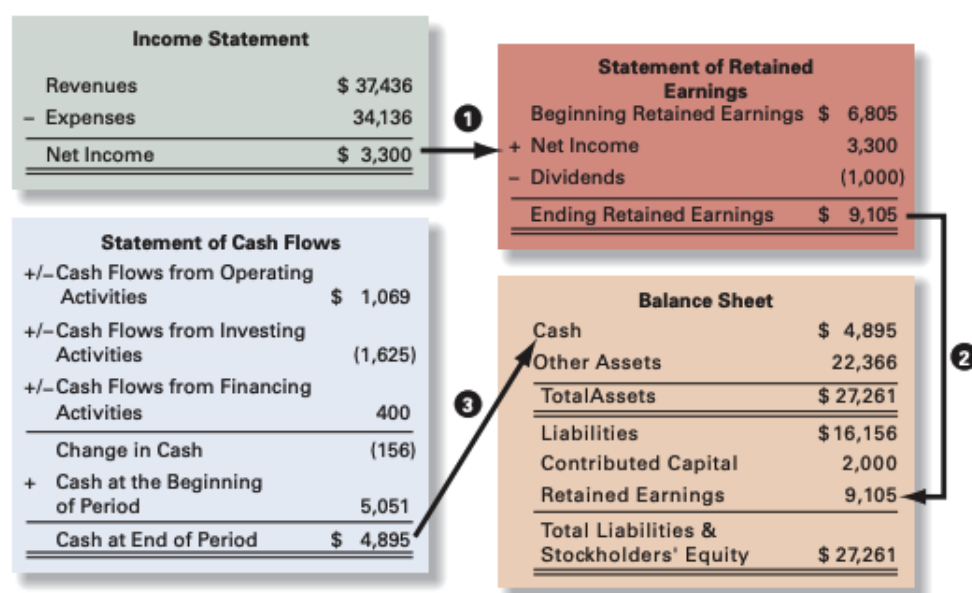


Figure 1.1: Financial statement relationships from “Financial Accounting” by Libby *et al.* [185]. All the statements are related and share specific line items that link them together.

The **Notes** to the **FS**, which can also be referred to as “Disclosure Notes” or just simply “Notes”, accompany the **FS** and are presented directly after the **FS**. Items in the **FS** that have accompanying **Notes** are numbered so that the reader can then consult the appropriate number in the **Notes** to get the additional information. These **Notes** provide critical additional information about items recognized in the financial statements [50, 167], as well as those that are not [148]. Disclosures may be required by the Generally Accepted Accounting

³Under the International Financial Reporting Standards, the balance sheet is referred to as the “Statement of Financial Position” and the income statement is referred to as the “Statement of Profit or Loss”.

Principles (GAAP) or other regulations, while other disclosures may be provided by management to facilitate user understanding [148]; there is no “one size fits all” approach to disclosures. Figure 1.2 shows an overlay of one note (Note 4 “Securities”) from the Royal Bank of Canada’s 2023 AR showing how the Notes are referred to in the FS and how the Notes are presented after the FS.

Consolidated Balance Sheets		As at	
(Millions of Canadian dollars)		October 31 2023	October 31 2022
Assets			
Cash and due from banks		\$ 61,989	\$ 72,397
Interest-bearing deposits with banks		71,086	108,011
Securities (Note 4)			
Trade receivables			
Investment, net of applicable allowance			
Assets purchased under reverse repurchase agreements			
Loans (Note 5)			
Retail			
Wholesale			
Allowance for loan losses (Note 5)			
Segregated fund net assets (Note 16)			
Other			
Customers' liability under acceptances			
Derivatives (Note 9)			
Premises and equipment (Note 10)			
Goodwill (Note 11)			
Other intangibles (Note 11)			
Other assets (Note 13)			
Total assets			
Liabilities and equity			
Deposits (Note 14)			
Personal			
Business and government			
Bank			
Segregated fund net liabilities (Note 16)			

Note 4 Securities (continued)		As at	
(Millions of Canadian dollars)		October 31, 2023	October 31, 2022
Gross gains and losses on securities at FVOCI (1), (2)			
		Cost/ Amortized cost	Gross unrealized gains
			Gross unrealized losses
			Fair value
Debt issued or guaranteed by:			
Canadian government	\$ 6,609	\$ 1	(\$ 351)
Federal	3,396	2	(650)
Provincial and municipal			
U.S. federal, state, municipal and agencies	75,326	343	(2,374)
Other GCCD government	6,200	1	(9)
Mortgage-backed securities	2,762	-	(61)
Asset-backed securities			
CDO	8,308	3	(46)
Non-CDO securities	444	2	(5)
Corporate debt and other debt	27,774	44	(95)
Equities	493	357	(8)
	\$ 131,312	\$ 753	(\$ 3,599)
			\$ 128,466
			\$ 95,630
			\$ 696
			(\$ 3,435)
			\$ 92,891

(1) Excludes \$98.83 million of held-to-collect securities as at October 31, 2023 that are carried at amortized cost, net of allowance for credit losses (October 31, 2022 - \$77.07 million).

(2) Gross unrealized gains and losses includes \$133 million of allowance for credit losses on debt securities at FVOCI as at October 31, 2023 (October 31, 2022 - \$193 million) recognized in income and other components of equity.

Allowance for credit losses on investment securities
The following tables reconcile the opening and closing allowance for debt securities at FVOCI and amortized cost by stage. Reconciling items include the following:

- Transfers between stages, which are presumed to occur before any corresponding remeasurement of the allowance.
- Purchases, which reflect the allowance related to assets newly recognized during the period, including those assets that were derecognized following a modification of terms.
- Sales and maturities, which reflect the allowance related to assets derecognized during the period without a credit loss being incurred, including those assets that were derecognized following a modification of terms.
- Changes in risk, parameters and exposures, which comprise the impact of changes in model inputs or assumptions, including changes in forward-looking macroeconomic conditions, partial repayments, changes in the measurement following a transfer between stages, and unwinding of the time value discount due to the passage of time.

Allowance for credit losses - securities at FVOCI (1)

Figure 1.2: An overlay of Note 4 (Securities) on the Consolidated Balance Sheet which refers the reader to the note, taken from The Royal Bank of Canada’s 2023 Annual Report [225].

Working with financial statements adds a layer of complexity for machine learning (ML). The independent and identically distributed (IID) assumption is fundamental to ML. IID refers to “where a collection of random variables are mutually independent and drawn from the same probability distribution” [40]. Depending on the ML task, achieving this type of independence can be a significant challenge due to the dependencies between statements. If, for example, Net Income and Retained Earnings are being used as features for a prediction task, then the IID assumption is violated, as Net Income (from the IS) is an input to calculating ending Retained Earnings, which is an input to the BS. Therefore, features must be carefully chosen to ensure that there are no underlying dependencies.

A popular approach to augmenting small datasets is to generate synthetic data that

closely approximates the distribution of the original data. Preserving the dependencies between statements in the augmentation process makes this a challenging task. If the relationships are not maintained, the generated synthetic statements will be unusable as they will not balance properly. Given that this task has never been done before, the scope has been narrowed to two of the FS: the IS and the BS.⁴

Our research focuses on two specific financial documents — the AR and the FS (including the Notes). The purpose of the AR, which shares a common aim between jurisdictions, is to discuss the company, its risks and financial health, and to provide audited financial statements either under the US GAAP or the International Financial Reporting Standards (IFRS) in Canada and Europe [328, 43, 102, 63].⁵

We note that there does not appear to be a universally accepted definition of synthetic data. Instead, many discussions point to criteria that data must meet in order to be considered “synthetic”, or how it can be created, rather than providing a definition. To illustrate, we look at several different discussions of criteria provided in the literature:

- “Synthetic data must meet two requirements: First, it must somewhat resemble the original data statistically, to ensure realism and keep problems engaging for data scientists. Second, it must also formally and structurally resemble the original data...” [246]. A concern, however, is the use of the word “somewhat”, which is a very broad term. A lemon “somewhat” resembles a lime in that they are both citrus fruit and are relatively the same shape, but they are distinct fruits. As well, we question the requirement for engagement as this suggests that there is an unspoken requirement that data should be interesting for data scientists to keep working on problems. We assert that the problem itself is paramount, regardless of whether the data is engaging or not.
- “A good synthetic dataset has two properties: it is representative of the original data, and it provides strong guarantees about privacy” [144]. The concern here is that there

⁴In augmenting the BS and IS we make two important assumptions: that the ending Retained Earnings balance is the same as Net Income and that no dividends have been declared; and that the cash balance on the IS is correct and is balanced with the cash at the end of the period on the Statement of Cash Flows.

⁵As the United Kingdom (UK) was formerly part of the European Union (EU) from January 1, 1973 until January 31, 2020, for the purpose of this research, we consider the UK part of the EU in a broad sense. Our dataset does not reach as far back as 1973, but any reports or financial statements prepared by UK companies after January 31, 2020 are governed by *Companies House*, an executive agency of the UK government. We do note, however, that the rules requiring the annual report, its contents, and the need for audited financial statements is substantially the same as that for the EU.

is no information in how the data must be representative of the original data (i.e., statistically, categorically, number of records). Also, why does synthetic data always have to be privacy preserving? If a company is creating synthetic data for its own model training purposes (which will only be used “in-house”), there is not necessarily a need for the data to preserve privacy. As well, if publicly available data is being used, the need for privacy is moot.

- “Synthetic data can be created from perturbations using accurate forward models (that is, models that simulate outcomes given specific inputs), physical simulations or AI-driven generative models” [54]. While it is helpful to know how synthetic data can be created, it still does not explain what synthetic data is.

We also note that there is often the assumption that there is already an understanding of what synthetic data is, and therefore, researchers do not take the time to explain or define the term in their work (e.g., [159, 201]). The only work that we were able to find that defined synthetic data as “Synthetic data is data that has been generated using a purposebuilt mathematical model or algorithm, with the aim of solving a (set of) data science task(s)” [158]. The authors then distinguish synthetic from real-world data as data created by events such as “financial transactions, satellite images, [and] medical tests” [158]. The implications of not having a widely accepted definition of synthetic data or assuming that there is a common understanding can quickly lead to confusion, while raising concerns (such as those we raised above). More importantly, it can also lead to a challenge in reproducibility of work as well as comparability of methods and results. While we raise this issue here for the term “synthetic”, we also draw attention to this type of issue in Project 1 (Section 3.2) for the term “bankruptcy”, as it also suffers from the same lack of common definition and assumed understanding, leading to experiments that are challenging to reproduce, as well as methods and results that are not comparable amongst researchers.

Therefore, our augmentation research problem comprises several facets. First, we aim to provide a more robust and broader definition of what augmentation is and its role in creating synthetic data by updating an existing definition of augmentation. Second, we will evaluate current methods for augmenting tabular FS data (Variational AutoEncoders, Generative Adversarial Networks, Synthetic Minority Oversampling Technique, and ChatGPT), and also present a new method whereby peers are used to augment small datasets before

being passed to generative algorithms for augmentation. And third, we will evaluate the use of ChatGPT to generate text for the synthetic FS created using Synthetic Minority Over-sampling Technique. Taking this approach addresses both sides of creating synthetic FS as it generates the tabular FS themselves as well as the contextualizing textual component, the Notes, to provide a complete FS package that can then be used as part of the AR.

1.3 Contributions

The research that we present here makes the following scientific contributions:

1. We investigate and demonstrate why only certain features of the FS can be passed to ML models for augmentation, and if complete FS are needed, the FS will need to be reconstructed after augmentation
2. A new approach to augmentation for FS is presented: peer augmentation. This approach is useful when the company’s dataset is too small to be feasibly passed to a generative algorithm. This approach is also useful if the company’s dataset does not have sufficient representation of features needed for analysis or for a downstream task. In both situations, peers are used as initial scaffolds to either provide a sufficient number of records to be passed to a generative algorithm, or to provide a broader feature set to the generative algorithm.
3. Three projects demonstrate ChatGPT’s ability to generate financial report text, FS Notes, and how ChatGPT struggles to tailor its responses to the needs and experience of its users in the financial domain.
4. A new application of semantic similarity is presented where it is used to evaluate text that has no “ground truth” text for comparison.
5. Work is presented on company authorship attribution (AA) which, to the best of our knowledge, has never been studied before. There is a good understanding of famous authors such as William Shakespeare, Mary Shelley, or George Washington, for example. This type of knowledge and understanding facilitates AA and authorship verification (AV) tasks. Yet, we do not have this same level of knowledge or understanding for companies. There is an unstated assumption that because the company’s name is on

the reports, the earnings announcement, or the [FS](#), then there is no doubt as to the author. Yet, there are many areas of reports or parts of the [FS](#) where the company's name is not mentioned. As well, now that Large Language Models ([LLMs](#)) can generate human-like text, it is imperative that we have a good understanding and knowledge to be able to do tasks such as company [AA](#) or [AV](#) to better use Artificial Intelligence (AI) tools in the creation of company communications or to combat fraud in a timely manner.

1.4 Publications

My publication record to date is provided below, listed chronologically starting with the most recent. I also have several papers under review, which have been included at the end.

Conferences

- **Taylor, S.** & Keselj, V. (2023, June). Don't Worry Accountants, ChatGPT Won't Be Taking Your Job...Yet. In Proceedings of the 36th Canadian Conference on Artificial Intelligence (CANAI)
- **Taylor, S.** & Keselj, V. (2022, December). Don't Blindly Use Data: Towards a Data Statement for Computational Financial Research. In Proceedings of the 1st International Workshop on Responsible AI and Data Ethics. IEEE BigData Conference 2022 (Big Data)
- Balfagih, A. , Keselj, V. and **Taylor, S.** (2022, May). N-gram and Word2Vec Feature Engineering Approaches for Spam Recognition on Some Influential Twitter Topics in Saudi Arabia. In 2022 ICISDM 6th International Conference on Information System and Data Mining (ICISDM)
- **Taylor, S.** & Keselj, V. (2021, September). Predicting the Distress of Financial Intermediaries using Convolutional Neural Networks. In 2021 IEEE 23rd Conference on Business Informatics (Vol. 2, pp. 71-77). (CBI)
- **Taylor, S.** & Keselj, V. (2020, July). e-Commerce and sentiment analysis: predicting outcomes of class action lawsuits. In Proceedings of the 3rd Workshop on e-Commerce and NLP (pp. 77-85) (ECNLP). ACL 2020 Conference. (ACL)

- Kosmajac, D., **Taylor, S.** & Saeidi, M. (2020, December). Dnlp@ fintoc'20: Table of contents detection in financial documents. In Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (pp. 169-173). COLING 2020. (COLING)
- **Taylor, S.** & Keselj, V. (2020). Using Extractive Lexicon-based Sentiment Analysis to Enhance Understanding of the Impact of Non-GAAP Measures in Financial Reporting. In Proceedings of the Second Workshop on Financial Technology and Natural Language Processing (pp. 40-46). IJCAI-PRICAI 2020. (IJCAI)

Journals

- Affordability of the Canada Food Guide — Current Challenges Amidst COVID-19, War in Ukraine, and Other World Events by **S.Taylor**, S. Charlebois, and J. Music. *Frontiers in Nutrition* (2023)
- The Local Food Paradox: A Second Study of Local Food Affordability in Canada by S. Charlebois, M. Ducharme, M. Morrison, J. Vezeau, **S. Taylor**. *Sustainability* (2023)
- N-gram and Word2Vec Feature Engineering Approaches for Spam Recognition on Some Influential Twitter Topics in Saudi Arabia by A. Balfagih, V. Keselj, and **S. Taylor**. *Journal of Advanced in Information Technology* (2022)

Books

- Contributing Author to “Building Relationships with Business Communication” by Linda MacDonald: MacDonald, L. (2021). *Building Relationships with Business Communication*. Dalhousie University. Available from <https://caul-cbua.pressbooks.pub/businesscommunication/>

Canada's Food Price Report

- Charlebois, S, Rankin, A, **Taylor, S.** Keselj, V, Abebe, G, Colombo, S, Kevany, K, Marignani, P, Music, J, Nason, R others. (2023, December) *Canada's Food Price Report*. Dalhousie University. www.dal.ca/sites/agri-food/publications.html

- Charlebois, S, Rankin, A, **Taylor, S**, Keselj, V, Abebe, G, Colombo, S, Kevany, K, Marignani, P, Music, J, Nason, R others. (2022, December) Canada's Food Price Report.
Dalhousie University. www.dal.ca/sites/agri-food/publications.html
- Charlebois, S, Gerhardt, A, **Taylor, S**, Kane, M, Keselj, V, Fitting, A, Foster, K, Kevany K. Columbo S, Music, J, Fiander, D, others. (2021, December) Canada's Food Price Report.
Dalhousie University. www.dal.ca/sites/agri-food/publications.html
- Charlebois, S, Gerhardt, A, **Taylor, S**, Kane, M, Keselj, V, Fitting, A, Foster, K, Kevany K. Columbo S, Music, J, Fiander, D, Kenny, T-A, Abebe, G, and others. (2020, December) Canada's Food Price Report.
Dalhousie University. www.dal.ca/sites/agri-food/publications.html

Currently Under Review

- Hey ChatGPT - is a Louis Vuitton bag an Investment? Evaluating LLM Readiness for Use in Financial Literacy and Education. Currently in Review and Resubmit Status with the Journal of Emerging Technology in Accounting.
- Economic Implications of Carbon Taxation and Effects on Food Affordability and the Agri-Foods Sector in Canada — a Scoping Review. Current under review with Science of the Total Environment.
- Implications of Carbon Taxing Policies on the Food Supply Chain in Canada under review with Environmental and Resource Economics.
- (Book) Chapter 19: Canada's Food Inflation Story. Currently under review with Future Food Systems.

1.5 Outline

The rest of this thesis is structured as follows:

Chapter 2: This chapter provides a general literature review of financial machine learning. We focus on four main areas of research in this review which are directly linked to our research: markets, distress/crises, service, and risk management. We identify important gaps in the literature and link them to our research projects.

Chapter 3: In this chapter, we present two research projects for tabular augmentation. First, we predict the distress of financial intermediaries using a Convolutional Neural Network (CNN) with a relatively small dataset. This dataset was constructed for this research and uses both quantitative and textual features for the companies in the dataset as well as the market at the time. Second, we build on the findings from the first project, focusing on finding more sophisticated ways of augmenting financial statement data by generating synthetic financial statements, which is needed when the original dataset is small. As part of this research, we developed a new method to [augment](#) financial statement data — peer augmentation. We also demonstrate how, for relatively small datasets, starting with the simple method of $\pm 10\%$ is needed before peer augmentation is possible.

Chapter 4: This chapter presents four projects that use ChatGPT for synthetic textual generation. The first three projects of this chapter evaluate how well ChatGPT generates synthetic financial text for the notes to the financial statements as well as for select parts of financial reports. These projects also serve as benchmarks as we use GPT-3 and the *davinci* fine-tune for GPT-3, as well as GPT-3.5 and GPT-4. The last project in this chapter evaluates how well ChatGPT tailors its textual responses based on the identified financial knowledge that its audience has, ranging from a non-financial user to a sophisticated financial user.

Chapter 5: Given ChatGPT’s ability to generate human-like text, with this last project and chapter, we foreshadow⁶ the need for company authorship attribution. To the best of our knowledge, the task of company authorship attribution has never been done before.

Chapter 6: In this final chapter, we give a general summary of the work presented in this thesis, discuss future work, and review important limitations identified throughout the thesis.

⁶We specifically use the term “foreshadow” to convey the need for a warning. While the term “anticipation” could also be used here, it does not convey the same urgency and gravity.

Chapter 2

Literature Review

Machine learning (ML) has been used in the financial domain for many decades [240]. Using recent (2021 and 2022) bibliometric reviews [5, 111], we find that financial machine learning (FML) break into six major categories: markets, pricing, distress/crises, service, social, and risk management. Within that, we find that there is a concentration of research related to the securities market, fraud, and sentiment (usually in combination with securities or fraud). These concentrations means that there is still a lot of gaps in the FML literature, particularly in our areas of research which focuses primarily on FS and financial reporting.

The majority of contemporary data augmentation research does not define or address what constitutes “data augmentation” (e.g., [262, 161, 259, 18, 341, 340]), as a common understanding is assumed. As well, many researchers discuss their methods, referencing the original data distribution, but stop short of first defining the goal of data augmentation. Augmentation is commonly understood to be the creation of synthetic data in an effort to create more or different training data, usually to improve performance in downstream tasks like classification or regression [313]. This approach only focuses on the “end goal”, which has not always been the case. Older approaches also tended to consider what data augmentation is on a more fundamental level: “The term data augmentation refers to methods for constructing iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables” [329]. Both points of view, however, do not incorporate other uses of synthetic data such as for the task of forecasting or “what if” scenarios, for example, that can expand understanding and knowledge of the way that changes in data can affect (positively, negatively, or neutrally) machine learning models. Therefore, we take a methodological approach to amend and update (shown in italics) the definition to incorporate a more contemporary view to define data augmentation as “The term data augmentation refers to methods for constructing iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables *for the improvement of downstream tasks and for expanding knowledge and understanding of datasets*”.

2.1 The First Generative Models

2.1.1 EM Algorithm

Data augmentation models have been around since the 1970s, starting with the expectation-maximization algorithm, better known as the *EM* algorithm [74]. Although we do not provide an in-depth analysis of each of the algorithms developed since the 1970s, we briefly touch on some early models where later models draw their inspiration.

The first algorithm is the EM algorithm which finds maximum likelihood estimates (MLE) in cases where there is missing data, or when there is latent data [202]. In the latter case, the latent data is considered to be *missing*, which facilitates the application of the EM algorithm [202]. The notion of there being incomplete data gives rise to two different sampling spaces: Y and X , where y is observed and x is latent [74]. There is also an assumption of a mapping $x \rightarrow y(x)$ from X to Y , where x represents the “complete-data”.

$$g(y|\phi) = \int_{X(y)} f(x|\phi)dx \quad (2.1)$$

Equation 2.1 shows how “[t]he complete-data specification $f(\dots|\dots)$ is related to the incomplete-data specification $g(y|\phi)$.” [74]. ϕ are parameters that the model depends on and that we are estimating to best fit the observed data. This means that the maximization occurs in finding a value ϕ to maximize $g(y|\phi)$ where y is an *observed* y while also using the sampling density of $f(x|\phi)$ to generate possible values of $g(y|\phi)$ [74].

2.1.2 The Boltzmann Machine

The Boltzmann machine facilitates an efficient search for “hypotheses” that will maximize input and constraints [97, 2, 136, 137]. Distributions in the form of Equation 2.2, where $E(X)$ is the energy function, is considered to be a Boltzmann machine [112].

$$\tilde{p}(X) = \exp(-E(X)) \quad (2.2)$$

This model ensures that a probability of zero (though they can reach very close to zero) will not result for X as $\exp(z)$ will be positive for all z [112].

2.1.3 Markov Random Field

A very related model is the Markov random field (MRF) (also known as Markov networks or undirected graphical models [149]). MRF suggests that a process is more likely to be influenced by events that are closer (in the “neighbourhood”) than those further away, where a family of neighbourhoods is formally expressed in Equation 2.3 where $\aleph(i)$ is the neighbour set of i , where $i \notin \aleph(i)$ [149].

$$\aleph = \{\aleph(i), i \in S\} \tag{2.3}$$

The use of the term “Boltzmann machine” has evolved over time. Originally, Boltzmann machines included models that contained latent variables as well as those that did not. That does not still seem to be the case. Contemporary use of “Boltzmann machine” now appears to only include models with latent variables; those without are (more often than not) referred to as log-linear models or Markov random fields [112]. MRFs have been used for data synthesis in recent literature. In one study, PrivMRF which provides flexibility in the choice of marginals (i.e., subsets that match each configuration of features [207]) to construct the MRF, which would then be used for synthesizing privacy protecting data [40]. The MRF has also been used to synthesize images, given the MRF’s ability to capture texture well [155, 183, 198, 129]. More recently, the Conditional Random Field, which is a specialized MRF model, has been used for text-to-speech systems to produce phrase break tags of recorded audio voices [194].

2.2 Contemporary Generative Models

In reviewing the literature, we find that no work has yet been done on generating synthetic data to augment FS. This is a gap in the literature that our work contributes to filling. There is, however, some work being done on generating synthetic financial data for other tasks. Therefore, as part of this literature review, we introduce several contemporary generative models and discuss work on generating synthetic data in the financial domain in general, as chapters three, four, and five provide literature reviews specific to the projects in those chapters.

2.2.1 Variational Autoencoders

A basic autoencoder is a neural network that is trained to produce output that is as close to the original input as possible. It does this by “learning mappings to and from” [247] the bottleneck. First, the representations of the input are encoded ($e(x)$) into specific points in the latent space (a representation of the compressed data [84]), yielding the encoded vector $z = e(x)$ [268]. The encoded vector is then decoded ($d(z)$) to attempt to re-create the original input [268]. In this way, the basic autoencoder is not considered to be a generative model as it simply reconstructs the original [247].

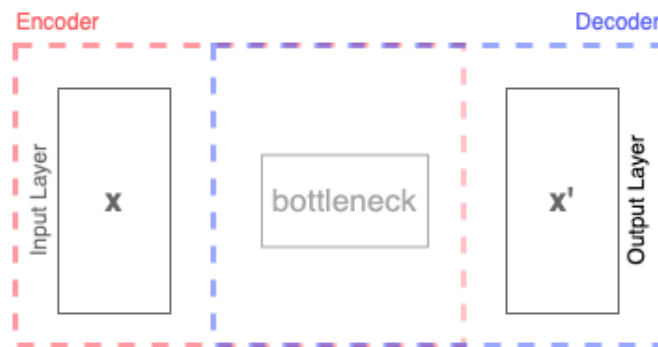


Figure 2.1: Basic Autoencoder Architecture

Although the VAE uses a basic autoencoder, the encoding is slightly different as it includes a mean and a standard deviation for each dimension of the latent space. In this way, the encoder “encodes as a distribution over the latent space” [268]. Equation 2.4 describes the VAE operation, where θ is parameters that $p_{decoder}$ depends on, and ϕ is parameters that $p_{encoder}$ depends on [112, 247]:

$$x' = p_{decoder_{\theta}}(p_{encoder_{\phi}}(x)) \approx x \quad (2.4)$$

In constraining the autoencoder to an approximation, the model is “forced to prioritize which aspects of the input should be copied” [112]. This process facilitates discovering interesting aspects of the inputs that would not have been uncovered without the constraint [112]. The architecture for a VAE can be seen Figure 2.2. In order for a model to be “generative”, it must define the distribution of the data to be able to probabilistically generate new data from the sampled data. As can be seen in Figure 2.2, the VAE learns the probability distributions (yellow box) before the bottleneck (z) which forces the VAE to prioritize the aspects to copy.

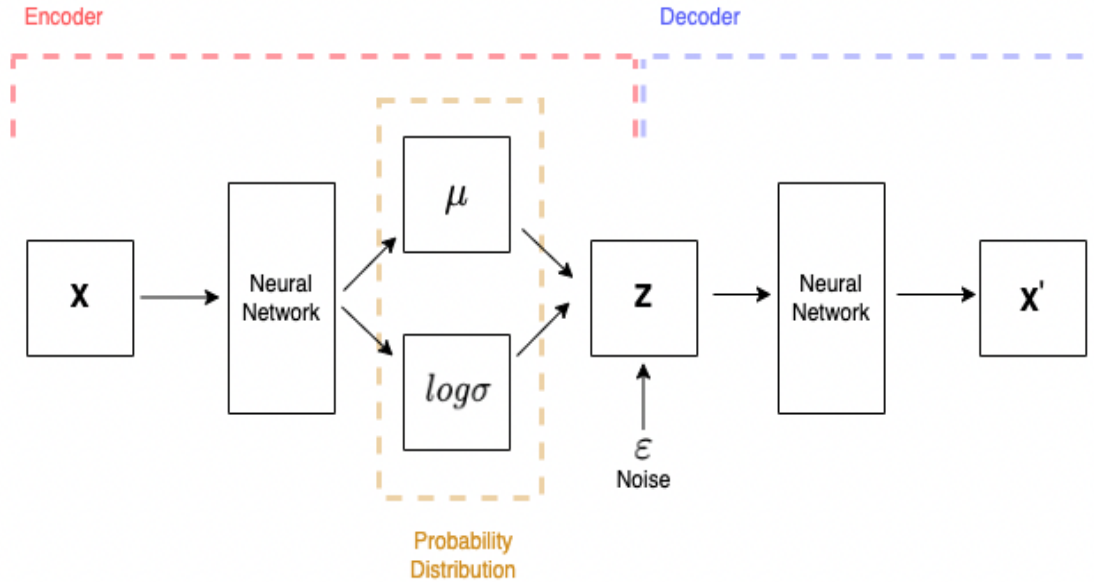


Figure 2.2: Variational Autoencoder Architecture (adapted), based on the work of Hand [247]

If using the squared error loss (also known as the L2 reconstruction error) for evaluation, this can be expressed as an optimization problem [247] as seen in equation 2.5:

$$\min_{\theta, \phi} \sum_{i=1}^n \|D_{\theta}(E_{\phi}(X_i)) - X_i\|^2 \quad (2.5)$$

where $\{X_i\}_{i=1 \dots n}$ is the dataset, and θ and ϕ are the parameters of the encoder and decoder, respectively. Therefore, the goal is to find parameters for both θ and ϕ that will minimize that error.

Given its architecture, the VAE has become a very popular algorithm for data synthesis, and researchers have created task-specific versions such as the Dynamical VAE for temporal tasks [109], the Ladder VAE to address training constraints created by layers of dependent stochastic variables [296], and Nonparametric VAEs for hierarchical representation learning to give flexibility in the latent representation space by blending tree-structured Bayesian non-parametric priors with VAEs [120].

2.2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) comprise two neural networks (a generator and a discriminator) which compete against each other to output synthetic data that cannot be

distinguished from the real data [114]. GANs were originally developed for image data [113]. Recently, GAN models have also been developed for tabular data which is structured data that consists of columns and rows, where the columns are the features, and the rows are the observations [349]. Tabular data is one of the most prevalent data structures used across many domains [103].

The GAN's task in the context of our research is as follows:

- The generator (G) is tasked with creating realistic tabular data
- The discriminator (D) is tasked with distinguishing the tabular data passed to it by the generator from the tabular data from the real dataset. The discriminator will output 1 for real data and 0 for fake data.

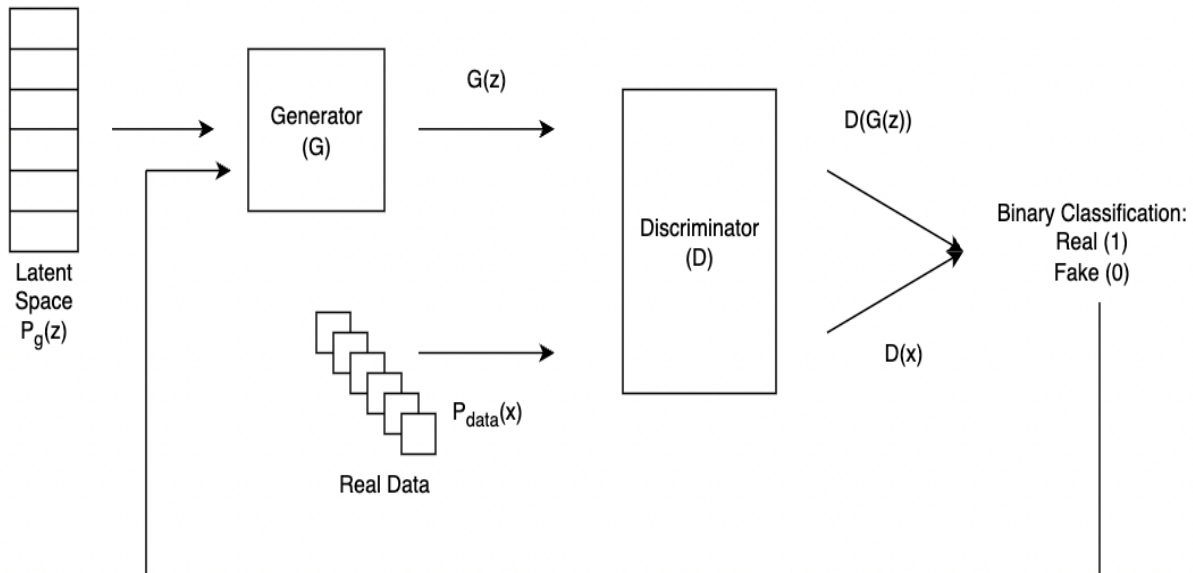


Figure 2.3: GAN Architecture

$G(z)$: generator output

$P_{data}(x)$: real data

$D(G(z))$: discriminator output when fake samples used

$D(x)$: discriminator output when real samples used

The training loop for the GAN follows a three-step process [114]:

1. The generator (G) is initialized with random weights and takes, as input, a sample from the latent space ($P_g(z)$), where the latent space is a representation of the compressed

Algorithm 1 Training steps for GAN

- 1: **for** each training step **do**
- 2: **for** k steps **do**
- 3: Sample m noise samples $\{z^{(1)}, z^{(2)}, \dots, z^{(m)}\}$ from latent space $P_g(z)$
- 4: Sample m noise samples $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ from real data $P_{data}(x)$
- 5: Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))] \quad (2.6)$$

- 6: **end for**
- 7: Sample m noise samples $\{z^{(1)}, z^{(2)}, \dots, z^{(m)}\}$ from latent space $P_g(z)$
- 8: Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z^{(i)}))) \quad (2.7)$$

- 9: **end for**
-

data [84]. From this, it generates a sample tabular data. This sample is then passed to the discriminator (D) to determine if the sample is real or fake. If the discriminator determines that the sample is fake, it outputs (0) and stochastic gradient descent is used to update the generator weights so that it improves its ability to fool the discriminator the next time it generates sample tabular data.

2. The generator (G) again produces a sample of tabular data, and in this step, the discriminator's weights (again using stochastic gradient descent) are updated to improve its ability to discriminate between real and fake.
3. The final step takes a sample of data from the real sample and is passed to the discriminator (D) to check if the discriminator is still able to distinguish real samples from fake samples.

During this training loop, if the discriminator determines the sample is fake, it will output (0). If it determines the sample is real, it will output (1). This training loop repeats until the discriminator is no longer able to tell the difference between the real and fake samples. At this point, the adversarial “game” is finished and the GAN will return the synthetic data along with the real data, as from the discriminator's point of view, it is all “real” data.

Formally, this can be expressed as an optimization problem which seeks to find a G that minimizes the cost function and a D that maximizes it:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [1 - \log D(G(z))] \quad (2.8)$$

The first term is the expectation of the log of the discriminator’s output when the input is from the real data distribution.

If $D(x)$ is 1 (or close to 1), then there is high confidence that the sample is from the real dataset.

The second term is the expectation of 1 minus the log of the discriminator’s prediction when the input is from the generator. This sets up the “adversarial” process in that the discriminator and the generator each want the opposite outcome. The generator’s goal is to drive the likelihood of $D(G(z))$ to 1 in order to “fool” the discriminator into classifying the sample as real, whereas the discriminator wants to drive the likelihood of $D(G(z))$ to 0.

$1 - \log D(G(z))$ transforms this term into a likelihood that the discriminator wants to maximize. Therefore, when both terms are added together, the generator (G) wants to minimize the equation and the discriminator (D) wants to maximize the equation. By making the second term into 1-log of the discriminator’s prediction, it transforms it from something that the discriminator wants to minimize into something that it wants to maximize. Therefore, the optimization problem is to find a G that minimizes the cost function and find a D that maximizes it.

There are hundreds of [GAN](#) models [135] that have been created since the first [GAN](#) model was introduced in 2014 [115]. It is a very popular deep learning method to generate synthetic records. Incomplete table synthesis [GAN](#) [242] (ITS-GAN) and Table [GAN](#) [53] are the closest work to ours in that these models seek to preserve relationships within tables. In the case of ITS-GAN, it tackles the problem of learning a model on an incomplete table and then generating a complete table, where the complete table will return better results than the incomplete table [242]. The model uses the same overall [GAN](#) architecture depicted in Figure 2.3, but the discriminator is fed two functional dependency error terms from autoencoders. The functional dependency for the first error term is determined for each row in the table as part of the “schema-level functional dependency”: $F : A \rightarrow B$, where F is the functional dependency, A is the first column and B is the second column. This

means that the values in column B are specifically determined by the values in column A. The functional dependency for the second error term is determined for the “record-level”, where $f : a \rightarrow b$, where f is the functional dependency, a is a value that specifically modifies b . The requirements for data validity is set out in the work in that the number of columns between the incomplete table and the new synthesized fully-complete table must be the same, the set of functional dependencies are satisfied in the new table, and that the new table must have the same column-wise average of for each column [242]. Table GAN uses a deep convolutional GAN to generate privacy-preserving tabular data [53]. Like ITS-GAN, Table GAN also uses the main architecture discussed in 2.3, but also incorporates a classifier to help prevent invalid data such as “gender = Male and disease = uterine cancer”. The classifier is added to the GAN architecture at the tail end of the discriminator right before the sigmoid activation function to determine if the record is real or synthetic.

GANs were originally developed for images, using Fréchet Inception Distance (FID) as the most common metric to evaluate how well the GAN has generated the synthetic images. Equation 2.9 provides the formula for FID [4]:

$$FID = d^2((m, C), (m_w, C_w)) = \|m - m_w\|_2^2 + Tr(C + C_w - 2(CC_w)^{\frac{1}{2}}) \quad (2.9)$$

where m and C are the mean and covariance of the feature vectors of the real images, m_w and C_w are the mean and covariance of the feature vectors of the the synthetic (generated) images, $\|m - m_w\|_2^2$ is the squared Euclidean distance between the means of the real and synthetic (generated) images feature vectors, and Tr is the sum of the elements on the main diagonal (top left to bottom right), also known as the “trace of a matrix” [306], of the sum of the covariance of the real and generated images which uses twice the square root of the product of the two covariances.

FID measures how far the means of the real and synthetic (generated) image feature distributions are in the model’s feature space. The smaller the FID, the better the synthetic (generated) image indicating that the synthetic image is more similar to the real image. FID has been adopted as the main evaluation metric for images, which provides a common comparator in the research. However, there is no commonly adopted measure in tabular generation, which makes it difficult to compare work across the research.

2.2.3 ChatGPT

ChatGPT’s capability to work with external input files is relatively new (September 2023), when users were granted access to attach files to the chatbox. By default, ChatGPT uses an 8-step framework for synthetic tabular generation as follows [51]:

1. Understand the data structure and content
2. Data preprocessing
3. Select synthetic data generation method
4. Implement the chosen method
5. Generate sythetic data
6. Post-processing and validation
7. Iterate if necessary
8. Ethical and legal considerations

Users can then customize this framework based on their needs. Based on ChatGPT’s understanding of the data structure and content, and the preprocessing it has performed in the background, it will select the method that it has determined is appropriate for the generation. The methods that it will select from are statistical, machine learning, and deep learning [51].¹

2.2.4 SMOTE

When the labelled classes in a dataset are not proportionately equal (or close to equal), the dataset is considered to be imbalanced [52]. Real world data is not balanced and, except in rare circumstances, cannot (and should not) be expected to be. The majority class is, more often than not, the “normally” occurring characteristics of the data, with the minority class comprising the “abnormal” [52]. A leading model to address class imbalance that has been

¹We also note that Gemini (formerly Bard) by Google was released on Feb 8, 2024 [6]. Although it can be used for textual generation, it cannot, at this time (February 2024), directly accept CSV files to produce tabular generation. As such, we leave the evaluation of Gemini to future work.

extensively used in computational finance is Synthetic Minority Over-sampling Technique (SMOTE) [52].²

As its name suggests, SMOTE over-samples the minority class by using k -nearest neighbours and line segments to create the synthetic data, as seen in Figure 2.4.³ To create the synthetic records, the algorithm goes through three steps which are then repeated until the number of required synthetic records are created:

1. The feature vector sample of the minority class (gray dots) and its nearest neighbours (k) (which are also in the minority class) are identified and the distance to each neighbour is calculated. In the illustration in Figure 2.4, we use $k=2$, so that means that distance between the feature vector and each each of its two nearest neighbours is calculated.
2. The difference calculated in step 1 is then multiplied by a random number between 0 and 1, which creates the new data points.

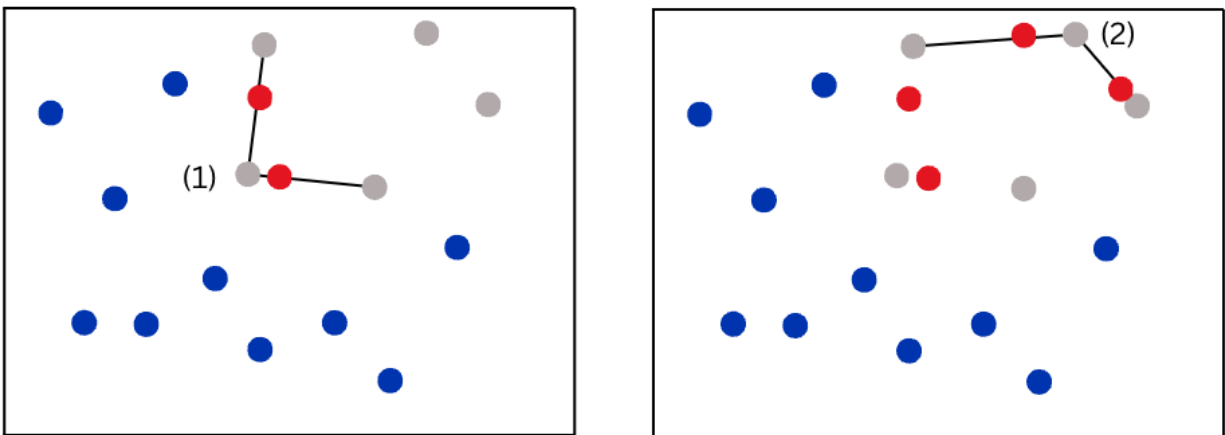


Figure 2.4: Example of creating synthetic data records with SMOTE, using k -nearest neighbours where $k=2$, the gray dots are the records of the minority class, and the red dots are the synthetic records created by SMOTE (adapted from [307]).

While there are manual techniques that can create a balanced dataset, using SMOTE (where the minority class is oversampled) in conjunction with undersampling the majority

²Variants such as “Borderline-SMOTE” [126] as “Safe-Level-SMOTE” [36] have also emerged.

³The original work used 5 nearest neighbours [52] but for the purpose of our illustration, we have only used 2 for ease of explanation.

class has been shown to achieve better classification results [52]. This method has proven quite popular and to work well. SMOTE, however, can only be used where there is a binary classification task with continuous data [1]. SMOTE also runs into problems when the minority class is too small to determine the probability distribution of the real data [1].

2.3 Consideration of State of the Art and Corporate Work

There are two important considerations when examining *State of the Art* (SOTA): the academic/research side and the corporate side. Academia focuses on “open” research. While the data may not always be available due to sharing constraints placed on by the source or by the type of data, or researchers may not want to share their data, the methodology is widely shared in conference papers, journal articles, and pre-prints. Well written methodologies provide sufficient detail to encourage and facilitate reproducibility.

We do not have the same visibility and access to information for corporate entities that are also in the machine learning domain. Since they are *for profit* entities, they do not have the same motivation and return on investment for sharing widely. Beyond giving a basic understanding of what products and services they offer, the rest is *black box*. And so, it is difficult to assess where companies are *truly* at in relation to SOTA.

That being said, as part of this research, we investigated several companies that specifically do machine learning as a service (MLaaS) that specifically mention that they cater to the financial domain:

- We Are Adaptive
- gretel
- Amazon Web Services (AWS)
- Microsoft Azure
- IBM Watson
- Google Cloud AI
- Kensho Technologies Inc

- ChAI
- RavenPack

We found that the main areas are customer behaviour and product/service recommendation, stock prediction, bankruptcy prediction, investment quality, investment recommendation, sentiment analysis, and credit risk. Therefore, it does not appear that, from the information we were able to review, that corporate work is being done on the augmentation of financial statements.

Chapter 3

Augmentation of Tabular Financial Statement Data

3.1 Chapter Introduction

ML is a very powerful tool as it can identify relevant patterns in the data [337] that we, as humans, may not be able to see, particularly if the dataset is large. **ML** has been used to perform tasks such as make recommendations, predictions, forecasts, identify cancerous cells, and analyze the sentiment of textual documents. In the last few decades, companies have incorporated **ML** into their operations, and data science (as a field and profession) has become “mainstream”.¹ However, **ML** is only usable if there is a sufficient amount of data for training, validation², and testing for the research question or task. If using 10-fold cross-validation, for example, a minimum of 10 instances or records is needed for this method to work. However, those 10 instances do not necessarily represent the optimal number of instances needed for the particular task, but may represent the available data. This means that a company may only have 10 years of audited **FS** and so only have 10 data instances (or records) available for use with **ML** (if they choose to simply use the audited statements). If the company wants to use **DL**, then the amount of data needed is considerably more. It has been suggested that in order for supervised **DL** to provide adequate performance³ that 5,000 labeled examples per category” are needed [112].

¹We note that while the terms **ML** and data science are often used synonymously, there is a difference between the two in that “data science brings structure to big data while machine learning focuses on learning from the data itself” [108]. For the purposes of this research as it involves both data structure and learning of the data, we will use the term **ML** to represent both tasks.

²We recognize that some researchers only do training and testing, without the use of validation. In those cases, there would need to be sufficient data for training and testing.

³The Oxford English Dictionary defines *adequate* as “Fully satisfying what is required; quite sufficient, suitable, or acceptable in quality or quantity.” [230]. Therefore, the term *adequate* is ambiguous and difficult to quantify as it can represent a range (and in some cases a very wide range) of performance, depending on the interpretation and audience. This means that what is deemed to be *adequate* performance should be identified and supported by the researcher, or it will be dictated the requirements of the task. For example, detecting cancer with a 0.692 accuracy using a TNM (tumor, node, metastasized [11]) staging system was reasonable performance until results using an artificial neural network returned a 0.730 accuracy, showing a statistically significant increase in the accuracy results [37].

Another important consideration is determining if the model or task can reasonably handle inaccurate data [171]. Researchers and practitioners alike have often been heard saying that *more data is always better*. While this may be true in some cases, it is not an absolute. The bigger the size of the data, the less likely it is that the researcher/user will be able to review all of the data to ensure its veracity. This means that there is a “tradeoff between quantity and quality” [171], with the hopes that any inaccuracies in the data will be overcome by the fact that there is a lot of data [205]. Domains like healthcare and finance, for example, require a high degree of precision which means that using problematic data could have detrimental consequences.⁴ In these cases, quality clearly beats quantity [98]. Therefore, in cases where only small data⁵ is available, researchers and users need to be able to [augment](#) datasets to use [ML](#) or improve their [ML](#) results. A more in-depth discussion Big and Small Data can be found in [Appendix B](#).

We present two projects in this chapter that explore current methods for augmentation and also present a new approach specific to augmenting [FS](#). Our research questions are as follows:

- Project 1: Can we predict the distress of large financial intermediaries in the US using a small dataset? This project uses a binary prediction of distressed or not distressed.
- Project 2: Building on the information learned in Project 1, can we find more sophisticated methods to [augment](#) financial datasets, with specific consideration to [FS](#)?

⁴We recognize that having a dataset that is completely free of inaccuracies is not always possible, but that researchers and users should be reviewing the data that they use and evaluating it against a set of standards. Any data that does not meet those standards should be removed, with supporting explanations as to why the data was removed.

⁵The term “big data” has been popularized in the last ten years but no *one* common definition exists. This term also takes on different meanings depending on the audience and context. “Big” means many things from the number of variables to the number of records (size), to whether it was collected by humans or by machines, to the size of the impact [98]. In a lot of respects, the term “small data” exists purely as a counter to “big data” [131], in an effort to distinguish between the two; in essence, all that is missing is “medium data”, and then the full compendium is represented.

3.2 Project 1: Predicting the Distress of Large US Financial Intermediaries using CNNs

3.2.1 Introduction

“Too big to fail” has become a well-known adage for US financial institutions. Over the past 15 years, the US has faced some very serious economic events — the 2008 financial meltdown, the 2019 Shadow banking⁶ crisis, the financial emergency brought on by COVID-19, and most recently the collapse of Silicon Valley Bank in 2023. These events have spurred renewed interest in predicting both financial distress and bankruptcy.⁷ Large US financial intermediaries⁸ are instrumental to the stability of both the US economy as well as the global economy as the world’s reserve currency is (currently) the US dollar. Therefore, it is very important to be able to predict the distress of large US financial intermediaries.

To recap, our research question for this project was: Can we predict the distress of large financial intermediaries in the US using a small dataset? To answer this question, we created a custom dataset that uses an extended feature set of financial ratios, sentiment analysis scores, the London Interbank Offered Rate (**LIBOR**), and they keywords “**Going Concern**” and “**Concentration Risk**”⁹. This dataset was used to predict financial distress¹⁰ for large US financial intermediaries using a **CNN**. We drew inspiration from contemporary bankruptcy research that indicated that **CNNs** had not been used for bankruptcy prediction [197]. Although this study concluded that **embeddings** returned stronger results [197],

⁶A shadow bank is a bank-like entity but which is not regulated as a bank. Please see Appendix C for more information on shadow banks and the US financial system.

⁷Using a basic Scopus search, over the period of 1969 to 2007, there were 209 documents returned for bankruptcy prediction and 33 documents for financial distress. From 2008 to current (2024), the number of documents returned for bankruptcy prediction jumped by 350% (209 to 943) and 1100% (33 to 396) for financial distress.

⁸Together, banks and shadow banks are known as financial intermediaries. Please see Appendix C for more information on intermediation.

⁹Going Concern and Concentration Risk have specific meanings in accounting. “**Going Concern** is the ability for a company to meet its financial obligations and continue business for the foreseeable future [217]. **Concentration Risk** is a risk that must be disclosed by the company if it poses a severe danger to the company [302]. This type of risk can manifest in many ways, but it most notably manifests itself in concentrations of the sources or types of credit that a bank extends. Please see Section 3.2.3.5 and Appendix C for more information on Going Concern and Concentration Risk.

¹⁰There is no universally agreed or consistent definition for **Financial distress**. Therefore, we define **Financial distress** using two conditions: when a company is not able to meet its financial obligations and when **Going Concern** has been flagged by the auditors. Please see Appendix C for more information on **Financial distress**.

we wanted to do a similar evaluation of the CNN using our custom-built dataset for financial intermediaries and then draw a comparison. To the best of our knowledge, predicting the distress of financial Intermediaries is the first of its kind and was published as part of the IEEE Conference on Business Informatics (CBI) in 2021.

3.2.2 Related Work

In our review of the literature, the majority of work in the area of bankruptcy and distress focuses on predicting bankruptcy by using companies that have been de-listed from stock exchanges. We have focused our research efforts on predicting financial distress rather than bankruptcy for two key reasons. Firstly, there are factors outside of the company’s control that may determine whether or not an intermediary *actually* goes bankrupt. A good example of such a factor is a government bailout, which many banks and shadow banks received in either the 2008 or 2019 crises, or both. And secondly, given the importance of the financial system on economic stability, if we can predict distress far enough in advance, then it provides the intermediary time to strengthen its financial position to avoid going further into distress, or into distress altogether.

CNNs have been used successfully in contemporary research for predicting financial distress. The research of Jan [156], Hosaka [142], and Mai *et al.* [197] spurred on our use of CNNs for this research. The work of Mai *et al.* [197], in particular, was of interest to us given that we wanted to use both textual and financial variables as part of our research. Mai *et al.* found success using sentiment analysis from the annual 10-K in conjunction with the CNN to predict the financial distress, with an accuracy of 71.2% [197].

Financial and mathematical models that have developed over the years tend to rely on assumptions such as independence and normality of variables which are not necessarily reflective of “real-world” finance [59, 147]. Machine Learning (ML), however, offers a great ability to work with unstructured data and the “non-linearity” [59] of the financial system. Due to this capability, many researchers have applied ML to the issue of predicting bankruptcy and financial distress, particularly with deep learning in recent years.

Research done by Jan focused on predicting financial distress of 344 Taiwanese listed and “Over-the-Counter”¹¹ companies over a period of 20 years (2000–2019) [156]. The data

¹¹ “Over-the-Counter” companies are those which sell through broker-dealers rather than listing their shares

consisted of financial and non-financial ratios [156]. The χ^2 automatic interaction detector (CHAID), which uses chi-square test statistics as well as p -value (originally proposed by Kass [163]), was used to select the most important variables in the dataset. The results were then fed into two models — a Deep Neural Network and a Convolutional Neural Network — to predict financial distress. The CHAID-DNN model returned an accuracy of 90.38%, and the CHAID-CNN returned an accuracy of 94.23% [85].

El-Bannany, Sreedharan, and Khder also focused on predicting financial distress using Multi-layer Perceptron (MLP), Long Short-Term Memory, and CNNs. They developed two datasets as part of their research [85]. The first was focused on Kuwaiti companies during the period from 2010 to 2017. They extracted attributes from company financial statements and used that to train their deep learning models. They found that the MLP produced the best results, with an accuracy of 93%. El-Bannany *et al.* also developed a second dataset, this time using companies from across the Gulf Cooperation Council (GCC). Countries included United Arab Emirates, Bahrain, Kuwait, Qatar, Saudi Arabia, and Oman. Again, they extracted financial attributes as their dataset and found that the highest accuracy was produced by the MLP, returning 84% [85].

Hosaka developed a novel approach where financial ratios were derived from the Consolidated Balance Sheets and Profit & Loss Statements of de-listed Japanese companies from the period January 2002 to June 2016. These ratios were then represented as grayscale images and fed into the GoogLeNet CNN [305] for training in order to predict bankruptcy. It achieved an accuracy of 92.00% [142]. Mai *et al.* also researched how to predict bankruptcy using three models: CNN, logistic regression, and Random Forest. Instead of using just financial attributes, they also included textual attributes as well. The dataset comprised information gathered from the annual 10-K filings (required by the US Securities and Exchange Commission — SEC) over the period of 1994–2014. In total, there were 11,827 firms in the dataset. The numerical attributes were gathered from accounting and equities data. Embeddings were created from the Management Discussion & Analysis (MD&A) section of the 10-K for the textual attributes. Initially, the numerical and textual attributes were handled separately by the CNN, but were then brought together by concatenating the final hidden layers of both models. The concatenation was then connected to a softmax layer to make the final prediction. Of the three models, the CNN returned the best results with an

on stock exchanges [48]

accuracy of 71.2% [197].

3.2.3 Methodology

Thankfully, financial intermediaries do not go bankrupt or fall into severe distress on a regular basis. This is a good thing economically, but it also means that data on such occurrences is scarce. Also, to the best of our knowledge, there are (currently) no existing publicly available datasets for predicting the distress of financial intermediaries. Therefore, using data from the 2008 financial crisis in the United States, we created a balanced dataset of financial intermediaries that included both financially distressed and financially strong entities.

3.2.3.1 Initial Selection of Financial Intermediaries

We consulted lists of financial intermediaries that had been bailed out by the US Treasury Department [60, 166]. The bailouts ranged in size from \$301 thousand to \$25 billion across 987 financial intermediaries. To select candidates for the dataset, an initial bailout threshold was set at \$1 billion. This was done for two main reasons: first to focus the research on large intermediaries, and second to help ensure that the intermediaries were registered with the [SEC](#), as under [SEC](#) rules, there are some exceptions where intermediaries do not have to register. As larger intermediaries also tend to be public companies, then registration with the [SEC](#) is required, which then creates filing requirements (for most public companies), one of which is the 10-K annual report.

The 10-K report provide comprehensive details on the company’s operations, financial health, risks [327] and includes the audit opinion as well as the [FS](#) and [Notes](#). One of the main sections in [10-K](#) is the Management Discussions & Analysis ([MD&A](#)) where Management speaks directly to the investor on its past and current performance, as well as its expected future performance. Its primary audience is the shareholder and, due to its content, is arguably the most tonal section of the 10-K. The [10-K](#) reports were then collected from the Electronic Data Gathering, Analysis, and Retrieval system ([EDGAR](#)) system [285] over the period of 2004–2007. We purposely did not collect those from 2008 when the meltdown occurred, as our goal was to predict distress prior to bankruptcy, and not bankruptcy itself. The [MD&A](#) was parsed out so that it could be used for sentiment analysis. Following previous work [308], we used an “extractive” sentiment approach whereby the sentiment was

calculated for the MD&A as written and then again with all of the Non-GAAP measures¹² removed (extracted). This provides important information on the change in sentiment once Non-GAAP measures are removed from the report.

Using a threshold of \$1 billion as a minimum bailout narrowed the list down to 48 candidate intermediaries, which were then cross-referenced with EDGAR, the SEC’s online filing tool [285]. Of the original 48 intermediaries, 3 were not registered and were removed. The list was then stratified based on repayment status for further consideration. Intermediaries that had not repaid were conditionally labelled as financially distressed. The reason behind this decision is that if the intermediary were strong, or even moderately strong, it would have repaid (either fully or partially) their bailout, particularly given the restrictions imposed as part of the bailout — the Troubled Asset Relief Program (TARP).¹³ Conversely, intermediaries that had either repaid a substantial portion of their bailout or repaid the full amount were conditionally labelled as financially strong (as the intermediary would not have been able to repay funds otherwise). 67% was used as the “substantial amount” threshold (based on the premise of a supermajority which is required to make large corporate decisions) rather than the simple majority of 50% as we did not believe that it demonstrated sufficient repayment to be classed as financially strong. This approach yielded 11 distressed and 34 strong intermediaries.

In the final phase of selection, we researched the remaining companies to determine if the condition label as either distressed or strong was correct. This was done by reviewing company announcements, traditional and social media news (directly from the company), and reviewing their financial results from the time that the label was assigned. From our research, we determined that the performance of 6 of the intermediaries had declined. We, therefore, concluded that these entities were not as strong as previously, and were removed. After confirming labels, this left 11 distressed and 28 strong intermediaries in the candidate pool. To avoid bias, intermediaries were randomly selected – 10 from each group (strong, distressed) to form the final dataset.¹⁴ Table 3.1 details the intermediaries included in our

¹²Non-GAAP measures are industry accepted measures that are outside of the accounting rules.

¹³There were a number of restrictions that financial institutions had to adhere to in order (such as restrictions on executive compensation and supervisory capital restrictions), and to receive TARP funds, which were overseen by a special master until the institution had repaid the funds and exited the program [324].

¹⁴An important aspect of the dataset to consider is that, as part of this research, we are not trying to detect, determine, or predict the prevalence of distressed or financially strong entities. If that were the case, then a balanced dataset would not be appropriate as prevalence is the number of intermediaries that fall

dataset.

Strong	Distressed
American Express	AIG
Bank of America	BearStearns
BlackRock	CountryWide
Capital One	Fannie Mae
CitiGroup	HSBC
JPMorganChase	Lehman Brothers
Wells Fargo	Merrill Lynch
Bank of New York	Sun Trust
KeyCorp	Bank of Wachovia
U.S. Bancorp	Washington Mutual

Table 3.1: Company Listing [309]

3.2.3.2 Sentiment Analysis

Sentiment analysis was conducted on the full [AR](#) and the [MD&A](#) in R using four well-known dictionaries: Harvard-IV [250], Loughran-McDonald [192], Henry [67] and the Quantitative Discourse Analysis Package ([QDAP](#)) [266]. We specifically chose these dictionaries over other alternatives such as Bidirectional Encoder Representations from Transformers ([BERT](#)) to capture two different end-user groups: financially sophisticated and the average investor [308]. As the audience for the [MD&A](#) is wide ranging such as investors who are financially savvy to those who are not (this group is termed the “average” investor), regulatory bodies, and credit granting institutions, using the dictionaries was instrumental to capturing the various sentiments that the different groups would likely exhibit to the information in the [MD&A](#). For example, to the average investor, the words “debt” and “taxes” will evoke negative emotions. But, to the financially savvy, these words are expected given that it is not common to find businesses that do not have any debt, and every business has to file tax returns even if there is no tax owing. Therefore, the words “debt” and “taxes” are interpreted very differently by the financially savvy and return a neutral sentiment rather than a negative sentiment.

into distress or go bankrupt relative to all of the financial intermediaries in the US. The use of a balanced dataset in this case would incorrectly bias the model towards financial distress and bankruptcy being more prevalent than it actually is. As our task is predicting if a financial intermediary will become distressed, then using a balanced dataset is appropriate.

Although there were numerous [BERT](#) sentiment models that had been already released at the time of this research, there was no [BERT](#) model that was specifically trained for finance as FinBERT was not released until several years later (in 2022). Therefore, in order to use the same approach for each group (financially savvy and the average investor), the decision was made to use the dictionaries. The dictionaries selected are representative of the two different groups: financially savvy (Loughran-McDonald and Henry, both of which are specifically targeted towards finance) and the average investor (Harvard-IV — a psychological dictionary — and the QDAP which is a collection of text analysis dictionaries).

3.2.3.3 LIBOR Rates

When US financial intermediaries need to borrow short-term funds, they can either do this from the (US) Federal Reserve or they can borrow from other financial institutions. The London Interbank Offered Rate ([LIBOR](#)) is the interest rate at which institutions will lend to each other [143]. Rising [LIBOR](#) rates indicate that there is increased risk in the interbank market, signalling possible financial distress [143]. Two [LIBOR](#) rates were chosen and were collected from the St. Louis Federal Reserve [264]: the averaged three month rate from the company's year end to the filing of the [10-K](#) with the SEC as well as the average over the entire year, for each of the years from 2004–2007.

We are cognizant of the fact that [LIBOR](#) will no longer be used after the end of 2021, as per regulators from both the United States and the United Kingdom, as well as the Intercontinental Exchange (ICE) Benchmark Administration, [LIBOR](#)'s administrator [61]. However, during the period of 2004–2007, [LIBOR](#) was the mainstream rate used for interbank loans, and remained so until 2017, when it was announced that the UK Financial Conduct Authority no longer believed that it could guarantee the stability of [LIBOR](#) beyond 2021. We have addressed this further in the limitations section.

3.2.3.4 Equity Ratio

The [Equity Ratio](#) (also known as the Equity to Assets Ratio) provides critical information on how much of the company is funding its assets. A ratio of less than 0.5 indicates that the company is using more debt than equity to fund its assets, and the lower the ratio, the greater the risk [47]. Therefore, given that we are predicting financial distress, we determined

that this would be an appropriate financial ratio to include in the dataset.

3.2.3.5 Critical Accounting Keywords

Two critical accounting keywords were also included in the dataset — [Going Concern](#) and [Concentration Risk](#). [Going Concern](#) is a “backbone” principle of accounting which assumes that a company’s operations will continue indefinitely [217]. It is evaluated by management and auditors on a regular basis, and is only discussed in the [10-K](#) filing if this status is in doubt. [Concentration Risk](#) also must be disclosed if it poses a severe danger to the company [302]. There are many types of [Concentration Risk](#), but for financial intermediaries, this risk manifests in elements such as the facilitation of credit (given that it is the bank’s main source of income [227]) or in the structure of an investment portfolio where the investment(s) are mainly held in either the same types of industries or companies [35], for example. To contextualize the importance of this risk, we offer the following example: In its 2004 Annual [10-K](#), Bear Stearns stated “*At November 30, 2004, the Company’s most significant concentrations are related to US government and agency inventory positions, including those of the Federal National Mortgage Association and the Federal Home Loan Mortgage Corporation*” [301]. Just over 3 years later, Bear Stearns collapsed.

3.2.3.6 Data Augmentation

As our dataset was very small (63 records), we experimented with augmentation to determine if it would enhance the model’s performance and address potential overfitting. Computer vision has traditionally used augmentation on its image datasets by first duplicating the original data and then applying one or more shifts to it such as rotation, flipping, shading, zooming, and distorting [20]. Using this as inspiration, we duplicated the dataset to ensure

the original data was not altered and then scaled it using typical sensitivity analysis¹⁵ percentages used in the financial industry of $\pm 10\%$.¹⁶ This practice follows an industry-accepted approach to forecasting where it is common to use $\pm 10\%$ [317].

Doing this, we generate a dataset of $2\times$ and $3\times$ the size, depending on if we are using both scaling shifts, or just one. Therefore, our original dataset of 63 records was augmented to 126 records with the addition of one shift (either $\pm 10\%$ of the original data), and 189 records with the addition of two shifts (both $\pm 10\%$ of the original data). We note that we only augmented the training set (not the testing set) in an effort to avoid evaluation bias. Applying sensitivity consistent with accepted practices of the financial domain, we believe that using $\pm 10\%$ to **augment** the dataset is realistic and does not distort or undermine the validity of the augmented data.

One-hot encoding was used for the keywords **Going Concern** and **Concentration Risk** where 1 indicated that the keywords were in the company’s **SEC** filing, 0 otherwise. Augmenting these encodings would not have made any sense as it would rendered the encodings useless¹⁷. Therefore, the original encodings (either 0 or 1) were used for the augmented data as well.

3.2.3.7 Dataset Summary

Table 3.2 provides an overview of the features included in the dataset.

3.2.3.8 Scaling

Often, machine learning models benefit from “standardization of the dataset” [248]. This is because many models are sensitive to the scale of the features or the features in the dataset use different scales [352]. This task is usually referred to as “standardization” or

¹⁵A sensitivity analysis — a “what if?” analysis — is commonly used in the financial domain to examine and evaluate fluctuations in the inputs. Typically, three main scenarios are run: the expected situation (using the original data) as well as the best and worst case scenarios, each of which use sensitivity data where the data has been developed using reasonable and informed assumptions. This translates into either specific amounts applied to certain items of the analysis, or, more commonly, a percentage applied to the analysis (either in part or in full). A good example of this is a financial forecast. The forecast is developed (which is the “expected” scenario), and then sensitivity is applied to look at what will happen to the overall results if expenses increase/decrease by $x\%$, sales increase/decrease by $y\%$, or foreign currency fluctuates by $z\%$, to give a few examples.

¹⁶5% and 8% are also common

¹⁷Using $\pm 10\%$ on the encodings would have returned encodings of 0.90 and 1.10.

Feature	Feature Count
Word count for complete SEC filing	1
Word count for MD&A section only	1
Harvard-IV General Inquirer dictionary sentiment scores (sentiment total, negativity, positivity) for the complete SEC filing	3
Harvard-IV General Inquirer dictionary sentiment scores (sentiment total, negativity, positivity) for the MD&A only	3
QDAP dictionary sentiment scores (sentiment total, negativity, positivity) for the complete SEC filing	3
QDAP dictionary sentiment scores (sentiment total, negativity, positivity) for the MD&A only	3
Henry dictionary sentiment scores (sentiment total, negativity, positivity) for the complete SEC filing	3
Henry sentiment scores (sentiment total, negativity, positivity) for the MD&A only	3
Loughran-McDonald dictionary sentiment scores (sentiment total, negativity, positivity, uncertainty) for the complete SEC filing	4
Loughran-McDonald dictionary sentiment scores (sentiment total, negativity, positivity, uncertainty) for the MD&A only	4
LIBOR 3-month average rate	1
LIBOR yearly average rate	1
Amount of Equity (in \$)	1
Amount of Assets (in \$)	1
Equity to Assets Ratio (calculated as Equity/Assets)	1
Keywords of Going Concern and Concentration Risk	2
Total	35

Table 3.2: Dataset Summary Information.

“normalization”. Scaling is one of the ways that datasets can be normalized, and there are several approaches to scaling offered in scikit-learn: for example, the Maximum Absolute Value Scaler (MaxAbsScaler), the Minimum-Maximum Scaler (MinMaxScaler) which scales data to a given range, and the Standard Scaler (StandardScaler) [248]. There are pros and cons to each approach and the decision should be made based on the data and the research question/task.

In looking at the financial machine learning literature, it appears that researchers predominantly use either the Standard Scaler or the Robust Scaler. The major drawback to using the Robust Scaler is that, along with removing the median, it uses quartiles to scale the data [248]. Using quartiles, however, would not have been appropriate for our dataset for two main reasons: first, the dataset is small and so using quartiles can “lead to less accurate representations of the true dataset” [99]; and secondly, outliers and extreme values can have a significant impact the quartiles and may skew the normalization [99]. Like the Robust Scaler, the Standard Scaler is also sensitive to outliers, but rather than using quartiles for the scaling, it uses the unit variance [248].¹⁸ The equation for the Standard Scaler is provided in equation 3.1.

$$z = \frac{x - \bar{x}}{s} \quad (3.1)$$

where \bar{x} is the mean of the training samples, s is the standard deviation of the training samples, and x is the original value of the feature [248].

3.2.3.9 Proposed Method

The majority of the current research in the literature applies CNNs to image data as there is an assumption that pixels next to and close to each other are related in image data — if a tabular dataset is arbitrarily constructed, this association may not be true [290]. The close association of the pixels, known as “spatial induction bias” is one of the factors that make CNNs so successful as this bias can be used to make inferences of the structure of the data [215] and thereby capture relationships and trends that we, as humans, cannot see. Therefore, it is important to ensure that the ordering of tabular data is carefully curated to provide association.

¹⁸A unit variance is where the values are divided by the standard deviation.

The model’s task was to provide a binary prediction for each financial intermediary of either distressed (1) or not distressed (0). The CNN was developed using Keras. Each layer — the convolution, pooling and dense (also called “fully-connected”) — has hyperparameters that can be tuned. The main hyperparameters that we focused on were the number of layers and filters, the activation function, and the pooling layers. In the convolutional layers, filters are applied, the number of kernels is chosen, and the activation function is assigned [299]. For our model, we experimented with using 3 and 4 convolutional layers with 32 filters in the first two layers (and 64 filters in the last two), all with kernel size 2, using the Rectified Linear Unit (ReLU) activation function as it is most widely used in the existing body of research. The pooling layers exist to reduce dimensionality, with *max pooling* or *average pooling* being two popular approaches, as it “pick[s] the maximum [or average] value after each pooling operation” [299]. One of the main purposes of pooling is to reduce the number of features that the model needs to learn. This is particularly useful if there is a large number of features in the dataset. Since our dataset only had 35 features, we did not think it was appropriate, or necessary, to reduce the number of features using the pooling layer. At this point, the output from the convolutional (and pooling layers, if used) will be flattened into order to create the flat structure required in order to be fed into the dense layer [299].

The last layer is the dense, or fully-connected, layer. Here, the final classification is done [299] and the activation function is usually changed to meet the requirements of the output. In our model, we are classifying the financial intermediaries as either distressed or not distressed, and so our activation function changes to sigmoid, given our binary classification task. Figure 1 provides a general approach to our CNN model for our research.

We used the standard train-test split of 80/20 with the Adam optimizer over 60 epochs. A common evaluation metric that is used in CNNs for financial data is accuracy. Therefore, in order to align with existing research, we also used accuracy.

Naïve Bayes (NB) is often used as a baseline to compare model results against. It is also a very powerful model on its own as well. The NB classifier is based on Bayes’ Theorem:

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Using this, we can determine the probability of A given B, on the assumption that each variable is independent of all other variables in determining the outcome of the dependent

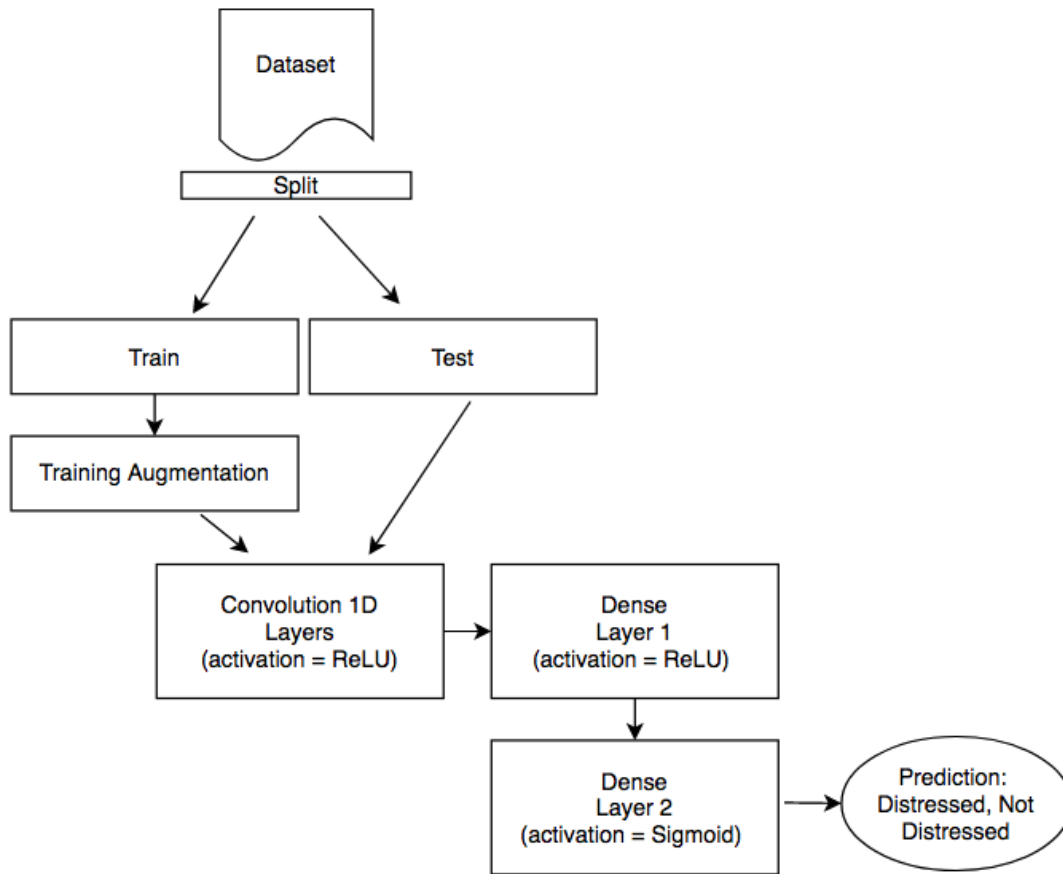


Figure 3.1: Convolutional Neural Network for this research. ReLU is rectified linear units.

variable [336]. When there are dependencies between the variables, however, NB does not perform as well, as the assumption of independence no longer *truly* holds [200].

However, only a few of the variables that we have included in the dataset are *truly* independent: the evaluation of the [Going Concern](#) assumption as this is on a per-company basis, and is conducted by management as well as the auditors in regards to the company's specific financial condition; and the concentration risk as this variable is also company specific. The remainder of the variables are interlinked and cannot be said to be independent. The sentiment analysis conducted on the [MD&A](#) is influenced by the company's internal managerial and financial position, but also by those of the industry through competition, responding to changes in industry and regulatory environments, as well as globally influenced factors such as disruptions in international trade, pandemics, and the like. To the best of our knowledge, the degree of interdependency has never been calculated at an industry level. In the context of our research, we would be able to calculate the degree of interdependency between financial intermediaries if we had access to internal company data. Without that information one *could* hypothesize on the degree of interdependency, but it would be extremely difficult to calculate it with any sense of accuracy, given that there are many different types of interdependency like financial (such as loans) and supply (for materials and human capital). Similarly, [LIBOR](#) is also not independent given that it is benchmark that responds to the lending activity of major global banks. When [LIBOR](#) rates go up, banks are charging other banks more for loans because there is an increase in the risk of defaulting on that lending [45]. Therefore, the rates are dependent on the economic situation that lending institutions are collectively facing and responding to.

Random Forest (RF) works in a very similar fashion to a decision tree. When you come to a decision point, you take the best branch until you reach the next decision point. This is done repeatedly until you have no more decisions to make. But, to choose the path at the decision point, RF uses the best between randomly chosen paths [184], which is done to help prevent overfitting [28].

Support Vector Machines (SVM) work to find the largest separation between the classes, which is known as the maximum margin hyperplane [336]. This is accomplished by “transforming the input using a nonlinear mapping” [336]. The instances with smallest distance to the maximum margin hyperplane are the support vectors which uniquely define the hyperplane [336]. There are a number of kernel functions that can be used for SVM. As part

of our research, we experimented used a Linear kernel ($k(x, y) = x^T y + c$) and the Gaussian Radial Basis function ($k(x, y) = \exp\left[-\frac{\|x-y\|^2}{2\sigma^2}\right]$) [298]. SVM typically works well for binary classification, which we have. Also, given that 50% of our dataset is atypical financial cases (which in a sense means that half of the dataset comprises outliers), it is the support vectors that define the hyperplane [336], so it is less sensitive to outliers.

3.2.4 Experiments

We conducted several experiments (outlined below) to provide a baseline and to evaluate the different augmentation shifts we proposed:

- No shifts to provide a baseline for accuracy
- One shift of +10%
- One shift of -10%
- Both shifts, applying $\pm 10\%$

3.2.5 Results and Discussion

Before evaluating the learning curve graphs, it is important to first establish what the training and validation curves represent. The training learning curve uses the training data and evaluates if, and how well, a model is learning [32]. The validation curve is evaluated on the validation set only and evaluates how well the model generalizes. By comparing the plots of both lines, we can see how the model is learning and generalizing, and diagnose symptoms of overfitting, underfitting, or if there is a good fit [32]. If a model is underfitting, it means that the model is struggling to learn (or cannot learn) the training data [32, 112]. Conversely, if the model is overfitting, it means that it has learned the training data too well and may have “memorized” the data [112]. There is evidence in the research of some models that are overfitting are still able to generalize well [17]. However, there is always a danger of a model that is overfitting will not be able to generalize well on new data it has not seen before [335]. A good fit is found if the gap between both the training and validation set is small and both the training and validation losses reach points of stability [32].

Initially, we used the original, unaugmented dataset with our CNN, which returned an accuracy of 82%, using 3 convolutional layers. Given the small size of the dataset, the model performed relatively well, as the gap between the training and validation loss is small in Figure 3.2, particularly at the end.

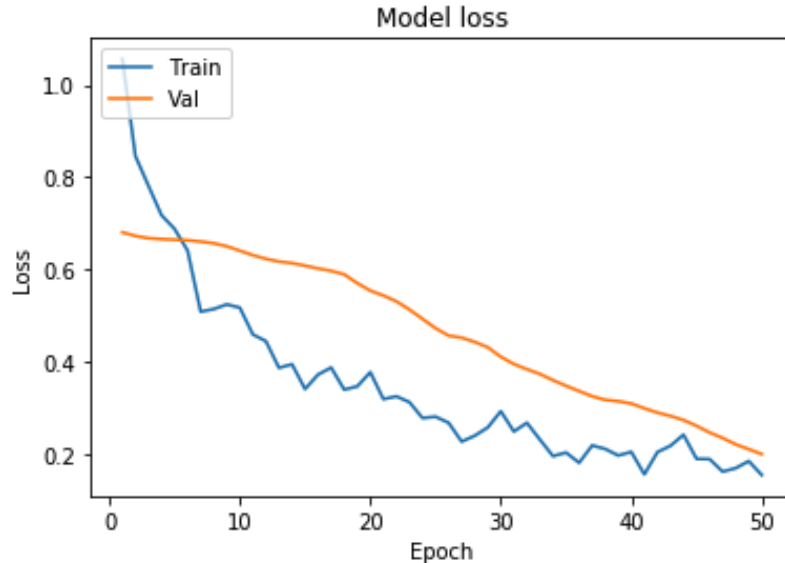


Figure 3.2: Model loss for the unaugmented dataset

By adding a 4th convolutional layer, we found that it slightly improved the model, returning an accuracy of 82.14%. We also see that the gap between the training and validation curves is smaller and the two intersect appear to be close to a point of stability. We then used the augmented dataset with 1 augmentation (+10%) with three convolutional layers. We found that the accuracy improved to 82.35%. In evaluating the loss, the model showed a good ability to generalize. With 4 convolutional layers, the accuracy improved again slightly to 83.31%, and still retained a good ability to generalize.

Using the dataset with 2 augmentations (both + and -10%), which we denote as the “fully augmented” dataset, with three convolutional layers, the model returned an accuracy of 88.00%. And then similar to our previous runs, adding the 4th convolutional layer improved the model slightly to 88.24%. The model loss shows a very small amount of overfitting at the end, as can be seen in Figure 3.3, but overall appears to be generalizing well.

In Table 3.3, the results of our CNN models have been summarized against those of Jan (CHAID-CNN) [85], Hosaka (Imaged Ratios) [305] and Mai *et al.* (Combined CNN) [197].

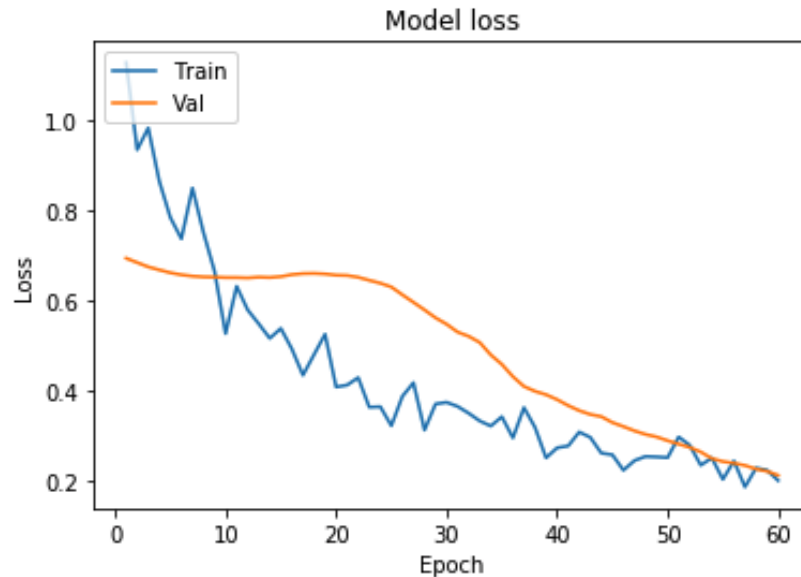


Figure 3.3: Model loss for the fully augmented dataset

Model	Accuracy
CNN (orig, 3 C layers)	82.00%
CNN (orig, 4 C layers)	82.14%
CNN (1 aug, 3 C layers)	82.35%
CNN (1 aug, 4 C layers)	83.31%
CNN (2 aug, 3 C layers)	88.13%
CNN (2 aug, 4 C layers)	88.24%
CHAID-CNN	94.23%
Imaged Ratio CNN	92.00%
Combined CNN	71.20%

Table 3.3: Comparative Convolutional Neural Network (CNN) Results. Bold font denotes the best results that our CNN achieved, showing its competitiveness.

As can be seen, our results are slightly lower than Jan and Hosaka. However, the size of the datasets used by Jan and Hosaka were much larger: Jan used data from 344 companies over 20 years [85] and Hosaka’s dataset comprised 7,520 images [305]. Yet, when compared to Mai *et al.*, our results are higher, even though they used a much larger dataset comprising 11,827 companies over an 11 year period [197]. Even though our dataset is small (consisting of 20 companies over 4 years), it would appear in comparison that our model performed well.

The results for Naïve Bayes were not as strong. Due to the outliers in the dataset, we encountered issues with zero probability. To address this, we used Laplace Smoothing which adds “one to each count” [200] in order to eliminate the zeros. NB usually works well on most datasets (including small ones), so we were surprised to find that the accuracy returned from the model was 76.46%. We believe that it returned less than optimal results given that fact that the bulk of the variables are not independent.

Comparatively, the accuracy of RF was 90.24%, which is a slightly higher accuracy than was returned by our CNN model (88.24%). Our dataset uses both categorical and numerical data, and RF is known to work well with this combination. As well, RF’s bootstrapping technique of selecting samples and replacing them (which then makes them available to be selected again in the future), makes the *Forest* very hardy. This robustness improves the classification, which is why we were not surprised at the high accuracy returned.

Finally, SVM using the Linear kernel returned an accuracy of 65.74% and the Gaussian Radial Basis kernel returned an accuracy of 73.66%. We initially chose two different kernels for our experiments as we were concerned that the data may not be as linearly separable as we initially thought, which turned out to be correct. The RBF kernel performed much stronger than the Linear kernel, which is inline with the work of Wu, Wang, and Shi in their financial distress research [339]. In both models, the number of support vectors was very high (≥ 50). Given the small size of the dataset, this suggested that the initial SVM models were overfitting quite a bit. To address this, we increased the c value to increase the penalty of misclassifications, which reduced the number of support vectors to 35 or 40, depending on which kernel was used, but still indicated that the model was overfitting. Even though we were using 10-fold cross validation, it is possible that given the highly unusual financial data being used, the model was not able to generalize well enough.

3.2.5.1 Post Hoc Experimentation

To address concerns that using $\pm 10\%$ is too simplistic and could be seen as too regular and simple to be generally useful, which could impact algorithmic learning (i.e., create the same effect as oversampling by essentially repeating some records multiple times), we introduced random perturbations over three ranges: $\pm 30\%$, $\pm 20\%$, and $\pm 10\%$, shown in Figure 3.4. That means that perturbations were applied to all features randomly across the specific range. For example, adding 10% to values (3,1,10) gives values of (3.3,4.4,1.1), but if we added randomized values into the range of 10%, we may get (3.15, 4.2, 10.71). See Figure 3.4 for the ranges. Random perturbations were applied across the dataset with the exception of the keywords [Going Concern](#) and [Concentration Risk](#). These two keywords were the exception because they needed to keep their original one-hot encoding values which would have been affected if we had applied the random perturbations. We also note that although we have included $\pm 10\%$, the key difference here is that previously, we had used a fixed $\pm 10\%$, whereas in these experiments, the value is randomly chosen within that range.

The same CNN model used for the original research was also used for the posthoc experiments. We took this approach for continuity and to also ensure that the results were comparable between experiment sets. In the original research, we experimented with three and four convolutional layers. For efficiency of the posthoc tests, we only experimented with three layers, given that the increase in accuracy in the original experiments only marginally increased when we used four layers. Therefore, using only three layers provided the main result with fine-tuning being provided by the fourth layer. Also, in line with the original experiments, we implemented early stopping. Therefore, some of the graphs show more epochs, while others show fewer.

For each range, we ran three separate random perturbations to determine what the effects would be on the validation loss. This resulted in nine individual perturbed datasets which were then passed to the CNN. As can be seen in Figures 3.5, 3.6, and 3.7, the results were mixed. Some of the random perturbations show a better ability to generalize when compared to the original research presented in Section 3.2.5 while others do not. Figures 3.5 (a) and (b) almost appear to be the same graph. We did overlay the two and there appears to be some very miniscule differences, suggesting that the randomization was not substantially different between the two. These graphs also show very large gaps between the training and

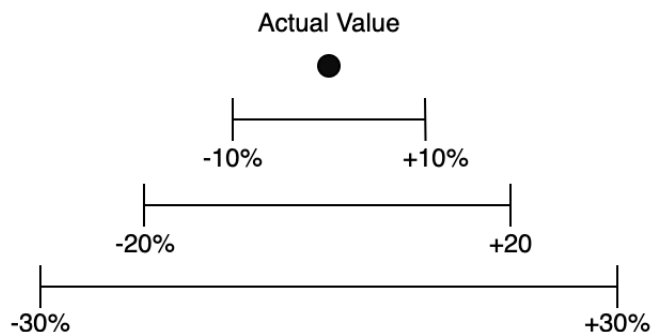


Figure 3.4: Range of random perturbations

validation loss, which are classic signs of overfitting. We also note that the curves of both graphs end with an upward tick, suggesting that if more epochs had been run, the gap would have continued to widen. The conclusion that we draw from these two random perturbations is that the model was not able to generalize well in these cases.

Using two evaluation criteria raised in Section 3.2.5 which indicates that if the gap between the training and validation set is small and both the training and validation losses reach points of stability, Figures 3.7(a) and (b) meet the criteria. Referring to Figure 3.7(a) the initial gap is wide but starts coming together around epoch 25, and converging right before epoch 40. Figure 3.7(b) follows mostly the same trajectory, but the extra bump ups in the learning curves for the validation set before it reaches a point of stability, converging with the training curve. These small jumps in the learning curve could be due to noise in the data, but could also show that there is some variance in the model, as well as some overfitting.

The remaining graphs (Figure 3.5(c); Figures 3.6 (a), (b), and (c); as well as Figure 3.7(c)) show promising results as, for the most part, the gaps between training and validation are small, but do not reach a point of stability, indicating that the generalization may be limited.

3.2.6 Conclusion, Future Work, and Limitations

The main purpose of this project was to apply a CNN to a small dataset of financial intermediaries to determine if this method could be used to predict financial distress. To that end, we used data from 20 financial intermediaries over four years (2004-2007), incorporating both financial and non-financial attributes, LIBOR, and two specific key words (Going

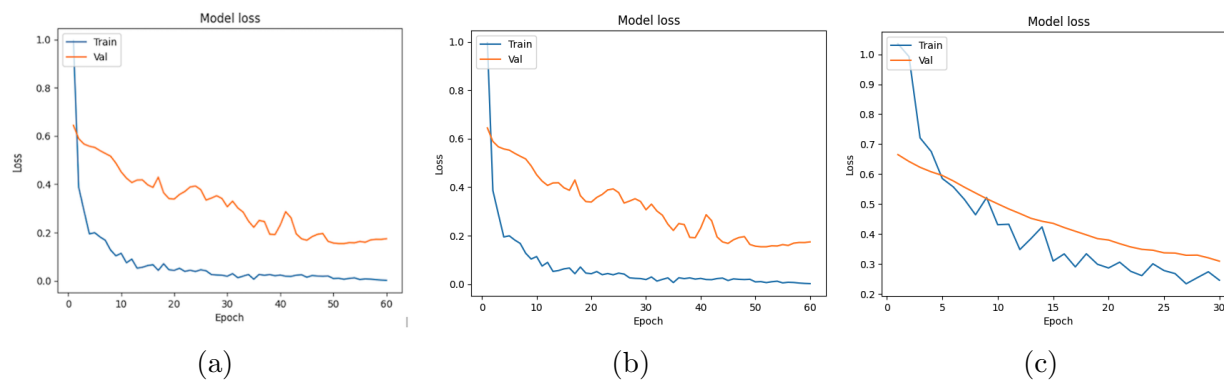


Figure 3.5: Model losses for three different random perturbations ((a) to (c)) across the range of $\pm 10\%$

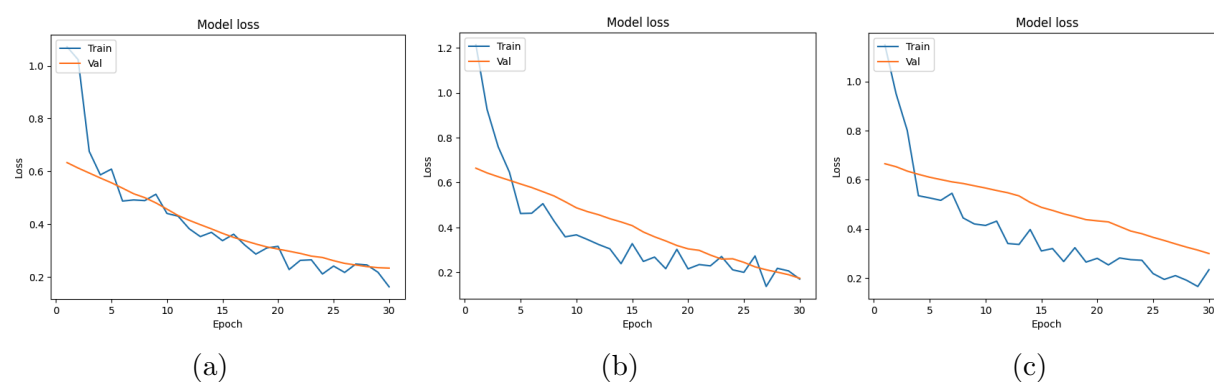


Figure 3.6: Model losses for three different random perturbations ((a) to (c)) across the range of $\pm 20\%$

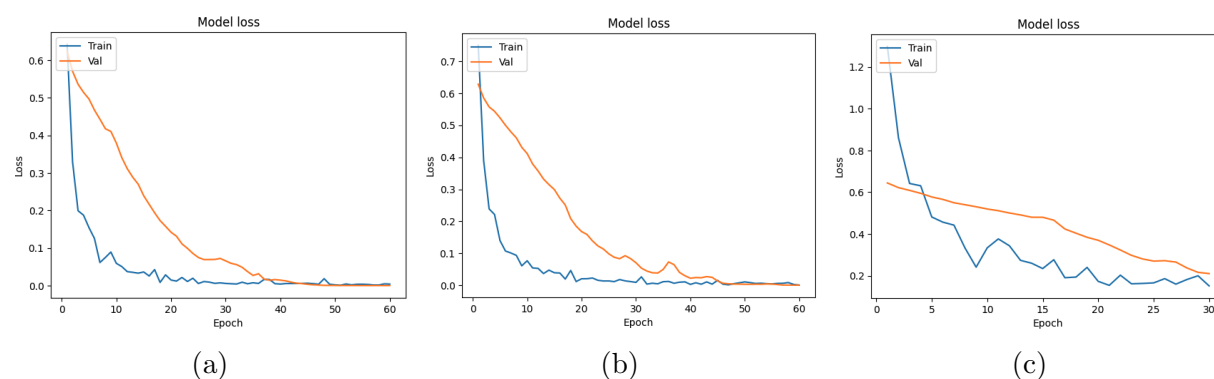


Figure 3.7: Model losses for three different random perturbations ((a) to (c)) across the range of $\pm 30\%$

Concern and **Concentration Risk**). While our dataset was small, our results are promising when compared against those produced by much larger datasets.

Although the financial crises over the past 15 years are not the first to rock the financial

system, the most recent ones have shown that when key intermediaries in the US financial system collapse, it can jeopardize the economy. Therefore, the ability to predict when intermediaries start heading towards financial distress is important. Once a large intermediary declares bankruptcy or collapses, little can be done to save it without government intervention. Yet, adequate warning of impending distress can facilitate company turnarounds, and enable them to strengthen financial positions to either avoid worsening current distress, or preventing them from entering distress situations entirely.

One of the most challenging aspects of financial research is availability of datasets to be used as benchmarks. For future work, we plan on creating a publicly available financial distress and bankruptcy dataset that can be used by researchers to facilitate more realistic comparability of results and make meaningful progress towards predicting entity distress earlier.

The foremost limitation of our work was the size of the dataset. We attempted to address this by using an augmentation technique inspired by computer vision. We also recognize that the companies that were included in the dataset are not fully representative of all of the different types of financial intermediaries that exist. For example, asset-backed commercial paper is both a money-market security and a corporate financing vehicle. In the latter situation, however, the asset-backed commercial paper is incorporated into the company's financial statements, making it difficult to analyze it separately as a "shadow banking" product.

As well, there are many different interpretations of what financial distress is. In some cases, financial distress is used interchangeably with bankruptcy, even though both are quite different (from a legal and accounting perspective). Therefore, the current body of research in this area may not be easily separable between financial distress and bankruptcy.

Finally, as we mentioned in our discussion of LIBOR (Section 3.2.3.3), the use of LIBOR is ending due to concerns on guaranteeing its stability beyond 2021. In our research, we addressed why we believe its inclusion in our dataset was appropriate. However, on a go-forward basis, this metric will no longer be appropriate. In its place, the US Federal Reserve's Alternative Reference Rates Committee has chosen the Secured Overnight Financing Rate (SOFR) as the US replacement for LIBOR. As such, on a go-forward basis, we recommend that researchers use the SOFR rate for research done with international data starting in 2022.

3.3 Project 2: Tabular Augmentation Strategies

3.3.1 Introduction

The term Big Data first appeared in the late 1990s¹⁹ and has been popularized ever since. One of the main issues with this term is that no common definition exists and this term takes on different meanings depending on the audience and the context. Big means many things from the number of variables, to the number of records (size), to whether it was collected by humans or machines, to the size of the impact [98]. In a lot of respects, the term Small Data has evolved and exists purely as a counter to Big Data [131], in an effort to distinguish between the two; all that is missing is Medium Data, and then the full compendium is represented. Without that common definition, however, it is clear that these terms are “relative rather than absolute” [343] and that context is needed to understand how these terms are being used.

Yet, an important distinguishing factor between Big and Small is often characterized by the question(s) that are being investigated and examined [171] — something that is not usually used to characterize Big Data. More and more, researchers are turning to Small Data as it can be collected under controlled conditions, provenance is more than likely known²⁰, and where quality is required (i.e., quality over quantity) [98]. The implications of this turn towards Small Data are important in terms of dataset size: if the research question/task is for a specific entity (whether it be a patient when $N = me$ [92] or it is for particular company), the dataset could be quite small by design or by the availability of data (e.g., the company has only been in business for 5 years); “Not all data is *big* data” [83].

However, every algorithm and model will have a lower-bound threshold that will either explicitly (through an error or warning) or implicitly (through poor results) dictate the least amount of records needed. The explicit scenario is easier to address in that the program will not run or will run with warnings due to the size of the data. For example, Random Forest using 10-fold cross validation needs at least 10 records, with one record belonging to each fold.²¹ The implicit scenario is more challenging to solve in that is intractable [31]:

¹⁹The term Big Data is often credited to John Mashey [203].

²⁰Provenance is very important in determining if a dataset is applicable to the research question or task, as well as in determining the veracity of data. If a researcher does not know what comprises their dataset, this can invalidate the models used and, ultimately, the results [310].

²¹Trying to use less than 10 records using 10-fold cross validation will result in an error that indicates that

a researcher may know that for their task, they are looking to achieve an accuracy of over 85%, but the amount and quality of the data needed to achieve that is hard to (and if not impossible) to determine ahead of time. But, if the amount of initial data available is small, there are limited options to address this problem: either collect more data if that is possible, and if not, create synthetic data to [augment](#) the available data.

Understanding that the methods used to [augment](#) the dataset in the previous project (see Section 3.2) was rudimentary (and was purposely designed as such), the project that we present here explores more sophisticated methods of augmenting the dataset. As discussed in Section 1.2.1, there is the added challenge of ensuring that the [FS](#) still balance after augmentation. We find, that in exploring the current methods for augmenting tabular data, we cannot simply give the dataset to the augmentation model. This means that synthetic [FS](#) data must be created using specific features, which can then be used “as is” for downstream tasks (i.e., using features only) or can be re-assembled into [FS](#) after the synthetic data has been created.

3.3.1.1 Related Work

One of the first tabular GAN models was Tabular GAN (also known as TGAN) that uses a Long Short-Term Memory with attention to generate data on a column-wise basis along with a multilayer perceptron for the discriminator [341]. Although TGAN was developed as a “general purpose” tabular generation model (the data that the research used for evaluation included census data, network intrusion detection data — the KDD99 dataset — and forest cover data) [341], it has been used in computational financial research for stock prediction [342] and automobile insurance fraud detection [122], for example. Conditional Tabular GAN (better known as CTGAN), which came shortly after TGAN, addressed challenges such as non-Gaussian distributions, the need to model both discrete and continuous data, as well as class imbalance in datasets [340]. In the computational financial literature, CTGAN is a very prevalent method to use for augmentation and has been used to determine optimal asset allocation [260], fraud [10, 245], and has also been used to generate synthetic data [41, 176, 331, 249, 140]. However, to the best of our knowledge, no work has been done on generating synthetic data for financial statements.

the number of folds cannot exceed the number of records.

TableGAN, released just before TGAN, is less well-known in the tabular generation research. The purpose of TableGAN is to generate data that is statistically similar to the data found in the original table while preserving privacy [242]. TableGAN uses a deep convolutional GAN (DCGAN) for the generator and discriminator, but also employs a classifier to help predict the labels of the synthetic data [242]. The classifier was added to improve the quality of the generated synthetic data, arguing that it can help prevent generating invalid data: “For instance, a record with gender = “Male” and disease = “Uterine Cancer” can be prevented” [242]. In external evaluation of TableGAN, however, researchers found that while it performed well with continuous variables but struggled when generating data for categorical values [29].

The most prevalent method that has been used to [augment](#) tabular datasets in the computational financial domain is SMOTE, having been used for financial distress [304], bankruptcy [295], fraud [347, 208, 350], financial risk [146], and stock splits [182]. Given that real-world data has a class imbalance more often than not, SMOTE has proven to be an effective algorithm to balance out the classes, while generating plausible data using its nearest neighbour approach. While SMOTE has been used to generate synthetic data for tasks that involve [FS](#) [347, 350, 322], those datasets comprise ratios based on the [FS](#), rather than the [FS](#) themselves. Our work appears to be the first use of SMOTE to generate synthetic feature data that can be re-assembled into balancing [FS](#).

3.3.2 Methodology

3.3.2.1 Dataset

Financial statements for National Commercial US Banks²² were collected from the Bloomberg Terminal [195]. For inclusion in the dataset, we used the following criteria:

- The bank’s SIC code had to be 6021;
- Registration with the [SEC](#) had to be current (as of July 8, 2022 when the data was collected) and not suspended;

²²The [SEC](#) labels each public company’s filing with a Standard Industrial Classification (SIC) code. The Office of Finance uses SIC codes ranging from 6021 (National Commercial Banks) to 6411 (Insurance Agents, Brokers, & Service). For this research, we focused on National Commercial Banks (SIC code 6021).

- The bank had to be a public company traded on a stock exchange;
- The bank had to have a 10-K filing requirement with the SEC that was not suspended; and²³
- For registrations identified as “current”, if the bank had not filed a 10-K with the SEC in over 2 years, the bank was not included in the dataset²⁴

As of July 8, 2022, there were 1,007 national commercial banks with SIC code 6021. We then reviewed all of the listed banks and found that only 147 met the criteria for inclusion in the dataset. As can be seen in Table 3.4, the minimum number of records (audited financial statements) collected from Bloomberg was 2 (company FSBC) and the maximum was 36 (company WAFD), with an average of ~ 21 years (rounded down from 21.81 as statements are either complete or not available). Therefore, more data was available for some companies over others. A complete table of all companies and the number of records collected is available in Appendix D.

Table 3.4: Tabular Dataset Descriptive Statistics

Total number of companies	147
Minimum	2
Maximum	36
Mean	21.81
Total records collected	3,207

Generating synthetic data for the FS is a non-trivial task given the relationships that must be maintained between the statements. Therefore, we have limited our work here to two statements — the BS and the IS — in order to investigate and demonstrate proof of concept, rather than the full set of four FS.

²³Under SEC rules, companies may have their requirement to file suspended. Over-the-counter (OTC) securities are traded through broker-dealers, rather than on the stock exchange. While OTCs may have regulatory requirements under the SEC, there are no requirements to file annual reports or 10-Ks [221].

²⁴We set the 2-year requirement as sometimes can be late in fulfilling their filing requirements. However, according to Oyeniyi *et al.*, late filings signal that there are some fundamental problems with the company beyond just the filing aspect of a Going Concern [239]. Therefore, to ensure the veracity of the dataset, we decided to not include filings that were late beyond two years.

3.3.2.2 Data Labelling

To label the dataset instances, we have used the cash ratio (also known as the “absolute liquidity” ratio).²⁵ This ratio, seen in Equation 3.2 only considers a company’s most liquid assets — its cash and marketable securities, to service its short term liabilities [272, 65]. In the United States, assets are listed in order of liquidity (i.e., how easily the asset can be converted to cash). Cash is the most liquid asset a company has, as it is already in the form of cash. The next most liquid asset is marketable securities as they can be sold (relatively quickly) on the securities markets. Therefore, these two items, providing that the company has these assets, are listed right at the top of the Assets section of the BS. Current Liabilities includes all of the liabilities (also on the BS) a company has that are expected to be paid/discharged within one year (see Figure 1.1).

$$\text{Cash Ratio} = \frac{(\text{Cash} + \text{Marketable Securities})^{26}}{\text{Current Liabilities}} \quad (3.2)$$

There is no exact interpretation of the cash ratio, but, generally, a ratio of at least 0.50 to 1.00 is preferred [65]. At 0.50 that means that for every dollar of liabilities, the bank has \$0.50 in cash; at 1.00, it means that for every dollar of liabilities, the bank has \$1.00 in cash. Below 0.50, the bank is considered a liquidity risk [65]. The labelling approach that we used in this research reflected the following ternary assessment:

- label 1: cash ratio < 0.17 ;
- label 2: cash ratio of ≥ 0.17 , but < 0.50 ; and
- label 3: cash ratio of ≥ 0.50

²⁵Typically, there are three liquidity ratios that are used to evaluate financial statements: the current ratio, the quick (or acid-test) ratio, and the cash ratio. Of the three, the cash ratio that we have used to label the data is the most stringent [65]. We believe that, of the three liquidity ratios, this is the most appropriate for our research, given the importance of financial institutions to the economy and the financial well-being of any country. As well, the most recent large US bank to collapse, Silicon Valley Bank, experienced a severe liquidity crisis [353], and forced the US government to intervene to save account holders. This further supports our use of the cash ratio for labelling the data.

²⁶Marketable Securities are included as they are considered “cash equivalents” as they are quickly convertible to cash [214]

3.3.3 Algorithms, Models, Experiments, Results, and Discussion

As we are evaluating a number of methods, we have grouped the discussion of the algorithms, models, experiments, results, and the discussion of the results into one section. Our aim in taking this approach is that it will provide more continuity of the discussion as we will discuss each algorithm, the experiments and results for each model before moving to the next algorithm. Once this is complete, we provide a summary overview and then discuss evaluation using the downstream task (classification) in the next subsection.

3.3.3.1 Parameters for Synthetic Data Validity

Fréchet Inception Distance (FID) is currently the most common metric used to evaluate generated synthetic images. This metric is discussed in Section 2.2.2. However, currently, there is no universally (or even largely) accepted method or metric to determine if generated synthetic tabular data is valid or not. Researchers have used methods such as Kullback-Leibler divergence, pairwise correlation difference, log-cluster, support coverage [110], likelihood fitness [340], Jensen-Shannon divergence, Wasserstein distance [351], and fidelity, utility, synthesis, and privacy [278]. Recently, a “universal” evaluation metric was proposed — “TabSynDex” — a single score based on the average of statistical properties, correlation, point-wise mean squared error, and support coverage [56]. The single score approach makes this metric easily comparable between datasets and models. However, the support coverage metric that has been used does not handle small samples well: if the number of synthetic data records is small or the number of samples per label is small, then the overall score can be greatly affected and return problematic results [56]. For our research, which can have very small data samples (either real or synthetic), this is a serious problem. Also, there is no evidence that “TabSynDex” was applied to financial datasets in the research [56]. Work has also been done on a “Resemblance” metric to determine how well the synthetic data resembles the original data. The main focus of “Resemblance” is privacy in healthcare. Therefore, there is no linked data in the same way that financial data is linked [134].

Again, we found ourselves in a very familiar position — research is moving ahead in other domains — in this case to develop standardized metrics for the evaluation of synthetic tabular data — just not in the computational financial domain. We draw attention to this as a gap in the research, which we consider in future research. Therefore, we use both a visual

inspection and the results of downstream tasks for evaluation. Based on a visual inspection, generated synthetic data was considered valid if:

- data generated for columns had positive values. This conforms with the way that Bloomberg outputs its data in that any values that should be added or subtracted are handled in the cell formulas rather than simply summing all values (both positive and negative), with some exceptions (discussed below); and
- data generated is within the normal range for that variable for each company. For example, for First Western Financial Inc ([ticker](#) MYFW), Net Interest Income ranges from \$24.47 million to \$56.51 million for company MYFW. Values for Net Interest Income outside of this range would not be considered valid for MYFW.²⁷

If synthetic data values are negative, that does not necessarily mean that they are not valid — more research will be needed to determine if the company has ever had negative values in the ground truth data, and if so, what the causes were to determine if the negative values in the synthetic data would be considered valid. Using Non-Operating Expense as an example:²⁸ negative values for this variable can occur for reasons such as lawsuits and reversal in inventory write-downs. Therefore, in accounting and financial domains, a domain expert would need to be consulted to determine if the negative values are valid or not.²⁹

3.3.3.2 Initial Testing

In theory, a user should be able to pass any dataset to a generation model (such as a GAN), in order to generate synthetic records. As discussed in Section 3.3.1.1, we were not able to find any previous work done on augmenting financial statements; the closest was the augmentation of linked tables using the primary key which is not close enough to capturing the [FS](#) relationships. Therefore, we conducted some simple initial testing to see how generation models would perform with linked data. Our suspicion was that the resulting

²⁷[ticker](#) MYFW is for First Western Financial Inc on the NASDAQ.

²⁸It is rare that Bloomberg uses negative numbers. Bloomberg’s convention is to use positive numbers and then use the calculation formula to either add or subtract the number.

²⁹The veracity required for tasks, as well as cost-benefit, will dictate whether investigation will be done for potentially anomalous records. We do caution researchers in the accounting and finance domain, however, as financial tasks have been shown to be very sensitive to incorrect data. So, researchers should consider the potential downstream effects of using potentially invalid data.

FS would be unusable as they would not balance and the shared fields (see Section 1.1) would not be correctly captured. Therefore, we conducted some initial tests with each algorithm (GAN, VAE, ChatGPT, and SMOTE) before beginning the augmentation task to determine how the algorithm behaves with linked data.

Figure 3.8 shows two years of sample results generated by GAN for First Busey Corp. (BUSE) As can be seen, Total Assets do not equal Total Liabilities & Equity which means that the BS itself is not balanced. Not only is the fact the BS is not balanced, it also means that the other statements will not balance either. (Refer back to Figure 1.1 for the relationships between statements.) The same type of problematic result was returned when the linked dataset of FS were passed to the VAE, and SMOTE. Therefore, we confirm our intuition that we cannot simply pass the FS to the generative model as these models are not designed to handle linked data and maintain the relationships as is found in the FS.

	Y1	Y2
Cash and Cash Equivalents	34.2	42.3
Interbanking Assets	1.1	1.12
ST and LT Investments	285.1	220.4
Total Commercial Loans	254.3	245.72
Total Consumer Loans	233.8	213.1
Other Loans	0	0
Total Assets	808.5	722.64
Liabilities		
Demand Deposits	72.1	70.32
Interest Bearing Deposits	637.2	643.11
Savings Deposits	385.4	465.43
ST Borrowings & Repos	22.23	0
LT Debt	0	0
Other Liabilities	0	0
Total Liabilities	1116.93	1178.86
Shareholders Equity		
Preferred Equity and Hybrid Capital	0	0
Share Capital	27.2	12.34
Treasury Stock	3.3	2.1
Other Equity	0	0
Total Equity	30.5	14.44
Total Liabilities & Equity	1147.43	1193.3

Total Assets does not balance to Total Liabilities & Equity which means that the Balance Sheet is not balanced.

Figure 3.8: Example of synthetic financial statements created for First Busey Corp (BUSE US) using a Generative Adversarial Network.

3.3.3.3 Features

Instead of using the full FS, we removed the linkages between the FS and used non-subtotal features, as well as non-total features, from the BS and IS and passed them to the generators.

We again used the VAE and GAN as preliminary testing models to determine the feasibility using non-total and non-subtotal features. Non-total features are those that are not providing totals. Cash and Cash Equivalents is an example of a non-total feature, whereas Total Assets is an example of a total-feature. The intuition behind taking a non-subtotal and non-total approach was that by using only these features (non-subtotal and non-total), the linkages within statements and across statements is removed. If the user needs FS, the synthetic data from the features can then be re-assembled and the subtotals and totals can be calculated using the synthetic data, thereby ensuring that the relationships between the statements as preserved.

Preliminary results using only features confirmed that this approach worked and that once the synthetic feature data was re-assembled, the BS and IS both balanced. Figure 3.9 shows the features that were used for generation of synthetic data. We also note that we have not included taxes, which is an important input to calculating Net Income. There are many factors that will influence the amount of taxes that a company pays (which are beyond the scope of this thesis). The amounts for the IS are therefore Income before Tax. The columns “Net Interest Income” to “Minority Interest” are from the IS and the columns “Total Assets” to “Minority/Non Controlling Interest” are from the BS. Labelling was done as discussed in Section 3.3.2.2.

Sample Income Statement and Balance Sheet Data for the Bank of New York Mellon											
12 Months Ending	+ Net Interest Income	+ Total Non-Interest Income	- Provision for Loan Losses	- Total Non-Interest Expense	- Non-Operating (Income) Loss	- Net Extraordinary Losses (Gains)	- Minority Interest	Total Assets	Total Liabilities	+ Minority/Non Controlling Interest	Label
12/31/1987	642	425	301	711	-49	0	0	23,065	21,851	0	2
12/31/1988	741	532	168	807	0	0	0	47,388	44,596	0	2
12/31/1989	1,253	901	783	1,325	-18	0	0	48,856	46,092	0	2
12/31/1990	1,319	936	423	1,399	0	0	0	45,390	42,504	0	2
12/31/1991	1,198	1,036	746	1,311	0	0	0	39,426	36,524	0	2
12/31/1992	1,367	1,183	443	1,519	0	0	0	45,210	41,480	0	2

Figure 3.9: Example of accounting data to be passed to the generator. Amounts are presented in millions. The first seven features are from the income statement and the last 3 features (excluding the label) are from the balance sheet.

3.3.3.4 Generative Adversarial Networks

We evaluated two popular GAN models: TGAN and CTGAN.³⁰ These two specific GANs were chosen as they have been used for financial research in the literature [342, 189, 260].

³⁰We recognize that FS can be both stationary and time series data, depending on the task. If we are considering the data to be stationary, then non-time series generation methods, such as TabGAN and CTGAN, would be acceptable. If we are considering the data as a time series for its predictive capabilities,

Therefore, following previous research in the computational financial domain, we evaluate both TGAN and CTGAN.

(i) TGAN

TabGAN is a library that is available for Python [19], based on the TGAN³¹ model [341, 340], which handles dataset rows independently and fits the data to a Gaussian Mixture Model. The current version is 1.2.1. For our preliminary experiments, we mainly used the default parameters with a few changes. The parameter “gen x times” which controls the amount of data generated was set to $3.1x$, rather than the default of $1.1x$. If we kept the default, in some experiments, no new data was synthesized, which is related to a sufficiency issue (discussed below). We also adjusted the quantile filtering.³² The default parameters are 0.001 and 0.999 for the bottom and top quantiles, respectively. These act as filters for postprocessing that fine-tune the generated data. By adjusting the quantiles to 0.002 and 0.998, we were able to correct, to some degree, for the sufficiency issue (discussed below), and have the GAN output more truly synthetic data.

TabGAN suffers from a “sufficiency” issue: if there is not enough training data, the GAN will output records from the original data as “new data”: 46 new records were generated, but 43 of those were duplicates of the original data. Only 3 were truly newly generated (red records in Figure 3.10). Although studies have shown that GANs do not memorize, Feng *et al.* have demonstrated that this is not the case and that GANs can, in fact, memorize [100]. As well, the results from our experiments support Feng *et al.*’s conclusion, as our results show a high level of replication.

We did not find this issue with the other GAN models that we have used for our experiments, which suggests that there is an issue with the adversarial pipeline. Due to the small amount of data that we are using for these early experiments, it may be that the training process does not have sufficient data and, as a result, is passing real data labelled as fake data to the discriminator. This would then pass the discriminatory process easily as

then a GAN that is specifically designed to capture the sequence of data points in the order of time would need to be used, such as TimeGAN. As the generation of tabular FS has never been done before and we are using classification as our downstream task, we leave capturing the temporal dependencies for future work.

³¹TGAN uses an LSTM with attention to generate data. Therefore, as the LSTM is already incorporated, we did not evaluate using an LSTM on its own.

³²A quantile “determines how many values in a distribution are above or below a certain limit” [300]. If the distribution is divided into four equal parts, each part is one quartile. If this distribution is divided into one hundred equal parts, each part is a percentile [300].

Year	Net_Int_Income	Non_Int_Income	Prov_For_Loan_Loss	Non_Int_Exp	Non_Op_Income	Extraordinary_Losses	Minority_Int	Total_Assets	Total_Liabilities	Shareholders_Equity	Class
2004	17150	12909	1717	17399	174	0	0	427849	389983	37866	1
2005	18504	14444	2383	19018	-1	0	0	481741	441081	40660	1
2006	19951	15740	2204	20742	0	0	0	481996	436120	45876	1
2007	20974	18416	4939	22824	0	0	0	575442	0	575442	1
2003	418744	220901	0	770030	1277516	0	0	45640580	9444334	36196246	0
2004	487089	379829	0	752091	1226680	0	0	53377835	9714680	43663155	0
2005	1232924	247816	0	1777463	1662628	0	0	67853063	91382114	-23529051	0
2006	3156830	194684	0	3592608	2274789	0	0	91382114	27362729	64019385	0
2007	4102000	534000	0	4433000	190000	0	0	96077	18356	77721	0
2003	483	5414	0	1063	1959	0	0	96077	18356	77721	0
2004	483	3259	0	1405	1535	0	0	76754	29425	47329	0
2005	760	5586	0	1495	1090	0	0	78838	31265	47573	0
2006	760	5586	0	1495	1090	0	0	78838	31265	47573	0
2007	760	5586	0	1495	1090	0	0	78838	31265	47573	0
2003	7116	5850	42	7408	5850	0	0	275178	255436	19742	0
2004	7629	4612	209	7535	4612	0	0	307918	286692	21226	0
2005	7902	5097	316	7620	5379	0	0	343573	316294	27279	0
2006	7305	6377	816	8807	4770	0	0	346288	319319	26969	0
2007	5070	6042	3107	10600	309	0	0	327913	303329	24584	0
2005	113546	104008	14431	14605	786788	16	25	5048245	128963	4919282	0
2006	357194	95913	3197	142170	602237	-37	27	2880885	2818248	62637	0
2003	6903	16944	786	231999	36427	-298	64	3495600	893502	2602098	1

Figure 3.10: TabGAN Sufficiency Issue

the “fake data” is actually “real data”. Through testing, we determined that the sufficiency of records needed is approximately 175 records. If less than 175 is used for training, the GAN increasingly passes off real data as new synthetic data at the end of the adversarial process. As part of our future work, we will be expanding our dataset, which we believe will help provide a better understanding of what is happening in the GAN pipeline.

(ii) CTGAN

CTGAN is a library that is available for Python [320, 340]. It builds on the TGAN model by being able to conditionally constrain the generator to create synthetic data based on the imbalance in the discreet (categorical) variables. The current version is 0.6.0. For our preliminary experiments, we used the default parameters outlined in the model with a few exceptions.

We also specified that the synthetic data that the GAN returns should be 20 records. The model default value for the number of synthetic records to create is 1000. For experimental purposes, we had the model synthesize 1000 records. Our expectations were that one of two scenarios would occur in synthesizing a set of records to almost 8 times the original: either the GAN would output “real” records as synthetic data, like TabGAN, or the GAN would only output records based on only a small portion of the latent space. To our surprise, neither of those scenarios occurred. CTGAN did output 1000 synthetic records, none of which were the original data labelled as “synthetic”. When we inspected the data, however, the numbers that it outputs for some of the features are extremely unlikely. These unlikely numbers are a sign of mode collapse which is where the GAN is only able to create a small variety of data

(based on using a small area of the latent space) that can fool the discriminator [116, 101] and be returned as “new” data. While negative net income (better known as “net loss”), for example, is not unheard of in real-world data, it is not a common occurrence for the banks in our dataset, particularly the largest banks such as JPMorgan Chase & Co. As can be seen in Figure 3.11, out of 15 synthetic records, it has created 12 with negative net interest income, five of which are in the trillions³³.

	Year	Net_Int_Income	Non_Int_Income	Prov_For_Loan_Loss	Non_Int_Exp	Non_Op_Income	Extraordinary_Losses	Minority_Int	Total_Assets	Total_Liabilities	Class
0	2004	-1222339	-23289	5793	-509182	808442	-314	31	44241434	2651967	0
1	2003	-1422004	-8905	18297	2296668	-54311	-5867	35	-21203481	2448086	0
2	2003	-526537	413811	9103	2387449	-126858	-2063	550	-7374376	3035482	1
3	2003	-32630	-156804	8607	1387431	-117885	-1202	55	-27964651	-5554481	1
4	2004	-2498062	-22432	24223	468908	-105534	-281	1527	-38940349	-180910	0
5	2007	-6344	-29127	839	376542	-107693	-8120	111	-37326795	4286138	0
6	2007	-547243	-46652	7529	106148	15221	-1552	116	-21702327	-1575139	0
7	2007	-169184	442486	-1136	-191235	-66104	-1964	26	-1569075	19497597	1
8	2004	111437	-33186	12712	756493	-36102	-1625	10	16996310	-1594290	1
9	2006	-90143	100	13082	406714	2653873	-3652	1672	-26310158	-1213570	1
10	2005	-490670	11242	949	-77053	-99190	-1144	759	-3338522	208498	0
11	2005	1529515	-35395	1342	4068388	-44617	-307	32	-41219723	-1645626	0
12	2004	-1350111	-11546	-1803	256069	-38687	-1607	86	5077993	-2744163	0
13	2004	-91760	-90276	12163	1741680	-14320	-1424	416	-36113823	-507534	0
14	2003	964094	13103	7248	1715770	73474	-901	83	-15523706	458594	1

Figure 3.11: CTGAN 1000 Synthetic Records

3.3.3.5 Variational AutoEncoders

VAEs are a popular method to use in the computational financial domain in the area of anomaly and fraud detection [297, 153]. Vector quantised VAEs have also been used with financial statement audits, but in the context of learning journal entries [281].³⁴ As such, we evaluate the use of the VAE for generating synthetic tabular data.

We used the Variational AutoEncoder implementation from Synthcity and used the TVAE (tabular VAE model) which is a conditional tabular VAE network [256]. The library was modelled after the conditional GAN [341], but implemented into a VAE.

³³Original data passed to the generator was presented in millions.

³⁴We also note that work has been done on “long financial reports” using conditional VAEs [333]. Colloquially, the term “financial reports” (in the domain of accounting and finance) has been used in place of the FS, as well as the corporate annual report. In reviewing this research, however, the authors are neither using the conditional VAE for the FS or for the annual report, but rather using (incorrectly) the term “long financial reports” to refer to longer financial news reports [333].

Year	Net_Int_Income	Total_Non_Int	Provision	Total_Non_Int_Exp	Non-Op	Extraordinary	Minority	Total_Assets	Total_Liabilities	Class
2021	19547.19582	20649.70906	5619.993507	-25931.03638	-1023.880662	646.468545	57.82991307	-106459.5041	-623573.1765	1
2008	15569.55508	25042.40824	-916.9177106	37116.62463	-86.72173652	286.0316618	-167.665174	147352.7451	1051270.452	1
2017	21378.24933	11470.05911	803.4963116	871.9741357	412.5512195	189.8236936	-28.07187023	927004.781	227942.5159	0
2014	-25988.77808	16651.30771	-1657.530863	43382.93193	84.88219133	301.8670841	168.0007466	1622294.045	-785862.6475	1
2007	18684.3151	-3923.518031	690.3639854	12952.13843	328.0289426	505.7744063	38.37749624	224285.6143	37195.04735	0
2019	13070.9859	-8585.37658	7290.543045	56160.78887	114.5661901	172.320385	27.62431592	1871592.504	262603.3772	1
2019	4800.585404	-87.23170124	3114.708574	25666.21498	560.9140604	-17.13432817	166.5363059	-773381.786	234818.395	0
2005	1364.095716	8253.763388	-3358.686134	4425.900489	-214.1470242	724.2951336	-4.112281273	690258.7368	104964.8402	2

Figure 3.12: Variational AutoEncoder Synthetic Records

3.3.3.6 ChatGPT

There is a growing body of literature on evaluating ChatGPT’s textual generation, but we were not able to find any research on evaluating its synthetic tabular generation. Therefore, to begin addressing this gap, we included ChatGPT into our evaluation of the available contemporary methods. For our dataset, ChatGPT determined that a column-wise approach³⁵ was the best to generate synthetic data (X_{new}) and used the equation found in Equation 3.3.

$$X_{new} = \mu + \sigma \times Z \quad (3.3)$$

where μ is the mean, σ is the standard deviation, and Z is a random number from a standard normal distribution.

Using the prompt in Figure 3.13, we passed our full tabular dataset to ChatGPT for it to use as the basis for generating synthetic tabular records. The result was similar to the other algorithms in that it contained a lot of invalid data. Figure 3.14 shows a sample of the synthetic records generated by ChatGPT. As can be seen, there are a lot of negative values, which are unlikely, particularly considering that all values are stated in millions. We also asked ChatGPT to determine peers for JPMorgan Chase & Co (JPM) and BayFirst National Bank (BAFN), without any further direction to determine how ChatGPT would respond. As it was given no criteria, ChatGPT assumed that every company in the dataset was a peer for both JPM and BAFN, as ChatGPT did not establish any criteria for itself. This is not in-line with its (ChatGPT’s) own framework discussed in Section 2.2.3 where its first task is to understand the data structure and the content. This suggests that there is a mismatch between ChatGPT’s use of the word “understanding” versus the expected

³⁵ChatGPT can do row-wise generation, but cautions users that while it is an efficient method, it may not capture the relevant relationships between columns or maintain the statistical characteristics of each column.

interpretation when asked to do a task: if a human performing the generation task did not understand or had questions on the parameters needed to complete the task, it is likely that they would ask for clarification. Therefore, it is likely that ChatGPT is using “understand” as an intransitive verb — having the power of comprehension — rather than as a transitive verb — to have thorough or technical acquaintance with or expertness in the practice of — the data [211].

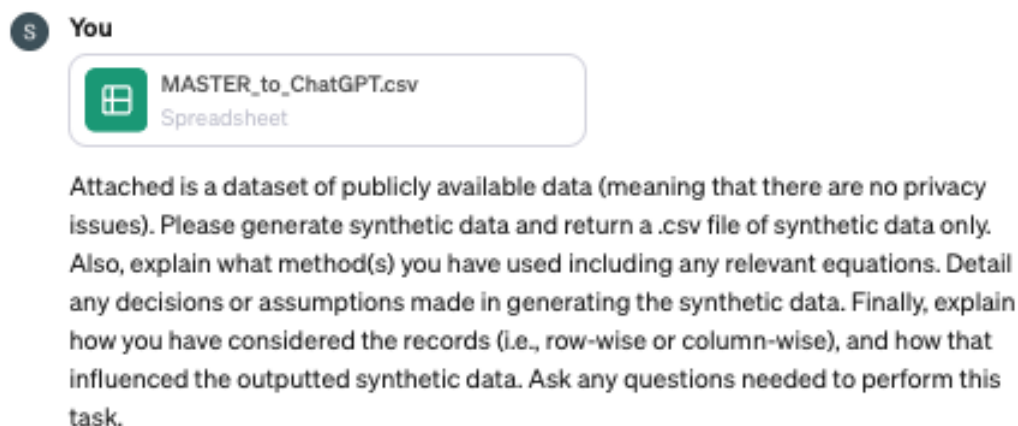


Figure 3.13: Prompt provided to ChatGPT to generate synthetic records for the full dataset.

DATE	Net_Int_Income	Total_Non_Int	Provision	Total_Non_Int_Exp	Non-Op	Extraordinary	Minority	Total_Assets	Total_Liabilities	Class
2017-03-06	2597.32008	-1169.399603	986.778968	6673.434579	115.827275	-24.58722091	58.9865509	345330.0849	61802.76783	1
1995-12-29	-7444.93599	7476.81847	1005.91577	4497.125994	-251.23903	108.8998205	-0.67944	-323059.5921	279435.1303	2
2019-01-18	-1116.844496	7549.383511	728.016511	1310.998095	59.6560887	-13.99360776	1.04714067	-202385.9775	2747.241189	1
1998-01-27	1925.130032	-355.7531227	-2962.1548	-1874.224392	130.234932	9.503026556	-97.661946	-44509.00757	375353.1056	3
2005-06-01	-4433.03665	1167.262517	530.127547	-4175.499239	-405.49175	-51.72168226	22.6523371	54628.58743	142699.3638	2
2018-09-05	2097.611385	-6894.124298	2285.22637	8450.376476	109.93874	136.7922172	-19.272108	290606.2669	-61659.64457	3
2012-04-17	-1055.555787	1023.416621	398.106347	999.6230019	-237.3974	-109.8009546	-12.214194	253344.7699	344996.2964	1
1998-12-27	4097.099005	377.8566042	1214.3538	3261.758133	-268.49526	-139.6646112	-9.2087316	193590.855	218549.3288	1
2016-04-17	-1433.51604	-81.03921255	1584.62541	8271.361756	273.248938	61.05141955	31.7205204	34500.95175	-58833.76054	3
2003-11-26	-3716.447414	4046.381799	2229.64303	3531.558362	-323.82065	-33.77538522	-8.8485826	322267.5964	91476.77362	2
2009-11-29	-8194.311914	-6713.850174	193.596211	14470.36624	-160.66319	-57.11964264	-40.062511	-352840.0187	46010.35297	1
1996-01-16	5323.671259	-517.8789104	9.06841945	1070.779973	-229.23667	52.95044681	63.1626549	-126891.7061	-147870.5063	2
2003-09-28	1402.409011	5896.575038	548.791267	-2526.358592	52.8179313	-178.9930852	2.84134563	-40522.64582	-17516.50271	2
2002-05-21	4709.836045	3707.601564	-1122.221	1138.861821	-44.606534	101.2215076	-28.906294	73361.45241	72541.21072	2
1997-07-11	-4320.704853	-10604.93428	567.923398	3603.407745	-17.106008	42.55191151	-8.2112964	346532.3344	-115574.0614	1
2004-12-11	-1156.743733	5034.858223	1505.99697	1229.559783	256.556242	-11.83194697	-44.014899	101891.6015	-113404.1951	3
2015-08-24	-4278.948906	421.3430119	951.619085	-6493.73716	82.4384079	53.43086537	-30.874962	51333.95665	-157005.0316	1
1995-03-31	4666.486006	225.1885674	58.5223876	-2770.898805	-170.96556	187.2916702	-36.76516	92759.33918	126195.6196	2
2019-09-24	4917.223717	575.2465358	-585.0817	13957.45464	-193.81966	112.2985706	47.6021291	329900.545	87611.49175	3

Figure 3.14: Sample of the full tabular synthetic generation provided by ChatGPT.

This experiment was followed by providing ChatGPT direction to use each column as key financial indicators and to use those to select peers for JPM (and output a .csv file with the names of the peers it had selected) and then return a .csv file with synthetic records

created using JPM’s data and the data for the peers that ChatGPT had selected. Providing these parameters improved ChatGPT’s ability to generate peers. From the financial metrics available, ChatGPT selected Net Interest Income, Total Non-Interest Income, Total Assets, and Total Liabilities, as these are common measures of a bank’s size and financial health, and used the average for each. We point out that this is a similar approach to ours when selecting peers. However, instead of using the average, we used a high-low approach and then added on a buffer, as explained in Section 3.3.3.7.

ChatGPT then calculated the Euclidean distance (see Equation 3.4).

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^N (q_i - p_i)^2} \quad (3.4)$$

where p and q are vectors in N-dimensional space representing the financial profiles of different companies. p represents the vectors of p_1, p_2, \dots, p_n for Company A and q represents the vectors of q_1, q_2, \dots, q_n for Company B. Each vector captures the values of each key metric, which in this case are Net Interest Income, Total Non-Interest Income, Total Assets, and Total Liabilities. p_i and q_i are the coordinates in the i th dimension, respectively, and $d(p, q)$ is the Euclidean distance between points p and q . The smaller the distance, the greater the similarity. As we did not give any cutoff threshold to ChatGPT in our instructions, ChatGPT chose the top 10 peers. The peers that it chose were:

- Wells Fargo
- Bank of New York Mellon
- PNC Financial Services Group Inc
- Mitsubishi UFJ Financial Group Inc
- Truist Financial Corp
- Fifth Third Bancorp
- Regions Financial Corp
- M&T Bank Corp

- Zions Bancorporation NA

Some of the peers that ChatGPT selected are correct such as Wells Fargo, PNC, Truist, and Fifth Third. However, banks such as Citibank and Bank of America, which are in the original data given to ChatGPT and are large banks, meaning that they are direct competitors to JPM, were not included in ChatGPT’s peer list. Upon visual inspection of the data, we also notice that the majority (see Figure 3.15) are not reasonable amounts. Of the 24 records that ChatGPT generated, only seven have positive numbers in all fields, which using the criteria set out in Section 3.3.3.1 means that the data for the remaining 17 is not valid, as the record includes negative numbers for one or more of the four financial metrics that ChatGPT used to generate the synthetic data. Using the fourth record as an example, and keeping in mind that the data passed to ChatGPT is in millions (‘000,000), this would mean that JPM had negative net interest income of 13.5 billion and had negative assets of just over 35 billion, which is not reasonable.

COMPANY	DATE	Net_Int_Income	Total_Non_Int	Provision	Total_Non_Int_Exp	Non-Op	Extraordinary	Minority	Total_Assets	Total_Liabilities	Class
JPM	1995-04-26	3161.003845	8936.848969	-300.1899138	-9894.581919	-43.15622894	-32.56441457	-29.74370115	-42197.7566	-84826.19753	2
JPM	1995-11-25	-3579.575686	-7105.046219	-556.7550762	1559.130896	-184.6577838	38.03696756	92.43157378	-309848.4147	-245853.7213	2
JPM	1996-08-03	885.4604101	6092.615464	1077.088789	7130.606344	-73.9020984	-160.3937153	54.29782349	141996.9223	85547.30115	3
JPM	1998-08-06	-13532.95584	1503.586122	-255.4605914	-116.8129268	-81.07084569	31.11245602	46.65908514	-350374.0135	215408.6297	2
JPM	1999-11-22	-1454.930724	-2303.447263	827.9067074	-631.640422	-39.75442034	-116.6184088	82.08432398	-140472.7894	-41022.71835	3
JPM	2001-07-06	2447.81767	-16.55778691	-1538.941489	1663.822913	-170.565163	56.40892281	21.96500524	406372.5146	-34236.82103	1
JPM	2002-08-27	-104.0927586	-1507.764655	-2531.369406	-5474.436177	78.27870357	22.71774177	-38.25332905	320825.6022	-151837.0472	2
JPM	2002-12-05	2045.248477	559.133259	362.3713682	-429.3441265	-111.0815105	-50.24537477	-36.0660924	-48303.86692	98475.73369	3
JPM	2003-06-02	-4844.012216	5178.90793	2555.100359	4970.749926	31.51156931	-43.34284345	40.91378589	101075.7977	176567.5461	1
JPM	2004-04-13	6429.767731	-2456.314892	-861.7079367	4425.087882	403.4425647	75.09719693	34.37732651	113462.6572	96995.99982	3
JPM	2005-01-16	-3538.950006	2142.633051	-659.2805629	5325.039394	-15.18818086	38.81623874	41.30394689	-164847.8954	110620.3987	2
JPM	2005-08-19	1245.699276	249.7081536	-1115.977894	-9944.60622	-26.08552204	-216.5200396	28.99320903	-108127.0649	256973.1638	1
JPM	2005-10-03	1607.011624	4803.402945	-293.4119964	-2566.615428	252.6095926	-111.0662953	-22.2349867	-207855.261	-88708.68532	3
JPM	2005-10-16	-3512.302935	-1784.306872	1190.017262	-14172.78553	72.70207055	13.20917672	-25.53071023	166775.5701	-174896.336	2
JPM	2008-02-23	2859.588534	-7687.730807	994.6796539	-486.6838222	-122.3745507	-17.07987081	-7.010542984	227575.7216	8236.628123	1
JPM	2009-12-01	-9800.514482	-4204.31509	860.1607153	15812.18438	16.96611042	-106.1896572	-35.74405226	393510.3274	222055.822	1
JPM	2010-04-23	5171.126603	1452.647045	804.6030637	-3644.120113	-22.57361699	113.4014572	19.70539368	47423.08047	-280685.1058	1
JPM	2013-07-30	911.3990648	1091.901379	629.2065906	3854.145307	112.5353994	-234.6545716	-40.70433219	-140154.5028	320343.1552	1
JPM	2013-10-07	5971.072248	4860.149649	57.42595095	1361.493879	-15.50518009	7.283504775	-45.66102045	74296.33186	31073.32079	3
JPM	2014-11-15	771.6694689	4599.776848	27.2642801	6414.512683	-103.7513996	129.0351739	75.57214178	-308149.6391	2182.433845	1
JPM	2016-05-26	935.3847827	-2177.609348	-471.9743454	9752.504814	-194.4492878	-65.96962296	-21.54015481	309956.3889	-234470.0307	2
JPM	2016-10-21	-3301.571633	-2696.152516	-1514.223767	-4778.397745	-180.4901333	83.39113172	47.33892022	176288.1609	215240.4499	2
JPM	2017-01-27	2524.491678	199.0405482	475.9091651	2049.786822	38.02011642	-246.990636	-47.92311196	248990.7671	29269.54326	3
JPM	2019-06-12	-517.5711088	8117.679219	-1887.192369	-717.7711424	55.39154422	32.93071275	-5.6411212	-79695.57969	-93613.21113	1

Figure 3.15: Visual Inspection of synthetic JPM data showing a lot of unlikely negative values.

Based on this, we asked ChatGPT to graph the distributions (see Figure 3.16) of Net Interest Income for the input file and its generated synthetic data to determine if the data that ChatGPT generated really did exhibit the same distribution as it claimed it would. Unsurprisingly, we find that the distributions are not the same. The shape of the distributions is different in that the original data (shown in blue) is skewed whereas the synthetic

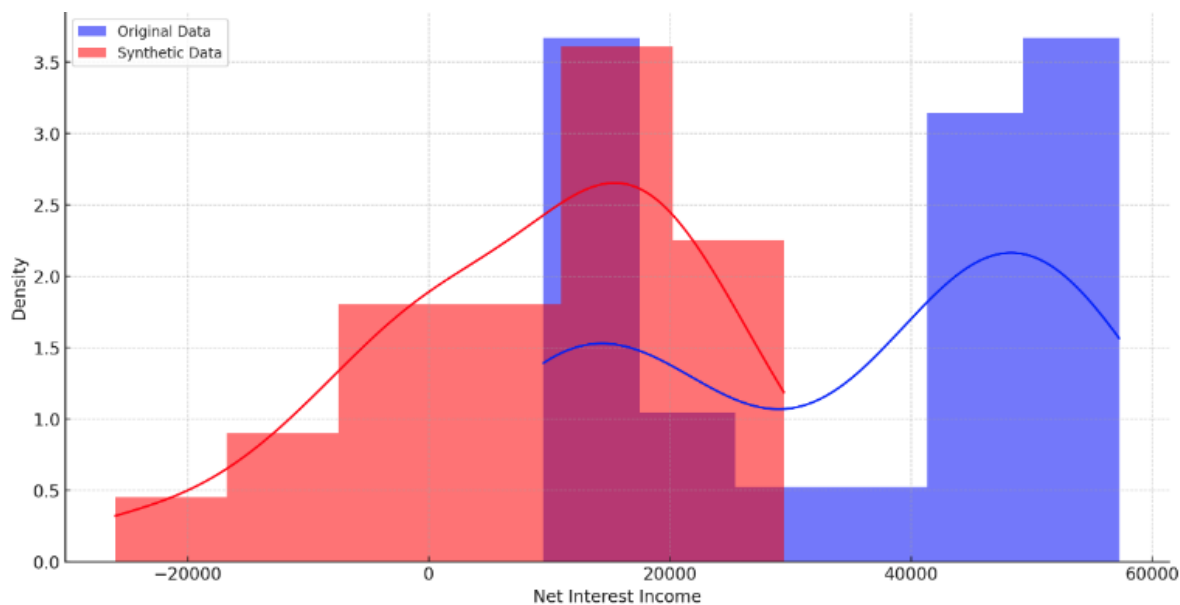


Figure 3.16: Distributions of Net Interest Income where the original data is shown in blue and the synthetic data is shown in red.

data appears more uniform. The range and spread are different in that the minimum and maximum values for the original data (shown in blue) are all positive which is not the case for the synthetic data (shown in red) where the majority of the data is negative. There is also a wider spread of the synthetic data (red) compared to the original data (blue). Finally, the central tendency for the original data (blue) is bimodal with two distinct peaks, whereas the synthetic data (red) has a peak that is less defined than the original data (blue). Therefore, the synthetic data does not appear to replicate the statistical distribution of the original data for Net Interest Income, and are therefore not statistically similar, even though ChatGPT originally indicated that the synthetic data that it generated was, in fact, statistically similar.

3.3.3.7 Synthetic Minority Oversampling Technique

In reviewing the literature, we could not find any discussion of the minimum number of records that SMOTE could viably generate synthetic data for. We did find one github post that indicated that each class had to have at least 2 records (or instances) [173]. We tested this comment using Washington Federal (ticker WAFD) which had 36 records, 10 of which

were label 2 and 1 was label 3, with the remainder (25) label 1. When we tried to SMOTE³⁶ the minority classes (labels 2 and 3), it raised an error for label 3 which only had one record. When we added an extra record to class 3 (meaning that there were two samples), we were able to run the algorithm. The intuition behind this is that in order to create the synthetic data, the minority class points have to have a nearest neighbour, as illustrated in Figure 2.4, to use the line between the feature vector and its nearest neighbour to create a synthetic data point. However, we also recognize that the number of records needed to create quality synthetic data will be influenced by more than just the number of samples in a class being 2 or more, as discussed in Section 3.1.

SMOTE can be used in binary and multi-class scenarios. In either case, the majority class will be used to determine the number of synthetic records that must be generated for the other class(es). In using SMOTE for our dataset, we found that, of all the algorithms, it provided the best synthetic data based on a visual inspection, but also in the downstream task (discussed below in Section 3.3.4). However, when we wanted to generate synthetic records for a specific bank (rather than the dataset as a whole), SMOTE struggled either because there were too few records in the minority class(es) or because there were limited records for the bank of interest. To address this challenge, we devised Peer Augmentation.

Augmentation with the Help of Peers

We borrow the idea of scaffolding from engineering, construction, and information modelling as inspiration on how to tackle data constraints. Scaffolding is a temporary structure that assists in a task [168]. For example, scaffolding on the side of a building is used to “assist mainly in the placement of bulk materials” [168]. Once the building is finished, the scaffolding is torn down, leaving only the building. In a similar fashion, we use peers to provide that scaffold for the dataset. If a dataset is too small or does not have sufficient features in its own data to be useful to a task, the data of peers can be used as an initial scaffold to [augment](#) the dataset before it is passed to the generative algorithm.

Figure 3.17 shows the process of peer augmentation. Depending on the researchers’ situation, peers may be new data that is yet to be retrieved, or it can be from within a larger dataset. Our experiments use the latter scenario, meaning that the peers will be selected from the existing dataset of 3,207 records. A critical component of this process is the peer

³⁶We purposely use “SMOTE” as a verb here to follow the Chawla et *al.* for when the algorithm is used with data [52].

selection process, as peers must be representative of the company's data. This means that the peer group's financial features should be closely resembling those of the company. If the resemblance parameters are too tight, then the resulting peer group will be too small to be of use to the augmentation process.³⁷ If the resemblance parameters are too lax, then the resulting peer group contains companies that are not truly peers and more like outliers. This distorts the distributions of the original data, as well as the resulting synthetic data. Therefore, peers must be carefully selected.

Using JPM as the target company³⁸, we used Random Forest to classify the original (un-augmented) dataset to determine the top two features in its classification and used those to determine peers. The Random Forest classifier indicated that Net Interest Income and Total Liabilities were the top two features in the classification of the original dataset. These two features give equal representation to both FS as Net Interest Income is from the IS and Total Liabilities is from the BS. For those two features, we used a “max-min” approach in that we determined maximum and minimum values for each feature, and then extended the range (top and bottom) by 5%. For example, if the range was \$1,000 – \$10,000 then the range extended by 5% was \$950 to \$10,500.³⁹ A hierarchical approach was taken and we started with Net Interest Income as this was the highest ranking feature, and then when the peers had been selected for Net Interest Income, we then refined the peer list based on the range for Total Liabilities, the second ranked feature.

Again, continuing with Figure 3.17, once the peers have been selected, the target company (bank) and its peers are passed to the algorithm to be used as the basis for synthetic generation. After the synthetic data is created, a visual inspection is performed to identify any anomalous or duplicate records, which are discarded or addressed. The remaining synthetic data is preserved, the peers are stripped out, and the data is added to the target company dataset. The process repeats until the researcher has determined the dataset is sufficiently large for their task or research question.

³⁷In initial tests, the resemblance parameters were too tight, returning only the bank of interest with no peers.

³⁸We also tried this procedure for five other companies: BAFN, CATY, EQBK, EWBC, and HMST, and achieved similar results

³⁹The extended minimum range was calculated as $1,000 \times 95\%$; the extended maximum range was calculated as $10,000 \times 105\%$. This process is similar to that depicted in Figure 3.4.

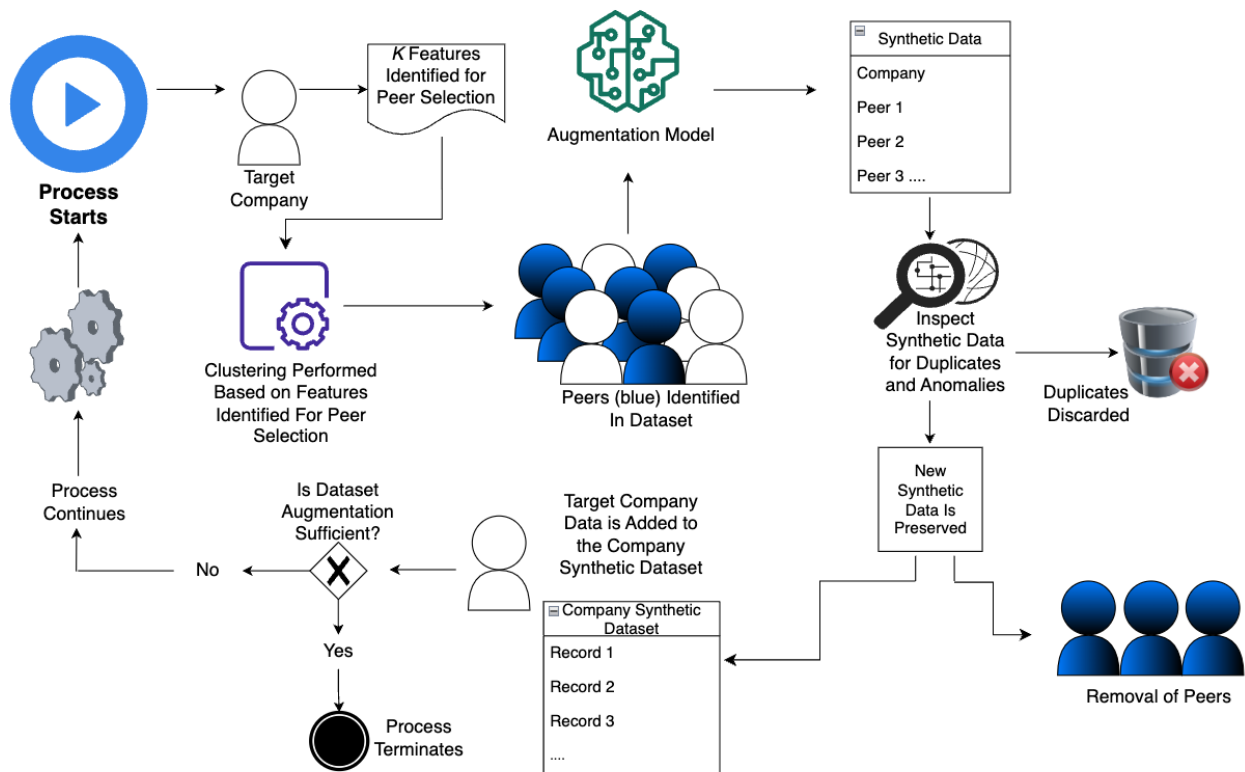


Figure 3.17: Peer augmentation process. A target company's peers are used to provide training data for the augmentation model. Once the synthetic samples are generated, peer data is stripped out, leaving only the synthetic samples for the target company.

3.3.4 Evaluation Using a Downstream Task

As there are no widely accepted methods for evaluating the quality of synthetic tabular data, downstream tasks are often used as an evaluation proxy. For our downstream task, we chose classification, and used accuracy (the most common metric, based on our review of the literature) as our metric. The initial data as well as the synthetic data was labelled into three classes using the cash ratio (see Section 3.3.2.2). Classification was done using the following standard algorithms: Naïve Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), and a Convolutional Neural Network (CNN). Each is briefly discussed below:

- NB: Naïve Bayes is based on Bayes’ Theorem: $P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$, where P is the probability, A and B are both events, and $P(A|B)$ is event A is conditional on other event, B . The Naïveté is introduced in making the assumption that all of the features are independent of each other which not only simplifies the learning, but also makes this algorithm very competitive against other classification algorithms [267]. In practice, NB is often used to provide a “baseline” for other algorithms to be compared against.
- RF: a tree-structured classification algorithm where each tree votes for the most popular class at input x . This algorithm is formally defined as $h(\mathbf{x}, \Theta_k), k = 1, \dots$ where Θ_k are “independently identically distributed random vectors” [22].⁴⁰
- SVM: input vectors are mapped to a high-dimensional feature space which are then separated by an optimal separation hyperplane. The maximized margin is the space between the optimal separation hyperplane and the closest points to either side of the hyperplane (positive and negative). These closest points are known as support vectors. The larger the separation, the better classification. The hyperplane is formally defined as $(w \cdot x) - b = 0$ with the optimal hyperplane as $(w \cdot x_i) - b \geq 1$ if $y_i = 1$ and $(w \cdot x_i) - b \leq -1$ if $y_i = -1$, where w is the weight vector, x is the input vector, and b is the bias [330]
- LR: This model classifies based on a set of predictive features [150]. Logistic regression is binary as the predicted value is restricted to two classes. For our research, we

⁴⁰If RF is used for regression tasks, then the results from the decision trees are averaged [152].

use multinomial logistic regression as we have three classes. The multinomial logistic classification model with k classes is formally defined as $\hat{y}(k) = \Theta_0^k + \Theta_1^k x_1 + \Theta_2^k x_2 + \dots + \Theta_n^k x_n$, where k is the number of classes, $\hat{y}(k)$ is the log odds of predicting the observation belongs to class k , and $\Theta_0^k + \Theta_1^k x_1 + \Theta_2^k x_2 + \dots + \Theta_n^k x_n$ are the model’s parameters which are estimated from the model’s data for class k . The softmax function would then be applied to convert the log odds calculated in the equation provided into probabilities [24].⁴¹

- CNN: A Convolutional Neural Network is a “network architecture that learns directly from data” and uses an input and output layer, as well layers such as the convolutional, pooling and activation layers in the hidden layers [180, 204, 151]. A layer of the CNN can be formally defined as $y = f(Wx + b)$ where y is the output, f is the activation function, W is the convolution matrix for the input x , and b is the bias [165].

We tested these algorithms under three separate conditions:

1. “Original Only” indicates that only original data (i.e., ground truth data taken from the company’s audit FS) was passed to the algorithm, with a train-test split of 80:20 used.
2. “Synthetic Only” indicates that only synthetic data (i.e., outputted data from the various generation algorithms — CTGAN, VAE, ChatGPT, and SMOTE) was passed to the algorithm, with a train-test split of 80:20 used.
3. “Combo Data” — indicates that combination data (i.e., ground truth and synthetic data) was combined and shuffled, passed to the algorithm, with a train-test split of 80:20 used.

All of the synthetic data was passed to the classifiers, even if, based on a visual inspection, we had determined it contained invalid data. In our paper “Don’t Blindly Use Data: Towards a Data Statement for Computational Financial Research”⁴² we argued that the veracity of

⁴¹While linear regression is also a “standard” machine learning, the values that it predicts are continuous rather than probabilistic. Therefore, linear regression would not be appropriate for our classification task [346, 321].

⁴²This paper was accepted and presented at the the First Workshop on Responsible AI and Data Ethics (RAIDE 2022) at the IEEE International Conference on Big Data

the dataset is a critical component to research otherwise the applicability of the results are questionable [311].

The results from the downstream task can be found in Table 3.5. Recalling that only the SMOTE (and peer SMOTE) data passed visual inspection, the increase in the results for the other algorithms (TGAN, CTGAN, VAE, and ChatGPT) is concerning. The increase in the performance of the classification (from the original data to the combination of original data and the synthetic data) shows an improvement in the classification model’s performance, which is correct. But, from a rational and reasonable point of view, the increase makes no sense, given that some of the data is fundamentally wrong. There are several possible explanations for this behaviour. Firstly, we selected the cash ratio to label the FS, as we believed that, given the limited information in the FS themselves (i.e., without any additional notes or context provided from the bank), this ratio would be indicative of the financial “stability” of a bank. We also tried other bases for the labels such as Net Income and Total Assets, but those values do not capture the absolute liquidity of a bank to cover unexpected demands on cash [88] — a position that Silicon Valley Bank found itself in and contributed to its demise. Based on testing, the cash ratio proved to be the best method that we found. We also note that, to the best of our knowledge, FS have not been previously labelled for this type of task, so we could not draw on the guidance of the literature in making our decisions on how to best label the FS. Secondly, the classifier is learning to predict the label, not the veracity of the FS. In many respects, these are two very different, but invariably linked tasks, resting squarely on the garbage-in-garbage-out principle. Here, we demonstrate that with problematic data come problematic results. And finally, these results firmly underscore the criticality of the data — a researcher must understand what is actually *in* the data, where its biases may be manifesting or hiding, and how that data could be influencing the results. We passed the problematic data to the classifiers, with full knowledge of what the issues were with that data. The results demonstrate how research can still provide good (or better) results even if the dataset is not carefully and appropriately considered. These types of misleading results demonstrate just how much the dataset informs the selection of the algorithms and models, as well as the results themselves.

The SMOTE data, which passed the initial visual inspection, shows improvements between the original data for NB and RF, but decreases for SVM, LR, and CNN. For SVM, due to computational cost, we were forced to use the RBF kernel, which may not be the

best choice for the data. We did try using the Linear kernel, but the amount of time that it took to run was very long. Over 15 different iterations of the data, this quickly became impractical. However, this impracticality may suggest that the combination data may not be well-suited to methods that create linear boundaries — such as SVM or LR. It was no surprise that the CNN results were poorer, given that Deep Learning requires a lot of data to learn and the size of the dataset (after augmentation) was only 6,414 records, which by Deep Learning standards, is still very small.

For the SMOTE with peer augmentation, the results show that this method of scaffolding can work well for small datasets. This is particularly supported by the increase in the results for the CNN which improved from 0.062 to 0.425, when comparing the original data to the combination data, understanding that improvement is still needed as the combination data results are still less than chance. With the exception of SVM, the rest of the algorithms show increases in the classification accuracy with the use of the synthetic data. While SVM did not show an increase in the results, it did not show a decrease either, which is important to note. Given the challenge with the kernel, discussed above, it may be that the results would be different using the Linear kernel.

3.3.5 Conclusion, Limitations, and Future Work

In this project, we have evaluated contemporary generative algorithms for the purpose of augmenting financial statement data — a task which, to the best of our knowledge, has never been done before in the domain of academia. We found that between TGAN, CTGAN, VAE, ChatGPT, and SMOTE that only SMOTE (and peer augmented SMOTE) produced valid synthetic data. We also demonstrated the importance of validating the synthetic data prior to use in a downstream task as invalid data can still contribute to improved results. The SMOTE results for NB and RF showed solid improvements between the classification for the original data and the SMOTED combination data. We also found that when datasets were small, well-chosen peers provided a good scaffold to [augment](#) the original size before using SMOTE.

Like the previous project, the main limitation that we encountered was the size of the dataset, which we addressed with peer scaffolding. However, to use Deep Learning, a more robust approach to scaffolding will be needed to improve the results to something greater

than chance. We also raise a potential limitation in the use of the cash ratio used to label the FS. To the best of our knowledge, this is the first time in the literature that the FS have been labelled. This gap in the literature has an important implication as we were unable to use the guidance of other researchers. While our research helps to address this gap, it may be that another approach to labelling the data would have been more appropriate. Finally, we enacted a self-imposed limitation by focusing only on the BS and IS, but those are only two of the four main FS. Therefore, as part of future work, we plan to expand our research to also include the Statement of Cash Flows and the Statement of Retained Earnings. We also plan to expand our downstream testing to determine if there is a better approach to labelling the FS.

	NB	RF	SVM*	LR	CNN
Original Only	0.596	0.654	0.581	0.603	0.638
TGAN					
Synthetic Only	0.586	0.600	0.591	0.602	0.585
Combo Data	0.595	0.653	0.589	0.564	0.500
CTGAN					
Synthetic Only	0.581	0.584	0.400	0.550	0.481
Combo Data	0.596	0.695	0.553	0.432	0.581
VAE					
Synthetic Only	0.333	0.500	0.500	0.500	0.432
Combo Data	0.589	0.617	0.569	0.568	0.579
ChatGPT					
Synthetic Only	0.635	0.601	0.637	0.457	0.638
Combo Data	0.643	0.615	0.718	0.731	0.661
SMOTE					
Synthetic Only	0.441	0.693	0.543	0.477	0.361
Combo Data	0.722	0.724	0.547	0.537	0.428
SMOTE with Peers**					
Original Only	0.818	0.800	0.900	0.900	0.062
Original with Peers	0.923	0.778	0.920	0.888	0.094
Synthetic Only	0.902	0.782	0.889	0.824	0.273
Combo Data	0.960	0.960	0.900	0.980	0.425

Table 3.5: Classification Results For Each Generation Algorithm Using Naïve Bayes, Random Forest, Support Vector Machines, Logistic Regression, and Convolutional Neural Networks.

* For SVM, we experimented with three kernels: linear, RBF, and Sigmoid. The best fit for the data was the RBF kernel. Linear did return some promising results, but the computational cost was very high (time-wise) and outweighed the use of the RBF kernel.

** The results provided for SMOTE with Peers are the averaged results from the companies tested. See Section 3.3.3.7 for details on companies tested.

Chapter 4

Augmentation of Textual Financial Data

4.1 Chapter Introduction

In the case of financial statements and the vast majority of financial reports, textual data is needed to contextualize the tabular values. In this chapter, we turn to generating textual data for the synthetic data generated in the previous chapter. However, before we can pass tabular data to ChatGPT for textual generation, we need to first evaluate how well ChatGPT is generating synthetic text for the financial domain. Doing this will help us establish our expectations for textual generation of the tabular data as well as understand where ChatGPT is in its evolution for the financial domain. Taking this approach contributes to the literature in several important ways. First, we provide benchmarking of ChatGPT that starts in the early days after its release and uses models that have since been retired (even though it is only just over 1-year old at the time of this writing) as well as current models. We benchmark across several tasks common to the accounting and finance domain — financial summary data, financial statement note disclosures, and financial literacy. And second, we engage ChatGPT in a generation task for the synthetic tabular data (created in the previous chapter) for which there is no ground-truth, and use a quasi-ground truth to evaluate against. To the best of our knowledge, we are the first to undertake all of these tasks, and the first to use quasi-ground truth as our comparator. We conduct our research over four projects (Projects 3, 4, 5, and 6) which, together, foreshadow the last project in the next chapter — the need for company authorship attribution.

In the interests of full disclosure and responsible use of AI, it should be noted that ChatGPT was used to generate text for comparative purposes only and was considered and used as a “data source” in the context of the research presented in this chapter. ChatGPT did not aid in the formulation of the research problems, the research itself, or in the writing.

4.2 Project 3: Fine-Tuning GPT-3 for Financial Summary Data

4.2.1 Introduction

At the time of this research, ChatGPT was relatively new. The options for fine-tuning GPT-3 were limited to *ada*, *babbage*, *curie*, and *davinci* — all first GPT-3 fine-tunable generation models. These models were retired on January 4, 2024, meaning that unless work was done on these legacy models while they were available, researchers cannot go back in order to create benchmarks. This is why benchmarking at the time the research is conducted is so valuable to research and to knowledge. The purpose of this experiment was to examine to what extent GPT-3 (using the *davinci* model) needed to be fine-tuned trained in order to increase the quality and similarity of its output for financial text. To do this, we evaluated ChatGPT against ground truth-text (human-written financial text) to determine similarity using Cosine Similarity, readability using the Flesch Reading Ease score, as well as sentiment using FinBERT.

4.2.2 Methodology

4.2.2.1 Dataset

It has been our experience that researchers in the computational financial domain do not often share datasets, even if they may indicate in a journal or publication that data is available upon request [310]. While there are some publicly available datasets, none matched our needs — we theorized that this is a common issue in the computational financial domain which leads researchers to create custom datasets for their research questions or tasks [310]. Therefore, for the research we present here, we have constructed a unique dataset to conduct our research using both human-written financial reports and text from ChatGPT.

We chose 5 large US national commercial banks to form the basis of our dataset. Banks selected were:

- JPMorgan Chase & Co (JPM)
- Bank of New York Mellon (BNY)
- Citigroup (Citi)

- Bank of America (BOA)
- Wells Fargo (WF)

To construct the human-written text dataset, the company’s annual reports were either downloaded from each bank’s website or from www.annualreports.com. Once downloaded, information was parsed out of the Financial Summary section of the annual reports to capture metrics such as Net Income, Net Interest Income, Total Book Value Per Share, Credit Losses, and Non-GAAP measures.¹ Table 4.1 provides information on how many prompt and completion pairs were used for fine-tune training for each bank. There are no regulations dictating what information must be presented in the Financial Summary section of the annual report. Therefore, there is some variation of what metrics were discussed year-over-year.² To keep the prompt and completion pairs consistent, the same information was parsed from each report, meaning that if a bank discussed a particular measure in years 1 and 3, but not in year 2, for example, a prompt and completion pair was not captured for year 2. Thus, some banks have more prompt and completion pairs than others.

From this set of pairs, we created six datasets to be used for fine-tune training: one for each bank and then one large dataset that included all 439 prompt and completion pairs which we termed “aggregated”. The purpose of taking this approach for the data was so that we would be able to determine if each bank needed to be individually fine-tune trained to improve the base model’s performance, or if an aggregated dataset could be used instead.

Bank	# of Pairs
JPMorgan Chase	116
Bank of New York	105
Citibank	95
Bank of America	42
Wells Fargo	81
Total	439

Table 4.1: Fine-Tuning Prompt and Completion Pairs

¹Non-GAAP measures are alternative metrics that do not conform to the Generally Accepted Accounting Principles (GAAP). Companies are allowed to use non-GAAP measures as long as they provide a reconciliation to the nearest GAAP measure.

²Some companies will provide a table to showcase the values in the financial summary, while others will provide a more textual summary, or flip between methods from one year to the next. That means that there will not necessarily be a textual passage related to these measures in each year’s annual report.

The prompt and completion pairs were initially inputted into a csv file and then formatted into the required JSONL³ format using EasyGPT [82]. In the JSONL, all prompts require a prompt suffix to indicate where the prompt ends (we used ###) as well as a stop sequence (we used “STOP”). OpenAI also recommends that a space be put between the prompt suffix and the completion; this was also added.

Following OpenAI’s best practices, the prompts were designed to be specific and contextual [289], but in order to mimic what an “average” financial user would likely input to GPT, we were less descriptive about the desired output. For example, one of the fine-tune training prompts for Bank of America was “How did Noninterest income change for BOA between 2020 and 2021?”. Here, the measure (Noninterest income) is specified, as well as the years (2020 and 2021), but the desired output format or length (such as text only with a maximum length of 20 words, for example) was not specified. We believe that this is more likely reflective of a prompt that the average financial user would input.

4.2.2.2 Fine-Tune Training Model Selection

The four models available at the time for fine-tune training were *ada*, *babbage*, *curie*, and *davinci*, ranging in speed, price, and capability. The default model was *davinci*, as it was the most capable model at the time. Each is briefly discussed below [318].

- *ada* - lowest cost, fastest running model. Does not handle nuanced text well. Needs considerable context for improving its performance.
- *babbage* - more capable than *ada* but does not always perform well. Well suited for tasks such as ranking and semantic search.
- *curie* - even more capable than *ada* or *babbage*. Balances speed and power, and can handle more nuanced text and tasks (e.g., sentiment and question-answer).
- *davinci* - the most capable model. Can understand and interpret the intent of text.

We selected the *davinci* model as it met our needs the best as it combined capability as well as the ability to understand (interpret) the intent of text. Given that financial text

³A JSONL, or JSON Lines, is a storage structure which will process the document one line at a record at a time [334]. It is the required format by OpenAI [235].

can be difficult to read and interpret, we believed that this model was the best option for our research. For evaluation, we compared the *davinci* base model against fine-tune trained models — one for each bank and one for the aggregate.

4.2.2.3 Cosine Similarity

We used Cosine similarity as it is the standard similarity measure for Information Retrieval. It is based on the word array vector representations of the texts and the cosine value of these two vectors [145, 333]. An important aspect of Cosine similarity that is particularly useful in our case is that text length is not a limiting factor and is therefore an appropriate measure for similarity of texts of different lengths [145]. The Cosine similarity formula between vectors x and y is given in Equation 4.1.

$$\text{Cosine}(x, y) = \frac{x \cdot y}{|x| \cdot |y|} \quad (4.1)$$

A threshold using an alpha (α) of 0.8 was used, where $0 \leq \alpha \leq 1$, following thresholds established in the literature for similar tasks [257, 292].

4.2.2.4 Prompt Sample Selection

To test both the base model and the fine-tuned models, 10 prompts were randomly selected from each bank’s prompt and completion pairs set, as well as from the aggregated set. This gave us 60 prompts in total that were used for testing — 10 for each of the 5 banks, as well as an additional 10 from the aggregate list. To mitigate the effects of any duplicate prompts between each individual bank’s prompt list and the aggregate, models were created and tested separately for each each bank as well as the aggregate.

4.2.2.5 Sentiment Analysis

There is already a considerable body of literature on sentiment analysis. However, as research broadens and deepens across domains, we are better able to understand and see the effects of text sentiment. Never before has this been more critical now that AI is generating text. Research was conducted that studied the effects of linking human-authored text to market outcomes, and indicated that in order for companies to be successful in their corporate

messaging, they need to re-frame how they talk about their business, with particular focus on their financial communication [46]; The initial pioneers of financial sentiment have been saying the same thing for over a decade [132, 192]. Now that AI has the ability to generate text on our behalf, it is important to evaluate how the sentiment changes, particularly given the aspect of corporate messaging management. To evaluate sentiment, we have used FinBERT [13] using the FinBERT library on Hugging Face [80]. FinBERT is a domain specific Bidirectional Encoder Representations from Transformers (BERT) that is trained on a large financial domain corpus and fine-tuned trained on the financial phrasebank [199].

4.2.2.6 Flesch Reading Ease Score

A chief complaint in the literature is that financial text is difficult to read (e.g. [193, 216, 57, 58]). We note that the Gunning Fog readability index (scores range from 0 to 20 to determine the grade level needed to read the passage of text [263]) is also very prominent in the research. In our review of the literature, there appears to be two distinct (virtually equal) “camps” — those who use Flesch Reading Ease Score (FRE) and those who use Gunning Fog. Both formulas consider syllables as part of their formula, but Gunning Fog specifically identifies and uses complex words (i.e., words that have three syllables or more) [7] in its formula, whereas the FRE does not. The word “liability” is five syllables, and would be considered a complex word, even though it is a relatively straightforward financial term. The financial term “capitalization”, however, is six syllables, and is a very complex term that has a specific meaning in accounting and a very different specific meaning in finance. Under Gunning Fog (unlike FRE), liability and capitalization would be treated equally, even though they represent two very different readability challenges. Therefore, we decided to use FRE over Gunning Fog, following prominent research done in the literature on the readability of the annual report [57] and for the domain of accounting [303].

Readability of the documents is calculated using the spaCy Readability library [212]. This library provides the readability scores for Flesch-Kincaid Grade Level, Flesch Reading Ease, and Dale-Chall. The Grade Level test focuses on the grade level needed in order to read the text. The purpose of our analysis is to examine how difficult the financial disclosures are for adults. Therefore, grade level is not within the scope of our research. Dale-Chall incorporates the percentage of difficult words in a sentence, along with the average length of the sentence [179]. As financial disclosures are not general text, we believe that this measure

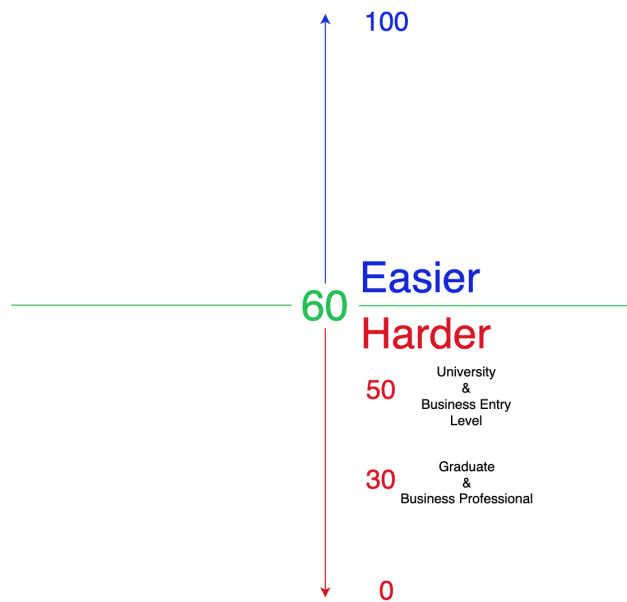


Figure 4.1: Flesch Reading Ease Scoring showing that 60 is the threshold between easy to read and hard to read.

would be biased, given the complexity of the subject matter.

The FRE scoring (see Figure 4.1) starts at 100 (very easy to read) and decreases as the text becomes more difficult; 60 is the threshold that divides easy to read and difficult to read [218]. Our focus was on texts which scored below 60, as that indicates that the text is no longer easy to read [218]. Texts that are below 50 are assessed as “university level”, and those below 30 are considered only readable by university graduates [218]. Given that some very well-known and successful entrepreneurs have no university degree (e.g. Bill Gates, Mark Zuckerberg, Steve Jobs, and Richard Branson [332]), we expanded this interpretation to indicate that scores below 50 are business entry-level, and scores below 30 are business professional-level, where professionals have several years of experience.

For both research experiments, we were interested in determining if ChatGPT increases the readability of financial text, with special attention paid to any ChatGPT-generated text that meets or exceeds the 60 threshold. For example, if Bank A’s human-authored explanation for revenue scores 45 (using the FRE) and ChatGPT’s score for the same explanation is 52, then this shows that using ChatGPT has improved readability (provided that ChatGPT’s explanation is correct). If, however, ChatGPT’s score is 60 (or above) for the same text (again assuming the explanation is correct), then that is of particular interest as it has

moved the text score from being difficult to read to easy to read, thereby making it accessible to a much larger group of users.

The equation for the Flesch Reading Ease (FRE) score is as follows:

$$\text{FRE} = 206.835 - 1.1015 \times \frac{\text{total words}}{\text{total sentences}} - 84.6 \times \frac{\text{total syllables}}{\text{total words}} \quad (4.2)$$

4.2.2.7 Statistical Significance Tests

As our data was non-parametric, we used Wilcoxon Signed Rank (WSR) to evaluate if the change in Cosine Similarity and Readability were statistically significant. In addition to the non-parametric requirement, WSR should be used when the “same subjects are evaluated under two different conditions” [280]. For our research, the same prompts and banks are being used between models (or conditions), meaning that WSR is the appropriate statistical measure. WSR was used to test statistical significance for both Cosine Similarity and Readability. For each, the null hypothesis was that there was no difference in the medians of the populations whereas under the alternative hypothesis, there was a difference. WSR was used to evaluate the following experiments:

1. Cosine Similarity between the human-authored text and the *davinci* base model, as well as the human-authored text and the fine-tune trained models. This was done for both the aggregate and each bank’s model.
2. Flesch Reading Ease Score between the human-authored text and the *davinci* base model and the aggregated fine-tune trained models for each bank.
3. Flesch Reading Ease Scores comparing the *davinci* base model and fine-tune models for the aggregate and each bank’s models.

As the human-authored text is “ground-truth” in all situations and comes from the published annual reports, no statistical evaluation of human vs. human text was done.

4.2.3 Experiments

4.2.3.1 Base Model Testing

Before conducting any fine-tuning, we tested the base model using the 60 sample prompts discussed in section 4.2.2.4, and passed them to the *davinci* model using OpenAI’s GPT-3 playground [238]. The resulting texts were then individually captured in plain text files⁴ to create the baseline dataset.

4.2.3.2 Fine-Tune Training

The *davinci* model was fine-tune trained using a total of 439 prompt and completion pairs. The “ground-truth” text was gathered from the Financial Summary section of each bank’s annual report, as discussed above in section 3.1.

Fine-tune training was done in stages in order to capture the different nuances of the training. To begin, the *davinci* base model was fine-tuned for each bank separately. This resulted in 5 *davinci* models — one for each bank. This was done to test how each bank’s fine-tuned model affected the text generation for each specific bank. Fine-tuning was also done for the aggregate.

4.2.4 Results and Discussion

Overall, fine-tune training shows an increase in the Cosine Similarity scores for all models (i.e., for each individual bank’s model as well as the aggregate model). However, when comparing the ground-truth text to the ChatGPT-generated text, we find that the Cosine Similarity is low when compared to the 0.8 threshold (see Section 4.2.2.3 for information on the threshold). Table 4.2 provides the high-level results.

There is an increase in the mean of the Cosine Similarity results from 0.269 (base model) to 0.504 (fine-tuned model) when using the aggregate model. This supports that aggregate fine-tune training containing prompt and completion pairs across banks, rather than just for one specific bank, can be used for fine-tune training. However, the larger gains are seen when the fine-tune training is done on a “per bank” basis, as the Cosine mean increases from 0.293 (base model) to 0.587 (fine-tune trained) overall when all banks are totalled. On

⁴UTF-8 encoding was used for each text file.

Table 4.2: Cosine Similarity Results

	Min	Max	Mean	Std Dev
Agg-Base	0.070	0.543	0.269	0.159
Agg-FT	0.134	0.921	0.504	0.230
JPM-Base	0.088	0.515	0.231	0.138
JPM-FT	0.030	0.918	0.518	0.323
BNY-Base	0.135	0.409	0.298	0.100
BNY-FT	0.362	0.790	0.602	0.138
Citi-Base	0.0074	0.589	0.318	0.158
Citi-FT	0.268	0.856	0.581	0.211
BOA-Base	0.108	0.460	0.341	0.112
BOA-FT	0.194	0.989	0.613	0.206
WF-Base	0.109	0.434	0.277	0.108
WF-FT	0.207	0.927	0.620	0.237
Per-Bank Total-Base	0.074	0.589	0.293	0.132
Per-Bank Total-FT	0.030	0.989	0.587	0.238

Note: “FT” denotes the fine-tuned models.

a “per-bank” basis, all of the means increased with fine-tuning, shown in Table 4.2. The largest increase was found for BNY. Bold font has been used to point out the means for the aggregated models, the total of the “per-bank” models, and BNY. Therefore, we conclude that, for our dataset, the better approach is to fine-tune train on a “per-bank” basis, rather than using an aggregate model.

When comparing the Cosine Similarity Scores to the benchmark of 0.8 (discussed in Section 4.2.2.3), we find that while none of the base model texts meet the threshold, seven of the fine-tune trained scores exceed 0.8. Of those seven, three are for WF, two are for JPM, and the aggregate and Citi both have one. This suggests that Wells Fargo’s texts are easier for the GPT-3 *davinci* model to use for training than that of other banks. This may be connected to the fine-tune training that brought WF’s sentiment much more inline with the human-authored sentiment, as compared to the base model, which is discussed in more detail below.

Although we followed previous work in identifying 0.8 as our threshold (discussed in Section 4.2.2.3), we did a side-by-side comparison for several texts that were over 0.60. Figures 4.2a and 4.2b shows one of these comparisons. In this case of the Citi side-by-sides, the Cosine Similarity of the Citi text was 0.65, meaning that it would not have met the

Citigroup reported net income for 2010 of \$10.6 billion, compared to a net loss of \$1.6 billion in 2009. Diluted EPS was \$0.35 per share in 2010 versus a loss of \$0.80 per share in 2009, and net revenues were \$86.6 billion in 2010, versus \$91.1 billion in 2009, on a comparable basis. On a reported basis, net interest revenue increased by \$5.7 billion, or 12%, to \$54.7 billion in 2010, generally as a result of the adoption of SFAS 166/167, partially offset by the continued run-off of higher-yielding assets in Citi Holdings and investments in lower-yielding securities. Non-interest revenues improved by approximately \$578 million, or 2%, to \$31.9 billion in 2010, primarily due to positive gross revenue marks in the Special Asset Pool in Citi Holdings of \$2.0 billion in 2010 versus negative revenue marks of \$4.6 billion in 2009, a \$11.1 billion gain in 2009 on the sale of Smith Barney, a \$1.4 billion pretax gain related to the public and private exchange offers consummated in July and September of 2009, and a \$10.1 billion pretax loss associated with the repayment of TARP and the exit from the loss-sharing agreement with the U.S. government in December 2009.

Citigroup's net income applicable to common shareholders for 2010 was \$10.6 billion, compared to a net income of -\$1.6 billion in 2009. This was driven by a decrease in non-interest expenses of \$13.2 billion, as well as an increase in net revenue of \$5.7 billion. Notable items in 2010 included a \$3.0 billion gain related to the sale of Smith Barney, a \$2.3 billion gain from the sale of a stake in Morgan Stanley Smith Barney, and a \$3.0 billion gain related to the sale of an insurance business. Additionally, Citigroup recorded a \$1.4 billion benefit from the sale of an equity investment in a joint venture.

(a) Citigroup Human-Authored

(b) Citigroup GPT-Generated

Figure 4.2: Side-by-Side comparison of text for Citigroup 2010

This figure shows a side-by-side comparison of the human-authored text and GPT-generated text in response to the prompt “Discuss Citigroup’s net income applicable to common shareholders and notable items for 2010”

threshold. However, in examining the text, we can see that there are a number of items that GPT got right (highlighted in gray), namely net income, the net loss (although GPT phrased it as “net income”, a negative net income is a net loss), net revenue was close (GPT discussed net revenue, whereas Citigroup discusses net interest revenue), and finally, the Smith Barney transaction (which was formerly called Morgan Stanley Smith Barney [220]). This demonstrates that 0.8 may be too constrictive, and that going forward, we may want to consider a slightly more relaxed benchmark.

In evaluating the statistical significance with WSR, there is a statistically significant difference (all have p-values less than 0.05) when comparing the medians of the Cosine Similarity of human-authored and GPT-3 base with human-authored and GPT fine-Tuned, for both the aggregate model and on a “per-bank” basis. The results can be found in Table 4.3.

In all cases we reject the null hypothesis (no difference in the medians of the populations due to treatments) and accept the alternative hypothesis (that there is a difference in the medians of the populations) due to treatments. If the null hypothesis were not rejected, it would mean that any differences in the populations would be due to chance, rather than the treatments (human vs. base and human vs. fine-tune). Therefore, as the results are statistically significant, we can conclude that there are consistent, systematic differences between the two treatments, resulting in a change in the medians.

	P-Value	Z-statistic
Agg	0.0371	2.0846
JPM	0.0137	2.4658
BNY	0.0020	3.0973
Citi	0.0137	2.4658
BOA	0.0020	3.0973
WF	0.0020	3.0973

Table 4.3: Wilcoxon Signed Rank test for Cosine

Table 4.4 provides the results from the Wilcoxon Signed Rank tests for Readability. “FT” denotes the fine-tuned models. Bold font indicates where there is a statistically significant change between conditions. Results show that human vs. base GPT and base GPT vs. fine-tuned GPT are statistically significant.

Comparisons	P-Value	Z-statistic
Human Vs. Base GPT	< 0.01	3.6007
Human Vs. JPM-FT	0.5703	-0.5676
Human Vs. BNY-FT	0.6250	-0.4888
Human Vs. CITI-FT	0.8457	0.1946
Human Vs. BOA-FT	0.2500	0.3834
Human Vs. WF-FT	0.7695	-0.293
Base GPT Vs. GPT-FT	< 0.01	-3.7165

Table 4.4: Wilcoxon Signed Rank test for Readability

In reviewing the results, in some cases ChatGPT is increasing the Readability, and in others, decreasing it. If clear and accessible communication is a company goal, then if AI (ChatGPT in this case) decreases the Readability scores, that raises two important questions: (1) is there still an advantage to using AI to generate text; and (2) is the AI-generated text capturing the correct message.

To answer the first question, this will depend on the company’s approach to communication. Many researchers have raised concerns regarding financial communication to shareholders and the use of obfuscation (e.g. [70, 105, 191]), particularly when the company faces tough times. While the SEC has a “plain English” requirement, which they define as

“[using] words economically and at a level the audience can understand” [286], there is obviously latitude that companies can take with this rule, while still abiding by it. Therefore, it will be up to each company’s core values as to how they approach Readability.

The second question is more straightforward, but requires domain expertise to assess whether the AI (ChatGPT in our research) is capturing the correct message. In looking back to our Citigroup example in Figures 4.2a and 4.2b, there is a difference in meaning and intention between net interest revenue versus net revenue, while there is only a word choice difference between net loss and net income showing a loss. Although side-by-side comparisons can be tedious and expensive, they do add value to the training process so we can understand why our results are being positively and negatively impacted.

Results show that while the Cosine Similarity increases after fine-tune training, the Readability scores decrease, which is concerning. This is also supported by the negative Z-statistic in Table 4.4. This suggests that in fine-tune training, we are increasing complexity, which decreases Readability and accessibility of the text. This will need to be further probed as part of future work.

Finally, sentiment was evaluated using FinBERT to determine how the base model and fine-tune training affect the tone of the financial discussions. Table 4.5 shows the results of the sentiment percentage scores for negative, neutral, and positive, for the different texts (human-authored, base *davinci*, and fine-tuned (FT))

Table 4.5: Sentiment Breakdown And Variance

	Neg %	Neut %	Pos %
Human	5.88	1.58	2.54
Base	2.04	4.57	3.39
FT	4.11	2.44	3.46
Total	1	1	1
Variance: Human to Base	3.84	-3.00	-0.84
Variance: Human to GPT-FT	1.77	-0.86	-0.92

With the exception of two fine-tunes (BNY and BOA), the overall sentiment variance between human and base is bigger than human and fine-tuned. From a sentiment and corporate messaging perspective, this shows the importance of fine-tune training, as it enables the model to better capture the true sentiment of the ground-truth (human-authored) text.

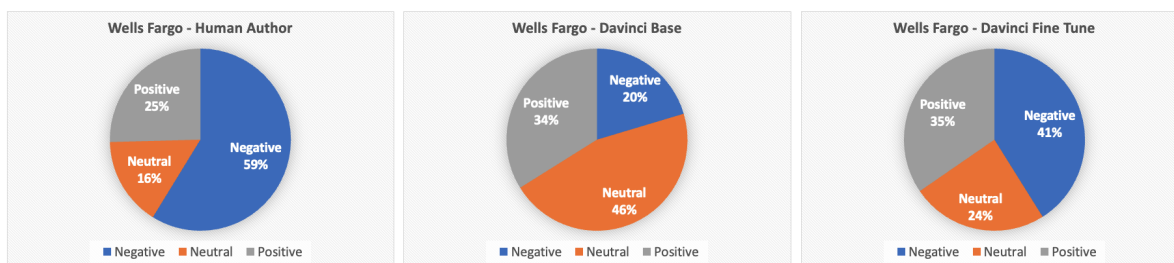


Figure 4.3: Sentiment Shifts Between Models For Wells Fargo

Sentiment shifts between models for Wells Fargo. On the far left, the ground-truth (human-authored) sentiment is shown. The middle pie chart shows the sentiment of the *davinci* base model, with the sentiment of the fine-tuned WF model on the far right.

Of the six models (aggregate, and five “per-bank”), Wells Fargo had the largest variance between human and base versus human and fine-tuned. Figure 4.3 shows the progression from the ground-truth (human-authored) to the fine-tuned model.

As can be seen in Figure 4.3, the negative sentiment experiences the biggest swings. In the ground-truth, it is 59%, which decreases to a mere 20% with the ChatGPT-generated text using the *davinci* base model. But, after fine-tune training, this is corrected and it increases to 42%, which is much closer to the original 59% than the base model’s 20%. Given the importance of sentiment to corporate communications, the sentiment aspect of any AI text generation task should be considered, as sub-optimal messaging released to the public can have important implications for any company.

Finally, we add our research to the body of literature that demonstrate that fine-tune training does not always guarantee a good outcome. We find that GPT-3 behaves in erratic and unexpected ways. As can be seen in Figure 4.4, after an experimental fine-tune training run, GPT-3 returned a paragraph full of questions, rather than responding to the prompt “Discuss the change in leadership between 2006 and 2007 for JPMorgan Chase”. We were unable to replicate this type of response to help us determine why GPT-3 would have returned a paragraph of questions. However, in examining the paragraph, it appears that this response may be based on a textbook chapter entitled “Chapter 10: JPMORGAN CHASE & Co. INTERNAL CONTROLS AND LEGAL RISKS”. Therefore, we theorize that our prompt is close to or identical to that used for this textbook.

Another interesting quirk that we encountered was GPT-3 generating text in Thai in

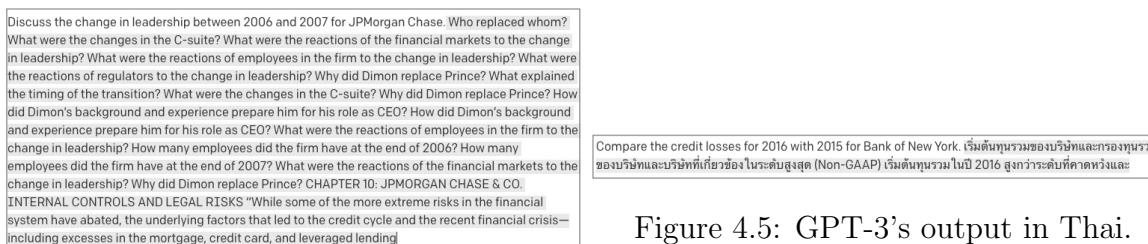


Figure 4.5: GPT-3's output in Thai.

Figure 4.4: GPT-3's output returning a paragraph of questions.

response to an English prompt (See Figure 4.5). We can tell by the English words “(Non-GAAP)” in the Thai sentence that the text is related to the financial domain. However, we do not know if GPT-3 answered the prompt correctly or not, as we do not read Thai. Again, we were not able to replicate this issue on subsequent training runs, and can offer no plausible explanation for this output.

4.2.5 Conclusion and Future Work

Crafting and curating the 10-K report is a very time consuming job, and is very text-heavy. It remains unclear, though, as to whether AI-generated text can (or should) be used without human review [107], given the potential consequences and market effects if incorrect financial text is publicly released. However, our research has shown that with a little fine-tune training, GPT-3 can be extremely useful to help draft such a laborious document.

In this research, we have provided an in-depth and thoughtful analysis of GPT-3's abilities to generate financial text for five major US banks. We focused on text gathered from the Financial Summary section of the bank's annual reports in order to benchmark GPT's performance using the base *davinci* model as well as fine-tuned trained models. We find that while Cosine Similarity is stronger and the sentiment of the text is better emulated in the fine-tune model text, Readability decreases. These results show an interesting trade-off, where accessibility appears to suffer for gains in similarity and sentiment. This has important implications for research going forward, as more analysis is needed to determine to what degree of trade-off is acceptable in order to maximize the fidelity and effectiveness of AI-generated text. This research could also help provide additional characteristics to help detect fraudulent generated text in real-time.

As part of future work, we would like to expand our sample set to include banks from Canada and Europe. We believe that this would provide insight into how the characteristics of Similarity, Sentiment, and Readability all work together in fine-tune training models for financial text. We also plan to further probe how AI-generated text increases complexity to decrease Readability and accessibility.

4.2.6 Limitations and Ethical Considerations

There are some important limitations and impacts that should be kept in mind for this research. First, while we used data gathered over 15 years, the number of banks that the text was retrieved from was small ($n = 5$). Therefore, GPT-3's ability to generate text may be improved if a different set of banks was used. Also, OpenAI has not provided details on the full extent of what GPT-3 was trained on. Therefore, it may be that banks in our sample set were either included or excluded from GPT-3's initial training. This could have a material effect on GPT-3's fine-tuning, as some banks may have a better "starting point" than others, depending on if they were included in *davinci's* training or not.

4.3 Project 4: Using ChatGPT to Generate Financial Disclosure Notes

4.3.1 Introduction

The release of ChatGPT has raised both interest and concern over its human-like text generation. Given its utility and range of generation abilities, this type of tool would be very useful in the creation of long regulatory documents such as the annual report for public companies, which is, on average, 186 pages [128]. While ChatGPT is widely trained and has shown its abilities to emulate styles and provide realistic answers to requests (with some error), this tool is still in its infancy. As such, we anticipate that the efficacy of this tool will evolve over time. This progression is of interest to both the scientific and professional communities. The purpose of this research is to examine and benchmark ChatGPT's current ability to generate the notes to the financial statements for public companies. This is important for several reasons. Textual financial statement data is scarce and the utility of ChatGPT to [augment](#) financial data should be evaluated. Also, ChatGPT's ability to emulate could be used to produce fraudulent or incorrect financial statement data, which may be difficult to

detect in time-sensitive situations. Therefore, the type of research and evaluation that we have presented in this paper can help address these concerns.

To conduct this research, we chose two note disclosures from the Annual Report for ChatGPT to generate for Canada’s five major banks over the period of 2019–2021. We find that when ChatGPT’s text is compared to that of human authors, the similarity is very low, with the highest at 0.42 for disclosure 1, and 0.32 for disclosure 2. We also find that not all texts are easily identifiable as either human-authored or Artificial Intelligence (AI)-authored. This was unexpected, but also revealed that there is more “boilerplate” text in the disclosure notes than expected. Finally, results also show that Readability is significantly improved at a statistical significance level of $p < 0.05$ for disclosure 1 when using ChatGPT. This also opens up an interesting avenue of research on how AI tools like ChatGPT can be used to help improve Readability in order to make complicated and difficult financial text more accessible to a wider population. This research was accepted, presented, and published as part of the Proceedings for Canadian AI 2023.

4.3.2 Background

In Canada, public companies are required to use the International Financial Reporting Standards (IFRS) when preparing interim and annual financial statements [119]. A critical part to the financial statements are the note disclosures, often referred to simply as “notes” (refer to Figure 1.2 for more information). These disclosures provide critical additional information about items recognized in the financial statements [50, 167], as well as those that are not [148].⁵ Disclosures may be required by GAAP or other regulations, while other disclosures may be provided by management to facilitate user understanding [148]; there is no “one size fits all” approach to disclosures.⁶

⁵An example of an item that is disclosed but not recognized in the financial statements is a contingent asset (IAS 37) when it is more likely than not that a cash inflow or benefit will be received. If considered virtually certain, however, then it would be reported as an asset as it “is no longer considered to be contingent” [104].

⁶Per International Accounting Standard (IAS) 1, “[t]he notes must present information about the basis of preparation of the financial statements and the specific accounting policies used, disclose any information required by IFRS that is not presented elsewhere in the financial statements and, provide additional information that is not presented elsewhere in the financial statements but is relevant to an understanding of any of them” [72].

4.3.3 Methodology

4.3.3.1 Dataset

The Annual Reports (ARs) were gathered from the System for Electronic Document Analysis and Retrieval (SEDAR)⁷ for three years over the period of 2019 – 2021 for Canada’s five major banks:

- Royal Bank of Canada (RBC)
- Bank of Nova Scotia (BNS)
- Canadian Imperial Bank of Commerce (CIBC)
- TD Bank (TD)
- Bank of Montreal (BMO)

Using the listing of the notes to the financial statements, we selected two financial statement note disclosures that the five banks had in common. We took this approach as many disclosures are specific to the company’s financial statements and may be different to those of another company. The note disclosures that were selected are: (1) Basis of Preparation and (2) Subordinated Debt. Both are required note disclosures under IFRS, specifically IAS 1 and IFRS 7 [72, 73]. They are referred to as “disclosure 1” and “disclosure 2”. For reproducibility purposes, Table 1 provides where each of these note disclosures can be found in each report. We also note that while disclosures may have a slightly different name between companies (e.g. *General Information* (RBC [274]) versus *Basis for Preparation* (CIBC [42])), the intent and information provided in the disclosures are substantially the same. This naming difference is also clearly outlined in Table 4.6.

The following questions were given to ChatGPT for generation:

For disclosure 1: “Write the <name of the disclosure> note disclosure for <insert bank name here> for the <insert year> Annual Report.” As discussed above, there is some variability in the name of this disclosure. For example, the Royal Bank uses “General Information” while CIBC uses “Basis of Preparation”. To account for this variability in

⁷SEDAR is the official regulatory filing repository for the Canadian Securities Administrators [44].

Table 4.6: Disclosure References by Bank and Year

Disclosure Name	Bank	Year (Page)
1: General Information	RBC	2019 (125), 2020 (132), 2021 (138)
1: Basis for Preparation	CIBC	2019 (108), 2020 (114), 2021 (122)
1: Statement of Compliance	BNS	2019 (148), 2020 (160), 2021 (158)
1: Nature of Operations	TD	2019 (132), 2020 (137), 2021 (139)
1: Basis of Presentation	BMO	2019 (142), 2020 (150), 2021 (151)
2: Subordinated Debentures	RBC	2019 (190), 2020 (200), 2021 (205)
2: Subordinated Indebtedness	CIBC	2019 (159), 2020 (162), 2021 (168)
2: Subordinated Debentures	BNS	2019 (204), 2020 (213), 2021 (212)
2: Subordinated Notes & Debentures	TD	2019 (188), 2020 (192), 2021 (188)
2: Subordinated Debt	BMO	2019 (176), 2020 (183), 2021 (616)

naming convention, the proper names (and casings) found in the bank’s [AR](#) were used. We did this to ensure that we queried ChatGPT using the most accurate information for the question in relation to each bank.

As an example, the question for RBC for 2019 was: “Write the *General Information* note disclosure for the Royal Bank of Canada for the 2019 Annual Report”, whereas the question for CIBC for 2019 was “Write the *Basis of Preparation* note disclosure for the Canadian Imperial Bank of Commerce for the 2019 Annual Report.”

For disclosure 2: “Write the note disclosure for <name of the note disclosure> for <insert bank name here> <insert year here> Annual Report, including a table of the debentures and any relevant footnotes.” We formulated the question this way as the note disclosures in the annual report contain a table and footnotes. We tried asking the question without mentioning the table and the footnotes, and ChatGPT would not include these elements in its generation without specific direction to do so.

Like disclosure 1, there is some variability in the naming convention such as “Subordinated Notes and Debentures” or “Subordinated Debt”, for example. And so we followed our approach from disclosure 1 and used the name of the disclosure (and its casing) as it was presented in the [AR](#). As an example, the question for TD Bank for 2019 was “Write the note disclosure for *Subordinated Notes and Debentures* for TD Bank’s 2019 Annual Report, including a table of the debentures and any relevant footnotes.”, whereas the question for BMO for 2019 was “Write the note disclosure for *Subordinated debt* for the Bank of Montreal’s

2019 Annual Report, including a table of the debentures and any relevant footnotes.”

While querying ChatGPT for the tables may seem extraneous given that they were ultimately removed (discussed below), we needed the tables to be generated in order to capture the footnotes. Also, we wanted to ensure that we used the exact same protocol for the human-authored text and the AI-generated text by ChatGPT in order to address any concerns of bias in the question posed to ChatGPT as well as any concern that we were not comparing truly “like” text.

The text of the note disclosures were then extracted from each report to be used for evaluation purposes. Tables for disclosure 2 were removed from both the bank’s annual report text, as well as ChatGPT’s text. Any footnotes, however, were retained, as any text outside of the tables should be included.

To assess the original and generated texts, two measures of similarity were used (Cosine and Jaccard); a stylometry analysis was conducted using Burrows’ Delta, and Readability was evaluated using the Flesch Reading Ease Score (FRE).

4.3.3.2 Cosine Similarity

As Cosine Similarity is the standard for determining similarity in the domain of Information Retrieval, we continued to use it in this project. Please see Section 4.2.2.3 above for information on this metric.

4.3.3.3 Jaccard Similarity

Jaccard similarity is another well-known measure that evaluates the similarity of two sets, U and V . Like Cosine Similarity between vectors with positive components (i.e., in the first quadrant), the range of Jaccard similarity is between 0 and 1, where 0 indicates that there is no overlap, and 1 indicates there is complete overlap. Therefore, if the result is closer to 0, it is said to be dissimilar, whereas results closer to 1 indicate a high similarity between the sets [16, 333]. The formula for the Jaccard similarity is

$$Jaccard(U, V) = \frac{|U \cap V|}{|U \cup V|} \quad (4.3)$$

4.3.3.4 Burrows’ Delta

Stylometry is the study of writing style where an author’s stylistic features which “includes lexical, syntactic, content, structural, and idiosyncratic features” [69]. This is predicated on the idea that no two authors will be exactly the same in the way that they use vocabulary, write phrases, or use punctuation, for example [178]. The most common task that stylometry has been applied to is authorship attribution [69, 222]. One of the main methods in stylometry is Burrows’ Delta, which has been shown to be very effective in authorship attribution tasks — even across translations where it can distinctively identify the original author from the translator [275]. Therefore, we follow the literature and use Burrows’ Delta for our experiments.

To calculate the Delta, a Z-score statistic must first be calculated, shown in Equation 4.4.

$$Z_i = \frac{C_i - \mu_i}{\sigma_i} \quad (4.4)$$

where z_i is the z-score for feature i , C_i is the observed frequency for feature i , μ_i is the mean of the means for feature i , and σ_i is the standard deviation for feature i [178]. The purpose of the Z-statistic is to normalize the word frequencies so that the “mean for each word is 0 and the standard deviation is 1” [96]. This is an important step because as per Zipf’s law [354, 355, 254], “The r th most frequent word has a frequency $f(r)$ that scales according to [Equation 4.5] for $\alpha \approx 1$ ” [251].⁸

$$f(r) \propto \frac{1}{r^\alpha} \quad (4.5)$$

where r is the frequency rank of the word, $f(r)$ is the frequency of the word in the corpus, and $\alpha \approx 1$ [251].

Therefore, the influence of the top-scoring words is reduced by first normalizing using the z-score. Otherwise, “the distance would barely be influenced by anything but a few top-scoring words” [96].

The Delta is then calculated, as seen in Equation 4.6.

⁸In a given corpus, the most frequent word appears approximately 2x as often as the second most frequent, 3x as often as the third most frequent word etc., as this will depend on the actual observed frequency of the words in the corpus. Therefore, Zipf’s law “states frequencies proportionally” [251].

$$\Delta_c = \sum_i \frac{|Z_{c(i)} - Z_{t(i)}|}{n} \quad (4.6)$$

where c is the disputed document, i is the test document, $Z_{c(i)}$ is the Z-score of word frequencies i in the disputed document, $Z_{t(i)}$ is the Z-score of the word frequencies of the test document, and n is the most frequent words in the corpus to be used as features [178]. We restrict n to the 30 most frequent words to be used as features, as we have determined, through experimentation, that this is the least number of words that can be used to achieve good results using Burrows’ Delta.

We note that previous research has raised the concern that the word frequency Z-scores of Burrows’ Delta are calculated using Manhattan distance (i.e., “the sum of the absolute differences between the two vectors” [33]), it uses mean and standard deviation for normalization, which would only make sense with a Gaussian distribution [14]. To address this “mismatch” between distance and normalization, two fixes were suggested — to use a Laplacian distribution (i.e., a symmetrical distribution which has a higher peak than a Gaussian distribution [226]) when normalizing (which was termed *Linear delta*), or use Euclidean distance (i.e., “square root of the sum of the squared differences between the two vectors” [33]) —rather than Manhattan — when normalizing (which was termed *Quadratic delta*) [14, 96]. However, related research found that a normal distribution (i.e. Gaussian) represented the data much better for their English language corpus [96]. Given that our datasets are in English, we continue to use the Delta as laid out by Burrows.

4.3.3.5 Flesch Reading Ease Score

For this research, we continue to use the FRE to assess Readability outlined in Section 4.2.2.6.

4.3.4 Experiments

Disclosures were grouped into disclosure 1 and disclosure 2 (as per Table 4.6), and any indication of the text author (company) was removed from the documents. For ease of passing the documents to the algorithm, documents were named as “disc1_text1”, “disc1_text2”, ... (for disclosure 1) and “disc2_text1”, “disc2_text2”, ... (for disclosure 2). For each

Table 4.7: Results for Cosine Similarity.

Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY	Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY
Disclosure 1				Disclosure 2			
RBC, 2019	0.23	N/A	N/A	RBC, 2019	0.30	N/A	N/A
RBC, 2020	0.23	0.97	0.88	RBC, 2020	0.31	0.95	0.97
RBC, 2021	0.19	0.96	0.91	RBC, 2021	0.31	0.92	0.97
CIBC, 2019	0.31	N/A	N/A	CIBC, 2019	0.29	N/A	N/A
CIBC, 2020	0.28	0.87	0.93	CIBC, 2020	0.28	0.94	0.99
CIBC, 2021	0.27	0.94	0.96	CIBC, 2021	0.29	0.94	0.99
BNS, 2019	0.42	N/A	N/A	BNS, 2019	0.29	N/A	N/A
BNS, 2020	0.42	0.95	0.99	BNS, 2020	0.32	0.79	0.99
BNS, 2021	0.42	0.95	0.99	BNS, 2021	0.30	0.99	0.97
TD, 2019	0.40	N/A	N/A	TD, 2019	0.30	N/A	N/A
TD, 2020	0.35	0.95	0.99	TD, 2020	0.32	0.79	0.99
TD, 2021	0.35	0.99	0.99	TD 2021	0.26	0.95	0.99
BMO, 2019	0.39	N/A	N/A	BMO, 2019	0.25	N/A	N/A
BMO, 2020	0.40	0.98	0.99	BMO, 2020	0.25	0.94	0.99
BMO, 2021	0.40	0.99	0.99	BMO, 2021	0.26	0.95	0.99
Mean	0.34	0.96	0.96	Mean	0.29	0.90	0.98
Median	0.35	0.96	0.99	Median	0.29	0.94	0.99
Std Dev	0.08	0.03	0.04	Std Dev	0.02	0.10	0.01

disclosure, there were a total of 30 texts — 15 that were human-authored and 15 that were generated by ChatGPT. Eight texts from each were randomly selected as the training set and labelled as either “human” or “GPT”. The remaining 14 documents were then labelled as “disputed” and made available for testing. Burrows’ Delta was used to identify the likely author of the disputed documents. We randomly selected four test documents from the disputed label (documents 22, 11, 1, and 30) for each disclosure and calculated the Delta for each.

4.3.5 Results and Discussion

4.3.5.1 Cosine Similarity and Jaccard Similarity

The results of the Cosine Similarity and Jaccard Similarity are found in Table 4.7 and Table 4.8, respectively. As expected, the similarity between the human-authored text and the ChatGPT generated text are very low using both measures.

The results indicate that the similarity never reaches 0.50 for either disclosure under both measures. The highest similarity is 0.42 and 0.32 for disclosures 1 and 2, respectively, using Cosine Similarity. The results for Jaccard are even lower, with the highest Similarity at 0.20 for disclosure 1 and 0.15 for disclosure 2. This supports that ChatGPT is not yet able to capture the bank’s *voice* when writing these disclosures, even for a straightforward

disclosure such as *The Basis of Preparation*. We do note, though, that similarity scores for disclosure 1 are higher than those for disclosure 2, indicating that ChatGPT’s performance is better for disclosure 1 than 2.

Another interesting finding from these results is that, like ChatGPT, the human-authored text also takes a very “boilerplate approach” for both disclosures as indicated by the high degree of similarity year-over-year (YoY). This raises an important question — how many times does a text have to be used before it is considered boilerplate? Current definitions suggests that text need only be used a few times before it is considered “boilerplate” as the Merriam-Webster dictionary defines boilerplate as “standardized text” [209] and the Cambridge Dictionary as “text that can be copied and used in legal documents or in computer programs, with only very small changes” [77]. We note that there is no indication from either as to the number of times that text can be reused before it is considered to be “boilerplate”.⁹ Previous research has indicated that business documents tend to reuse text [283]. Given that ARs are very long, the analysis that is done in our work here opens up a new avenue of research in examining important questions like “Are investors more or less likely to miss key information if disclosures take a boilerplate approach?” or, “Is there a better way to communicate the changes from year-to-year of a (mostly) boilerplate report to investors?”

As well, interest in this area has been growing over the past five years, with research looking at boilerplate detection itself, but also targeted removal of boilerplate from either documents or web corpora [279, 181, 27]. Given what we have shown in our results, removal of boilerplate could have unintended consequences, resulting in the removal of large swaths of text from regulatory documents or corpora.

4.3.5.2 Burrows’ Delta

The results of Burrows’ Delta are found in Table 4.9. In this table, the lowest score between human, GPT, and the disputed category indicates the likely author, shown in bold font. If the lowest score points to the disputed category, it means that the Delta is having trouble distinguishing between human and GPT.

Documents were blinded when given to the Burrows’ Delta calculation. For reference,

⁹We note that in two of the three jurisdictions researched in this thesis — Canada, the United States, and Europe — that the AR is a legal document in the US and the European Union [325, 95] but not in Canada (though it is customary to do so) [315].

Table 4.8: Results for Jaccard Similarity

Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY	Bank, Year	Human Vs. GPT	Human YoY	ChatGPT YoY
Disclosure 1				Disclosure 2			
RBC, 2019	0.16	N/A	N/A	RBC, 2019	0.14	N/A	N/A
RBC, 2020	0.15	0.95	0.74	RBC, 2020	0.13	0.89	0.95
RBC, 2021	0.15	0.92	0.81	RBC, 2021	0.14	0.84	0.94
CIBC, 2019	0.15	N/A	N/A	CIBC, 2019	0.13	N/A	N/A
CIBC, 2020	0.14	0.67	0.84	CIBC, 2020	0.13	0.84	0.97
CIBC, 2021	0.13	0.76	0.91	CIBC, 2021	0.12	0.85	0.97
BNS, 2019	0.20	N/A	N/A	BNS, 2019	0.13	N/A	N/A
BNS, 2020	0.20	0.86	0.98	BNS, 2020	0.13	0.90	0.94
BNS, 2021	0.20	0.88	0.98	BNS, 2021	0.14	0.96	0.95
TD, 2019	0.19	N/A	N/A	TD, 2019	0.14	N/A	N/A
TD, 2020	0.17	0.86	0.97	TD, 2020	0.13	0.59	0.96
TD, 2021	0.17	0.92	0.99	TD 2021	0.15	0.39	0.96
BMO, 2019	0.17	N/A	N/A	BMO, 2019	0.11	N/A	N/A
BMO, 2020	0.18	0.86	0.98	BMO, 2020	0.11	0.87	0.96
BMO, 2021	0.18	0.97	0.98	BMO, 2021	0.11	0.87	0.96
Mean	0.17	0.86	0.92	Mean	0.13	0.80	0.95
Median	0.17	0.87	0.98	Median	0.13	0.86	0.96
Std Dev	0.02	0.09	0.09	Std Dev	0.01	0.17	0.01

Table 4.9: The Results of Burrows' Delta

Disclosure	Test Document	Human	GPT	Disputed
Disclosure 1	22	3.65	2.90	3.11
Disclosure 1	11	2.42	2.97	2.36
Disclosure 1	1	2.33	3.55	3.04
Disclosure 1	30	3.88	2.82	3.14
Disclosure 2	22	1.59	0.19	1.43
Disclosure 2	11	1.79	2.11	1.33
Disclosure 2	1	1.24	2.36	1.54
Disclosure 2	30	2.32	1.35	2.10

documents 1–15 are human-authored and documents 16–30 are AI-generated (ChatGPT). Therefore, when we reviewed the four test cases (8 documents total for each disclosure) Burrows' Delta was able to correctly identify 6 documents — for documents 22, 1, and 30 for both disclosure 1 and disclosure 2. The results did identify an interesting test case, however, for test document 11. This document is more similar to the stylometry found in the disputed documents, which is a mix of both human and AI authored texts. We found this very interesting and using our master key list of all documents and authors, we determined that the text for document 11 for both disclosure 1 and disclosure 2 were authored by TD Bank. So, we extended our testing to evaluate all of TD's texts. Using the same blinded

protocol, we calculated Burrows' Delta for documents 11 (included in the original test set), as well as documents 10 and 12 for both disclosures. The results were the same — these documents were more similar to the disputed category. We did notice that the Cosine results show an incredibly high similarity YoY for these disclosures for TD Bank (range of 0.94 to 0.99), suggesting a “boilerplate” approach, which is also found in the AI-generated disclosures of ChatGPT. This may be why the Delta cannot fully distinguish TD's disclosures as either human or GPT.

4.3.5.3 Flesch Reading Ease

The results from the FRE evaluation can be found in Table 4.10. As discussed in the methodology section, we are interested in texts that are below 60 as those texts are considered difficult to read. Overall, we find that the ChatGPT texts are slightly easier to read as those texts have a mean of 25.22 and 45.88 for disclosures 1 and 2 respectively, whereas the human-authored texts have a mean of 24.29 and 42.08. This has important implications for the readability of the annual report. While ChatGPT's text is currently lacking in similarity for both disclosures, the increase in the Readability is a good outcome. Keeping in mind that readability below 30 is considered to be university grad/professional-level, the majority of the human-authored texts fall in that category, primarily for disclosure 1. ChatGPT's text is more accessible and measured at university level/business entry-level, and get very close to the main threshold of 60 for disclosure 2, which is the very beginning of the “difficult to read” category.

We also compared the statistical significance of the standard deviations using the F-test. The difference of standard deviations between the human-authored and ChatGPT-generated texts for disclosure 1 is statistically significant with a p-value of 0.02 and an F-statistic of 3.66, at a significance level of 0.05. The difference for Disclosure 2 was not statistically significant, with a p-value of 0.4991 and an F-statistic of 1.446.

4.3.5.4 Side-By-Side Comparison of Text

While Similarity, Stylometry, and Readability can provide a lot of information, it is also important to do a side-by-side comparison of the texts to see what ChatGPT emulates well and what it does not. It is also key to see what textual components are left out or added in

Table 4.10: Results for Flesch Reading Ease

Doc - Human	Disclosure 1	Disclosure 2	Doc - GPT	Disclosure 1	Disclosure 2
text1	23.53	47.81	text16	32.70	41.14
text2	26.52	51.81	text17	33.63	41.60
text3	22.87	52.17	text18	33.63	40.05
text4	19.01	43.98	text19	19.84	41.21
text5	22.25	43.70	text20	20.60	41.21
text6	25.95	46.37	text21	21.54	41.21
text7	39.65	34.56	text22	21.78	40.93
text8	39.95	32.31	text23	21.78	43.01
text9	39.95	31.89	text24	21.78	43.01
text10	15.50	29.33	text25	25.53	48.43
text11	14.00	42.28	text26	25.53	48.43
text12	14.00	35.38	text27	25.53	48.43
text13	21.02	46.84	text28	24.84	56.50
text14	21.58	45.46	text29	24.84	56.60
text15	21.58	47.41	text30	24.84	56.50
Mean	24.49	42.08	Mean	25.22	45.88
Median	22.25	43.89	Median	24.84	43.01
Std Dev	8.79	7.46	Std Dev	4.60	6.21

by ChatGPT.

For disclosure 1 — Basis of Presentation — the Bank of Montreal’s 2019 note disclosure is provided alongside ChatGPT’s generated text in Figure 1, where (a) presents the note disclosure as written by BMO and (b) provides ChatGPT’s generated note disclosure for BMO. Boxes have been added to draw attention to missing information as well as additional information that is not normally provided as part of the disclosure. It is clear that ChatGPT “understood” the query, as it has begun the note disclosure with “Basis of Presentation”. It did not provide all of the required information as this note disclosure is usually several pages long and discusses the accounting policies used in much more detail. As such, we truncated the bank’s original text to include only the general information when analyzing the similarity, stylometry, and readability in an effort to compare “like” text based on ChatGPT’s limitations.

ChatGPT does capture a number of relevant information points such as bank name, what products and services the bank offers, and that the statements have been prepared under IFRS. It does miss some important information such as the bank’s charter, the fact that it is a public company, the bank’s head office address, that it is traded on the Toronto Stock Exchange (TSX) and New York Stock Exchange (NYSE), and that the consolidated financial statements were authorized for issue by the Board (along with the date of authorization). (See blue boxes in figure (a)).

We also noticed that ChatGPT included text that does not appear in any of the five banks’ “Basis” note disclosures — a boilerplate advisory that the disclosure is not a substitute for reading the bank’s Annual Report in its entirety (even though the note disclosure is *in* the [AR](#)), and a “For more information” notice that includes the bank’s website. We found this very interesting as ChatGPT has added this in. Given that the number of Canadian [ARs](#) is far less ubiquitous than American [ARs](#), we consulted the United States’ largest bank JPMorgan Chase & Co and reviewed its “Basis of Presentation” note disclosure [160]. This extra information was not present in JPMorgan’s note disclosure for 2017–2021 either. This raises an interesting question as to why ChatGPT is providing the boilerplate advisory (when the note is part of the [AR](#)), and why it is including the website, when this information is not normally provided as part of this disclosure.

ChatGPT’s performance for disclosure 2 was a good attempt, but it only provided “dummy” information, and generated the same dummy information for every bank, when

NOTES TO CONSOLIDATED FINANCIAL STATEMENTS

Note 1: Basis of Presentation

Bank of Montreal ("the bank") is a chartered bank under the *Bank Act (Canada)* and is a public company incorporated in Canada. We are a highly diversified financial services company, providing a broad range of personal and commercial banking, wealth management and investment banking products and services. The bank's head office is at 129 rue Saint-Jacques, Montreal, Quebec. Our executive offices are at 100 King Street West, 1 First Canadian Place, Toronto, Ontario. Our common shares are listed on the Toronto Stock Exchange ("TSX") and the New York Stock Exchange.

We have prepared these consolidated financial statements in accordance with International Financial Reporting Standards ("IFRS") as issued by the International Accounting Standards Board ("IASB"). We also comply with interpretations of IFRS by our regulator, the Office of the Superintendent of Financial Institutions Canada ("OSFI").

Our consolidated financial statements have been prepared on a historic cost basis, except for the revaluation of the following items: assets and liabilities held for trading; financial assets and liabilities measured or designated at fair value through profit or loss ("FVTPL"); financial assets measured or designated at fair value through other comprehensive income ("FVOCI"); financial assets and financial liabilities designated as hedged items in qualifying fair value hedge relationships; cash-settled share-based payment liabilities; defined benefit pension and other employee future benefit liabilities; and insurance-related liabilities.

These consolidated financial statements were authorized for issue by the Board of Directors on December 3, 2019.

(a)



Write the Basis of Preparation note disclosure for the Bank of Montreal for the 2019 Annual Report.



Basis of Preparation Note Disclosure for Bank of Montreal
2019 Annual Report



The Bank of Montreal (BMO) is a leading North American financial institution that offers a range of financial products and services to personal, commercial, corporate, and institutional customers in Canada, the United States, and around the world.

The financial statements included in this 2019 Annual Report have been prepared in accordance with International Financial Reporting Standards (IFRS) as issued by the International Accounting Standards Board (IASB). The financial statements have been prepared on a going concern basis, which assumes that BMO will continue in operational existence for the foreseeable future and will be able to realize its assets and discharge its liabilities in the normal course of business.

BMO's functional and presentation currency is the Canadian dollar. The financial statements have been prepared on the historical cost basis, except for certain financial instruments, such as investments in available-for-sale financial assets and derivatives, which are measured at fair value.

This basis of preparation disclosure is provided to give an overview of BMO's activities and is not a substitute for reading the bank's 2019 Annual Report in its entirety. The 2019 Annual Report includes additional information about BMO's financial results, business activities, risk factors, and other important disclosures.

For more information about BMO, please visit the bank's website at www.bmo.com.

(b)

Figure 4.6: Comparison of Note Disclosure 1. (a) BMO's Basis of Presentation note disclosure. The Text highlighted shows the information that ChatGPT did not capture in its generation. (b) ChatGPT's Basis of Presentation note disclosure for BMO. The text highlighted in gray shows the information that ChatGPT included in its generation which is not normally provided as part of this disclosure.

Note 15 | Subordinated indebtedness

The debt issues included in the table below are outstanding unsecured obligations of CIBC and its subsidiaries and are subordinated to the claims of depositors and other creditors as set out in their terms. Foreign currency denominated indebtedness funds foreign currency denominated assets. All redemptions are subject to regulatory approval.

Terms of subordinated indebtedness

Interest rate %	Earliest date redeemable		Par value	Carrying value ⁽²⁾	Par value	Carrying value ⁽²⁾	Denominated in foreign currency	Par value	Carrying value ⁽²⁾
	Contractual maturity date	At greater of Canada Yield Price ⁽¹⁾ and par							
5.75 ⁽³⁾	July 11, 2024 ⁽⁴⁾	January 26, 2021 ⁽⁷⁾	\$ 32	\$ 32	\$ 35	\$ 35	TT\$175 million	\$ 35	\$ 35
3.42 ⁽⁵⁾⁽⁶⁾	January 26, 2026	April 4, 2023	1,500	1,525	1,500	1,568		1,000	1,001
3.45 ⁽⁵⁾⁽⁶⁾	April 4, 2028	June 19, 2024	25	37	25	40		1,500	1,568
8.70	May 25, 2029 ⁽⁴⁾	July 21, 2025	1,500	1,484	1,500	1,535		1,500	1,535
2.95 ⁽⁵⁾⁽⁶⁾	June 19, 2029	April 21, 2026	1,000	976	1,000	1,000		1,000	1,000
2.01 ⁽¹⁰⁾	July 21, 2030	May 15, 2021	200	196	200	214		200	214
11.60	January 7, 2031	July 27, 1990	1,000	976	—	—		—	—
1.96 ⁽¹¹⁾	April 21, 2031	August 20, 1991	150	146	150	160		150	160
10.80	May 15, 2031		25	39	25	44		25	44
8.70	May 25, 2032 ⁽⁴⁾		25	40	25	45		25	45
8.70	May 25, 2033 ⁽⁴⁾		25	42	25	48		25	48
8.70	May 25, 2035 ⁽⁴⁾		47	47	47	59		47	59
Floating ⁽¹²⁾	July 31, 2084		14	14	14	17		14	17
Floating ⁽¹²⁾	August 31, 2085		5,643	5,554	5,561	5,766		5,561	5,766
			(15)	(15)	(54)	(54)		(54)	(54)
			\$ 5,628	\$ 5,539	\$ 5,507	\$ 5,712		\$ 5,507	\$ 5,712

Subordinated indebtedness sold short (held) for trading purposes

- (1) Canada Yield Price: a price calculated at the time of redemption to provide a yield to maturity equal to the yield of a Government of Canada bond of appropriate maturity plus a pre-determined spread.
- (2) Carrying value includes unamortized discounts and accretion of premiums. Indebtedness reflects the impact of interest rate hedges in an effective hedge relationship.
- (3) Subsidiary of CIBC FirstCaribbean, and guaranteed on a subordinated basis by CIBC FirstCaribbean.
- (4) Not redeemable prior to maturity date.
- (5) Debentures are also subject to a non-viability contingent capital (NVCC) provision, necessary for the Debentures to qualify as Tier 2 regulatory capital under Basel III. As such, the Debentures are automatically converted into common shares upon the occurrence of a Trigger Event as described in the capital adequacy guidelines. In such an event, the Debentures will be converted into common shares at a price of \$5.00 per share (subject to adjustment in certain events as defined in the relevant prospectus supplements).
- (6) Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 2.57% above the three-month Canadian dollar bankers' acceptance rate.
- (7) On January 26, 2021, we redeemed all \$1.0 billion of our 3.42% Debentures due January 26, 2026. In accordance with their terms, the Debentures were redeemed at par.
- (8) Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 1.00% above the three-month Canadian dollar bankers' acceptance rate.
- (9) Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 1.18% above the three-month Canadian dollar bankers' acceptance rate.
- (10) Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 1.28% above the three-month Canadian dollar bankers' acceptance rate.
- (11) Interest rate is fixed at the indicated rate until the earliest date redeemable at par by CIBC and, thereafter, at a rate of 0.56% above the three-month Canadian dollar bankers' acceptance rate.
- (12) Interest rate is based on the six-month US\$ LIBOR plus 0.25%.
- (13) US\$6 million (2020: US\$21 million) of this issue was repurchased and cancelled during 2021.
- (14) US\$2 million (2020: US\$4 million) of this issue was repurchased and cancelled during 2021.
- (15) US\$2 million (2020: US\$4 million) of this issue was repurchased and cancelled during 2021.

(a)

Figure 4.7: Comparison of Note Disclosure 2. (a) CIBC's Basis of Presentation note disclosure. (b) ChatGPT's Basis of Presentation note disclosure for CIBC. The text highlighted in gray shows the information that ChatGPT included in its generation which is not normally provided as part of this disclosure.

(b)

Write the note disclosure for Subordinated Indebtedness for the Canadian Imperial Bank of Commerce's 2021 Annual Report, including a table of the debentures and any relevant footnotes.

Subordinated Indebtedness

The Canadian Imperial Bank of Commerce (the "Bank") has issued the following subordinated debt instruments (the "Debentures"). The Debentures are unsecured obligations and rank subordinate to all senior obligations of the Bank and rank equally among themselves.

The following table sets forth certain information with respect to the Bank's outstanding Debentures as at October 31, 2021.

Issue Date	Maturity Date	Coupon Rate	Outstanding Principal Amount
January 15, 2020	January 15, 2040	5.25%	\$90,000,000
January 15, 2021	January 15, 2041	5.50%	\$190,000,000
January 15, 2022	January 15, 2042	5.75%	\$285,000,000

Footnotes:

1. The Debentures are subject to redemption at the option of the Bank on or after January 15, 2025.
2. Interest on the Debentures is payable semi-annually on January 15 and July 15 of each year.
3. The Debentures are listed on the Toronto Stock Exchange.

This information is intended as a general summary only and should not be considered a complete description of the terms of the Debentures. For a complete understanding of the terms of the Debentures, reference should be made to the relevant trust indenture and other related documents filed with Canadian securities regulatory authorities.

queried. Although the table was stripped out for the analysis, it was necessary to get ChatGPT to generate the footnotes. ChatGPT also included an advisory that the debenture information was only a summary and that for a complete understanding, the trust indentures and other related documents should be referred to. While Canadian banks do provide supplementary documents on their trust indentures that users can consult, the note disclosure is required to provide all of the relevant information on the debentures. In CIBC's footnotes (Figure 2, (a)), the relevant high level information is provided in the footnotes and there is no reference to debenture documents filed with regulators included in its footnotes (or for any of the other four banks' disclosures). Again, we also checked JPMorgan Chase & Co's subordinated debt disclosure, and there is no mention of referring to supplementary debenture regulatory filings [160]. Therefore, we conclude that ChatGPT has added this advisory on its own.

4.3.6 Conclusion, Future Work, and Limitations

This research has provided a benchmark for ChatGPT's current abilities to write financial statement note disclosures. This benchmark is important as it identifies where the gap is between the generated text and the desired/needed text. It also identifies areas where, for use in the financial world, ChatGPT needs to be further trained. Our research highlights that note disclosures are *currently* challenging for ChatGPT. We also draw attention to the fact that while ChatGPT is not able to fully provide the necessary output just yet, its text is much more readable (and therefore accessible) than that of its human counterparts, particularly for disclosure 1. These results create exciting opportunities for further research.

A big question that we would like to address in future research is how LLMs (Large Language Models, like ChatGPT) can be used to help improve the readability of financial statement note disclosures. The original intent and purpose of the stock market was to make it open to everyone — professional and lay-person alike. As financial reporting and market regulations have evolved, the original intent is becoming more difficult to achieve, as the communications from company to shareholder (or potential shareholder) have become exceedingly long with very complicated text. The potential ability of LLMs to distill the message to one that is more universally accessible is an area of research that is worth exploring.

The high similarity results for both human and ChatGPT’s texts on a YoY basis demonstrate that both use a “boilerplate” approach for the note disclosures. Contemporary research is interested in the detection and removal of boilerplate in web corpora. An unintended consequence of premature boilerplate removal, however, could render financial reports unintelligible, as key parts of the reports could be removed. Also, given the scarcity of financial statement text as well as the opportunities for fraudulent or incorrect text to be generated, more work is needed on how LLMs can **augment** financial reporting text, and how fraudulent/incorrect text can be detected quickly.

There are several important limitations of this research: we only evaluated the text from Canada’s five major banks (RBC, CIBC, BNS, TD, and BMO). It may be that ChatGPT’s performance is better in generating the disclosures of other Canadian banks. Also, we only selected two note disclosures. Although these are required (and fairly standardized) note disclosures, it may be that ChatGPT’s performance is better when generating other note disclosures. Finally, ChatGPT was trained on the common crawl web corpora which consists of 12 years of common crawl data [30]. That means that for each of the 5 banks, there are only 12 annual reports that ChatGPT has seen. This could have a material effect on the outcome of its generation.

4.4 Project 5: ChatGPT for Financial Literacy

4.4.1 Introduction

As discussed in previous projects, ChatGPT’s access is widely available (with some exceptions)¹⁰. ChatGPT is a “sibling” of InstructGPT which was intended to have users provide “instruction in a prompt and [InstructGPT would] provide a detailed response” [232]. ChatGPT caters to those who want to use its technology through removing coding barriers by providing an easy-to-use web interface, with the option to also use the application programming interface (API). This expands the potential user pool as it facilitates users typing in their question or the task that they would like ChatGPT to do into a “question box”, and ChatGPT will respond accordingly. An important aspect, however, of this approach to user-interface and user-experience (commonly known as UI-UX) opens up the possibility of

¹⁰A notable exception is China. Access to ChatGPT is unavailable due to a geo-lock (i.e., preventing access due to geographic location, in this case China) imposed by the Chinese government.

ChatGPT being used in ways that were not part of the original conception, whereby it is being used to learn new concepts, expand one's education, and also provide financial advice. The objective of this project is to evaluate ChatGPT's efficacy at communicating financial concepts to different audiences, ranging from an unknown audience, to a non-financial audience, to financial financial experts. This evaluation is particularly important to a general audience who may be looking to by-pass financial experts as well as to students who are learning fundamental financial principles.

This paper was authored in collaboration with Ms. Samantha Taylor, Dr. Shannon Lin, and Dr. Vlado Kešelj. We received some very positive reviews from the Journal of Emerging Technology in Accounting and it is currently in Revise and Resubmit status.¹¹

4.4.2 Methodology

4.4.2.1 User Definitions

To communicate effectively, accountants first need to determine the level of financial understanding the audience has, based on either interacting with or by making reasonable assumptions about the audience. An accounting instructor teaching a fourth-year advanced accounting class could reasonably expect that students in the class have a solid foundational understanding of accounting and finance. A tax accountant, on the other hand, completing a new client's taxes could reasonably assume that the client is not overly familiar with tax concepts and may need extra explanations and guidance. When doubt exists, the typical approach is to make communication understandable to a general audience. There is no universal definition of the different types of financial and non-financial users. The Cambridge Dictionary defines the term "user" as "someone who uses a product, machine, or service" [78]. Even when applied to a financial situation, it is a very wide definition, encompassing anyone who uses (or interacts with) financial information. The Handbook of International Education Pronouncements simply refers to "users" and does not elaborate on what defines users [154]. A review of the publicly available case studies published by the Chartered Professional Accountants of Canada (CPA) also does not define users. In fact, CPA candidates writing responses to cases are expected to identify the users and infer, from

¹¹All authors have given permission for this paper to be included in the thesis. Confirmation can be provided upon request.

the case facts, the level of financial understanding, to tailor their responses appropriately. For example, if the user in a case is determined to be the audit partner, a CPA candidate will not have to explain Generally Accepted Accounting Principles (GAAP). But if the user is a client who does not currently use any accounting framework, the candidate is then expected to provide a brief explanation of GAAP so the client understands the importance of using an accounting framework and how that will impact their accounting. Further investigation in the area reveals that there is some identification of who financial users could be, but no formal parameterized definition. Section 407 of the Sarbanes–Oxley Act addresses who can be identified as an “audit committee financial expert” [326, 138]. Originally, only those with “accounting experience” could serve in this role. However, in the final ruling, the US. Securities and Exchange Commission (SEC) determined that “people actively engaged in industries such as investment banking and venture capital investment... [and] professional financial analysts” could also be designated as financial experts, given their experience with financial statements. Hoitash et al. [138] examined different types of experts: Accounting Financial Experts (AFE), Supervisory Financial Experts (SFE), and User Financial Experts (UFE). They define this last category as “Individuals with experience performing extensive financial statement analysis or evaluation (e.g., financial analysts, investment bankers)”, in the context of expertise for service on the audit committee. To investigate the broad range of users, and to evaluate how well ChatGPT “understands” different user needs, we have selected six different users:

- Financially unsophisticated user: a user who lacks basic knowledge of accounting and finance.
- Non-financial user: a user who has a basic knowledge of accounting and finance.
- Financial user: a user who has a good grasp of accounting and finance, either through training, experience, or combination of the two.
- Financially sophisticated user: a user who is an accounting and financial professional which is achieved through extensive training and a designation, or who has extensive experience in a corporate position involving financial leadership.
- No audience orientation identified: This type of “user” is when ChatGPT is not provided any information on the user(s). In this type of situation, humans would tailor

their communication to be more broad, more general, and typically high-level, understanding that if more information was revealed on the knowledge and understanding of the user(s) in the course of communicating with the user(s), the communication approach would change as new information became available or was revealed.

- General audience: a collective group of users where there is a wide range of knowledge and experience with accounting and finance, ranging from a financially unsophisticated user to a financially sophisticated user.

4.4.2.2 Prompts

ChatGPT was prompted to explain basic financial reporting, audit, and finance questions. As both of ChatGPT’s LLMs can learn from a user’s previous conversations¹², two separate user accounts were used (one for GPT-3.5 and one for GPT-4) to ensure there was no “lookback contamination”.¹³ The dataset comprised ChatGPT’s answers (using both LLMs) to seven different prompts for each selected financial term. Each prompt uses a standard format. If the term is being explained to a user group, it follows the format of: “Explain <insert term here> to a <insert user here>”. If the term is being explained on its own and not in relation to a user group, the prompt follows the format of “Explain <insert term here>”. This format was consistently used for both GPTs to ensure the veracity as well as the comparability of the data. To ensure coverage between accounting and finance we chose five terms each from financial reporting, auditing, and finance:

Financial Reporting

- Net Income
- Net Revenue
- Non-Interest Income
- Provision for Loan Losses
- Subordinated Debt

¹²If the conversations are held in the same window, and up to 4,096 tokens [15] which equates to approximately 3,000 words [234, 90] for GPT-3.5, and 8,000 tokens for GPT-4 at the time of this research.

¹³Some of our prompts and the resulting answers were less than 3,000 words between the prompt and completion.

Audit and Assurance

- Generally Accepted Auditing Standards (GAAS)
- Material Misstatement
- Professional Judgment
- Professional Skepticism
- Reasonable Assurance

General Financial

- Investment
- Capital
- Assets
- Liabilities
- Financial Statements

These terms were selected because they are deemed fundamental terms that one would come across in financial statements, financial communications, as well as financial education.

4.4.2.3 Dataset

Responses generated by ChatGPT were saved in individual plain text files using UTF-8 encoding. Summary statistics for the dataset are in Table 4.11. To ensure the consistency of the text's presentation to the algorithms, spaces between paragraphs were removed so that all the text was saved in one large paragraph. Numbered lists were kept (with spaces removed between numbered items). Bullet point lists were also kept in the same fashion as the numbered lists, but bullets were removed. Unlike numbered lists, bullet points did not have ending punctuation. Therefore, to ensure that the sentences were grammatically correct, periods were added at the end of every bullet point list item. This also served to delineate bullet points. The text itself, however, was not altered in any way.

Table 4.11: Financial Literary Dataset Summary Statistics

GPT-3.5	Number of Documents	Min Words	Max Words	Mean Words	Std Dev Words
Financially Sophisticated	15	160	272	214.4	37.35
Financially Unsophisticated	15	94	256	172.1	48.45
Financial User	15	174	309	224.7	39.98
General Audience	15	121	310	180.8	60.30
No Audience In Mind	15	115	239	180.6	37.85
Non-Financial User	15	132	223	184.6	31.46
Define Term	15	160	270	197.5	33.73
GPT-4	Number of Documents	Min Words	Max Words	Mean Words	Std Dev Words
Financially Sophisticated	10	153	331	260.8	53.62
Financially Unsophisticated	13	143	278	219.5	48.66
Financial User	11	127	316	257	56.51
General Audience	12	129	251	158.1	40.05
No Audience In Mind	13	133	265	202.7	44.61
Non-Financial User	13	139	279	212.1	47.90
Define Term	13	131	314	212.8	57.14

While data collection was straightforward using GPT-3.5, it proved difficult using GPT-4 due to user demand for GPT-4. At the time of the research, GPT-4 was limited to 25 messages every three hours. We also struggled with GPT-4 returning incomplete answers. In some cases, we were able to coax GPT-4 to finish the answer by providing feedback that the response was unhelpful as it was unfinished. In most cases, GPT-4 would generate another incomplete response. Markovski recommended specifying the desired length in the prompt (e.g., “A list of ten science fiction books”) [344]. While this can work well for a list setting, it does not work well for explaining financial terms; if a maximum length is specified, while GPT-4 might provide a complete textual response, it could still return an incomplete explanation, having left out pertinent information to meet the length requirement. As well, GPT-4 is limited (at the time of this research) to 8,000 tokens (pieces of words) [344, 233], which equates to $\sim 32,000$ characters in total, between prompt and response. It is important to also note that characters include alpha numeric characters (letters and numbers) as well as spaces between words and punctuation. Therefore, to specify the maximum length, a user would need to determine the length of the prompt and then subtract that amount from the overall 32,000-character length, which is impractical. As such, we did not specify a response length in our research.

4.4.3 Experiments and Hypotheses

We used Cosine Similarity — a commonly used metric in information retrieval — to measure the similarity of the text. This metric uses the frequency of words to calculate the similarity of text [66, 293]. The text of each GPT answer was first tokenized using the word tokenizer in the Natural Language ToolKit (NLTK) [190] in Python and saved in list format. Each list was then added to a different set. The sets were vectorized and the Cosine Similarity was then calculated. This formula is given in Equation 1. Cosine Similarity is not sensitive to the text length, meaning that this measure of similarity is very useful in the context of our research because we are not specifying response text length up front [145, 333].

To determine the similarity of explanations given by ChatGPT for different users, a similarity threshold was established. Following the work of Qurashi et al [257] and Singh et al. [291], we used use an alpha of 0.8, where $0 \leq \alpha \leq 1$ [219]. In the context of this financial literacy research, the threshold was used to determine group matches, discussed below.

Following the user definitions described in Section 4.4.2.1, we separated the users into three main groups.

- Group 1: financially unsophisticated and non-financial users;
- Group 2: financially sophisticated and financial users; and
- Group 3: no audience orientation, general audience, and the definition of a financial term.

If the text meets the similarity threshold within the groups, but not with other groups, then we can conclude that ChatGPT is properly tailoring responses to the different users. A “match” denotes that two groups’ users have exceed the similarity threshold of 0.8 for the answer to a given prompt. For example, if the text for explaining Net Income to a financially sophisticated user and a financial user is 0.8763, then this would be considered a match that meets the 0.8 threshold. Conversely, if the text for explaining Net Income to a financially sophisticated user and a financially unsophisticated user is 0.4361, this would not be considered a match because the calculated similarity is below the required threshold of 0.8.

Hypothesis 1. *Similarity of Financially Unsophisticated and Non-Financial User will meet or exceed the similarity threshold of 0.8, as defined above.*

Hypothesis 2. *Similarity of Financially Sophisticated and Non-Financial User will meet or exceed the similarity threshold of 0.8, as defined above.*

Hypothesis 3. *Similarity of No Audience Orientation, General Audience, and Term Definition will meet or exceed the similarity threshold of 0.8, as defined above.*

The Flesch Reading Ease Score (FRE) was also used to evaluate the readability of the text for the different users. Using the spaCy Readability package in Python [212], we calculated the FRE for each answer provided by ChatGPT. The FRE has been a standard measure to evaluate readability for decades and has been used frequently in research in the accounting domain (e.g., [303, 243, 57]). Most end-users who will be using documents and terminology from the accounting and finance domain will be adults, and to be consistent with previous research, we opted to use the FRE measure.

Following Taylor and Keselj [312], we use an expanded interpretation of the general education level because a university degree does not dictate success in the financial domain. Taylor and Keselj [312] pointed to Bill Gates, Mark Zuckerberg, Steve Jobs, and Richard Branson [177] as extremely successful and commonly understood to be financially savvy, but who do not have university degrees. As such, “university level” is also equated with “early business professional” and “graduate level” is equated with “seasoned business professional”.

Hypothesis 1. *Readability of Financially Unsophisticated and Non-Financial User will be > 60 , based on the discussion found in Section 4.2.2.6.*

Hypothesis 2. *Readability of Financially Sophisticated and Non-Financial User will be < 50 , based on the discussion found in Section 4.2.2.6.*

Hypothesis 3. *Readability of No Audience Orientation, General Audience, and Term Definition will > 50 , based on the discussion found in Section 4.2.2.6.*

4.4.4 Results and Discussion

4.4.4.1 GPT-3.5

The Cosine Similarity shows that GPT-3.5 does not differentiate between users well, and that the 0.8 similarity threshold is rarely met. Results have been provided in Appendix E.

Group 3 (general, no audience, define term) had the most matches: Net Revenue, Net Interest Income, Material Misstatement, Reasonable Assurance, Capital, and Financial Statements. Seven of the eight matches were between *no audience in mind* and *define term*. The other was between *general audience* and *define term*. Group 1 (financially unsophisticated, non-financial user) had matches for Subordinated Debt and Material Misstatement. Group 2 (financially sophisticated, financial user) only had two matches for GAAS and Capital. The remaining matches raised concerns:

- Net Revenue — general audience matched with non-financial user
- Net Interest Income — financially sophisticated matched with no audience and “define term”.
- Provision for Loan Losses — financial user matched with “define term”
- GAAS — financial user matched with no audience in mind.
- Reasonable Assurance — financially sophisticated user matched with no audience and define term.
- Capital — general audience matched with non-financial user.
- Financial Statements — financial user matched with general audience.

These are problematic because they match across groups — conflating the needs of the users in the different groups. These results suggest a lack of understanding as to how financially sophisticated users and financial users are different from a broader audience for certain financial terms. It also shows an assumption that simply requesting a definition of the term suggests prior knowledge of accounting and finance. Human authors, however, would take the opposite approach, understanding that without any information on the questioner, it is best to assume they have no prior experience in this domain and make the explanation more general.

The FRE results (Appendix E) show that all but two of the explanations are considered difficult to read, with the majority (58%) in the very difficult or extremely difficult category (i.e. below 30 and below 10, respectively). Given that five of seven users are either weak in their knowledge of accounting and finance, have none (financially unsophisticated and non-financial user) or are part of the general audience (no audience in mind, general audience, or define term), these readability results support that GPT-3.5 is not ready to be effectively teaching foundational knowledge.

We also detected that some of the answers GPT-3.5 gave were misleading or incorrect in

reviewing transcripts. For an accounting and finance professional, the errors are funny, but not material to their understanding and comprehension of the subject matter. For a less knowledgeable user — whether it be the average investor or a student learning accounting — these errors are extremely problematic and, in some cases, lead to a significant misunderstanding and misapplication of the term. Net Income, for example, is one of the most fundamental terms in accounting. GPT-3.5 indicated that:

Net Income is a financial term that represents the amount of money a person or a business earns after all expenses have been paid. It is the difference between the total amount of money earned from selling goods or services and the total amount of money spent to produce those goods or services.

This shows that GPT-3.5 has conflated earnings and cash, which, in a financial context, are not the same thing. GPT-3.5 used this analogy in all its explanations, with the one exception of “no audience orientation in mind”. GPT-3.5 also indicated that Net Income was “subject to corporate income tax.” When we asked for clarification, GPT-3.5 indicated that Net Income should be calculated first and then taxes applied, which is incorrect; rather, taxes are considered in the calculation of Net Income.

We also find that GPT-3.5 is biased towards information from the United States. While it understood that Canada, for example, has Generally Accepted Accounting Principles (GAAP) comprising Accounting Standards for Private Entities (ASPE) and International Financial Reporting Standards (IFRS), and the U.S. has US GAAP and IFRS, it could not make the same “intelligent leap” for the Generally Accepted Audit Standards (GAAS). GPT-3.5 was quite insistent that only the U.S. has GAAS and other countries had audit standards, but that they were not GAAS. Using the “chain-of-thought” method [55], we prompted GPT-3.5 on associations with GAAP between countries, finally arriving at the fact that Canada does have GAAS comprising the Canadian Audit Standards (CAS).

Given that bookkeeping software has not replaced more jobs in accounting and finance than it already has is due largely to the need to exercise professional judgment in these fields. Some of the Accounting Standards, for example, have areas of flexibility in the rule where professional judgment is needed. We asked GPT-3.5 whether it could exercise professional judgment and it said that it could not. We then tested the extent that GPT-3.5 could provide insight and come close to exercising professional judgment. One of the criteria in ASPE 3065

(Leases) addresses the Present Value of the Minimum Lease Payment (PVMLP) where the determination of “substantially all” is “usually 90 percent or more”. As Leases is a common topic taught at the university level and covered in a publicly accessible CPA Canada Core 1 Self-Assessment Entrance Exam Solution, we prompted GPT-3.5: “If I calculate the present value of the minimum lease payment to be 89%, is this criterion met?” GPT-3.5 returned the following:

If the present value, at the beginning of the lease term, of the minimum lease payments, excluding any portion thereof relating to executory costs, is equal to 89% of the fair value of the leased property, then this criterion would not be met. The criterion requires that the present value of the minimum lease payments be equal to “substantially all” of the fair value of the leased property, which is typically considered to be 90% or more. Therefore, if the present value is 89%, it falls short of the 90% threshold and the criterion would not be considered met.

It is important to note that a student would likely take GPT-3.5’s answer above as definitive because they “don’t know what they don’t know” (a phrase often attributed to Socrates and later, Donald Rumsfeld). Learning to read and interpret Standards takes time, experience, and guidance. In this case, students tend to latch on to the 90% and use it as a “bright line” or unambiguous threshold that is either met or not met when, in reality, it is not, because of the words usually and substantially all.

The above demonstrates GPT-3.5’s inability to emulate professional judgment. A human accountant would conclude that 89% would still meet the requirements of substantially all because of its relative proximity to the usual 90% requirement. It is also important to note that if GPT-3.5 is implemented in education, these types of errors and challenges with regards to the exercise of professional judgment could materially hinder learning. Furthermore, the inability for an LLM like GPT-3.5 to exercise professional judgment could also create marking errors where students are not awarded correct grades because the LLM does not and cannot account for the student’s correct exercise of professional judgment and the variation of that judgment between student responses.

4.4.4.2 GPT-4

Although there is some improvement in the results when compared to GPT-3.5, the results for Cosine similarity are still strongly indicative that GPT-4 is not differentiating well between groups (Appendix E). Group 2 (financially sophisticated, financial user) had the most matches (seven in total), which is an improvement over GPT-3.5 two matches. Matches were identified for Net Income, Net Interest Income, GAAS, Material Misstatement, Professional Judgment, Professional Skepticism, and Reasonable Assurance. Group 3 (general, no audience, define term) had six matches, all of which were between no audience and define term for Net Income, Provision for Loan Losses, Subordinated Debt, Material Misstatement, Professional Skepticism, and Reasonable Assurance. Group 1 (financial unsophisticated and non-financial user) only had two matches for GAAS and Reasonable Assurance. Similar to GPT-3.5, there were concerns in the remaining matches:

- Net Revenue — financially sophisticated matched with define term
- Provision for Loan Losses — financially sophisticated matched with no audience and define term; financial user matched with no audience and define term
- Subordinated Debt — financial user matched with no audience and define term
- Material Misstatement — both financially sophisticated and financial user matched with no audience and define term

Again, these mismatches are considered problematic as they match across groups rather than within groups.

Readability was also tested using the FRE Score. We observe a marked improvement in readability of the available responses. Fifty-eight percent of GPT-3.5's responses were in the very difficult to read or extremely difficult to read categories (below 30 and below 10). GPT-4's responses, however, show a 10% improvement, as only 48% of the available responses are in the very difficult or extremely difficult to read categories (Appendix E).

In reviewing the transcripts, we noticed that while GPT-3.5's error of taxes after the calculation of net income had been resolved in GPT-4, the problem of referring to earnings as money persisted, as did the U.S. bias problem. We also tested GPT-4's ability to exercise professional judgment using the same scenario of the PVMLP that we used for GPT-3.5. While it did a better job of explaining the standard, it still interpreted the 90% as a bright-line threshold: "Since your calculated present value of the minimum lease payments is 89%,

it is close to, but does not meet, the 90% threshold.” It did, however, recommend that the user use professional judgment by indicating “You may need to use your professional judgment to determine if this is considered “substantially all” of the fair value of the leased property based on the specific facts and circumstances of the situation.”

An often-asked question on the internet is whether a Louis Vuitton bag is an asset or investment (e.g., [231, 314]). The well-accepted financial interpretations here are important to ground the GPTs answers. Per IFRS, an asset is defined as “a resource controlled by the entity as a result of past events and from which future economic benefits are expected to flow to the entity” [71]. As well, the traditional definition of an investment is to generate profit: invest money to make money. Based on these established definitions, purses, on average, are not assets or investments. Therefore, the more applicable categorization of a purse is consumption. While it is possible to sell items for more than original cost, that is not the expected outcome.

We prompted both GPTs on whether a Louis Vuitton bag was an investment. GPT-3.5 indicated that it is not really an investment, nor a “sound investment strategy”, whereas GPT-4 indicated that yes, it is an investment in a non-traditional sense. The transcripts show there is a differentiation in the semantic interpretation of the question relating to the definitions of investment. GPT-3.5 is much more “traditional” than GPT-4 which appears to recognize the evolution of language and that investment has also taken on a colloquial understanding in the sense that taking certain actions now can influence one’s ability/likelihood to earn more on a personal level in the future. For example, “investing” in an education – here the expectation is that with a better education, one will earn more, thereby providing future “dividends.”

The wording of a question is very important as it can inadvertently influence the outcome of the answer. Both GPT-3.5 and GPT-4 are very agreeable in their answers, even if the LLM begins with agreeing but then does not actually hold a clear position in its answer which we saw with the Louis Vuitton bag question. Rather than asking the question in a positive way if education was an investment, we took a more negative approach by asking “Would you agree that investing in an education is not really an investment? Please discuss.” Initially GPT-3.5 responded saying that it can be seen as an investment and even went on to discuss if it would be a worthwhile investment, whereas GPT-4 looked at both sides of the argument as to whether education was an investment which depends on a person’s perspectives and

experiences.

Finally, we questioned both ChatGPT (using GPT-3.5 and GPT-4) on its respective abilities to audit firms or provide accounting services. While both LLMs (eventually) indicated “no”, GPT-3.5 first addressed the complexities and training required to provide both audit and accounting services, whereas GPT-4’s response largely pointed to getting professional advice.

4.4.5 Conclusion and Future Work

While these large language models are robust and can perform impressive tasks, our research demonstrates that this is not always the case for the accounting and finance domain; being partly right is not sufficient for those who are looking to acquire or strengthen their financial knowledge. Indeed, we have a long way to go until the “technology is perfected” [23]. Net Income is one of the most fundamental concepts in accounting, and our conversation with both GPTs identifies concerns in the way that it explains and contextualizes this principle. We also show that both GPT-3.5 and GPT-4 are incapable, by their own admissions and by demonstration, of exercising professional judgment or professional skepticism. Using the example of the PVMLP, we demonstrated that GPT-3.5 was not able to correctly interpret the standard and apply it to a situational question, concluding that 89% was significantly below the threshold of 90%. While GPT-4 fared better in its answer as it indicated to the user that they should exercise their own professional judgment, the fundamental purpose of our question was designed help a learner understand how to use professional judgment in this type of scenario — which GPT-4 could not do, nor reasonably explain, nor provide any real guidance to the user. This demonstrates that GPT-3.5 or GPT-4 cannot be relied on to explain and contextualize accounting scenarios, nor to mark assignments/exam questions.

Finally, there are several ethical issues that warrant consideration regarding the use of AI in the accounting domain. Given the “abilities” of LLMs like GPT-3.5 and GPT-4, the implementation of this technology is very attractive, particularly since financial documents such as the Annual Report to shareholders tends to be very long, 186 pages on average[128]. In fact, having the ability to generate this text using AI seems to be a “perfect” fit. However, as demonstrated, the current version is unfit for this task as it has made significant errors in the explanation of one of accounting’s most fundamental principles – Net Income. Regardless

of whether the end-user is financially savvy or not, there are significant implications as text from annual reports is often directly quoted in lawsuits [270]).

Education had some of the earliest adopters of GPT-3.5 either through student use [319] or by instructors incorporating it into the classroom [186, 228]. While students and instructors may find this a useful (and fun!) resource, it is only a good resource if it provides the correct answer every time [316]. If AI is not returning the right answer consistently, how do students know? GPTs can serve as a useful tool to someone who knows the correct answer and can distinguish it from the incorrect ones, but not for someone who wishes to learn the correct answer. Returning to our example of the blunder with Net Income, students new to accounting may take ChatGPT's answer at face value — that net income is first calculated and then taxes are applied, and that net income is the money you have when all is said and done. Socrates' "You don't know what you don't know" is a highly applicable adage in this type of situation.

ChatGPT's hyped and reported "feats" on news, social media, and in academic research (e.g., [175]) have lent it a high degree of credibility that may not (yet) be warranted, and ours is not the only research to find serious problems with this type of AI. Alkaissi and McFarlane raised the concern of AI "hallucinations", which occur when AI generates something that seems real but is not based on real-world input [8]. While hallucination is uncommon, it does occur. Alkaissi and McFarlane tested GPT-3.5 by asking it to write short essays on common topics and found that "ChatGPT provided confident responses that seemed faithful and non-sensical when viewed in light of the common knowledge in these areas" [8]. Frieder et al. found that GPT-3.5's mathematical abilities were far below expectation and concluded that students would be better off cheating off peers than using GPT-3.5 [106].

ChatGPT's performance is highly attributable to the training data used. GPT-3.5 was trained on the common crawl data [30] which includes "petabytes of data collected over 12 years of web crawling [and] contains raw web page data, metadata extracts and text extracts" [62]. That means that the common crawl contains everything — including "undesirable content" [196] — but also content that is incorrect, false, or misleading.

Returning to the level of trust that has been afforded GPT, there is an expectation that this type of AI tool would be trained on carefully vetted data, which is not the case. In fact, in this context, ChatGPT is just a fancier version of the internet. Rather than Googling for

the answer and then scrolling to find a website that “looks right”, the AI tool is giving us an answer that may or may not be reliable, which OpenAI fully admits and warns users about on its blog, as well as on its user interface [237]. Incorrect responses are understandable and expected as LLMs evolve and mature, these errors do not create problems for financially sophisticated users; they do, however, create significant issues and concerns for financially unsophisticated users. This group does not have the same ability as financially sophisticated users to differentiate between accurate and inaccurate responses.

We end our discussion by highlighting some limitations. When our research was conducted, both GPT-3.5 and GPT-4 were trained until September 20, 2021 [236, 237]. Therefore, resources added after that date to the web corporas (datasets) or other undisclosed datasets that OpenAI used to train these models, were excluded. Additionally, unless OpenAI developers decide to train the models further on user-identified errors or gaps, the GPT models will not improve their responses to those questions across the platform. Finally, a severe limitation is that GPT-3.5 is easily fooled. We easily taught GPT-3.5 a non-existent financial metric: EBITDACR — Earnings before interest, tax, depreciation, amortization, and cryptocurrency expenses. We created this metric based on a popular Wall Street meme — Earnings before interest, tax, depreciation, amortization, and coronavirus riots [75]. When we asked GPT-3.5 if it was sure that EDITDACR was earnings before interest, tax, depreciation, amortization, and cryptocurrency expenses and not coronavirus riots, it concluded that crypto expenses was correct, even though no such metric exists. GPT-4 was more resistant to the new metric, but when we provided more information (all fabricated), it was more than happy to explain what EBITDACR was, which is concerning.

4.5 Project 6: Using Raw Accounting Data To Augment Financial Text Using ChatGPT

4.5.1 Introduction

Augmentation for FS and reports is a two-step process as discussed in section 1.1 where both the quantitative data and the qualitative data are interlinked and must be augmented together. Therefore, it is important to [augment](#) text and [Notes](#) to accompany the tabular data. The project that we present in this project takes selected synthetic tabular data from Project 2 (see Section 3.3) and generates text for it. As we have demonstrated in the previous

projects in this chapter, and as other researchers and users have demonstrated, ChatGPT is very capable of generating human-like text. However, just because ChatGPT can generate text does not mean that the text that it generates is appropriate, correct, or usable for its specific purpose. The challenge here, however, is that for financial text, the *specific purpose* depends on who is reading the text and what they are using it for, as the audience for the [Notes](#) and the financial reports is very broad. This research project also has the added challenge that, for the synthetic tabular data, there is no ground truth text. Historically, creating ground truth has involved humans to provide text or labels. This practice was very labour intensive, and in the era of human-like AI text generation, this approach may no longer as practical as it once was.

Therefore, we give ChatGPT some of the tabular data generated in Project 2 (Section 3.3) and using in-context training, ask it to generate text for two financial metrics to be used in a financial report. A challenge in doing this is that popular and time-tested evaluation metrics such as Rouge [187], Bleu [241] and Bleurt [287] require ground-truth text for training and comparison, which we do not have. Therefore, we created a program in Python to confirm the quantitative numbers in the text against the tabular data to check ChatGPT’s use of the numbers. We, then, used semantic similarity to evaluate the text against quasi-ground truth from previous years’ financial reports. This type of task is important for rigorous evaluation of AI-generation (ChatGPT specifically here) but also to determine the level of trust that we can place on AI at this time for financial text.

4.5.2 Related Work

The most common metrics used in Natural Language Processing for evaluating text are Recall-Oriented Understudy for Gisting Evaluation ([ROUGE](#)) and bilingual evaluation understudy ([BLEU](#)), both which need ground truth text as comparators. The purpose of ROUGE is to evaluate text summaries against a ground-truth (or collection of ground-truth) documents. There are multiple ROUGE metrics such as ROUGE-N, ROUGE-L, and ROUGE-W. Equation 4.7 shows how ROUGE-N uses ngram recall to evaluate a candidate text against ground-truth [187].

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}_{\text{match}}(gram_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} \text{Count}(gram_n)} \quad (4.7)$$

where n is the n-gram length, $gram_n$ is the maximum number of n-grams co-occurring in the candidate summary and the set of reference summaries denoted by $Count_{match}(gram_n)$ [187]¹⁴

BLEU is a related measure in that it uses precision, rather than recall [187]. BLEU uses both a sentence brevity penalty (BP) calculated in Equation 4.8. The full mathematical expression for BLEU is found in Equation 4.9.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \quad (4.8)$$

“where c is the length of the candidate translation and r is the effective corpus reference length” [241].

$$BLUE = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (4.9)$$

where BP is the brevity penalty, outlined above in Equation 4.8, N is the maximum n-gram size when calculating the precision score, w_n are weights that have been assigned to each n-gram precision score, and $\log p_n$ is the logarithm of the modified precision score for each n-gram size. Together, the brevity penalty (BP) is combined with the precision score $\exp(\sum_{n=1}^N w_n \log p_n)$ to compute the overall BLEU score [241].

4.5.3 Methodology

4.5.3.1 Dataset

We used two datasets for this project: the synthetic tabular data that was generated in Project 2 (see Section 3.3) for the basis of the textual augmentation as well as the textual data from Project 3 (Section 4.2) to serve as the quasi-ground truth. We use the term quasi-ground truth as the textual passages in the dataset were not written specifically for the synthetic data, but rather for the data for which the reports were written. The passages are similar to what we asked ChatGPT to generate and is therefore ground-truth, but not specific to the synthetic data. As the quasi-ground truth text from Project 3 was limited to 5 major US banks — JPMorgan Chase & Co (JPM), Bank of New York Mellon (BNY), Citigroup

¹⁴As “the denominator of the equation is the total sum of the number of n-grams occurring” for the reference text, this measure is recall-related measure [187].

(Citi), Bank of America (BoA), and Wells Fargo (WF) — we parse out the synthetic data from Project 2 for those specific banks for the project that we present here.

4.5.3.2 Semantic Similarity Using Word Embeddings

In previous projects, we used Cosine Similarity and Jaccard Similarity to evaluate the equivalencies of texts. In those tasks, we were investigating how close the generated text was to ground truth text to determine how ChatGPT behaved in generating the text. In this project, the text that ChatGPT generated did not need to “match” any ground truth text. Therefore, we determined that semantic similarity would be most appropriate as the synthetic text should have either the same meaning or the opposite meaning (depending on the situation), rather than have the same words.

Sentences are considered to be semantically similar if they have the same meaning and elicit (essentially) the same response [345]. If the sentences do not have the same meaning and do not elicit (essentially) the same response, then they are considered to be semantically dissimilar, as seen in Figure 4.8.



Figure 4.8: Illustration of semantic similarity and semantic dissimilarity [345].

To capture the meaning of the text, we used FinBERT — an open-source NLP model specifically created for finance [174]. Embeddings are vector representations of the text (either words or phrases). Here, we used sentence embeddings which provided us with a vector representation of the sentence [174]. This is useful when you need a fixed-size representation of a variable-length text [188]. We used cosine similarity to measure the cosine of the angle between the two vectors, sentence 1 and sentence 2. The cosine similarity ranges

from -1 to 1 . 1 means that the vectors are identical. 0 means the vectors are orthogonal or uncorrelated, and -1 means completely dissimilar. We also considered the Distributional Hypothesis — words that occur in the same contexts share semantic meaning. The more overlap in contexts, the more similar the words are considered [276].

4.5.3.3 Readability

Readability is also an important aspect of financial text, whether human-written or AI-generated. Therefore, we used the Flesch Reading Ease Score, discussed in Section 4.2.2.6, to evaluate the readability of the generated synthetic text.

4.5.4 Experiments

The original dataset for Project 3 consisted of 439 prompt and completion pairs, as per Table 4.1. We selected 20% of the pairs (88 pairs) to be used as the quasi-ground truth, the breakdown of which is found in table 4.12. This file was provided to ChatGPT and used as part of in-context training for style and information on how each bank sets up its responses.

Bank	# of Pairs
JPMorgan Chase	23
Bank of New York	21
Citibank	19
Bank of America	9
Wells Fargo	16
Total	88

Table 4.12: Fine-Tuning Prompt and Completion Pairs

4.5.5 Results and Discussion

As can be seen in Figure 4.10, even though ChatGPT was given an in-context training file, it defaulted to the exact same pair of messages for each bank. Using the similarity threshold of 0.8 [257, 291] that we have used in other experiments, we find that only two of the responses exceed that threshold. We also find that, when comparing to the ground-truth, ChatGPT’s responses are much “fancier” than the bank’s human-authored text for

this measure, an example of which is Figure 4.9. Ignoring the values for the moment, the text itself is semantically similar in that it discusses the measure (net interest income) and the time frame (2017). None of the training text or the ground-truth text mention maximizing shareholder value or prudent interest income management.

<p>Net interest revenue totaled \$3.3 billion in 2017 compared with \$3.1 billion in 2016, an increase of 5%.</p>	<p>In 31-Dec-17, our net interest income stood at \$30.39 billion, demonstrating our commitment to maximizing shareholder value through prudent interest income management.</p>
---	---

Figure 4.9: On the left is the ground-truth human-authored text from Bank of New York Mellon and on the right is the text generated by ChatGPT for the same time period using the synthetic data created in Project 2.

Another important aspect of the text generation, however, is that ChatGPT has defaulted to a pair of responses and applied those for every bank as seen in Figure 4.10.

Bank of New York Mellon

"For the year ending 31-Dec-16, Bank of New York Mellon reported a net interest income of \$29.34 billion, reflecting our robust financial management strategies."
 "In 31-Dec-17, our net interest income stood at \$30.39 billion, demonstrating our commitment to maximizing shareholder value through prudent interest income management."

Bank of America

"For the year ending 31-Dec-05, Bank of America reported a net interest income of \$14.56 billion, reflecting our robust financial management strategies."
 "In 31-Dec-12, our net interest income stood at \$22.67 billion, demonstrating our commitment to maximizing shareholder value through prudent interest income management."

Citigroup

"For the year ending 31-Dec-19, Citigroup reported a net interest income of \$50.94 billion, reflecting our robust financial management strategies."
 "In 31-Dec-10, our net interest income stood at \$21.23 billion, demonstrating our commitment to maximizing shareholder value through prudent interest income management."

JPMorgan Chase & Co

"For the year ending 31-Dec-23, JPMorgan Chase & Co reported a net interest income of \$98.34 billion, reflecting our robust financial management strategies."
 "In 31-Dec-08, our net interest income stood at \$16.32 billion, demonstrating our commitment to maximizing shareholder value through prudent interest income management."

Wells Fargo

"For the year ending 31-Dec-11, Wells Fargo reported a net interest income of \$22.50 billion, reflecting our robust financial management strategies."
 "In 31-Dec-09, our net interest income stood at \$20.75 billion, demonstrating our commitment to maximizing shareholder value through prudent interest income management."

Figure 4.10: ChatGPT's generated text for Q1

In the instructions, ChatGPT was told to use the in-context training file as reference, which it acknowledged. Yet, it does not look like it has actually done that, as its generation is the same for every bank, even though the in-context training instances it was given are different. Using the same example of Net Interest Revenue in the training document, JPMorgan Chase wrote "We generated earnings of \$11.9 billion on revenue of \$38.3 billion — a record year for our business" whereas Wells Fargo wrote "Our revenue increased 6% from the previous year." So, even though the question of discussing the Net Interest Revenue is

Bank	Q1 Semantic Similarity	Bank	Q2 Semantic Similarity
GPT-3.5		GPT-4	
BNY 1	0.74	BNY 1	0.76
BNY 2	0.69	BNY 2	0.73
BoA 1	0.77	BoA 1	0.62
BoA 2	0.79	BoA 2	0.56
Citi 1	0.63	Citi 1	0.74
Citi 2	0.78	Citi 2	0.69
JPM 1	0.82	JPM 1	0.76
JPM 2	0.82	JPM 2	0.78
WF 1	0.77	WF 1	0.75
WF 2	0.66	WF 2	0.78
Mean	0.74	Mean	0.72
Median	0.77	Median	0.74
Std Dev	0.07	Std Dev	0.07

Table 4.13: Results for Semantic Similarity Using GPT-3.5 and GPT-4

still the same between banks, there are many examples that ChatGPT was given showing a wide range of responses. Therefore, vacillating between two responses, regardless of bank, shows a lack of understanding in how to complete the task. The semantic similarity results are found in Table 4.13.

The readability results were very interesting. We have shown in previous projects that ChatGPT’s text is usually more readable, and therefore more accessible to a wider audience. That is not the case here. ChatGPT’s use of elevated and verbose language (“fanciness”) is causing it to have lower scores than the human-authored text. Referring back to Figure 4.10, Bank of New York Mellon’s text is very straightforward in that it indicates the total for 2017 and then provides a comparison to 2016, making BNY’s text readable and strong. ChatGPT’s text, however, discusses “demonstrating commitment to maximizing shareholder value” and “prudent interest income management”. ChatGPT’s use of grandiose language clutters up the text, making it harder to understand and changes the messaging moving the focus from the results to BNY’s suggested dedication to maximizing shareholder wealth by taking well-advised and sensible steps to managing interest income. This change in messaging is also reflected in a drop of the positivity and negativity, and an increase of the neutrality of the sentiment (using FinBERT) between the two text passages, shown in Table 4.14. Adding in the extra (and potentially superfluous) words makes it harder to read and harder

for the audience to understand. To borrow a salient example from a colleague¹⁵ “the fowl ambled over the thoroughfare” is a very fancy way of *trying* to say “the chicken crossed the road” which also creates problems for readers unfamiliar with a more arcane style of writing: fowl is not as commonly used anymore and can mean chicken but can also be used as a collective term for birds; ambled is a very slow and relaxed way of walking and does not necessarily mean in a direct fashion; and thoroughfare is not a commonly used word anymore and can mean either road or main road. Using elevated language can be helpful in the right context, provided the perceived or interpreted meaning is the same as the intended meaning. Here, using elevated language only serves to obscure the message; the same can be said for ChatGPT’s generated text for BNY, seen in Figure 4.9.

Author	Positivity	Negativity	Neutrality
Human	0.909	0.072	0.018
ChatGPT	0.802	0.008	0.190

Table 4.14: Change in Sentiment Scores

As will be discussed in Project 7 (Section 5.1), companies (banks included!) invest a lot into branding, and part of branding is the way that a company communicates to public. Word (lexical) choice can have a significant effect on branding [133, 252, 127], as can tone [252, 127]. Therefore, more work is needed to evaluate and determine the extent to which ChatGPT is able to emulate a company’s style in keeping with its brand and communications approach.

4.5.6 Conclusion, Future Work, and Limitations

In this project, we took the SMOTED tabular data (synthetic only) generated in Project 2 and passed it to ChatGPT for generation. FS have limited corporate use as well as research use without the contextualizing information that is either provided in the Notes or in the financial report text. Therefore, the main goal of Projects 2 and 6 were to provide synthetic FS along with contextualizing synthetic text. Although we demonstrated that this end-to-end approach is very useful, we also showed that ChatGPT is not yet ready to take on this type of task. Even though we provided it with in-context training, it was not able to distinguish the different styles of the banks sufficiently to replicate each bank’s

¹⁵Many thanks to Dr. Kala Hirtle for this excellent example.

own style, and used its own grandiose style instead which impeded readability, which can affect understandability and accessibility — two aspects of communication that the financial domain already struggles with. The fact that ChatGPT is not helping these two aspects, but rather hindering them, means that more work is needed before ChatGPT should be brought in to help on this task.

The main challenge with this Project was that there is no universally accepted approach to evaluating text without ground-truth. As ChatGPT and other LLMs became more mainstream, we must work towards establishing ways to evaluate text where there is no ground-truth. To address this challenge, we used Semantic Similarity and evaluated the ChatGPT text to quasi-ground truth — financial text taken from each bank’s human written reports for the real data. The intuition behind this was that the text does not have to have the same words, but it must have the same or opposite meaning, depending on the situation. Even though the semantic similarity never reached the threshold of 0.8, the majority of the scores were in the range of 0.7 to 0.79, showing that it was approaching the threshold. For future work, we would like to expand this project to include notes to the financial statements, as well as other jurisdictions such as Canada and Europe.

The main limitation we had in this Project was the size and focus of the sample (5 US banks). It may be that ChatGPT would have performed better if a larger sample was used, or samples from other jurisdictions were used. The size of the in-context training sample could also have been a limitation, which we plan to evaluate and improve on in future work.

Chapter 5

Foreshadowing the Need for Company Author Attribution

5.1 Project 7: Company Authorship Attribution and Verification

5.1.1 Introduction

The study of authorship has been an active area of research for many decades and has created a robust body of knowledge. This field is very broad and comprises work done on attribution, verification, and profiling, among others. Attribution and verification are two highly related areas; attribution seeks to identify an unknown text’s most likely author¹, whereas verification seeks to verify that the stated author is the most likely author. Profiling identifies demographics such as gender, age, location, and native language [277, 273, 76], for example, which can then be used to better understand the author or use these demographics in attribution or verification tasks. Initial work done on attribution, verification, and profile focused on one author. In recent years, however, there has been an increase in interest in identifying authors in collaborative works, particularly with the rise of collaborative writing platforms such as *Wikipedia* and *Google Docs*, for example. In all cases, whether doing tasks with single or multi-authored documents, the true author labels are available to the researcher, and are only hidden or labelled as “unknown” for the purposes of the experiment.

However, many corporate documents, like the [AR](#), do not list the author. Rather such corporate documents are seen as being authored by a “figurehead” author, meaning that the company is the “author” (in name only), and the real authors of the report are never identified. As a very brief review, the [AR](#) is a vehicle by which public companies update investors, stakeholder, and regulatory bodies, on their performance during the past year and layout plans for the future. The AR has many contributing authors such as the Chief Executive Officer, the Chairperson of the Board, and firm management [162]. While author

¹While the [AR](#) contains the company’s name, there are many sections in the [AR](#) that do not contain the company’s name. Therefore, whether the task is attribution or verification will depend on the text passage under consideration.

profile demographics (e.g. gender, age, location, and native language) can be helpful in author attribution tasks [76], the only demographic information that we have access to is where the specific author of a section is identified — the Chairperson’s letter ². This also assumes that the author (the Chairperson in our example) was the *true* author of the letter, rather than having an aide draft the letter, and then the Chairperson amending the draft. Yet, in most cases, the letter is drafted on their behalf, and then amended by the Chairperson [86]. Also, a company’s writing “team” can change from one year to the next, as “up to 20 people” [86] can be involved, or the company could use an external firm to write the report on their behalf [86]. This introduces the potential for inconsistency in authors, particularly if the author(s) are unknown.

Unlike other types of documents and works (e.g., peer reviewed or non-peer reviewed research, fiction or non-fiction work, or traditional or social media articles), corporate documents tend to reuse a lot of text and information from one year to the next. The amount of text reuse can be seen as either a feature, a genre-specific peculiarity, or both. While companies will use a writing team — either internal or external — they usually do not start the document from a “scratch”; instead, they take last year’s report (or a previous year’s report) and then make amendments to it [86]. Figure 5.1 shows a sample page derived from comparing RBC’s 2012 and 2013 Annual Reports. The black text shows text that remains unchanged between the 2012 and 2013 reports. The blue text shows, through track changes, which text was changed for the 2013 report, using the 2012 as its “starting point”. We did this comparison using the Python library Aspose.Words [20]. Companies invest significant amounts of resources into branding ³ and are highly aware of the importance of “[a]ll the elements of corporate brand, from tone and personality, functional and emotional benefits, core message and end goal, to its reputation [213]. A company’s AR is one of the manifestations of the brand, and each company (banks are no different) will go to great lengths to distinguish itself from that of competitors.

There are several caveats that need to be kept in mind when reviewing the related work and experiments below. First, we recognize that, unlike our company situation discussed above, co-authors of research or literary works are specifically named, thereby affording the

²The Chairperson’s Letter is an integral part of the AR and is usually one of the first documents presented in the AR, for example.

³In reviewing the income statements from the 15 banks in the data, money spent on advertising ranged from \$173 million to \$3.9 billion. Branding is highly enmeshed with marketing.

performance and non-GAAP measures section. (4) Effective the first quarter of 2013, we calculate capital ratios and Assets-to-capital multiple using the Basel III framework. Capital ratios presented above are en an, "all-in," basis. Capital ratios and Assets-to-capital multiple in 2012 were calculated using the Basel II framework. Basel III and Basel II are not directly comparable. The CET1 ratio is a regulatory measure under the Basel III framework and is not applicable (n.a.) for 2012. For further details, refer to the Capital management section. (5) Includes \$31.2 billion (2013 , \$32.6 billion, 2012 , \$38.4 billion) of securitized mortgages and credit card loans. (6) Defined as dividends per common share divided by the average of the high and low share price in the relevant period. (7) Average amounts are calculated using month-end spot rates for the period. n.m. not meaningful Management's Discussion and Analysis Royal Bank of Canada: Annual Report 2014 11 About Royal Bank of Canada Royal Bank of Canada (RY on TSX and NYSE) is Canada, "s largest bank, and one of the largest banks in the world, based on market capitalization. We are one of North America, "s leading diversified financial services companies, and provide personal and commercial banking, wealth management services, insurance, investor services and capital markets products and services on a global basis. We employ approximately 78,000 full- and part-time employees who serve more than 16 million personal, business, public sector and institutional clients through offices in Canada, the U.S. and 38 other countries. For more information, please visit rbc.com. Our business segments are described below. Personal & Commercial Banking operates in Canada, the Caribbean and the U.S., and comprises our personal and business banking operations, as well as our auto financing and retail investment businesses. Wealth Management serves affluent, high net worth and ultra-high net worth clients from our offices in key financial centres mainly in Canada, the U.S., the U.K., Channel Islands, continental Europe, and Asia with a comprehensive suite of investment, trust, banking, credit and other wealth management solutions. We also provide asset management products and services directly to institutional and also to individual clients through our distribution channels and third-party distributors. Insurance provides a wide range of life, health, home, auto, travel, wealth and reinsurance products and solutions. We offer insurance products and services through our proprietary distribution channels, comprised of the field sales force which includes retail insurance branches, our field sales representatives, call centres and online, as well as through independent insurance advisors and affinity relationships in Canada. Outside Canada, we operate in reinsurance markets globally. Investor & Treasury Services serves the needs of institutional investing clients by providing asset servicing, custodial, advisory, financing and other services to safeguard assets, maximize liquidity and manage risk in multiple jurisdictions around the world. We also provide short- term funding and liquidity management for RBC. Capital Markets provides public and private companies, institutional investors, governments and central banks with a wide range of products and services. In North America, we offer a full suite of products and services which include corporate and investment banking, equity and debt origination and distribution, and structuring and trading. Outside North America, we offer a diversified set of capabilities in our key sectors of expertise such as energy, mining and infrastructure and we are now expanding into industrial, consumer and health care in Europe. Our business segments

Created with an evaluation copy of Aspose.Words. To discover the full versions of our APIs please visit: <https://products.aspose.com/words/>

Figure 5.1: A comparison of RBC's 2012 and 2013 annual reports using the Python library Aspose.Words. The blue text shows, through track changes, what text has changed between 2012 and 2013. The black text shows the text that has not changed between the two years.

opportunity to review the work of those co-authors where they were the sole authors. Second, due to the regulatory nature of the AR, the format and presentation will be very similar between companies. Third, to reduce the burden of writing the reports, which, as previously identified in Section 4.3.1, is very long (on average ≈ 186 pages in length in Canada and the US [128]), companies tend to re-use a lot of text year-over-year; this is not a practice that authors in other situations (like the ones outlined above) would normally follow. And fourth, regulatory documents tend to have boilerplate advisories for readers. The MD&A, for example, begins with such an advisory warning readers that actual results may be different than the expected or anticipated results discussed as part of the MD&A. The beginning of the Bank of New York Mellon’s boilerplate advisory at the start of its MD&A is as follows:

The Company’s actual results of future operations may differ from those estimated or anticipated in certain forward-looking statements contained herein for reasons which are discussed below and under the heading “Forward Looking Statements...” [21]

We note here that there is a philosophical and functional question of what the difference is between a writer and an author, given the fact that a company’s writing team (internal or external) is not necessarily *creating* the material, but rather reporting on and discussing what has already occurred in financial *life* of the company, as well as discussing ancillary, promotional, or forward-looking issues. While this question is interesting, it is out of the scope of the thesis and will not be addressed. We do, however, recognize the dichotomy, and for the purposes of this thesis, will continue to use the term “author” to refer to any writer involved in the creation of the AR.⁴

While a tremendous amount of work has been done using the AR as its base document such as sentiment analysis (e.g., [125]), evaluating market reactions to the tone in the annual report (e.g., [348]), or predicting stock market returns (e.g., [124]), for example, to the best of our knowledge, no work has yet been done on authorship attribution, verification, or profiling of the AR, which represents a major gap in the research. The research that we present in this chapter contributes to filling that gap.

⁴We also recognize the argument made by other researchers indicating that level of contribution as well as active and timely contribution are needed to *qualify* as a co-author [139]. As this is outside of the scope of our research, we will be focusing on the text itself and how it can be identified as *Company A* versus *Company B*, for example, rather than making any attempts to quantify co-author contribution to determine if their contribution qualifies them for co-authorship.

As well, the implications of the lack of known authors for the AR is profound in that it presents a very unique research problem — one that is a cross between single author and multi-author attribution where only the “figurehead author” is known. In this context, we, therefore, have to approach this research as a single author attribution knowing that, in reality, it is a multi-author attribution that comprises both writing and non-writing authors (all of which we will likely never know), a strong tendency to reuse significant portions of text year-over-year, and where the text must take a collaborative approach so that it reads with “one voice” (even though it has been authored by many). This combination makes it a very interesting and important research problem.

5.1.2 Problem Statement

For this work, we explore three common approaches to authorship attribution: topic modelling (using Latent Dirichlet Allocation [25]), stylometry (using Burrows’ Delta [38]), and distance (using Common N-gram distance [164]), to determine which is the most effective approach for the AR. As well, we circle back to the work that we presented for financial statement generation and use ChatGPT (using GPT-4) to generate text for the synthetic statements produced as part of the research in Chapter 3.

5.1.3 Methodology

5.1.3.1 Data Collection and Processing

ARs from Canada, the US, and Europe were used to conduct this research. For the purposes of this research, we refer to Europe in a “continental” sense. Therefore, the United Kingdom is included as part of “continental” Europe, as it has been both in ⁵ and out ⁶ of the European Union over the time period of the dataset. Five large and well-known banks were selected from each jurisdiction, presented in Table 5.1 and reports were downloaded directly from the bank’s website. The number of annual reports gathered for each bank depended on availability. Some banks only provide, for example, the last 10 or 20 years, whereas Wells Fargo makes reports available back to 1969 on its website. Table 5.1 indicates the date ranges of the reports included for each bank.

⁵Britain joined the European Union January 1, 1973

⁶Britain’s departure from the European Union (BREXIT) took effect January 31, 2020.

Canada (# of Reports)	US (# of Reports)	Europe (# of Reports)
RBC (23)	JPM (20)	Barclay’s (33)
CIBC (13)	BNY (15)	Deutsche Bank (16)
BNS (21)	WF (28)	Sberbank (10)
BMO (23)	Citi (14)	Erste Bank (15)
NBC (8)	USB (22)	Rabo Bank (12)

Table 5.1: Bank Listing

As some of the older reports are scans of original paper copies and not PDFs created directly from the source document, optical character recognition (OCR) was needed. To convert the PDFs to text, we used PDF2GO’s OCR engine, which upon visual inspection, worked fairly well as the OCR engine returned clean text and removed the pictures and graphics. We did notice that the tables sometimes converted strangely, particularly if the document used a two-column format but the table spanned both columns. As our focus was on the text itself, rather than the contents of the tables, we determined that the poor conversion of the tables would not present an issue in the context of our research. We did try several well-known tools such as PyTesseract, OCRmyPDF (which uses the Tesseract engine), and Camelot, but the conversion from PDF to text did not work well, particularly if there was handwriting or other markings (such as a “Received” stamp) on the pages.⁷ The text documents for each bank were then compiled into separate comma separated value (CSV) reference corpora.

ARs gathered were either original English or were translations of the original language into English. Minimal text cleaning was performed as we wanted to preserve the text that each bank (or the translator for the bank) created. This means that any ambiguity of meaning, incorrect word usage in the translations, differences of spelling between jurisdictions (such as Canadian/British spelling versus American spelling or any Cyrillic text) were kept. As well, text cleaning and formatting can be a lengthy process. With the large-scale availability of LLMs, where LLMs could be used for fraudulent purposes, time for lengthy text cleaning may not be available. Therefore, we took a “minimalist” approach to text cleaning.

While Figure 5.1 is helpful in supporting that there is a considerable amount of text re-use in the ARs, it also raised the concern that the inclusion of textual features (features) such

⁷Other well-known tools for PDF to text conversion are PyPDF and PDFMiner, neither of which have OCR capabilities.

as names, phrases, addresses, and phone numbers could artificially inflate the algorithms’ performance. Therefore, we ran two sets of experiments for each algorithm: *named* and *unnamed*. Named refers to the original text “as is” with no removal of any features. Unnamed refers to the text once the features have been removed. A full listing of all of the features removed for each bank can be found in Appendix I.

5.1.4 Experiments

Three experiments, using Latent Dirichlet Allocation (LDA), Burrows’ Delta, and Common n-Gram distance (CNG), were used to classify the most likely author for a corpus of ARs. As mentioned in Section 5.1.1, the *true* authors of the “disputed” ARs are not known. Therefore, for the purposes of this research, we attributed these ARs to the “figurehead author” (i.e., company). Classification accuracy was used to evaluate the efficacy of each algorithm as it is the predominant metric used in the literature.

Our main corpus contained 15 banks from three jurisdictions. However, due to the size of the reports, Burrows’ Delta (discussed in Section 5.1.4.2) created a bottleneck as it takes a considerable amount of time to run, particularly for the European texts, which are extremely long (512 pages on average⁸). The time complexity is $\mathcal{O}(t \times (l + 2n))$ where l is the average lengths of the texts, t is the number of texts, and n is the number of common words used in the analysis. The number of common words was fixed to 500, meaning that n becomes a constant. l and n both create a linear increase as these inputs grow meaning that the time requirement will scale proportionally with the value of l and n . Due to the large size of the corpus, we broke it into three: one for each jurisdiction (Canada, Europe, and US).

5.1.4.1 Experiment 1 - Latent Dirichlet Allocation

For this experiment, we used 30 documents from our corpus and generated the top 12 topics for each document, with each topic being represented by 12 words along with its distribution. The distributions for each document were then added to a document-by-topic (DBT) matrix [9] and cosine similarity was used for evaluation purposes. Figure 5.2 provides a partial sample of the DBT matrix. Overall, the matrix itself is sparsely populated as some

⁸This was calculated by taking the average of all of the pages of the European reports.

topics, such as equity was only generated for Deutsche Bank. Other topics, such as asset, risk, and financial were generated for the vast majority of banks.

As previously discussed, minimal text cleaning was done to preserve the text in as close to its published format as possible. Using the spaCy Python library with Gensim, each document was then tokenized, stop words were removed, and bigrams were created for each text. Lemmatization was then done, using the part-of-speech (pos) tags for noun, adjective, verb, and adverb, to create a bag-of-words. Given the size of each text, we disabled the parser and named entity recognition features in the spaCy library, as these are computationally expensive and, if kept on, will cause memory to run out. Using a DBT matrix implementation, the space complexity is $\mathcal{O}(D \times K)$, where D is the number of documents and K is the number of topics.

Coherence scores were generated using the Topic coherence pipeline [258] for topics ranging from 5 to 12 to determine the optimal number of topics per document. The coherence score is heavily related to the Normalized Pointwise Mutual Information (NPMI) in Equation 5.1:

$$NPMI(w', w^*) = \frac{\log \frac{P(w', w^*) + \epsilon}{P(w')(w^*)}}{-\log(P(w', w^*) + \epsilon)} \quad (5.1)$$

where w is used to represent words, $P(w', w^*)$ is the joint probability of observing both w' and w^* , Pw' and Pw^* are the marginal probabilities of observing w' and w^* , and ϵ is a small constant used to prevent a log of zero (undefined) in the case where w' and w^* do not occur together.

A key parameter for this model is the coherence measure. Topics learned from a document or a corpus should, theoretically, “provide sufficient information to determine the subject area and interpretation of a topic, and distinguish one topic from another” [223]. The coherence score measures semantic coherence of the topic words to determine the extent to which high scoring words learned from the document or corpus “belong” together [288, 223]. Theoretically, the stronger the coherence score (based on the scoring system for each coherence algorithm), the more the topics belong together thereby providing that “distinguishing” element from one topic to the next. For example, suppose that we have the following two topics, each containing high scoring words:

- Topic 1: car, wheels, roof, headlights, windshield
- Topic 2: house, cat, storm, school, baseball

In evaluating topic 1, the words are all related as these are elements that every car has. Topic 2, however, has some related words such as house and cat, and school and baseball, but the association is not as strong between those words as it is for topic 1. Therefore, topic 1 would score much more strongly than topic 2, indicating that topic 1 is a better topic.

There are a number of different coherence measures available. UCI, “ u_{mass} ”, and “ C_v ” are the measures used most often in the literature:

- UCI measures coherence based on pointwise mutual information [223] which “represents the relation between words and documents” [79]. A major drawback with UCI, however, is its bias towards infrequently used words in the document or corpus, as it assigns very high scores to these rarely used words [68]. Given that financial reports often have infrequently used words (such as “[Concentration Risk](#)” and “[Going Concern](#)”), using this measure would likely introduce bias into the coherence score through the heavier weighting given to rare words.
- u_{mass} measures coherence based on word pair co-occurrence [271]. Often, ϵ is added to this measure to ensure that $\log(0)$ (undefined) is avoided [87]. A requirement to use the u_{mass} measure is a training corpus which is used to calculate the co-occurrence [123]. As we did not have a training corpus, we elected not to use this measure.
- c_v measures coherence using normalized pointwise mutual information. Word vectors are then aggregated into a topic vector which is then used as a reference against which the cosine similarities of each topic will be compared, in order to calculate the coherence score [89]. Although this measure is not as fast as u_{mass} , it has shown to have better topic coherence results [269], leading us to select this coherence measure.

Coherence scores were calculated for 5 to 12 topics for each bank. The optimal number of topics was identified by the highest coherence score for each bank. The coherence scores for each bank (known and unknown) can be found in Appendix H. We performed two experiments — one that includes the bank’s name and any prominent information such as address,

value	loan	credit	group	capital	total	market	share	asset	report	risk	rate
	0.008		0.009			0.008	0.008	0.013	0.009	0.019	
	0.008		0.014				0.007	0.008		0.015	
		0.008	0.014				0.008	0.011	0.01	0.016	
0.01			0.014		0.009			0.01	0.007	0.016	
			0.015				0.01	0.013	0.008	0.019	
			0.012	0.007	0.007		0.008	0.007	0.008	0.017	
0.009		0.006	0.012					0.012	0.009	0.02	0.008
	0.007		0.012				0.008	0.01	0.007	0.017	
0.008	0.008	0.008	0.012						0.008	0.017	0.008
0.008			0.016	0.007			0.007	0.01	0.009	0.022	
0.007			0.013		0.008		0.007	0.01		0.021	
0.007		0.008	0.014					0.013	0.008	0.017	
			0.012				0.009	0.011		0.018	0.007
0.008			0.015				0.008	0.011	0.01	0.018	
		0.007	0.013				0.008	0.013		0.017	
			0.012		0.008	0.007		0.011	0.008	0.014	
0.007			0.016				0.008	0.008	0.007	0.018	0.008
0.008	0.009	0.007	0.011					0.01	0.008	0.02	
0.008			0.013				0.007	0.011	0.009	0.017	
0.007		0.008	0.01				0.009	0.009	0.008	0.02	
0.009		0.011						0.013		0.013	

Figure 5.2: Partial sample of the DBT matrix, showing that it is densely populated for some features, and sparsely populated for other features.

city, self-named programs etc. and one that does not include the bank’s name. We have termed these “Named” and “Unnamed”. Appendix J provides the topic scores for each topic for each bank.

5.1.4.2 Experiment 2 - Burrows’ Delta

To evaluate stylometry, we used the Burrows Delta (Delta) statistic [38], following the work of Laramée [178]. This measure is used to identify the author (or likely author) of an unidentified text and equally weights features to arrive at the result [178]. The entire corpus (Canada, Europe, and the US) was used for this experiment. Due to the size of the documents and the running time required, each jurisdiction was evaluated separately. Training documents were allocated to the “known” group and testing documents were allocated to the “disputed” group. The “disputed” group contains documents whose authorship is in question. Test cases were run individually as Burrows’ Delta is very computationally expensive in the memory needed and the time that it takes to run all of the calculations. The space complexity of Burrows’ Delta is $\mathcal{O}(n)$, meaning that the amount of memory required grows linearly with the size of the input.

Burrows’ Delta is calculated as follows:

$$\Delta_c = \sum_i \frac{|Z_{c(i)} - Z_{t(i)}|}{n} \quad (5.2)$$

where $Z_{c(i)}$ is the Z-score for feature i in the unidentified candidate document c , $Z_{t(i)}$ is the Z-score for i in the test case document, and n is the most frequent words in the corpus which are then used as features [178].

5.1.4.3 Experiment 3 - Common N-gram Distance

The final experiment was conducted using Common N-gram Distance (CNG) implemented in Python. This algorithm works by finding “a relatively small set of frequent n-grams.” [164]. The distance is then calculated between the profile and the unknown (or “testing”) document to determine similarity (or dissimilarity) [39], thereby identifying the most likely author for a given text [157].

CNG is calculated as follows [164]:

$$d_x(P_1, P_2) = \sum_x \left(\frac{f_{P_1}(x) - f_{P_2}(x)}{\frac{f_{P_1}(x) + f_{P_2}(x)}{2}} \right)^2 \quad (5.3)$$

“where $f_{P_i}(x)$ is the normalized frequency of an n-gram x in the profile P_i , $i = 1, 2$, where $f_{P_i}(x) = 0$ whenever x does not appear in the profile P ” [157].

5.1.5 Results and Discussion

5.1.5.1 Experiment 1: Latent Dirichlet Allocation

The cosine similarity was extremely high for all four experiments, ranging from 0.98 to almost 1, which indicates that the reports are highly similar. This is not surprising given the regulatory nature of these reports. While companies do have a certain amount of latitude and discretion about what information they provide regulators and investors, securities exchanges have stringent rules for what must be discussed in the financial reports. The implications of these rules mean that companies within the same industry — banks in this case — tend to discuss the same types of topics in that their key financial measures, risks, customers, and opportunities for growth, for example, are largely the same, if not exactly the same.

This “sameness” can make it difficult for machine learning to distinguish between companies (banks) on the topic level as there is little room for individuality in topics; that individuality would be much more apparent and present in the stylometry of the text as well as the branding behind the text — something that a set of topics cannot capture.

Even though the cosine similarity of the topics was high, we further evaluated the classification on two levels. First, we examined the classification at the jurisdiction level (i.e., Canada, Europe, and the US). As each jurisdiction’s regulations and laws are different, it is important that the algorithm be able to identify the origin of the AR. Second, we examined the classification at the individual bank level (which we termed the “company” level) to determine if the algorithm could correctly identify the correct company as the author. Our reasoning behind this multi-level approach was to determine that if the algorithm was effective at correctly classifying the jurisdiction, was it also effective at correctly classifying the company as the figurehead author.

At the jurisdiction level, we found that using topics alone worked well in that the algorithm could correctly determine, with a high degree of accuracy, if the report was written by authors in Canada, Europe, or the US. Results for the classification accuracy by jurisdiction is presented in two tables — one with the top features removed (Table 5.2) and one with the bottom features removed (Table 5.3). See Section I.2 for information on topics removed. Interestingly, removing the names and identifying information (see Appendix I), improved the jurisdiction-level classification results in both cases (top and bottom features removed). This suggests that the presence of textual features such as name, address, and company programs, for example, may serve as a distracton for algorithm, which is an interesting find.

Although the classification results at the jurisdiction-level were high, the algorithm struggled at the company level seen in Tables 5.4 and 5.5. At the company level, some classification results benefit from having no identifying features, whereas others do not. Results also depend on if the top or bottom features have been removed. Looking at Scotiabank, for example, the classification was higher using un-named when the top features are removed, but low for un-named when the bottom features are removed — 71% vs. 43%. As discussed in Section 5.1.1, the regulatory nature of the AR influences the format and presentation of the reports. This also affects what topics companies *must* address in the AR in that companies are required to speak to items such as liquidity and capital resources as well as

	Canada	Europe	US
Named			
Y	32	42	41
N	9	1	8
Total	41	43	49
Accuracy	78%	98%	84%
Un-named			
Y	36	36	43
N	2	2	0
Total	38	38	43
Accuracy	95%	95%	100%

Table 5.2: Classification accuracy by jurisdiction for “named” and “un-named” with top features removed

	Canada	Europe	US
Named			
Y	38	43	49
N	3	0	0
Total	41	43	49
Accuracy	93%	100%	100%
Un-named			
Y	37	41	52
N	1	1	0
Total	38	42	52
Accuracy	97%	98%	100%

Table 5.3: Classification accuracy by jurisdiction for “named” and “un-named” with bottom features removed

	Canada	Europe	US
Named			
	BMO 6/6	Barclays 4/5	BNY 1/10
Accuracy:	100%	80%	100%
	CIBC 1/8	DB 0/6	Citi 6/10
Accuracy:	13%	0%	60%
	NBC 6/8	Erste 6/10	JPM 6/7
Accuracy:	75%	60%	86%
	RBC 5/7	Rabo 7/9	USB 4/5
Accuracy:	71%	78%	80%
	Scotia 8/12	Sberbank 13/13	WF 4/5
Accuracy:	67%	100%	80%
Un-named			
	BMO 5/8	Barclays 8/8	BNY 7/12
Accuracy:	63%	100%	58%
	CIBC 2/9	DB 6/12	Citi 4/5
Accuracy:	22%	50%	80%
	NBC 5/7	Erste 4/4	JPM 10/10
Accuracy:	71%	100%	100%
	RBC 0/7	Rabo 8/8	USB 11/11
Accuracy:	0%	100%	100%
	Scotia 5/7	Sberbank 5/6	WF 5/5
Accuracy:	71%	83%	100%

Table 5.4: Classification accuracy by company for “named” and “un-named” with top features removed

	BMO 3/6	Barclays 5/15	BNY 10/10
Accuracy:	50%	33%	100%
	CIBC 6/8	DB 0/6	Citi 9/10
Accuracy:	75%	0%	90%
	NBC 5/8	Erste 6/10	JPM 5/7
Accuracy:	63%	60%	71%
	RBC 7/7	Rabo 9/9	USB 2/5
Accuracy:	100%	100%	40%
	Scotia 8/12	Sberbank 13/13	WF 8/17
Accuracy:	67%	100%	47%
Un-named			
	BMO 5/8	Barclays 10/15	BNY 5/7
Accuracy:	63%	67%	71%
	CIBC 8/9	DB 5/5	Citi 6/9
Accuracy:	89%	100%	67%
	NBC 4/7	Erste 7/7	JPM 15/15
Accuracy:	57%	100%	100%
	RBC 2/7	Rabo 7/7	USB 1/8
Accuracy:	29%	100%	13%
	Scotia 3/7	Sberbank 8/8	WF 13/13
Accuracy:	43%	100%	100%

Table 5.5: Classification accuracy by company for “named” and “un-named” with bottom features removed

	Canada	Europe	US
Named	85%	91%	86%
Unnamed	95%	91%	87.5%

Table 5.6: Classification accuracy by jurisdiction for “named” and “un-named”.

a company’s expected future results [49]. Therefore, companies are required to discuss a required set of topics, and address similar types of topics in the information provided beyond the “required”. The implications of these requirements and practices support why removing the top features increased the classification as it was these features that were common to each company which only served to create distractions for the algorithm in its classification task. Yet, other banks like Sberbank were not as affected by either the named vs. un-named, or the top vs. bottom features. The lowest result that Sberbank had was 83%, which was the misclassification of one record; the rest of Sberbank’s results had 100% classification. The almost contradictory nature of these results is something that we plan to investigate in future work.

5.1.5.2 Experiment 2: Burrows’ Delta

The Burrows’ Delta algorithm performed well at the jurisdictional level and returned fairly high results, ranging from 85% to 91% accuracy. Results for both named and unnamed can be seen in Table 5.6. Similar to our findings for LDA, results improved for both Canada and the US when the name and identifying features were removed (See Appendix I for full list of terms removed for each bank). Canada’s classification accuracy increased 10% and the US’ increased 1.5%.

In reviewing the granular results, we observed that Burrows’ Delta did not uniformly struggle across all banks, but rather with specific banks. In Canada, in the named reports, Scotiabank proved the most difficult for Burrows’ Delta to determine the likely author. Burrows’ Delta correctly identified seven of the 11 Scotiabank reports, but identified the “disputed” category as the most likely author for the other four. This struggle continued once the identifying features were removed.

5.1.5.3 Experiment 3: CNG Distance

Results from all three jurisdictions show that word n-grams provide stronger results than character n-grams. Results for each jurisdiction are provided in Tables 5.7 to 5.12 when n is the number of word n-grams or character n-grams, and ℓ is the profile length.

Starting with Canada, the best combination, using word n-grams, is $(n, \ell) = (3, 200)$, for both named and unnamed, which returns 100% accuracy. The best results (95% accuracy) for character n-grams are for (10, 60) named and (12, 500) unnamed. For Europe, word bi-grams perform best, and return 100% accuracy for (2, 60) named and (2, 100) unnamed. The best combinations for character n-grams are (8, 500) and (11, 500) each with an accuracy of 100% and 97.7%, respectively. Unlike Canada and Europe, however, the best results for the US using named word n-grams is 81.4% accuracy. Interestingly, unnamed word n-grams achieve the best results with 100% accuracy at (2, 100). This indicates that removing the name and address, along with common identifiers such as acronyms and the CEO's name, improves performance. This suggests that names and addresses may add noise to the text, making it harder to find common n-grams, particularly if these common identifiers are sparsely used, or not uniformly used. Named character n-grams also return less accuracy than either Canada and Europe. The best combination is (10, 100) with an accuracy of 84%, which is far below the other jurisdictions (Canada at 95% and Europe at 100%). However, we again saw performance improve with the removal of names, as accuracy increased to 98% for (14, 500). It is also important to note that for some companies, the removal of the common identifiers changed the top most common n-grams in the profile. Banks tend to name programs available to customers after themselves to increase brand awareness and recognition. In our corpus, for example, some of Citigroup's programs include Citigold, Citi Quick Link, and Citi Simplicity. Therefore, removal of these types of common identifiers forces the algorithm to find other common n-grams instead.

Given that the annual reports are tightly structured, particularly in the United States, it was interesting to find differences in the results for the algorithm between jurisdictions. Canada and Europe, while slightly different in their actual results, exhibited the same behaviour with word n-grams performing best, followed by a small drop in performance when the name and other identifying features were removed from the text. However, we did not see that in the results for the US; in fact, it was the opposite as performance improved for

both word and character n-grams when the name was removed.

L	N-gram size														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Canada - Named - Word n-grams														
30	0.50	0.88	0.93	0.93	0.95	<u>0.98</u>	<u>0.98</u>	<u>0.98</u>	0.88	0.73	0.70	0.65	0.68	0.60	0.60
60	0.50	0.88	0.95	0.88	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.98	0.88	0.85	0.88	0.75	0.77	0.80	0.88
100	0.43	0.90	0.95	0.95	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.95	0.95	0.98	0.95	0.98	0.98	0.98
200	0.45	0.90	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.98	0.98	0.98	0.95	0.95	0.95	0.93
	Canada Unnamed - Word n-grams														
30	0.53	0.85	0.90	0.88	0.90	0.98	<u>1.00</u>	0.98	0.75	0.65	0.63	0.68	0.58	0.53	0.73
60	0.43	0.80	0.90	0.88	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.95	0.78	0.83	0.85	0.75	0.78	0.78	0.78
100	0.43	0.88	0.90	0.90	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.95	0.95	0.98	0.95	0.98	0.95	0.98
200	0.45	0.88	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.98	0.98	<u>1.00</u>	0.95	0.95	0.95	0.93

Table 5.7: Canada — word n-grams

Although we could not run the full set of annual reports all together as one large jurisdiction, we did perform post-hoc tests, taking two banks from each jurisdiction and ran them all together in order to test inter-jurisdictional performance on a smaller scale. We randomly selected two banks from each jurisdiction to form our corpus: NBC, Scotia, Erste, Sberbank, Citi, and USB. For the post-hoc experiments, we focused on jurisdictional classification only, as the size of the corpus was restricted to just 6 companies. This was done to address the issue with Burrows’ Delta being computationally expensive in both time and space complexity. We also focused on “named” only, and continued to use classification accuracy as the classification metric. Our results indicated that with the smaller corpus, the classification accuracy of LDA dropped to 70% for Canada, 83% for Europe, and 60% (the

L	N-gram size														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Europe Named - Word n-grams														
30	0.34	0.96	0.98	0.96	0.91	0.80	0.77	0.84	0.80	0.77	0.50	0.53	0.50	0.36	0.36
60	0.46	<u>1.00</u>	0.98	<u>1.00</u>	0.89	0.89	0.98	0.98	0.93	0.71	0.66	0.77	0.55	0.48	0.48
100	0.50	<u>1.00</u>	<u>1.00</u>	0.96	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.96	0.93	0.91	0.71	0.71	0.68	0.57	0.57
200	0.50	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.98	<u>1.00</u>	0.98	0.93	0.96	0.86	0.80	0.75	0.89	0.86
	Europe Unnamed - Word n-grams														
30	0.27	0.96	0.96	0.96	0.89	0.82	0.82	0.89	0.82	0.61	0.71	0.71	0.50	0.41	0.41
60	0.39	0.98	0.96	<u>1.00</u>	0.86	0.96	0.98	0.96	0.91	0.73	0.66	0.61	0.59	0.43	0.43
100	0.50	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.98	0.98	0.98	0.96	0.91	0.89	0.73	0.64	0.61	0.80	0.80
200	0.50	<u>1.00</u>	<u>1.00</u>	0.98	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.93	0.96	0.86	0.84	0.73	0.71	0.75

Table 5.8: Europe — word n-grams

L	N-gram size														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	US Named - Word n-grams														
30	0.51	0.77	0.79	<u>0.81</u>	0.72	0.72	0.77	0.79	0.77	0.74	0.63	0.65	0.67	0.54	0.40
60	0.72	0.79	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	0.79	0.79	0.79	0.79	0.79	0.79	0.72	0.65	0.51
100	0.74	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	0.79	0.81	0.79	0.74	0.74	0.70	0.72	0.65
200	0.77	<u>0.81</u>	0.79	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	<u>0.81</u>	0.74	0.74	0.74	0.79	0.74
	US Unnamed - Word n-grams														
30	0.66	0.96	0.94	0.92	0.86	0.88	0.94	<u>0.98</u>	0.96	0.94	0.84	0.80	0.72	0.66	0.66
60	0.80	0.96	0.98	0.98	0.96	<u>1.00</u>	0.96	0.98	0.98	0.98	0.90	0.90	0.84	0.82	0.76
100	0.84	<u>1.00</u>	0.96	0.96	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.98	<u>1.00</u>	0.92	0.92	0.88	0.84	0.72
200	0.94	<u>1.00</u>	0.94	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.92	0.92	0.92	0.92	0.86

Table 5.9: US — word n-grams

biggest drop amongst the three jurisdictions) for the US. Burrows' Delta struggled significantly for Canada and Europe, dropping to 50% for both Canada and Europe (from 85% and 91%, respectively), yet increasing the US from 86% to 100%. Finally, the CNG algorithm was still able to achieve 100% classification for each jurisdiction. As part of future work, we plan to extend the post-hoc tests by experimenting with different combinations of companies across jurisdictions. We also plan to revisit the implementation of Burrows' Delta to see if the algorithm can be further optimized in order to reduce the time and space complexities.

L	N-gram size														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Canada Named - character n-grams														
30	0.10	0.10	0.10	0.43	0.5	0.58	0.55	0.63	0.78	0.70	<u>0.95</u>	0.93	0.80	0.83	0.85
60	0.10	0.10	0.35	0.43	0.45	0.60	0.63	0.63	0.80	<u>0.95</u>	0.88	0.88	0.88	0.88	0.88
100	0.10	0.10	0.33	0.43	0.48	0.65	0.65	0.73	<u>0.95</u>	0.88	0.88	0.88	0.88	0.88	0.88
200	0.10	0.10	0.28	0.38	0.45	0.48	0.65	0.90	0.88	0.88	0.88	0.88	<u>0.93</u>	0.85	0.90
500	0.10	0.10	0.10	0.33	0.48	0.63	0.80	0.88	0.88	0.90	0.93	0.93	<u>0.95</u>	<u>0.95</u>	<u>0.95</u>
	Canada Unnamed - character n-grams														
30	0.10	0.10	0.10	0.48	0.50	0.63	0.50	0.53	0.63	0.58	0.83	0.68	<u>0.88</u>	0.75	0.83
60	0.10	0.10	0.25	0.48	0.63	0.55	0.60	0.70	0.73	0.65	0.73	<u>0.88</u>	<u>0.88</u>	0.83	0.83
100	0.10	0.10	0.25	0.43	0.43	0.63	0.75	0.50	0.83	<u>0.88</u>	<u>0.88</u>	<u>0.88</u>	<u>0.88</u>	0.85	<u>0.88</u>
200	0.10	0.10	0.18	0.38	0.45	0.55	0.68	0.90	0.88	0.88	0.88	0.88	0.88	0.85	<u>0.90</u>
500	0.10	0.10	0.10	0.30	0.48	0.63	0.73	0.88	0.88	0.93	0.93	<u>0.95</u>	0.93	0.93	0.93

Table 5.10: Canada — character n-grams

L	N-gram size														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Europe Named - character n-grams														
30	0.11	0.11	0.11	0.16	0.43	0.93	0.93	<u>0.98</u>	0.84	0.84	0.84	0.77	0.84	0.82	0.77
60	0.11	0.11	0.14	0.27	0.68	0.93	<u>0.96</u>	0.89	0.89	0.84	0.84	<u>0.96</u>	0.93	0.93	<u>0.96</u>
100	0.11	0.11	0.11	0.27	0.50	0.93	0.91	0.93	0.93	0.96	0.98	<u>1.00</u>	0.98	0.96	0.96
200	0.11	0.11	0.11	0.23	0.55	0.75	0.93	0.96	0.96	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	0.98	0.96	0.96
500	0.11	0.11	0.11	0.30	0.50	0.2	0.96	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>	<u>1.00</u>
	Europe Unnamed - character n-grams														
30	0.11	0.11	0.11	0.14	0.30	0.43	0.82	<u>0.91</u>	0.82	0.84	0.84	0.82	0.91	0.89	0.89
60	0.11	0.11	0.14	0.18	0.46	0.73	0.86	0.89	0.91	0.84	0.84	<u>0.96</u>	0.93	0.93	0.93
100	0.11	0.11	0.11	0.18	0.32	0.82	0.84	0.93	0.93	0.91	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>
200	0.11	0.11	0.11	0.14	0.50	0.66	0.91	0.91	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>	0.93	0.93	0.93
500	0.11	0.11	0.11	0.21	0.48	0.50	0.89	0.96	0.92	0.96	<u>0.98</u>	<u>0.98</u>	0.96	0.96	0.96

Table 5.11: Europe — character n-grams

L	N-gram size														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	US Named - character n-grams														
30	0.32	0.34	0.40	0.74	0.64	0.72	0.74	0.72	0.78	<u>0.82</u>	<u>0.82</u>	<u>0.82</u>	0.80	0.76	0.74
60	0.22	0.36	0.54	0.72	0.76	0.80	0.74	<u>0.82</u>	<u>0.82</u>	<u>0.82</u>	<u>0.82</u>	0.78	0.78	0.80	0.80
100	0.22	0.34	0.56	0.72	0.78	0.78	0.80	0.80	0.82	<u>0.84</u>	0.80	0.78	0.78	0.80	0.80
200	0.22	0.36	0.52	0.76	0.78	0.80	0.78	<u>0.82</u>	0.80	0.80	0.80	0.80	0.80	0.80	0.80
500	0.22	0.36	0.42	0.68	0.80	0.82	0.80	0.80	0.78	0.80	<u>0.84</u>	0.80	<u>0.84</u>	<u>0.84</u>	<u>0.84</u>
	US Unnamed - character n-grams														
30	0.30	0.22	0.32	0.64	0.66	0.84	0.92	0.84	0.94	0.86	0.92	<u>0.94</u>	<u>0.94</u>	0.88	0.86
60	0.22	0.34	0.50	0.72	0.90	0.94	0.84	0.94	0.92	<u>0.96</u>	<u>0.96</u>	0.92	0.92	0.94	<u>0.96</u>
100	0.22	0.34	0.52	0.60	<u>0.96</u>	0.92	0.94	0.92	<u>0.96</u>	0.94	0.94	0.94	0.92	0.94	0.92
200	0.22	0.36	0.50	0.72	0.88	0.88	0.94	0.94	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>	0.94	0.94	0.94
500	0.22	0.34	0.36	0.62	0.88	0.88	0.96	0.94	0.94	0.94	0.96	0.94	0.96	<u>0.98</u>	<u>0.98</u>

Table 5.12: US — character n-grams

Chapter 6

Conclusion and Future Work

For decades, there has been a growing body of research in financial machine learning (FML) related to securities; predicting stock prices and algorithmic trading, for example, are very popular topics. Yet, computational financial research (CFR) on financial statement data and textual financial data remains thin, particularly at the *company* level, leaving significant gaps in the literature. As well, with the rapid advancement of AI in 2023, the FML gaps have only widened as the use of LLMs to predict the stock market have further increased interest in securities research. The research in this thesis contributes to filling the FML gaps in CFR for financial statement data and textual financial data.

6.1 Achieved Objectives

1. Predicting distress of financial intermediaries using CNN models

Our first project examined how CNNs can be used to predict the distress of financial intermediaries using a small dataset. As discussed in Chapter 3, it is not always possible to have a large amount of company-level data, and using an aggregated dataset of multiple companies is not always appropriate. Therefore, in a world of ubiquitous financial data, we faced an interesting problem — the scarcity of financial statement data at the company-level. Therefore, to [augment](#) company-level data, we drew inspiration from computer vision, which uses data transformations such as flipping and rotating to [augment](#) image datasets, and applied a shift of $\pm 10\%$ to our data to [augment](#) it. The dataset was originally 63 records, which was augmented to 126 records (using one shift of either plus or minus 10%) or 189 records (using both shifts of plus and minus 10%). Using four convolutional layers, we were able to achieve an accuracy of 88.24% for predicting the distress of financial intermediaries, which, when compared to similar research done by contemporary peers, was a promising result, particularly considering that our dataset was considerably smaller than those used by other researchers.

When working with deep learning, having *more* data is always a preferable situation, as training requires a considerable amount of data. However, this research shows two important aspects in relation to that *need*: firstly, what is conventionally thought of a “big data” is not necessarily needed for deep learning. We were able to reasonably use a deep learning CNN model and achieve good results with using a fairly small dataset. And secondly, it is not always possible to have *big* data for training at the company-level. This means that in order to employ deep learning models, researchers and practitioners must find ways to appropriately [augment](#) the data. This need became the catalyst for the second project.

2. Augmentation strategies using DL

Although the rudimentary shifts of $\pm 10\%$ returned good results, we wanted to explore established ML approaches to data augmentation such as GANs, VAEs, and SMOTE that could automate the augmentation process. Financial statements are the backbone of any company’s financial health, and since there is a major gap in the FML literature with respect to using financial statements in CFR, we focused on finding the best method to [augment](#) company financial statements. As discussed in Chapter 3, however, we are unable to pass a dataset of financial statements to an ML model due to the dependency between statements. Simply passing the data to a GAN, for example, returns (essentially) *junk* data as these models do not capture the necessary balancing figures in the creation of the synthetic data. We demonstrated that only specific features could be used as a basis of augmentation and that the financial statements, if needed, would need to be “re-assembled” after augmentation. While feature-based learning and augmentation are already accepted practices, augmentation of financial statements using “automatic” frameworks has never been done (to the best of our knowledge), meaning that we first had to test passing the full dataset to the models and establish why this approach should not be used. We evaluated several well-known algorithms that can be used to create synthetic data, finding that SMOTE worked the best, but not under every condition. If the dataset was too small or there were too few samples in the minority class(es), then the SMOTE algorithm cannot be used. To address this issue, we introduced Peer Augmentation, which used the data of peers to scaffold the creation of sufficient initial data where peer scaffolding can be used as a stand-alone method or in an ensemble with SMOTE.

3. Fine-Tuning GPT-3

This project was the first of five projects that examined the use of ChatGPT for financial text. GPT-3's *davinci* model — the most capable fine-tunable model that Open AI had at the time of the research — was fine-tuned using prompt and completion pairs of financial text for five large well-known banks in the US to investigate how fine-tune training would affect cosine similarity, sentiment, and reading ease of the ChatGPT generated text, comparing it to the human-written text. We found that fine-tune-training positively affected the cosine similarity with a statistical significance of $p < 0.05$. We also found statistically significant increases in readability when comparing human to the base *davinci* model, but more importantly when comparing the base *davinci* model to the fine-tune trained *davinci* model.

The most important find of this research, however, was discovering that without fine-tune training, the sentiment that the base *davinci* model generated was not in-line with the sentiment of the human-written text. Given the effects that textual stylistic choices have on branding, changing the sentiment of the messaging could have unintended consequences. After the model had been fine-tune trained, however, we showed that the sentiment, while not perfectly aligned, was much closer aligned, showing the importance of fine-tune training.

4. Using ChatGPT to Generate Note Disclosure⁴

While **FS** can “stand on their own” to a certain degree without the **Notes**, one can only glean very superficial information by looking at the **FS** themselves. The **Notes** are there to provide important context on the values in the **FS** to help the reader understand. Therefore, this project evaluated how well ChatGPT could generate two common **Notes** — basis of preparation and subordinated debt. We found that at the time of the research, that generating **Notes** was a very challenging task for ChatGPT to complete, as it was not able to provide the necessary output at the time, as important information was left out or more concerningly, ChatGPT was adding in information that is not usually contained in the **Notes**. We did, however, find that readability increased for ChatGPT's text, making it much more accessible to readers of the annual report and the **FS**.

5. ChatGPT For Financial Literacy

This project helped evaluate the extent to which ChatGPT can help students and individuals improve their financial literacy. Given that everyone is affected by finance — by payroll, taxes, and retirement benefits, for example — having basic financial knowledge is important. For students who aspire to become accountants, financial analysts, work in securities, or

become business leaders, it is important that they get good foundational and upper-level training that is both factually and technically correct, and training that will extend their knowledge and understanding. For humans, this can be accomplished through education and experience. As AI moves forward, training and experience are more in flux than ever. Massive Open Online Courses, better known as MOOCs, have made training more accessible, and “work from home” strategies have allowed people to enter (or continue in) the workforce from *home locations*. ChatGPT has already been used and evaluated in the classroom — to both acclaim and disappointment in its abilities. Therefore, we wanted to evaluate how well ChatGPT understands its users and if (and how) it can adapt its generation.

The experiments that we ran for this project were done in early 2023 using GPT-3.5 and GPT-4. Overall, we found that ChatGPT does not “understand”¹ the needs of its users well in that it cannot adapt its text well for its audience. This can impede (human) user understanding, leading to incorrect learning and understanding. Based on the available literature, we believe that we are the first to bring forward this issue. This lack of “understanding” and tailoring needs to be more fully investigated and better understood to determine if the base corpora used to train the models is the root cause, that there is not sufficient human reinforcement learning being used, or that (human) users need to better craft their prompts to get better answers from ChatGPT.

6. Generating Financial Text using Synthetic Data

In this project, we used ChatGPT to create synthetic textual data for the synthetic tabular data created in Project 2. A key aspect of this task, however, was the lack of ground-truth text for comparison. To the best of our knowledge, we are the first to consider this task for text. To evaluate the AI-generated text, we used an ensemble method consisting of checking the accuracy of the figures in text against the tabular input, and then used “pseudo ground-truth” which uses the text of a human-written document that is similar to the textual task we are asking ChatGPT to complete. Therefore, we used semantic similarity as our metric as the text does not have to use the same or similar words as the pseudo ground-truth text, but it does have to have the same meaning. We found that while only two of the ChatGPT generated text passages exceeded the 0.8 threshold, the remainder were very close

¹We use quotation marks here to denote the idea of “understanding” fully comprehending that AI does not have the same type of understanding as humans do. It seems, however, that at this time, the domain of AI does not yet have a comparable term for AI-understanding versus that of humans. Therefore, we use the best approximate, which is “understanding”.

as the majority (12 of the remaining 18) were in the range of 0.7 to 0.79. A very interesting find, however, was that readability decreased. This was extremely surprising given that in the previous experiments, readability usually increased when using ChatGPT. We attribute this drop in the readability to ChatGPT’s attempt at being “fancy”, citing the example of ChatGPT using phrases such as “prudent interest income management” when none of the examples that were given to ChatGPT as part of its in-context training used such phrasing.

7. Company Author Attribution

After evaluating current and proposed methods for tabular augmentation for financial statements, we then assessed authorship attribution for ChatGPT. The purpose of this research was to establish a baseline so that when we provided ChatGPT with financial inputs from the synthetic financial statements created in Chapter 3, we could determine how accurate ChatGPT’s generated segments of textual financial reports were.²

In some respects, determining if ChatGPT authored the segments or a human wrote them is an *artificial* problem as research that flagged financial fake news³ has not gained widespread attention. Given the date of the research (2018, with lawsuits from the SEC dating back to 2014 and 2017) [172] we can conclude that LLMs were not involved (GPT-1 was released in 2018). However, since the release of ChatGPT in 2022, this tool has made tremendous leaps forward, meaning that it is only a matter of time until ChatGPT (“unknowingly” and “unwittingly”) used to create financial fake news, which will in-turn, harm the market. Therefore, having a documented and well-established baseline of authorship is needed.

In our experiments for this project, we used a myriad of tools to evaluate the text generated by ChatGPT against its human counterparts, as well as comparing ChatGPT against itself on a year-over-year (YoY) basis. Due to the limitations of ChatGPT’s generative capabilities (the research was done largely in the first half of 2023), we focused on short text including selections of results from the management discussion and analysis section of the AR as well as note disclosures. Like the previous chapter, one of the main goals for this research was to establish and document baselines for similarity between human-authored and

²We focused on segments as it is not yet possible to provide ChatGPT with a set of financial statements and other supporting data and have it generate a full annual report. We anticipate, however, that with the continued evolution of AI, that this will soon (in several years’ time) become a possibility. But, for now, ChatGPT is limited to 3,000 words.

³Part of the dataset used for this research was from an SEC investigation.

ChatGPT-authored work, authorship attribution, readability, and sentiment. Our findings show that ChatGPT cannot yet emulate the style and voice of the human-authored bank text and without fine-tune training cannot generate text that has a comparable sentiment. Using the Flesch Reading Ease Score (FRE), we did find that the ChatGPT-generated text was more readable than the human-authored text. This is encouraging, as readability is an area of communications struggle for the financial domain.

An unanticipated, but extremely helpful, byproduct of this research was discovering the amount of text re-use in both the human-authored text and the ChatGPT-authored text. While it is intuitive that humans would use a “control-find” strategy to update the reports from year-to-year, this research reinforces this supposition. Furthermore, when considering the interest in detecting and removing boilerplate text from web corpora, this could create significant problems for financial text given that there, to the best of our knowledge, no understanding of what constitutes boilerplate in the financial domain, and how many times text can be reused before it is considered boilerplate. Therefore, our research has uncovered an important area of future research.

6.2 Future Work

As this thesis spanned seven projects, we have developed a number of items for future work. We will discuss each in descending order of importance.

Publicly Accessible Datasets - In the computational financial domain, researchers do not often share datasets. This impedes progress of the literature and also makes results between researchers difficult to compare, as often researchers are using custom datasets that are not publicly available or shared (even when their research paper indicates that data will be made available upon request). We would like to create a publicly available computational financial dataset repository with data statements so that researchers can then access the data and also know important aspects such as what is in the dataset, how it was constructed, and what decisions were made for inclusion and exclusion. We believe that this would be a significant contribution to research.

Readability - In some of the Projects, readability was enhanced when using ChatGPT, but not in others. Given that communication is a significant challenge for the financial domain, we would like to do more research as to what affects ChatGPT’s use of words and style that

results in text that is more or less readable. This will also help to better understand what ChatGPT’s own style is, which will be foreshadowed in the last project will be important.

Expand ChatGPT testing - as ChatGPT evolves, it is acquiring more capabilities. It will be important to not only continue to benchmark its abilities, but also expand the range of testing to include things like analyst reports, earnings reports, and financial press releases, in order to more fully capture the various communication tasks in the financial domain.

Gemini - Google’s Gemini was just released (February 8, 2024). It offers a “ChatGPT-like” type of interface, and feels very reminiscent of ChatGPT. However, in some initial testing that we have done, its output (for the prompts that we have passed to it) is vastly different, even when benchmarking against ChatGPT’s very early performance. While we (as users) are not back to the beginning in terms of learning how to use an LLM, the fact that we cannot use the same prompt and get a certain level of expected output means that we have to learn how to fine-tune our prompts and expectations for Gemini. This opens up a very exciting avenue of research; it will be interesting to see if users self-select into camps similar to those of PC and Mac users.

6.3 Coda

A large focus of this thesis has been evaluating the evolving quality of ChatGPT’s textual generation, but also its recent capabilities for tabular generation. We focused on ChatGPT because of its ubiquity and ease of access to the general public. In just over a year since its public release in November 2023, it has been quickly adopted and become a main stream *business partner* technology to help improve productivity, streamline tasks, and act a knowledge source. Even with this breakthrough, or “inflection moment” [130], there is still much to learn. We end with a quote about AI from Fei-Fei Li — renowned and world class computer scientist [64]:

“We should recognize human intelligence is very, very complex. It’s emotional, it’s compassionate, it’s intentional, it has its own blind spots, it’s social. When we develop tomorrow’s AI, we should be inspired by this level of nuance.”

References

- [1] Shaza M Abd Elrahman and Ajith Abraham. A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1(2013):332–340, 2013.
- [2] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.
- [3] Adam Hayes. Stock Symbol (Ticker Symbol): Abbreviation for a Company’s Stock. <https://www.investopedia.com/terms/s/stocksymbol.asp>, 2023. Last Accessed: 2024-02-26.
- [4] Kailash Ahirwar. *Generative adversarial networks projects: Build next-generation generative models using TensorFlow and Keras*. Packt Publishing Ltd, 2019.
- [5] Shamima Ahmed, Muneer M Alshater, Anis El Ammari, and Helmi Hammami. Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, 61:101646, 2022.
- [6] Google AI. Gemini advanced language model. Internal Google Research Project, 2024. Accessed February 11, 2024.
- [7] Aymen Ajina, Mhamed Laouiti, and Badreddine Msolli. Guiding through the fog: does annual report readability reveal earnings management? *Research in International Business and Finance*, 38:509–516, 2016.
- [8] Hussam Alkaissi and Samy I McFarlane. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2), 2023.
- [9] Alvin Chen. Enc2045 computational linguistics. https://alvinntnu.github.io/NTNU_ENC2045_LECTURES/intro.html, n.d. Last Accessed: 2023-24-9.
- [10] Ayed Alwadain, Rao Faizan Ali, and Amgad Muneer. Estimating financial fraud through transaction-level features and machine learning. *Mathematics*, 11(5):1184, 2023.
- [11] American Cancer Society. Cancer Staging. <https://www.cancer.org/cancer/diagnosis-staging/staging.html>, 2022. Last Accessed: 2024-01-8.
- [12] J Anuradha et al. A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia computer science*, 48:319–324, 2015.
- [13] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

- [14] Shlomo Argamon. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, 2008.
- [15] Paul Armstrong. It will be much better , when chat gpt can remember about all previous chats. <https://community.openai.com/t/it-will-be-much-better-when-chat-gpt-can-remember-about-all-previous-chats/152017/3>, 2023. Last Accessed: 2023-05-12.
- [16] Valerio Arnaboldi, Andrea Passarella, Marco Conti, and Robin I.M. Dunbar. Chapter 5 - evolutionary dynamics in twitter ego networks. In Valerio Arnaboldi, Andrea Passarella, Marco Conti, and Robin I.M. Dunbar, editors, *Online Social Networks*, Computer Science Reviews and Trends, pages 75–92. Elsevier, Boston, 2015.
- [17] Devansh Arpit, Stanislaw Jastrzkbki, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.
- [18] Insaf Ashrapov. Tabular gans for uneven distribution, 2020.
- [19] Insaf Ashrapov. tabgan 1.2.1. <https://pypi.org/project/tabgan/>, n.d.
- [20] Aspose. aspose-words 23.9.0. <https://pypi.org/project/aspose-words/>, 2023. Last Accessed: 2023-24-9.
- [21] Bank of New York Mellon. 2008 Annual Report. <https://www.bnymellon.com/us/en/investor-relations/annual-reports-and-proxy.html>, n.d. Last Accessed: 2023-16-1.
- [22] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.
- [23] Bill Gates. The Age of AI Has Begun. <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun.>, 2023. Last Accessed: 2023-03-31.
- [24] Ekaba Bisong and Ekaba Bisong. Logistic regression. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, pages 243–250, 2019.
- [25] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [26] Public Company Accounting Oversight Board. Consideration of an Entity's Ability to Continue as a Going Concern. <https://pcaobus.org/oversight/standards/auditing-standards/details/AS2415>, N.D. Last Accessed: 2021-06-2.
- [27] Joy Bose. Extraction of relevant images for boilerplate removal in web browsers. *arXiv preprint arXiv:2001.04338*, 2019.
- [28] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

- [29] Bauke Brenninkmeijer, A de Vries, E Marchiori, and Youri Hille. On the generation and evaluation of tabular data using gans. *PhD diss., Radboud University*, 2019.
- [30] TB Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners advances in neural information processing systems 33. 2020.
- [31] Jason Brownlee. How much data is required for machine learning. <https://machinelearningmastery.com/much-training-data-required-machine-learning/>, 2019. Last Accessed: 2024-01-20.
- [32] Jason Brownlee. How to use learning curves to diagnose machine learning model performance. <https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/>, 2019.
- [33] Jason Brownlee. 4 Distance Measures for Machine Learning. <https://machinelearningmastery.com/distance-measures-for-machine-learning/>, 2020. Last Accessed: 2024-01-20.
- [34] Jason Brownlee. Tune hyperparameters for classification machine learning algorithms. [http://www.morningstar.com/insights/2019/05/02/focused-portfolio](https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/#\protect\protect\leavevmode@ifvmode\kern+.2222em\relax~\protect\protect\leavevmode@ifvmode\kern+.2222em\relaxtext\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{T\global\mathchardef\accent@spacefactor\spacefactor}\let\beginngroup\let\typeout\protect\beginngroup\def\MessageBreak{\Omega(Font)}\let\protect\immediate\write\m@ne{LaTeXFontInfo:oninputline181.}\endgroup\endgroup\relax\let\ignorespaces\relax\accent22T\egroup\spacefactor\accent@spacefactorhe%20most%20important%20hyperparameter%20for,\perhaps%20just%20the%20odd%20numbers\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{&\global\mathchardef\accent@spacefactor\spacefactor}\let\beginngroup\let\typeout\protect\beginngroup\def\MessageBreak{\Omega(Font)}\let\protect\immediate\write\m@ne{LaTeXFontInfo:oninputline181.}\endgroup\endgroup\relax\let\ignorespaces\relax\accent95&\egroup\spacefactor\accent@spacefactortext\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{I\global\mathchardef\accent@spacefactor\spacefactor}\let\beginngroup\let\typeout\protect\beginngroup\def\MessageBreak{\Omega(Font)}\let\protect\immediate\write\m@ne{LaTeXFontInfo:oninputline181.}\endgroup\endgroup\relax\let\ignorespaces\relax\accent22I\egroup\spacefactor\accent@spacefactort%20may%20also%20be%20interesting,\the%20composition%20of%20the%20neighborhood, 2020.</p>
<p>[35] Alex Bryan. Insights. <a href=), 2019. Last Accessed: 2021-08-17.

- [36] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 475–482. Springer, 2009.
- [37] Harry B Burke, Philip H Goodman, David B Rosen, Donald E Henson, John N Weinstein, Frank E Harrell Jr, Jeffrey R Marks, David P Winchester, and David G Bostwick. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79(4):857–862, 1997.
- [38] John Burrows. ‘delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287, 2002.
- [39] Matthew Butler and Vlado Kešelj. Financial forecasting using character n-gram analysis and readability scores of annual reports. In *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*, pages 39–51. Springer, 2009.
- [40] Yuezhou Cai and Wenjiang Wu. Synthetic financial evaluation by a method of combining dea with ahp. *International Transactions in Operational Research*, 8(5):603–609, 2001.
- [41] Handenur Caliskan, Omer Faruk Yayla, and Yakup Genc. A comparative analysis of synthetic data generation with vae and ctgan models on financial credit loan offer data. In *2023 8th International Conference on Computer Science and Engineering (UBMK)*, pages 212–217. IEEE, 2023.
- [42] Canadian Imperial Bank of Commerce. Investor Relations. <https://www.cibc.com/en/about-cibc/investor-relations.html>, n.d. Last Accessed: 2023-11-2.
- [43] Canadian Securities Administrators. Access Rules & Policies. <https://www.securities-administrators.ca/resources/access-rules-policies/>, n.d. Last Accessed: 2023-08-26.
- [44] Canadian Securities Administrators. SEDAR Homepage. https://www.sedar.com/homepage_en.htm, n.d.
- [45] Antonio Victor Canales Saavedra. How is sofr different from libor?, 2021.
- [46] Sean Cao, Wei Jiang, Baozhong Yang, and Alan L Zhang. How to talk when a machine is listening?: Corporate disclosure in the age of ai. Technical report, National Bureau of Economic Research, 2020.
- [47] CFI Team. Equity ratio, n.d.
- [48] CFI Team. Over-the-Counter(OTC). <https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/over-the-counter-otc/>, n.d. Last Accessed: 2024-05-12.

- [49] CFI Team. What is MD&A. <https://corporatefinanceinstitute.com/resources/valuation/mda-management-discussion-analysis/>, n.d. Last Accessed: 2024-05-12.
- [50] Chartered Professional Accountants of Canada. Understanding Reports on Financial Statements. <https://www.cpacanada.ca/-/media/site/operational/rg-research-guidance-and-support/docs/01878-rg-understanding-reports-on-financial-statements-jan-2020.pdf?1a=en&hash=BE02C8C0BDC4B7FB908004C113D9538AE266FBB9>, 2020. Last Accessed: 2023-16-1.
- [51] ChatGPT. Generating synthetic data from tabular csv. <https://www.openai.com/>, 2024. Personal communication via OpenAI's ChatGPT.
- [52] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [53] Haipeng Chen, Sushil Jajodia, Jing Liu, Noseong Park, Vadim Sokolov, and VS Subrahmanian. Faketables: Using gans to generate functional dependency preserving tables with bounded real data. In *IJCAI*, pages 2074–2080, 2019.
- [54] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.
- [55] Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *J. Legal Educ.*, 71:387, 2021.
- [56] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, Mukund Lahoti, and Pratik Narang. A universal metric for robust evaluation of synthetic tabular data. *IEEE Transactions on Artificial Intelligence*, 2022.
- [57] Mark Clatworthy and Michael John Jones. The effect of thematic structure on the variability of annual report readability. *Accounting, Auditing & Accountability Journal*, 14(3):311–326, 2001.
- [58] Mark A Clatworthy and Michael John Jones. Differential patterns of textual characteristics and company performance in the chairman's statement. *Accounting, Auditing & Accountability Journal*, 19(4):493–511, 2006.
- [59] Laura Cleofas-Sánchez, Vicente García, AI Marqués, and Jose Salvador Sánchez. Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing*, 44:144–152, 2016.
- [60] CNN Money. Bailed out banks. <https://money.cnn.com/news/specialstorsysupplement/bankbailout/>, n.d. Last Accessed: 2021-06-2.

- [61] The Alternative Reference Rates Committee. Progress Report: The Transition from U.S. Dollar LIBOR. <https://www.newyorkfed.org/medialibrary/Microsites/arrc/files/2021/USD-LIBOR-transition-progress-report-mar-21.pdf>, March 2021. Last Accessed: 2021-06-2.
- [62] commoncrawl.org. Common Crawl - Want to use our data? <https://commoncrawl.org/the-data/>, n.d. Last Accessed: 2023-03-18.
- [63] Companies House - UK Government. Companies House accounts guidance. <https://www.gov.uk/government/publications/life-of-a-company-annual-requirements/life-of-a-company-part-1-accounts>, 2023. Last Accessed: 2023-08-26.
- [64] McKinsey & Company. Author Talks: Dr. Fei-Fei Li sees ‘worlds’ of possibilities in a multidisciplinary approach to AI. 2023.
- [65] Corporate Finance Institute. Liquidity ratio, 2020.
- [66] Julián D Cortés. What is the mission of innovation?—lexical structure, sentiment analysis, and cosine similarity of mission statements of research-knowledge intensive institutions. *Plos one*, 17(8):e0267454, 2022.
- [67] Comprehensive R Archive Network (CRAN). DictionaryHE. <https://rdr.io/cran/SentimentAnalysis/man/DictionaryHE.html>, 2019. Last Accessed: 2019-05-17.
- [68] Dan Jurafsky. Lecture #3, vector semantics [powerpoint slides], n.d.
- [69] Edwin Dauber, Rebekah Overdorf, and Rachel Greenstadt. Stylometric authorship attribution of collaborative documents. In *Cyber Security Cryptography and Machine Learning: First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017, Proceedings 1*, pages 115–135. Springer, 2017.
- [70] Angela K Davis and Isho Tama-Sweet. Managers’ use of language across alternative disclosure outlets: Earnings press releases versus md&a. *Contemporary Accounting Research*, 29(3):804–837, 2012.
- [71] Deloitte. Conceptual Framework for Financial Reporting 2018. <https://www.iasplus.com/en/standards/other/framework>, 2018. Last Accessed: 2023-03-31.
- [72] Deloitte. IAS 1 — Presentation of Financial Statements. <https://www.iasplus.com/en/standards/ias/ias1>, n.d. Last Accessed: 2023-16-1.
- [73] Deloitte. IFRS 7 — Financial Instruments: Disclosures. <https://www.iasplus.com/en/standards/ifrs/ifrs7>, n.d. Last Accessed: 2023-16-1.
- [74] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

- [75] designdot. EBITDACR. <https://www.redbubble.com/i/mug/EBITDACR-Earnings-Before-Interest-Taxes-Depreciation-Amortization-Coronavirus-Riots-by-designdot/49408436.9Q0AD>, 2021. Last Accessed: 2023-03-18.
- [76] Caio Deutsch and Ivandr e Paraboni. Authorship attribution using author profiling classifiers. *Natural Language Engineering*, 29(1):110–137, 2023.
- [77] Cambridge Dictionary. boilerplate. *Cambridge Dictionary*. Accessed Feb 24, 2024 [Online].
- [78] Cambridge Dictionary. User. *Cambridge Dictionary*. Accessed Aug 26, 2023 [Online].
- [79] Yuxin Ding and Shengli Yan. Topic optimization method based on pointwise mutual information. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings Part III 22*, pages 148–155. Springer, 2015.
- [80] Dogu Araci. Finbert. <https://huggingface.co/ProsusAI/finbert>, n.d. Last Accessed: 2023-5-23.
- [81] Dun & Bradstreet. What You Need to Know About Business Bankruptcies in the US. <https://www.dnb.com/resources/business-bankruptcies.html#definitions>, n.d. Last Accessed: 2023-20-12.
- [82] EasyGPT. EasyGPT. <https://easygpt.io>, n.d. Last Accessed: 2023-5-23.
- [83] EduPristine. Problems of Small Data and How to Handle Them. <https://www.edupristine.com/blog/managing-small-data>, 2016. Last Accessed: 2021-06-2.
- [84] Ekin Tiu. Understanding latent space in machine learning. <https://towardsdatascience.com/understanding-latent-space-in-machine-learning-de5a7c687d8d>, 2020. Last Accessed: 2022-11-5.
- [85] Magdi El-Bannany, Meenu Sreedharan, and Ahmed M Khedr. A robust deep learning model for financial distress prediction. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(2):170–175, 2020.
- [86] Mahmoud El-Haj. Analysing uk annual report narratives using text analysis and natural language processing. In *Glasgow IR Group Seminars*, 2015.
- [87] Eleonora Fontana. Hyperparameters tuning — topic coherence and lsi model, 2020.
- [88] Douglas J Elliot. Bank liquidity requirements: An introduction and overview. 2014.
- [89] Emil Rijcken. c_v topic coherence explained, 2023.

- [90] ENGGtalks. Tech+ 10K Token Credit. <https://www.enggtalks.com/product/tech-10k-token-credit/?v=3e8d115eb4b3#:~:text=As%20a%20general%20guideline%2C%20one,0.75%20words%20for%20English%20text.&text=It's%20important%20to%20note%20that,tokens%20or%20approximately%203000%20words.>, 2023. Last Accessed: 2023-05-12.
- [91] Samuel J Enumah and David C Chang. Predictors of financial distress among private us hospitals. *Journal of Surgical Research*, 267:251–259, 2021.
- [92] Deborah Estrin. Small data, where n= me. *Communications of the ACM*, 57(4):32–34, 2014.
- [93] Lawrence J. Gitman et al. 132 U.S. financial institutions. <https://opentextbc.ca/businessopenstax/chapter/u-s-financial-institutions/>, 2018. Last Accessed: 2021-11-10.
- [94] European Central Bank. CompNet firm-level based dataset: User guide for researchers 4th round of data collection . https://www.ecb.europa.eu/home/pdf/research/compnet/CompNet-database-user_guide-round4.pdf, 2016. Last Accessed: 2024-01-05.
- [95] European Commission. Financial Reporting. [https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/financial-reporting_en#:~:text=Regulation%20\(EC\)%20No%201606%2F,international%20accounting%20standards%20\(IAS\).](https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/financial-reporting_en#:~:text=Regulation%20(EC)%20No%201606%2F,international%20accounting%20standards%20(IAS).), 2013. Last Accessed: 2024-02-20.
- [96] Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16, 2017.
- [97] Scott E Fahlman, Geoffrey E Hinton, and Terrence J Sejnowski. Massively parallel architectures for al: Netl, thistle, and boltzmann machines. In *National Conference on Artificial Intelligence, AAAI*, 1983.
- [98] Julian J Faraway and Nicole H Augustin. When small data beats big data. *Statistics & Probability Letters*, 136:142–145, 2018.
- [99] FasterCapital. Quartile Normalization: Adjusting Data for Statistical Analysis. <https://fastercapital.com/content/Quartile-Normalization-Adjusting-Data-for-Statistical-Analysis.html#Potential-Caveats-and-Limitations-of-Quartile-Normalization>, 2023. Last Accessed: 2024-01-12.
- [100] Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix M Martinez. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6701–6710, 2021.

- [101] Alvaro Figueira and Bruno Vaz. Survey on synthetic data generation, evaluation methods and gans. *Mathematics*, 10(15):2733, 2022.
- [102] Finance - European Commission. Financial Reporting, n.d. Last Accessed: 2023-08-26.
- [103] Joao Fonseca and Fernando Bacao. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1):115, 2023.
- [104] IFRS Foundation. Ias 37 provisions, contingent liabilities and contingent assets. <https://www.ifrs.org/issued-standards/list-of-standards/ias-37-provisions-contingent-liabilities-and-contingent-assets/#~:text=Contingent%20assets%20are%20not%20recognised,longer%20considered%20to%20be%20contingent.,2023>.
- [105] Richard M. Frankel, Sarah McVay, and Mark T. Soliman. Street earnings and board independence. 2004.
- [106] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.
- [107] Nicolas Garneau, Eve Gaumont, Luc Lamontagne, and Pierre-Luc Déziel. Evaluating legal accuracy of neural generators on the generation of criminal court dockets description. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 73–99, 2022.
- [108] Gauri Mathur. Data science vs. machine learning. <https://www.ibm.com/blog/data-science-vs-machine-learning-whats-the-difference/>, 2023. Last Accessed: 2024-01-8.
- [109] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.
- [110] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40, 2020.
- [111] John W Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32:100577, 2021.
- [112] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [113] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

- [114] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [115] Ian J Goodfellow. On distinguishability criteria for estimating generative models. *arXiv preprint arXiv:1412.6515*, 2014.
- [116] Google. Real world gans - common problems. <https://developers.google.com/machine-learning/gan/problems>, n.d.
- [117] Myron J Gordon. Towards a theory of financial distress. *the Journal of Finance*, 26(2):347–356, 1971.
- [118] Gary Gorton, Andrew Metrick, Andrei Shleifer, and Daniel K Tarullo. Regulating the shadow banking system [with comments and discussion]. *Brookings papers on economic activity*, pages 261–312, 2010.
- [119] Government of Canada. International Financial Reporting Standards (IFRS). <https://www.canada.ca/en/revenue-agency/services/tax/businesses/topics/international-financial-reporting-standards-ifrs.html>, n.d. Last Accessed: 2023-16-1.
- [120] Prasoon Goyal, Zhiting Hu, Xiaodan Liang, Chenyu Wang, and Eric P Xing. Nonparametric variational auto-encoders for hierarchical representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5094–5102, 2017.
- [121] Greg Buchak and Gregor Matvos and Tomasz Piskorski. Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks. <https://www.fdic.gov/analysis/cfr/bank-research-conference/annual-17th/papers/15-piskorski.pdf>, 2017. Last Accessed: 2023-20-12.
- [122] Rohan Yashraj Gupta, Satya Sai Mudigonda, and Pallav Kumar Baruah. Tgans with machine learning models in automobile insurance fraud detection and comparative study with other data imbalance techniques. *International Journal of Recent Technology and Engineering*, 9(5):236–244, 2012.
- [123] Alfiuddin R Hadiat. *Topic Modeling Evaluations: The Relationship Between Coherency and Accuracy*. PhD thesis, 2022.
- [124] Petr Hajek, Renata Myskova, and Vladimir Olej. Predicting stock return volatility using sentiment analysis of corporate annual reports. In *The Essentials of Machine Learning in Finance and Accounting*, pages 75–96. Routledge, 2021.
- [125] Petr Hájek, Vladimir Olej, and Renata Myskova. Forecasting corporate financial performance using sentiment in annual reports for stakeholders’ decision-making. *Technological and Economic Development of Economy*, 20(4):721–738, 2014.

- [126] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.
- [127] Roderick P Hart, Jay P Childers, and Colene J Lind. *Political tone: How leaders talk and why*. University of Chicago Press, 2013.
- [128] Sarah Harvey. Advice for managing the length of annual reports. <https://www.fm-magazine.com/news/2017/dec/managing-the-length-of-annual-reports-201717989.html>, 2017. Last Accessed: 2023-16-1.
- [129] Anselm Haselhoff, Christian Nunn, Dennis Müller, Mirko Meuter, and Lutz Roese-Koerner. Markov random field for image synthesis with an application to traffic sign recognition. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1407–1412. IEEE, 2017.
- [130] Melissa Heikkilä. Ai is at an inflection point, fei-fei li says. 2023.
- [131] Eric B Hekler, Predrag Klasnja, Guillaume Chevance, Natalie M Golaszewski, Dana Lewis, and Ida Sim. Why we need a small data paradigm. *BMC medicine*, 17(1):1–9, 2019.
- [132] Elaine Henry. Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, 45(4):363–407, 2008.
- [133] Lorena Pérez Hernández. Cognitive tools for successful branding. *Applied linguistics*, 32(4):369–388, 2011.
- [134] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Standardised metrics and methods for synthetic tabular data evaluation. *Authorea Preprints*, 2023.
- [135] hindupuravinash. The gan zoo. <https://github.com/hindupuravinash/the-gan-zoo>, 2018.
- [136] Geoffrey E Hinton, Terrence J Sejnowski, and David H Ackley. *Boltzmann machines: Constraint satisfaction networks that learn*. Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, 1984.
- [137] Geoffrey E Hinton, Terrence J Sejnowski, et al. Learning and relearning in boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317):2, 1986.
- [138] Udi Hoitash, Rani Hoitash, and Jean C Bedard. Corporate governance and internal control over financial reporting: A comparison of regulatory regimes. *The accounting review*, 84(3):839–867, 2009.

- [139] David R Holmes Jr, Patricia K Hodgson, Rick A Nishimura, and Robert D Simari. Careers in cardiovascular medicine. *Circulation*, 120:906–913, 2009.
- [140] Dong-Suk Hong and Cheol Baik. Generating and validating synthetic training data for predicting bankruptcy of individual businesses. *Journal of Information & Communication Convergence Engineering*, 19(4), 2021.
- [141] Charles Horngren, Walter Harrison, Linda Bamber, W. Morley Lemon, and Peter Norwood. *Accounting Vol. 1, 6th Canadian edition*. Pearson Canada, 2004.
- [142] Tadaaki Hosaka. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert systems with applications*, 117:287–299, 2019.
- [143] David Hou and David R Skeie. Libor: Origins, economics, crisis, scandal, and reform. *FRB of New York Staff Report*, 1(667), 2014.
- [144] Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. Synthetic data for social good. *arXiv preprint arXiv:1710.08874*, 2017.
- [145] Anna Huang et al. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.
- [146] Xun Huang, Cheng-Zhao Zhang, and Jia Yuan. Predicting extreme financial risks on imbalanced dataset: A combined kernel fcm and kernel smote based svm classifier. *Computational Economics*, 56:187–216, 2020.
- [147] Yu-Pei Huang and Meng-Feng Yen. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing*, 83:105663, 2019.
- [148] J Hughes and A Fisher. Reading Financial Statements - What Do I Need To Know? <https://www.cpacanada.ca/-/media/site/business-and-accounting-resources/docs/reading-financial-statements----what-do-i-need-to-know.pdf?la=en&hash=3BDE48F69C73D2C4C022935CBC6404261DE764F9>, 2014. Last Accessed: 2023-16-1.
- [149] Oliver Ibe. *Markov processes for stochastic modeling*. Newnes, 2013.
- [150] IBM. Multinomial logistic regression. <https://www.ibm.com/docs/en/spss-statistics/29.0.0?topic=regression-multinomial-logistic>, n.d. Last Accessed: 2024-02-5.
- [151] IBM. What are convolutional neural networks. <https://www.ibm.com/topics/convolutional-neural-networks>, n.d. Last Accessed: 2024-02-5.
- [152] IBM. What is random forest. <https://www.ibm.com/topics/random-forest>, n.d. Last Accessed: 2024-02-5.

- [153] Bekkouch Imad Ibrahim, Dragoş Constantin Nicolae, Adil Khan, Syed Imran Ali, and Asad Khattak. Vae-gan based zero-shot outlier detection. In *Proceedings of the 2020 4th International Symposium on Computer Science and Intelligent Control*, pages 1–5, 2020.
- [154] International Accounting Education Standards Board. <https://www.iaesb.org/iaesb/publications/2019-handbook-international-education-standards>. <https://www.iaesb.org/iaesb/publications/2019-handbook-international-education-standards>, n.d. Last Accessed: 2024-05-12.
- [155] AK Jain and M Tuceryan. The handbook of pattern recognition and computer vision, chapter 11, texture analysis, 1992.
- [156] Chyan-long Jan. Financial information asymmetry: Using deep learning algorithms to predict financial distress. *Symmetry*, 13(3):443, 2021.
- [157] Magdalena Jankowska. Author style analysis in text documents based on character and word n-grams. 2017.
- [158] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data—what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.
- [159] Indu Joshi, Marcel Grimmer, Christian Rathgeb, Christoph Busch, Francois Bremond, and Antitza Dantcheva. Synthetic data in human analysis: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [160] JPMorgan Chase & Co. INVESTOR RELATIONS - Annual Report & Proxy. <https://www.jpmorganchase.com/ir/annual-report>, n.d. Last Accessed: 2023-16-1.
- [161] Ehsan Kamaloo, Mehdi Rezagholizadeh, and Ali Ghodsi. When chosen wisely, more data is what you need: A universal sample-efficient strategy for data augmentation. *arXiv preprint arXiv:2203.09391*, 2022.
- [162] Taeyoung Kang, Do-Hyung Park, and Ingoo Han. Beyond the numbers: The effect of 10-k tone on firms’ performance predictions using text analytics. *Telematics and Informatics*, 35(2):370–381, 2018.
- [163] Gordon V Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127, 1980.
- [164] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.
- [165] Fahad Khan, Xiaojun Yu, Zhaohui Yuan, and Atiq Ur Rehman. Ecg classification using 1-d convolutional deep residual neural network. *Plos one*, 18(4):e0284791, 2023.

- [166] Paul Kiel and Dan Ngyuen. Bailout Recipients. <https://projects.propublica.org/bailout/>, February 2021. Last Accessed: 2021-06-2.
- [167] Donald E Kieso, Jerry J Weygandt, Terry D Warfield, Irene M Wiecek, and Bruce J McConomy. *Intermediate Accounting, Volume 1*. John Wiley & Sons, 2019.
- [168] Kyungki Kim and Jochen Teizer. Automatic design and planning of scaffolding systems using building information modeling. *Advanced Engineering Informatics*, 28(1):66–80, 2014.
- [169] Paul D Kimmel, Jerry J Weygandt, and Donald E Kieso. *Accounting: Tools for business decision making*. John Wiley & Sons, 2018.
- [170] Zainul Kisman and Dian Krisandi. How to predict financial distress in the wholesale sector: Lesson from indonesian stock exchange. *Journal of Economics and Business*, 2(3):569–585, 2019.
- [171] Rob Kitchin and Tracey P Lauriault. Small data in the era of big data. *GeoJournal*, 80:463–475, 2015.
- [172] Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. Fake news: Evidence from financial markets. *Available at SSRN*, 3237763, 2019.
- [173] György Kovács. Minimum number of rows in a class #35, 2020.
- [174] Abhijeet Kumar. Finbert embedding. https://github.com/abhijeet3922/finbert_embedding, 2020.
- [175] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.
- [176] Kevin Kuo. Generative synthesis of insurance datasets. *arXiv preprint arXiv:1912.02423*, 2019.
- [177] L. Woods. 15 Rich Influencers Who Didn’t Need a College Degree. https://ca.news.yahoo.com/15-rich-influencers-didn-t-170000601.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMw2NKcy3W4aC3NvcON4LM70v_j7YLS0iYJRS1wT8PPnH3gnTfpiPjLaYZVi1h0_gWffYHtkQthM1b2coVoZuHNmlT4B5CGZCFKJ5L1hNW7kJARowDWvc3DRvRjC5SKDC0t\--U7uvalYyXjuhcWuMdJToZ_A3-zemmIt58C7QT1e, 2020. Last Accessed: 2023-03-01.
- [178] François Dominic Laramée. Introduction to stylometry with python. *The Programming Historian*, 7, 2018.
- [179] James R Layton. A chart for computing the dale-chall readability formula above fourth grade level. *Journal of Reading*, 24(3):239–244, 1980.

- [180] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.
- [181] Jurek Leonhardt, Avishek Anand, and Megha Khosla. Boilerplate removal using a neural sequence labeling model. In *Companion Proceedings of the Web Conference 2020*, pages 226–229, 2020.
- [182] Ang Li, Mark Liu, and Simon Sheather. Predicting stock splits using ensemble machine learning and smote oversampling. *Pacific-Basin Finance Journal*, 78:101948, 2023.
- [183] Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.
- [184] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2/3:18–22, 2002.
- [185] Robert Libby, Daniel Short, and Patricia Libby. *EBOOK: Financial Accounting*. McGraw Hill, 2014.
- [186] M. Lieberman. What Is ChatGPT and How Is It Used in Education? <https://www.edweek.org/technology/what-is-chatgpt-and-how-is-it-used-in-education/2023/01>, 2023, n.d. Last Accessed: 2023-01-26.
- [187] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [188] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.
- [189] Mikael Ljung. Synthetic data generation for the financial industry using generative adversarial networks, 2021.
- [190] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.
- [191] B. Lougee and C. Marquardt. Earnings informativeness and strategic disclosure: An empirical examination of “pro forma” earnings. *The Accounting Review*, 79:769–795, 2004.
- [192] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.
- [193] Tim Loughran and Bill McDonald. Measuring readability in financial disclosures. *the Journal of Finance*, 69(4):1643–1671, 2014.
- [194] Johannes A Louw and Avashlin Moodley. Speaker specific phrase break modeling with conditional random fields for text-to-speech. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6. IEEE, 2016.

- [195] Bloomberg L.P. Annual financial statement analysis, 1997 - 2021 in US. <https://www.bloomberg.com/company/>, N.D. Last Accessed: 2022-8-15.
- [196] Alexandra Sasha Luccioni and Joseph D Viviano. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*, 2021.
- [197] Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. Deep learning models for bankruptcy prediction using textual disclosures. *European journal of operational research*, 274(2):743–758, 2019.
- [198] Sarvesh Makthal and Arun Ross. Synthesis of iris images using markov random fields. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE, 2005.
- [199] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.
- [200] Manning, Christopher D. and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
- [201] Miro Mannino and Azza Abouzied. Is this real? generating synthetic data that looks real. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 549–561, 2019.
- [202] Martin Haugh. The em algorithm. http://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf, 2015. Last Accessed: 2022-11-5.
- [203] John R Mashey. Big data and the next wave of {InfraStress} problems, solutions, opportunities. In *1999 USENIX annual technical conference (USENIX ATC 99)*, 1999.
- [204] MathWorkds. What is a convolutional neural network. [https://www.mathworks.com/discovery/convolutional-neural-network.html#:~:text=A%20convolutional%20neural%20network%20\(CNN,objects%2C%20classes%2C%20and%20categories.,n.d.](https://www.mathworks.com/discovery/convolutional-neural-network.html#:~:text=A%20convolutional%20neural%20network%20(CNN,objects%2C%20classes%2C%20and%20categories.,n.d.) Last Accessed: 2024-02-5.
- [205] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.
- [206] Matthew Mayo. Centroid Initialization Methods for k-means Clustering. [https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html#\protect\protect\leavevmode@ifvmode\kern+.2222em\relax~\protect\protect\leavevmode@ifvmode\kern+.2222em\relaxtext\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox@tempboxa\hbox{k\global\mathchardef\accent@spacefactor\spacefactor}\let\beginngroup\let\typeout\protect\beginngroup\def\MessageBreak{\Omega\(Font\)}\let\protect\immediate\write\m@ne{LaTeXFontInfo:oninputline1074.}\endgroup\endgroup\relax\let\ignorespaces\relax\accent22k\egroup\spacefactor\accent@spacefactor%](https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html#\protect\protect\leavevmode@ifvmode\kern+.2222em\relax~\protect\protect\leavevmode@ifvmode\kern+.2222em\relaxtext\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox@tempboxa\hbox{k\global\mathchardef\accent@spacefactor\spacefactor}\let\beginngroup\let\typeout\protect\beginngroup\def\MessageBreak{\Omega(Font)}\let\protect\immediate\write\m@ne{LaTeXFontInfo:oninputline1074.}\endgroup\endgroup\relax\let\ignorespaces\relax\accent22k\egroup\spacefactor\accent@spacefactor%)

- 2Dmeans%2B%2B%3A%20As%20spreading\protect\protect\leavevmode@ifvmode\kern+.1667em\relaxprobability\%20proportional%20to%20the%20squared, 2022. Last Accessed: 2024-01-20.
- [207] Ryan McKenna and Terrance Liu. <https://differentialprivacy.org/synth-data-1/>.
- [208] Cuizhu Meng, Li Zhou, and Bisong Liu. A case study in credit fraud detection with smote and xgboost. In *Journal of Physics: Conference Series*, volume 1601, page 052016. IOP Publishing, 2020.
- [209] Merriam-Webster. Boilerplate. *Merriam-Webster*. Accessed Nov 10, 2021 [Online].
- [210] Merriam-Webster. Intermediation. *Merriam-Webster*. Accessed Nov 10, 2021 [Online].
- [211] Merriam-Webster. Understand. *Merriam-Webster*. Accessed Feb 10, 2024 [Online].
- [212] Michael Holtzsch. spacy readability. https://github.com/mholtzsch/spacy_readability, 2019. Last Accessed: 2023-16-1.
- [213] Lenka Mikáčová and Petra Gavlaková. The role of public relations in branding. *Procedia-Social and Behavioral Sciences*, 110:832–840, 2014.
- [214] Wayne H Mikkelson and M Megan Partch. Do persistent large cash reserves hinder performance? *Journal of financial and quantitative analysis*, 38(2):275–294, 2003.
- [215] Benjamin R Mitchell et al. *The spatial inductive bias of deep learning*. PhD thesis, Johns Hopkins University, 2017.
- [216] Kevin Moffitt and Mary B Burns. What does that mean? investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. *AMCIS 2009 Proceedings*, page 399, 2009.
- [217] Maurice Moonitz. Basic postulates of accounting; accounting research study no. 01, 1961.
- [218] Moraine Park Technical College. What Flesch Reading Ease Score Should My Content Have? <https://www.morainepark.edu/help/>, n.d. Last Accessed: 2023-16-1.
- [219] Vânia Moreira Costa, Nuno A De Sá Teixeira, Ana Cordeiro Santos, and Eduardo Santos. When more is less in financial decision-making: financial literacy magnifies framing effects. *Psychological Research*, 85(5):2036–2046, 2021.
- [220] Morgan Stanley Wealth Management. Morgan Stanley Smith Barney is Now Morgan Stanley Wealth Management. https://www.morganstanley.com/press-releases/morgan-stanley-smith-barney-is-now-morgan-stanley-wealth-management_7a78aa1d-036a-4fbf-9df7-1e73387a1c8a, 2012. Last Accessed: 2023-5-14.
- [221] C. Murphy. Over-the-Counter (OTC): Trading and Security Types Defined. <https://www.investopedia.com/terms/o/otc.asp>, 2022. Last Accessed: 2023-10-3.

- [222] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36, 2017.
- [223] David Newman, Edwin V Bonilla, and Wray Buntine. Improving topic coherence with regularized topic models. *Advances in neural information processing systems*, 24, 2011.
- [224] OCC. Office of the comptroller of the currency. <https://www.occ.treas.gov/about/index-about.html>, n.d. Last Accessed: 2022-11-9.
- [225] Royal Bank of Canada. Royal bank of canada 2023 annual report. https://www.rbc.com/investor-relations/_assets-custom/pdf/ar_2023_e.pdf, 2023.
- [226] University of Michigan Statistics Online Computational Resource. Ap statistics curriculum 2007 laplace, 2020.
- [227] Office of the Comptroller of the Currency. Concentrations of Credit, Version 2.0. <https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/concentrations-of-credit/pub-ch-concentrations.pdf>, October, 2020. Last Accessed: 2021-08-17.
- [228] Erik Ofgang. What is ChatGPT and how can you teach with it? Tips & tricks. <https://www.edweek.org/technology/what-is-chatgpt-and-how-is-it-used-in-education/2023/01>, 2023, 2022. Last Accessed: 2023-01-26.
- [229] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization, 2017.
- [230] OED Online. adequate, adj., 2011.
- [231] Open for Vintage. A Beginner’s Guide To Investing In Vintage Louis Vuitton Handbags. <https://www.openforvintage.com/en-ca/blogs/news/a-beginners-guide-to-investing-in-vintage-louis-vuitton-handbags>, 2022. Last Accessed: 2023-03-01.
- [232] OpenAI. Introducing ChatGPT. <https://openai.com/blog/chatgpt>, 2022. Last Accessed: 2024-02-20.
- [233] OpenAI. What Is the Difference between the GPT-4 Models? <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>, 2023. Last Accessed: 2023-03-01.
- [234] OpenAI. What are tokens and how to count them? <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>, 2024. Last Accessed: 2024-02-12.
- [235] OpenAI. Fine tuning. <https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset>, n.d.

- [236] OpenAI. GPT-4. <https://openai.com/research/gpt-4>, n.d. Last Accessed: 2023-03-18.
- [237] OpenAI. Introducing ChatGPT. <https://openai.com/blog/chatgpt>, n.d. Last Accessed: 2023-03-18.
- [238] OpenAI. Playground. <https://platform.openai.com/playground>, n.d. Last Accessed: 2023-5-23.
- [239] Oyeniyi, T. and Tortoriello, R. and Ma, Li and Fruin, Paul. Late to File: The Costs of Delayed 10-Q and 10-K Company Filings. <https://www.spglobal.com/marketintelligence/en/documents/sp-capital-\iq-quantamental-late-to-file-the-costs-of-delayed-10-q-and-10-k-company-filings.pdf>, 2015. Last Accessed: 2023-10-3.
- [240] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *Applied Soft Computing*, 93:106384, 2020.
- [241] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [242] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.
- [243] Lee D Parker. Social and environmental accountability research: A view from the commentary box. *Accounting, Auditing & Accountability Journal*, 18(6):842–860, 2005.
- [244] Frank Partnoy. Financial derivatives and the costs of regulatory arbitrage. *J. corP. l.*, 22:211, 1996.
- [245] Tushar Patil. *Credit Card Fraud Detection Using Conditional Tabular Generative Adversarial Networks (CT-GAN) and Supervised Machine Learning Techniques*. PhD thesis, Dublin, National College of Ireland, 2021.
- [246] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE, 2016.
- [247] Paul Hand. Variational autoencoders. http://khoury.northeastern.edu/home/hand/teaching/cs7150-summer-2020/Variational_Autoencoders.pdf, 2020. Last Accessed: 2022-11-5.
- [248] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Preprocessing data - scikit-learn. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

- [249] José-Manuel Peña, Fernando Suárez, Omar Larré, Domingo Ramírez, and Arturo Cifuentes. A modified ctgan-plus-features based method for optimal asset allocation. *arXiv preprint arXiv:2302.02269*, 2023.
- [250] J. Zvi Namenwirth Philip J. Stone, Robert F. Bales. Welcome to the General Inquirer Home Page. <http://www.wjh.harvard.edu/~inquirer/Home.html>. Last Accessed: 2019-06-21.
- [251] Steven T Piantadosi. Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.
- [252] Christine S Pitt, Kirk A Plangger, Elsamari Botha, Jan Kietzmann, and Leyland Pitt. How employees engage with b2b brands on social media: Word choice and verbal tone. *Industrial Marketing Management*, 81:130–137, 2019.
- [253] Harlan D Platt and Marjorie B Platt. Predicting corporate financial distress: reflections on choice-based sample bias. *Journal of economics and finance*, 26(2):184–199, 2002.
- [254] E. Prokosch. *Language*, 9(1):89–92, 1933.
- [255] Guo-Jun Qi and Jiebo Luo. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2168–2187, 2020.
- [256] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023.
- [257] Abdul Wahab Qurashi, Violeta Holmes, and Anju P Johnson. Document processing: Methods for semantic text similarity analysis. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE, 2020.
- [258] Radim Rehurek. Topic coherence pipeline. <https://radimrehurek.com/gensim/models/coherencemodel.html>, n.d. Last Accessed: 2023-24-9.
- [259] Amirarsalan Rajabi and Ozlem Ozmen Garibay. Tabfairgan: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction*, 4(2):488–501, 2022.
- [260] Domingo Ramirez, Jose-Manuel Peña, Fernando Suárez, Omar Larré, and Arturo Cifuentes. A machine learning plus-features based approach for optimal asset allocation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 549–556, 2023.
- [261] Sebastian Raschka. A short chronology of deep learning for tabular data, 2022.
- [262] Suman Ravuri and Oriol Vinyals. Seeing is not necessarily believing: Limitations of biggans for data augmentation. 2019.

- [263] Readable. What is a Gunning Fog Index readability score? <https://readable.com/readability/gunning-fog-index/#:~:text=The%20Gunning%20Fog%20formula%20generates%20a%20grade%20level%20between%200%20and%2020.>, n.d. Last Accessed: 2021-08-17.
- [264] St. Louis Federal Reserve. St. Louis Federal Reserve. <https://www.stlouisfed.org/>, N.D. Last Accessed: 2021-08-17.
- [265] Annelise Riles. Managing regulatory arbitrage: a conflict of laws approach. *Cornell Int'l LJ*, 47:63, 2014.
- [266] Tyler Rinker. Package ‘qdapDictionaries’. <https://cran.r-project.org/web/packages/qdapDictionaries/qdapDictionaries.pdf>, 2018. Last Accessed: 2019-06-21.
- [267] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. Citeseer, 2001.
- [268] Joseph Rocca. Understanding Variational Autoencoders (VAEs). <https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73>, 2019. Last Accessed: 2024-01-20.
- [269] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.
- [270] Jonathan L Rogers, Andrew Van Buskirk, and Sarah LC Zechman. Disclosure tone and shareholder litigation. *The Accounting Review*, 86(6):2155–2183, 2011.
- [271] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*, 2014.
- [272] Stephen A Ross, Randolph Westerfield, and Bradford D Jordan. *Fundamentals of corporate finance*. Irwin New York, NY, USA, 2014.
- [273] Paolo Rosso, Francisco Rangel, Irazu Hernández Farías, Leticia Cagnina, Wajdi Zaghouni, and Anis Charfi. A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass*, 12(4):e12275, 2018. e12275 LNCO-0720.R1.
- [274] Royal Bank of Canada. Investor Relations. <https://www.rbc.com/investor-relations/>, n.d. Last Accessed: 2023-11-2.
- [275] Jan Rybicki and Magda Heydel. The stylistics and stylometry of collaborative translation: Woolf’s night and day in polish. *Literary and Linguistic Computing*, 28(4):708–717, 2013.

- [276] Magnus Sahlgren. The distributional hypothesis. *Italian Journal of linguistics*, 20:33–53, 2008.
- [277] Yunita Sari, Mark Stevenson, and Andreas Vlachos. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th international conference on computational linguistics*, pages 343–353, 2018.
- [278] Timur Sattarov, Marco Schreyer, and Damian Borth. Findiff: Diffusion models for financial tabular data generation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 64–72, 2023.
- [279] Roland Schäfer. Accurate and efficient general-purpose boilerplate detection for crawled web corpora. *Language Resources and Evaluation*, 51:873–889, 2017.
- [280] Stephen W Scheff. *Fundamental statistical principles for the neurobiologist: A survival guide*. Academic Press, 2016.
- [281] Marco Schreyer, Timur Sattarov, Anita Gierbl, Bernd Reimer, and Damian Borth. Learning sampling in financial statement audits using vector quantised variational autoencoder neural networks. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8, 2020.
- [282] scikit learn. Kmeans. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>, n.d.
- [283] Tyler A Scott, Nicholas Marantz, and Nicola Ulibarri. Use of boilerplate language in regulatory documents: evidence from environmental impact statements. *Journal of Public Administration Research and Theory*, 32(3):576–590, 2022.
- [284] SEC. Regulation of insurance products. <https://www.sec.gov/oig/reportspubs/aboutoigaudit242finhtm.html>, 1996. Last Accessed: 2021-11-10.
- [285] Securities and Exchange Commission. EDGAR. <https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm>, N.D. Last Accessed: 2021-08-17.
- [286] Securities and Exchange Commission. A Plain English Handbook - How to create clear SEC disclosure documents . <https://www.sec.gov/pdf/handbook.pdf>, n.d. Last Accessed: 2023-5-23.
- [287] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.
- [288] Shashank Kapadia. Evaluate topic models: Latent dirichlet allocation (lda), 2019.
- [289] Jessica Shieh. Best practices for prompt engineering with OpenAI API. <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api>, 2023. Last Accessed: 2023-5-23.

- [290] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.
- [291] Ksh Nareshkumar Singh, S Dickeeta Devi, H Mamata Devi, and Anjana Kakoti Mahanta. A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *International Journal of Information Management Data Insights*, 2(1):100061, 2022.
- [292] Rishi R Singh, Deepika Koundal, and Rajeev Tiwari. Linguistic approach for authentic authorship. 2021.
- [293] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.
- [294] Mario Situm. Recovery from distress and insolvency: A comparative analysis using accounting ratios. In *Proceedings of the 6th Global Conference on Managing in Recovering Markets, GCMRM*, pages 589–606, 2015.
- [295] Salima Smiti and Makram Soui. Bankruptcy prediction using deep learning approach based on borderline smote. *Information Systems Frontiers*, 22:1067–1083, 2020.
- [296] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.
- [297] Ahyun Song, Euseong Seo, and Heeyoul Kim. Anomaly vae-transformer: A deep learning approach for anomaly detection in decentralized finance. *IEEE Access*, 2023.
- [298] Cesar Souza. Kernel Functions for Machine Learning Applications. <http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/>. Last Accessed: 2019-06-21.
- [299] Ravisutha Sakrepatna Srinivasamurthy. *Understanding 1D Convolutional Neural Networks Using Multiclass Time-Varying Signalss*. PhD thesis, Clemson University, 2018.
- [300] Statista. Definition quantile. <https://www.statista.com/statistics-glossary/definition/356/quantile/>, n.d.
- [301] Bear Stearns. 10-K Form, 2004.
- [302] Reva B Steinberg and Judith Fellner Weiss. New rules on disclosure of certain significant risks and unc. *The CPA Journal*, 65(3):16, 1995.
- [303] Gerard Stone and Lee D Parker. Developing the flesch reading ease formula for the contemporary accounting communications landscape. *Qualitative Research in Accounting & Management*, 10(1):31–59, 2013.

- [304] Jie Sun, Hui Li, Hamido Fujita, Binbin Fu, and Wenguo Ai. Class-imbalanced dynamic financial distress prediction based on adaboost-svm ensemble combined with smote and time weighting. *Information Fusion*, 54:128–144, 2020.
- [305] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [306] Marco Taboga. Trace of a matrix. <https://www.statlect.com/matrix-algebra/trace-of-a-matrix>, n.d.
- [307] Travis Tang. Smote: Oversampling for class imbalance, 2023.
- [308] Stacey Taylor and Vlado Keselj. Using extractive lexicon-based sentiment analysis to enhance understanding of the impact of non-gaap measures in financial reporting. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 40–46, 2020.
- [309] Stacey Taylor and Vlado Keselj. Predicting the distress of financial intermediaries using convolutional neural networks. In *2021 IEEE 23rd Conference on Business Informatics (CBI)*, volume 2, pages 71–77. IEEE, 2021.
- [310] Stacey Taylor and Vlado Keselj. Don't blindly use data: Towards a data statement for computational financial research. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5100–5109. IEEE, 2022.
- [311] Stacey Taylor and Vlado Keselj. Don't blindly use data: Towards a data statement for computational financial research. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5100–5109. IEEE, 2022.
- [312] Stacey Taylor and Vlado Keselj. Don't worry accountants, chatgpt won't be taking your job...yet. In *Proceedings of the Canadian Conference on Artificial Intelligence - Forthcoming - June 2023*, 2023.
- [313] Igor V Tetko, Pavel Karpov, Eric Bruno, Talia B Kimber, and Guillaume Godin. Augmentation is what you need! In *International Conference on Artificial Neural Networks*, pages 831–835. Springer, 2019.
- [314] The Beauty Junkee. Are luxury bags an investment? <https://thebeautyjunkee.blogspot.com/2022/06/are-luxury-bags-investment.html>, 2022. Last Accessed: 2023-03-01.
- [315] Thompson Reuters Practical Law. Annual Report. [https://ca.practicallaw.thomsonreuters.com/8-570-0163?transitionType=Default&contextData=\(sc.Default\)&firstPage=true#:~:text=There%20is%20no%20specific%20requirement,The%20required%20annual%20financial%20information.,](https://ca.practicallaw.thomsonreuters.com/8-570-0163?transitionType=Default&contextData=(sc.Default)&firstPage=true#:~:text=There%20is%20no%20specific%20requirement,The%20required%20annual%20financial%20information.,) n.d. Last Accessed: 2024-02-20.

- [316] H Holden Thorp. Chatgpt is fun, but not an author, 2023.
- [317] Tim Vipond. What is sensitivity analysis. <https://corporatefinanceinstitute.com/resources/financial-modeling/what-is-sensitivity-analysis/>, n.d. Last Accessed: 2024-05-12.
- [318] S. Tingris and B. Kinsella. Exploring gpt-3 (1st ed). <https://www.perlego.com/book/2875049/exploring-gpt3-pdf> , 2021.
- [319] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15, 2023.
- [320] MIT Data to AI Lab. ctgan 0.6.0. <https://pypi.org/project/ctgan/>, n.d.
- [321] Giovanni Tripepi, KJ Jager, FW Dekker, and Carmine Zoccali. Linear and logistic regression analysis. *Kidney international*, 73(7):806–810, 2008.
- [322] Miloš Tumpach, Adriana Surovičová, Zuzana Juhaszova, Anton Marci, and Zuzana Kubaščíková. Prediction of the bankruptcy of slovak companies using neural networks with smote. *Ekonomický časopis*, 68(10):1021–1039, 2020.
- [323] Muhammad Fahim Uddin, Navarun Gupta, et al. Seven v’s of big data understanding big data to extract value. In *Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education*, pages 1–5. IEEE, 2014.
- [324] U.S. Department of the Treasury. About TARP. <https://home.treasury.gov/data/troubled-assets-relief-program/bank-investment-programs>, n.d. Last Accessed: 2024-02-26.
- [325] U.S. Government. Securities Exchange Act of 1934. <https://www.govinfo.gov/content/pkg/COMPS-1885/pdf/COMPS-1885.pdf>, 1934. Last Accessed: 2023-11-2.
- [326] U.S. Securities and Exchange Commission. Standards Relating to Listed Company Audit Committees. <https://www.sec.gov/rules/2003/04/standards-relating-listed-company-audit-committees>, 2003. Last Accessed: 2024-05-12.
- [327] U.S. Securities and Exchange Commission. Form 10-K. <https://www.sec.gov/files/reada10k.pdf>, n.d. Last Accessed: 2023-08-26.
- [328] U.S. Securities and Exchange Commission. Form 10-K. <https://www.sec.gov/files/form10-k.pdf>, n.d. Last Accessed: 2023-08-26.
- [329] David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

- [330] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- [331] Narongchai Visitpanya and Taweesak Samanchuen. Synthesis of tax return datasets for development of tax evasion detection. *IEEE Access*, 2023.
- [332] Ben Walker. 16 Millionaires Who Made Their Fortunes Without a College Degree. <https://financebuzz.com/millionaires-without-a-college-degree>, 2022. Last Accessed: 2023-16-1.
- [333] Jiapeng Wang and Yihong Dong. Measurement of text similarity: a survey. *Information*, 11(9):421, 2020.
- [334] Ian Ward. Json lines. <https://jsonlines.org/>, n.d.
- [335] Daniela Witten and Gareth James. *An introduction to statistical learning with applications in R*. springer publication, 2013.
- [336] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, fourth edition, 2017.
- [337] Jason Wittenbach, Brian d’Alessandro, and C Bayan Bruss. Machine learning for temporal data in finance: Challenges and opportunities. *arXiv preprint arXiv:2009.05636*, 2020.
- [338] Karen Hopper Wruck. Financial distress, reorganization, and organizational efficiency. *Journal of financial economics*, 27(2):419–444, 1990.
- [339] Chong Wu, Lu Wang, and Zhe Shi. Financial distress prediction based on support vector machine with a modified kernel function. *Journal of intelligent systems*, 25(3):417–429, 2016.
- [340] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.
- [341] Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*, 2018.
- [342] Mengfei Xu, Jiejie Chen, Ping Jiang, Baiyan Hu, and Honggang Yang. Stock market price prediction based on tgan. In *2023 9th International Conference on Control Science and Systems Engineering (ICCSSE)*, pages 246–251. IEEE, 2023.
- [343] Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9(1):42, 2023.

- [344] Y. Markovski. Controlling the length of completions. <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>, 2023. Last Accessed: 2023-03-10.
- [345] Yinfei Yang and Chris Tar. Advances in semantic textual similarity. <https://blog.research.google/2018/05/advances-in-semantic-textual-similarity.html>, 2018.
- [346] Yasas Sandeepa. Regression for classification - hands on experience. <https://towardsdatascience.com/regression-for-classification-hands-on-experience-8754a909a298>, 2021. Last Accessed: 2024-02-5.
- [347] Huanzhuo Ye, Lin Xiang, and Yanping Gan. Detecting financial statement fraud using random forest with smote. In *IOP Conference Series: Materials Science and Engineering*, volume 612, page 052051. IOP Publishing, 2019.
- [348] Liafisu Sina Yekini, Tomasz Piotr Wisniewski, and Yuval Millo. Market reaction to the positiveness of annual report narratives. *The British Accounting Review*, 48(4):415–430, 2016.
- [349] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.
- [350] Zhihong Zhao and Tongyuan Bai. Financial fraud detection and prediction in listed companies using smote and machine learning algorithms. *Entropy*, 24(8):1157, 2022.
- [351] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.
- [352] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. ” O’Reilly Media, Inc.”, 2018.
- [353] Yuhao Zhou and Cristián Bravo. Expert insight: What canada can learn from the collapse of silicon valley bank, 2023.
- [354] George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 1936.
- [355] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 1949.

Appendix A

Financial Statement Overview

Financial statements also have specific recording uses, which are also important to consider:

1. The Income Statement (**IS**) is a period specific (ex. monthly, quarterly, or yearly) record of profitability. It details revenues and expenses so that users, both internally and externally, can see and understand the context of the changes. The naming convention of the income statement therefore reflects the period; the **IS** for the end of the year would be titled “Income Statement for the year ending December 31st, 2022” (if December 31st was the company’s year end). Once the year is over, all of the balances are transferred to retained earnings, and the next year’s income statement will start “fresh”.
2. The Balance Sheet (**BS**) is a cumulative statement that reflects all of the changes to a company’s financial position since inception. In that sense, the **BS** “lives forever”. The naming convention of the **BS** also reflects the ongoing nature of the **BS** and is stated to be an “as at” date, rather than for the period ending. The **BS** for the end of the year would be titled “Balance Sheet as at December 31st, 2022” (if the balance sheet was being prepared as of that date).
3. The Statement of Cash Flows (**SCF**) details the inflows and outflows of cash for the company for a specific period. The naming convention therefore reflects that and uses the same type of naming as the **IS**. This statement is very important to understand how the company has used its cash and how much cash it has. A common misunderstanding is that *Net Income* represents the amount of cash that the company has earned; however cash and income are two different things. Revenue can be recognized as you have done the work and sent out an invoice, but until that invoice get paid, all you have is *accounts receivable* but no cash. Therefore, your **IS** will show revenue, but your statement of cash flow (and your balance sheet) will be reflective of the fact that the company is still

waiting to get paid. The **SCF** is both a “continually living” statement like the **BS**, but is also meant to be period specific. When the period is over, the changes in the **SCF** are shown in the **SCF** which is reconciled to “cash at the beginning of the period”. The change in cash is added to “cash at the beginning of the period” to calculate “cash at the end of the period”. This end of period amount then becomes the beginning of the period amount at the start of the next period. This reconciliation can be seen in Figure 2.1.

4. Like the **BS** and the **SCF**, the Statement of Retained Earnings reflects all of the earnings (or losses) that a company has accumulated over time since inception. It also reconciles the amount of retained earnings that will be reflected on the balance sheet, as seen in Figure 2.1. Like the **SCF**, it is meant to show the changes in earnings for a specific period, and therefore has a similar naming convention and approach (i.e. beginning retained earnings and ending retained earnings). Some sources may indicated that Net Income comes over to the **BS** as retained earnings. As long as no dividends have been declared, then that is true [169].

Appendix B

Big and Small Data

A common approach to contextualizing “big data” in computer science is to use the “5 Vs”: volume, velocity, variety, value, and veracity [12]. This model has also been extended to 7 Vs to include validity and volatility [323]. Yet another definition also includes exhaustivity, meaning that all available records have been collected where $n = \text{all}$; fine-grained; relational; and flexible [171]. The caveat with the last definition is that there may be some instances where “big data” does not have all of these characteristics due to applicability. An example is provided for sensor data — the requirement of variety will not be met as this type of data is limited, but will, however, meet the other “big data” requirements [171]. Conversely, small data is often characterized by the fact that it has limited amounts of the Vs and additional characteristics of big data, both discussed above. Small data also tends to be characterized by the question(s) that are being examined or trying to be answered [171] — something that is not usually used to characterize big data.

However, these terms are “relative rather than absolute” [343] and that context is needed in order to understand the usage of these terms. As well, if exhaustivity means “all” (or perhaps substantially all, given the context), that also complements Estrin’s formulation where $n = \text{me}$ [92], meaning that the data is centred on one patient. We can extend this interpretation to be, for example, one company, one domain, or one industry. As there is no universally accepted definition of big data versus small data, for the purpose of this research, we define big data as following the 5 Vs of volume, velocity, variety, value, and veracity. We also include in our definition exhaustivity of scope where $n = \text{all}$. Data that does not meet those criteria will be characterized as “small”.

Financial data is both big and small data at the same time and depends largely on access and context to determine its appropriate designation. Every minute of every day, somewhere financial data is being generated. This may seem impossible and counter-intuitive, but different countries, religions, and cultures celebrate different holidays at different times of the year, and weekends are not on the same days depending on the country. In North

America and Europe, weekends are Saturday-Sunday, whereas weekends in Arab countries are Thursday-Friday. Also, some shops are open 24-hours a day, around the world.

Therefore, at any given time on any given day, transactions are occurring somewhere — whether online or in person. This means that financial data meets the all of the requirements to be considered “big data”. However, no one has access to all of the financial data that is created. This means that, in reality, we only have access to “small” financial data belonging to specific contexts but also availability. In the context of this research, financial data is also “small” due to inaccessibility of data, we must rely solely on publicly available data. As well, the computational financial research community does not tend to share data due to a myriad of reasons, including that they are using proprietary data. Therefore, for our research, we have had to construct an entirely new dataset which will be discussed in our methodology section.

Another important point that should be addressed is the common philosophy that in order to be useful and find meaningful answers, more is better [205, 98]. This philosophy seems reasonable given that the more *data* a model has, the better informed it is. However, this is based on several important assumptions (1) that the lack of accuracy will be overcome by the size of the data [171], meaning that problematic data will be compensated for by the (assumed) amount of good data in the dataset [255]; and (2) that the data has a high veracity, meaning that it was well curated, is accurate, and can be trusted.

When working with any data, big or small, a major task is determining what data should be used, what features of that data will provide the best insight, and what type(s) of data is most appropriate to answer the research question(s) or hypotheses [131]. Usually, this task requires a very good understanding of the research question(s) or hypotheses, the domain, as well as what is *in* the data itself [131]. When researchers work with big data, it makes this last point even more challenging. It also makes it more challenging to identify bias or potential bias, as it becomes much more understated and elusive the larger dataset. Unstated bias makes the model more vulnerable and can lead to very prominent unintended biases [131].

Deep learning requires much more data than classical (statistical) machine learning in order to be effective. Deep learning learns different things at different levels. Having a lot of data, particularly diverse data, gives the neural network a lot of examples to work with and learn from. This can also help researchers better understand what stimulates neurons [229]. The advantage of deep learning for vision tasks is that it can learn the most important and

useful features of images [261] - a core component of the computer vision pipeline. Tabular datasets, however, typically already contain the necessary features [261], as they have been identified by the dataset author(s), meaning that there is likely no (great) advantage to using deep learning for tabular data [261].

Even with the heightened awareness, particularly in the fields of medicine and health informatics, and finance, for example, of the importance of “ $n = me$ ” or “ n of 1”, the state-of-the-art models are still focusing on deep learning. Research has shown that “big” tabular data is not needed to use deep learning. Therefore, as part of this work, we research the minimum viable thresholds for various downstream tasks using financial statement data, how generating synthetic data can help meet that threshold, and how using synthetic data affects the results of the downstream tasks, using both deep and statistical machine learning.

Appendix C

US Financial System

The US financial system comprises banks and *shadow* banks. Banks are regulated by the Office of the Comptroller of the Currency (OCC), and those regulations exist to help ensure financial stability of the country [224]. Shadow banks, on the other hand, include “bank-like” entities such as mutual funds, insurance companies, investment banks, and other products like asset-backed commercial paper [118]. A key characteristic of shadow banks, however, is that they “do not take deposits which frees them from a large amount of regulatory oversight” imposed on regular banks [121]. This lack of oversight creates a phenomenon known as “regulatory arbitrage” [265] where “financial transactions [are] designed specifically to reduce costs or capture profit opportunities created by different regulations or laws” [244]. While regulatory arbitrage is not within the scope of this thesis, it is important to understand that regulatory asymmetry exists amongst entities which “threatens the [financial] well being of national economies” [265] — in this case, the US. Shadow banks are not regulated as *banks*, but may be regulated by other sectors. For example, insurance companies may be regulated by the Department of Insurance for the particular state in which they operate.

C.1 Intermediation

Intermediation is the “act of coming between”, as defined by the Merriam Webster dictionary [210]. In a common financial scenario, there are lenders (credit card companies, as an example) and borrowers (credit card holders), with the bank acting as the intermediary. Intermediaries facilitate the extension of credit and also deal in securities [93] such as pension funds and life insurance (which is considered a “security” under the US Securities and Exchange Commission’s (SEC) rules [284]. Together, banks and shadow banks are known as “financial intermediaries”. The research presented in this chapter focuses on domestic US financial intermediaries that were bailed out by the US Treasury Department during the 2008 financial crisis.

C.2 Going Concern and Concentration Risk

Going concern — the ability for a company to meet its obligations and continue business for the foreseeable future — is a critical accounting assumption that directly affects the way that the financial statements are presented. When the company is evaluated (by management and then separately by the auditors) and is determined to be a *going concern*, the financial statements are prepared on what is known as a “going concern” basis. If the company is determined that it cannot meet or is unlikely to meet this assumption, then the financial statements are (usually) prepared on a “liquidation basis” as the company will have to liquidate as it is no longer able to meet its commitments. Under the Public Company Accounting Oversight Board’s (PCAOB) auditing standard AS 2415, auditors are required to flag any doubt regarding the going concern assumption in their audit report [26], which is included in the 10-K/Annual Report.¹

Concentration Risk also must be disclosed by the company in their Management Discussion & Analysis (in the CAR) if it poses a severe danger to the company [16]. There are many types of Concentration Risk, but for financial intermediaries, this risk manifests in elements such as the facilitation of credit (given that it is the bank’s main source of income [17]) or in the structure of an investment portfolio where the investment(s) are mainly held in either the same types of industries or companies [18], for example. To contextualize the importance of this risk, we offer the following example: In its 2004 Annual 10-K, Bear Stearns stated “*At November 30, 2004, the Company’s most significant concentrations are related to US government and agency inventory positions, including those of the Federal National Mortgage Association and the Federal Home Loan Mortgage Corporation*” [19]. Just over 3 years later, Bear Stearns collapsed.

C.3 Financial Distress

The term “financial distress” is vague as it is often used generally to address any pre-bankruptcy state. While there are many distress stages that a company goes through before declaring bankruptcy (i.e., it is rare for companies to go bankrupt unexpectedly overnight with no warning), these phases are less clearly established. In fact, many researchers assume

¹We note that the terms 10-K and Annual Report are often used synonymously, even though there are some differences in content and requirement

that there is a common understanding of the term “financial distress” and, therefore, do not define it. Broadly, financial distress is when a company is no longer able to “meet its financial obligations” [91], but it can also be interpreted as the financial decline of a company [170, 253], a failed firm [294] or where a firm cannot pay the “interest and principal on its debt” [117]. Surprisingly, in some research, financial distress is also used synonymously with bankruptcy [338]. Our research uncovered a terminology challenge which has not been resolved in the literature. Rather than address the lack of agreement, researchers have opted to state their own definitions of financial distress. However, it is important to note that bankruptcy is a legal state; financial distress is not. US Code Title 11 is the bankruptcy code which lays out the different types and requirements for bankruptcy. There are six different types (more commonly referred to as “Chapters”) of bankruptcy [81]. Corporate bankruptcies are usually filed under Chapters 7, 11, and 15. Bankruptcy may be initiated by either the company itself (voluntary bankruptcy) or the company’s creditors (involuntary bankruptcy).

Appendix D

Tabular Dataset Company Listing

The following table outlines the companies and the number of records (years of financial statements) collected for each company, using its ticker symbol (symbol on the stock exchange). For ease of reading, it has been organized in ascending order based on the number records collected. *Note that the data was collected July 7 and 8th, 2022.*

Table D.1: Tabular Dataset — Companies and Records

Company	Records	Company	Records	Company	Records	Company	Records	Company	Records	Company	Records
FSBC	2	HTH	10	FCCO	20	EBMT	24	FMNB	27	WSFS	32
MNSB	3	FFWM	10	EBTC	20	HBIA	25	IBOC	27	INDB	32
PFHD	3	ISTR	10	FIBK	20	CCNE	25	CZNC	27	THFF	32
BAFN	3	CUBI	11	BOTJ	20	EWBC	25	PNBK	28	FBP	32
FINW	3	INBK	11	FMBM	20	MCBC	25	GABC	28	TFC	32
BRBS	3	MSBI	12	GLBZ	20	CFFI	25	FRAF	28	WABC	32
CFB	4	IBTX	12	MBCN	20	FCBC	25	LKFN	28	WSBC	32
MCBS	5	NBHC	12	CSBB	20	EGBN	25	BANR	28	RF	32
MYFW	6	EMYB	13	SMBK	20	MBWM	25	NWBI	29	PRK	32
MRBK	6	HMST	13	VABK	21	CWBC	25	FFIC	29	CPF	35
BSVN	6	SFST	14	CIZN	21	WFC	26	FFIN	30	PNC	35
BWB	6	MOFG	15	JUVF	21	FHB	26	FBNC	30	TRMK	35
LBC	6	MVBF	17	BMRC	22	CBAN	26	CCBG	30	BK	35
MBIN	7	FFNW	17	BKSC	22	BHB	26	SASR	30	BOH	35
CBFV	7	UBOH	17	CHMG	22	FUSB	26	SFNC	30	CBSH	35
HVBC	7	LMST	18	MUFG	22	HBNC	26	CATY	30	FITB	35
MCB	7	HOMB	18	PNFP	22	CNOB	26	COLB	31	CVBF	35
BWFG	8	CVCY	18	PACW	23	HTBK	26	UMBF	31	ZION	35
HONE	8	CZFS	18	FKYS	23	COFS	27	VLY	31	MTB	35
BFST	8	OPHC	19	OPOF	23	HTLF	27	GBCI	31	SNV	35
CSTR	8	FBIZ	19	SHBI	23	BUSE	27	GSBC	32	ONB	35
BCML	9	MPB	19	LARK	24	ISBA	27	HWC	32	WAFD	36
FBK	9	PFIS	19	EFSC	24	CIVB	27	CTBI	32		
EQBK	9	TCFC	19	UBFO	24	LBAI	27	IBCP	32		
LOB	10	EFST	19	JPM	24	NKSH	27	NBTB	32		

Appendix E

Financial Literacy Results

E.1 Cosine Similarity Results

The results for Cosine Similarity for GPT-3.5 and GPT-4 are presented in this section. In the tables, boldface indicates scores that met or exceeded the 0.8 similarity threshold.

E.2 Readability Results

We also present the Readability results for both GPT-3.5 and GPT-4 from the experiments conducted using the Flesch Reading Ease Score. Boldface in these results indicate scores that have met or exceeded 60, which is the dividing threshold between easier to read and harder to read.

Table E.1: Cosine Similarity Results for GPT-3.5

Net Income							
	fin_soph	fin_uns.	fin_user	gen_aud	no_aud	non-fin	def_term
fin_soph	1						
fin_unsoph	0.4727	1					
fin_user	0.4789	0.4598	1				
gen_aud	0.4114	0.488	0.425	1			
no_aud	0.5527	0.3879	0.482	0.4279	1		
non-fin	0.4076	0.5213	0.3795	0.5458	0.3825	1	
def_term	0.5589	0.488	0.5141	0.5023	0.6343	0.4273	1
Net Revenue							
fin_soph	1						
fin_unsoph	0.5078	1					
fin_user	0.7748	0.4358	1				
gen_aud	0.5188	0.4548	0.4756	1			
no_aud	0.6584	0.4837	0.6323	0.5806	1		
non-fin	0.4302	0.4289	0.4048	0.8049	0.4029	1	
define_term	0.6411	0.4464	0.5815	0.583	0.8371	0.3628	1
Net Interest Income							
fin_soph	1						
fin_unsoph	0.6616	1					
fin_user	0.777	0.616	1				
gen_aud	0.491	0.4715	0.5035	1			
no_aud	0.9114	0.6821	0.7553	0.5141	1		
non-fin	0.5864	0.7061	0.5683	0.4717	0.6062	1	
define_term	0.8809	0.6699	0.7692	0.4693	0.8215	0.618	1
Provision for Loan Loss							
fin_soph	1						
fin_unsoph	0.3829	1					
fin_user	0.6135	0.3645	1				
gen_aud	0.5547	0.4556	0.7147	1			
no_aud	0.6535	0.3927	0.7407	0.6808	1		
non-fin	0.4336	0.4559	0.5005	0.704	0.6316	1	
define_term	0.6272	0.4	0.818	0.7308	0.7854	0.549	1
Subordinated Debt							
fin_soph	1						
fin_unsoph	0.6581	1					
fin_user	0.6464	0.5972	1				
gen_aud	0.4592	0.5051	0.5027	1			
no_aud	0.7709	0.6703	0.6868	0.5228	1		
non-fin	0.6176	0.868	0.6197	0.4459	0.625	1	
define_term	0.6464	0.6351	0.7547	0.4808	0.6868	0.6197	1

Table E.2: Cosine Similarity Results for GPT-3.5 (Continued)

Generally Accepted Audit Standards (GAAS)							
	fin_soph	fin_uns.	fin_user	gen_aud	no_aud	non-fin	def_term
fin_soph	1						
fin_unsoph	0.4943	1					
fin_user	0.8090	0.4346	1				
gen_aud	0.5140	0.4799	0.4930	1			
no_aud	0.7994	0.4570	0.8725	0.5327	1		
non-fin	0.4721	0.5840	0.4327	0.4532	0.4656	1	
def_term	0.4912	0.4866	0.4541	0.5178	0.5096	0.4634	1
Material Misstatement							
fin_soph	1						
fin_unsoph	0.4744	1					
fin_user	0.7380	0.4979	1				
gen_aud	0.4431	0.4196	0.4742	1			
no_aud	0.6737	0.4746	0.6547	0.4795	1		
non-fin	0.4507	0.9346	0.4696	0.4127	0.4686	1	
define_term	0.7136	0.4673	0.7991	0.4995	0.8122	0.4623	1
Professional Judgment							
fin_soph	1						
fin_unsoph	0.4467	1					
fin_user	0.4467	0.4653	1				
gen_aud	0.4550	0.4736	0.3576	1			
no_aud	0.4315	0.4680	0.3463	0.3746	1		
non-fin	0.3715	0.4065	0.3141	0.4689	0.4018	1	
define_term	0.5075	0.4478	0.4776	0.4274	0.4986	0.4178	1
Professional Skepticism							
fin_soph	1						
fin_unsoph	0.4063	1					
fin_user	0.4479	0.5145	1				
gen_aud	0.4646	0.4668	0.5346	1			
no_aud	0.4489	0.4446	0.4939	0.5239	1		
non-fin	0.4348	0.4519	0.4556	0.4435	0.4576	1	
define_term	0.4237	0.3759	0.4546	0.4715	0.4111	0.4511	1

Table E.4: Cosine Similarity Results for GPT-4

Net Income							
	fin_soph	fin_uns.	fin_user	gen_aud	no_aud	non-fin	def_term
fin_soph	1						
fin_unsoph	0.3411	1					
fin_user	0.9264	0.3663	1				
gen_aud	0.4939	0.4348	0.5188	1			
no_aud	0.6035	0.3465	0.5626	0.6087	1		
non-fin	0.2666	0.3874	0.2869	0.3129	0.2611	1	
def_term	0.5762	0.3672	0.5462	0.6686	0.9136	0.2831	1
Net Revenue							
fin_soph	1						
fin_unsoph	0.3744	1					
fin_user	0.7565	0.3940	1				
gen_aud	0.5547	0.5317	0.5898	1			
no_aud	0.7867	0.3940	0.8066	0.6054	1		
non-fin	0.2744	0.5673	0.2813	0.4086	0.2995	1	
define_term	0.9196	0.3837	0.7332	0.5622	0.8602	0.2888	1
Net Interest Income							
fin_soph	1						
fin_unsoph	0.3833	1					
fin_user	0.9486	0.3761	1				
gen_aud	0.5132	0.4305	0.4901	1			
no_aud	0.7147	0.4869	0.6797	0.5643	1		
non-fin	0.4085	0.4906	0.4095	0.6751	0.4662	1	
define_term	0.7744	0.4577	0.7555	0.5687	0.8554	0.4382	1
Provision for Loan Loss							
fin_soph	1						
fin_unsoph	0.3584	1					
fin_user	0.7514	0.4012	1				
gen_aud	0.4468	0.6860	0.4739	1			
no_aud	0.8088	0.3935	0.8298	0.4935	1		
non-fin	0.3610	0.7393	0.4026	0.6712	0.3850	1	
define_term	0.8088	0.4154	0.8703	0.5237	0.9140	0.4278	1
Subordinated Debt							
fin_soph	1						
fin_unsoph	0.3813	1					
fin_user	0.6619	0.4130	1				
gen_aud	0.4310	0.5022	0.4341	1			
no_aud	0.6768	0.3779	0.8521	0.4091	1		
non-fin	0.3629	0.5566	0.4071	0.4489	0.3822	1	
define_term	0.6600	0.3852	0.8536	0.4082	0.9654	0.4070	1

Table E.5: Cosine Similarity Results for GPT-4 (Continued)

Generally Accepted Audit Standards (GAAS)							
	fin_soph	fin_uns.	fin_user	gen_aud	no_aud	non-fin	def_term
fin_soph	1						
fin_unsoph	0.4792	1					
fin_user	0.9241	0.4837	1				
gen_aud	0.3816	0.4244	0.3977	1			
no_aud	0.5545	0.3918	0.5600	0.3212	1		
non-fin	0.4242	0.8373	0.4152	0.4119	0.3664	1	
def_term	0.7619	0.5319	0.7900	0.4146	0.5251	0.4560	1
Material Misstatement							
fin_soph	1						
fin_unsoph	0.3226	1					
fin_user	0.9641	0.2866	1				
gen_aud	0.4883	0.4112	0.4619	1			
no_aud	0.9321	0.3260	0.9247	0.4872	1		
non-fin	0.3292	0.7410	0.2999	0.4024	0.3395	1	
define_term	0.9131	0.3212	0.8990	0.4904	0.9613	0.3349	1
Professional Judgment							
fin_soph	1						
fin_unsoph	0.4895	1					
fin_user	0.8196	0.4930	1				
gen_aud	0.3616	0.4177	0.3766	1			
no_aud	0.5075	0.4760	0.4973	0.4119	1		
non-fin	0.4162	0.4640	0.4059	0.4513	0.4510	1	
define_term	0.4926	0.4491	0.4956	0.4129	0.7889	0.4621	1
Professional Skepticism							
fin_soph	1						
fin_unsoph	0.4564	1					
fin_user	0.9393	0.4625	1				
gen_aud	0.4239	0.4500	0.4413	1			
no_aud	0.5610	0.4352	0.5853	0.4186	1		
non-fin	0.4287	0.7166	0.4298	0.4065	0.5703	1	
define_term	0.5795	0.4538	0.5914	0.4268	0.9388	0.5962	1

Net Income	Readability Score	# of Words	Generally Accepted Audit Standards (GAAS)	Readability Score	# of Words	Capital	Readability Score	# of Words
fin_soph	21.72	242	fin_soph	10.24	192	fin_soph	20.28	232
fin_unsoph	55.64	208	fin_unsoph	38.43	135	fin_unsoph	33.36	185
fin_user	25.10	200	fin_user	5.61	226	fin_user	26.15	255
gen_aud	65.05	237	gen_aud	31.65	163	gen_aud	38.72	182
no_aud	23.86	184	no_aud	4.69	172	no_aud	16.99	141
non_fin	75.38	167	non_fin	39.68	132	non_fin	40.76	183
define_term	35.28	182	define_term	31.72	236	define_term	20.53	186
Net Revenue			Material Misstatement			Investment		
fin_soph	30.74	220	fin_soph	-12.61	172	fin_soph	20.11	230
fin_unsoph	39.07	115	fin_unsoph	24.55	193	fin_unsoph	55.52	189
fin_user	24.44	235	fin_user	4.35	265	fin_user	23.50	230
gen_aud	49.02	217	gen_aud	12.91	136	gen_aud	35.74	230
no_aud	23.25	166	no_aud	7.14	146	no_aud	23.45	177
non_fin	54.55	159	non_fin	22.38	195	non_fin	54.65	209
define_term	24.56	178	define_term	4.38	193	define_term	29.99	232
Net Interest Income			Professional Judgment			Assets		
fin_soph	19.78	218	fin_soph	4.32	181	fin_soph	32.46	202
fin_unsoph	34.09	204	fin_unsoph	22.91	180	fin_unsoph	47.37	161
fin_user	19.34	222	fin_user	17.68	191	fin_user	25.18	224
gen_aud	32.72	126	gen_aud	25.37	199	gen_aud	68.09	187
no_aud	18.23	211	no_aud	8.21	167	no_aud	23.55	161
non_fin	28.50	218	non_fin	25.20	217	non_fin	68.61	165
define_term	18.10	184	define_term	15.16	165	define_term	30.52	177
Provision for Loan Losses			Liabilities					
fin_soph	21.85	272	fin_soph	0.94	160	fin_soph	25.94	195
fin_unsoph	52.02	256	fin_unsoph	3.73	94	fin_unsoph	34.09	205
fin_user	30.24	309	fin_user	9.00	174	fin_user	31.13	243
gen_aud	45.49	310	gen_aud	21.08	163	gen_aud	33.77	136
no_aud	32.57	239	no_aud	8.20	115	no_aud	24.33	130
non_fin	49.39	223	non_fin	6.50	154	non_fin	49.82	145
define_term	30.80	270	define_term	15.19	160	define_term	24.52	163
Subordinated Debt			Financial Statements					
fin_soph	40.58	257	fin_soph	11.27	230	fin_soph	17.78	264
fin_unsoph	41.88	184	fin_unsoph	19.16	152	fin_unsoph	25.02	216
fin_user	38.57	191	fin_user	10.05	234	fin_user	22.92	249
gen_aud	46.15	121	gen_aud	23.96	136	gen_aud	25.90	254
no_aud	35.95	175	no_aud	23.96	231	no_aud	12.94	203
non_fin	41.34	205	non_fin	25.31	176	non_fin	33.81	232
define_term	46.30	195	define_term	4.57	212	define_term	18.06	260

Net Income	Readability Score	# of Words	Generally Accepted Audit Standards (GAAS)	Readability Score	# of Words	Capital	Readability Score	# of Words
fin_soph	36.62	322	fin_soph	24.84	331	fin_soph	INC	
fin_unsoph	77.01	203	fin_unsoph	48.15	274	fin_unsoph	65.08	226
fin_user	36.17	315	fin_user	25.71	316	fin_user	17.72	204
gen_aud	40.79	192	gen_aud	16.99	141	gen_aud	48.28	142
no_aud	38.95	210	no_aud	17.77	178	no_aud	26.27	113
non_fin	63.08	179	non_fin	49.36	279	non_fin	62.29	212
define_term	38.63	169	define_term	31.88	314	define_term	19.96	119
Net Revenue			Material Misstatement			Investment		
fin_soph	14.33	153	fin_soph	-3.32	278	fin_soph	INC	
fin_unsoph	75.37	147	fin_unsoph	36.03	252	fin_unsoph	77.5	200
fin_user	11.53	139	fin_user	-1.24	290	fin_user	INC	
gen_aud	39.72	129	gen_aud	10.18	139	gen_aud	INC	
no_aud	18.12	133	no_aud	-2.38	262	no_aud	33.59	195
non_fin	77.68	156	non_fin	33.72	239	non_fin	72.29	204
define_term	17.23	131	define_term	-5.23	255	define_term	39.86	275
Net Interest Income			Professional Judgment			Assets		
fin_soph	26.66	254	fin_soph	20.41	286	fin_soph	INC	
fin_unsoph	54.01	195	fin_unsoph	48.63	207	fin_unsoph	52.45	218
fin_user	16.26	240	fin_user	21.37	279	fin_user	INC	
gen_aud	57.39	251	gen_aud	17.23	130	gen_aud	68.86	142
no_aud	35.35	191	no_aud	29.05	166	no_aud	33.25	71
non_fin	63.9	210	non_fin	47.15	228	non_fin	72.01	229
define_term	35.28	211	define_term	24.89	157	define_term	40.68	193
Provision for Loan Losses			Professional Skepticism			Liabilities		
fin_soph	30.11	196	fin_soph	-0.38	276	fin_soph	INC	
fin_unsoph	58.47	143	fin_unsoph	26.54	278	fin_unsoph	INC	
fin_user	24.31	185	fin_user	0.28	287	fin_user	INC	
gen_aud	56.24	175	gen_aud	33.83	127	gen_aud	INC	
no_aud	31.38	161	no_aud	15.94	220	no_aud	INC	
non_fin	63.87	139	non_fin	28.96	261	non_fin	INC	
define_term	32.68	164	define_term	18.38	227	define_term	INC	
Subordinated Debt			Reasonable Assurance			Financial Statements		
fin_soph	28.16	248	fin_soph	13.62	264	fin_soph	INC	
fin_unsoph	53.62	255	fin_unsoph	37.57	241	fin_unsoph	INC	
fin_user	21.15	257	fin_user	16.31	262	fin_user	INC	
gen_aud	49.45	170	gen_aud	28.5	127	gen_aud	INC	
no_aud	22.02	241	no_aud	25.77	265	no_aud	INC	
non_fin	56.89	175	non_fin	40.4	255	non_fin	INC	
define_term	23.15	238	define_term	23.13	262	define_term	INC	

Appendix F

Selection of Peers

Before developing the approach of using peers to help with the generation of synthetic data, we experimented with two well-known clustering algorithms — k-means and k-nearest neighbour. K-means is an unsupervised clustering method that we thought could be helpful to find peers. Given the size of the data (3,207 records in our dataset), an unsupervised method would be ideal, as it could identify peers faster than a supervised or doing it by hand. Initially, k-means randomly chooses centroids. However, after the first pass, centroids are then chosen using the data “based on a probability to the squared distance away from a given point’s nearest existing centroid” [206], using a row-wise approach [282]. This makes sense given that records in a dataset are usually organized by rows. However, when looking for peers for our task, row-wise does not find a bank’s true competitive peers based on size, but rather the distribution of each record, which can skew the peer results.

We also tried k-nearest neighbour with the labelled data (labelled following Section 3.3.2.2) to see if it would improve the peer identification results. There are three popular choices for calculating the distance between the data point and the training set: Minkowski, Euclidean, and Manhattan. The distance metric (parameter p) has a direct effect on the “neighbourhood” [34]. As Euclidean is the most common, we selected that as our distance metric. Although k-nearest neighbour worked better than k-means, it still did work well for finding peers based on size.

Appendix G

ChatGPT Transcripts for Synthetic Generation

Below are the transcripts for the synthetic data generation for the entire dataset, for JP-Morgan Chase & Co (JPM) and Bayfirst Financial Corp (BAFN).

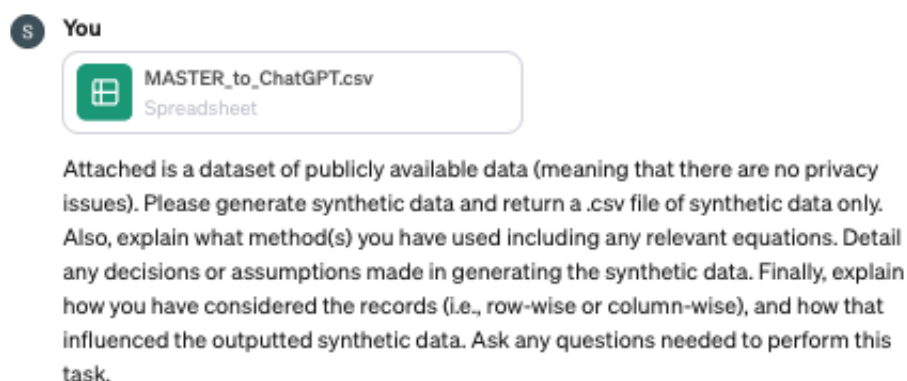


Figure G.1: Prompt given to ChatGPT to generate synthetic data for the entire dataset.



ChatGPT

The dataset contains financial data for various companies, with 3,207 entries and 17 columns. The columns include a mix of numerical and categorical data, such as company names, dates, and various financial metrics like net interest income, total assets, and liabilities.

Approach for Generating Synthetic Data

1. **Column-Wise Generation:**
 - I will generate synthetic data for each column independently. This method is simpler and more computationally efficient but may not fully capture complex inter-column correlations.
2. **Handling Different Data Types:**
 - **Numerical Data:** For columns like `'Net_Int_Income'`, `'Total_Non_Int'`, etc., I'll use random sampling with statistical constraints based on the original data's distribution.
 - **Categorical Data:** For the `'COMPANY'` column, I'll replicate the existing company names as they are specific identifiers.
 - **Date Data:** For the `'DATE'` column, I'll generate new dates that are within the range of the original dates.
3. **Assumptions and Decisions:**
 - **Preserving Statistical Properties:** The mean and standard deviation of the synthetic numerical data will be kept close to the original data.
 - **Handling Null Values:** I noticed some null values in the `'Minority'` column. I will impute these using the mean of the column for simplicity.
4. **Relevant Equations:**
 - For a numerical column X , synthetic data X_{new} is generated as:

$$X_{new} = \mu + \sigma \times Z$$
 where μ and σ are the mean and standard deviation of X , and Z is a random number from a standard normal distribution.

Figure G.2: Response 1 for ChatGPT

Generating the Synthetic Data

I will now proceed to generate the synthetic data based on the above approach. The output will be a .csv file containing only the synthetic data. [↗]

The synthetic data has been generated and saved as a CSV file. You can download it using the link below:

[Download the Synthetic Data CSV File](#)

Methodology and Assumptions:

1. **Numerical Data:** I generated synthetic values for numerical columns using a normal distribution based on the mean and standard deviation of each column. This approach preserves the overall statistical properties like mean and variance.
2. **Categorical Data:** For the 'COMPANY' column, I retained the existing company names, as changing them would lose the context of the dataset.
3. **Date Data:** For the 'DATE' column, I generated random dates within the range of the original dataset's dates.
4. **Null Value Handling:** The 'Minority' column had some missing values, which I imputed using the mean of the column.
5. **Row-wise vs. Column-wise Approach:** I considered each column independently for synthetic data generation. This method is computationally efficient and suitable for maintaining individual column statistics. However, it might not fully capture complex inter-column correlations.
6. **Relevant Equations:**
 - For a numerical column X , synthetic data X_{new} is generated as: $X_{new} = \mu + \sigma \times Z$, where μ and σ are the mean and standard deviation of X , and Z is a random number from a standard normal distribution.

This approach ensures the synthetic dataset has similar statistical properties to your original data while creating entirely new data points. [↗]



Figure G.3: Response 2 for ChatGPT

Appendix H

Coherence Scores

Table H.2: Coherence Scores - Unknown

NAMED	Barclays	BMO	BNY	CIBC	Citi	DB	Erste	JPM	NBC	Rabo	RBC	Sberbank	Scotia	USB	WF
# of Topics															
5	0.30096	0.28359	0.33279	0.30486	0.31291	0.29763	0.33584	0.31997	0.34386	0.32335	0.29419	0.27996	0.31197	0.31247	0.32817
6	0.28324	0.29138	0.33099	0.31492	0.31411	0.30871	0.33872	0.31659	0.34590	0.32237	0.29723	0.27369	0.31567	0.31422	0.33358
7	0.29024	0.28976	0.33370	0.30710	0.31379	0.30500	0.34027	0.31999	0.33735	0.32765	0.29230	0.27485	0.30467	0.31084	0.32777
8	0.28536	0.28905	0.33101	0.31232	0.31997	0.30343	0.34549	0.32614	0.32543	0.32961	0.29439	0.27495	0.31102	0.31382	0.33591
9	0.29291	0.28798	0.33065	0.30760	0.31795	0.29874	0.33285	0.32098	0.33846	0.33791	0.29765	0.26425	0.31486	0.30906	0.33284
10	0.29647	0.29205	0.33112	0.31017	0.31965	0.30796	0.33372	0.32053	0.33431	0.31866	0.29896	0.27099	0.31064	0.31071	0.32911
11	0.29459	0.28282	0.33507	0.30289	0.32647	0.29811	0.34367	0.31978	0.34347	0.32774	0.29473	0.27263	0.30608	0.31605	0.32631
12	0.29155	0.29639	0.33186	0.30876	0.32097	0.29657	0.34389	0.31542	0.33902	0.31399	0.29678	0.27406	0.30940	0.31269	0.32985
13	0.29850	0.28515	0.33038	0.31550	0.30675	0.30302	0.34706	0.31931	0.33698	0.31991	0.29487	0.28223	0.31438	0.31128	0.32299
14	0.29483	0.28571	0.33345	0.30574	0.31671	0.30123	0.35177	0.32147	0.34445	0.32491	0.29870	0.27924	0.31024	0.30640	0.33561
15	0.29936	0.28685	0.33294	0.30882	0.32035	0.29957	0.34172	0.31760	0.33960	0.31830	0.29902	0.27686	0.31023	0.31101	0.32880
UNNAMED															
# of Topics															
5	0.30399	0.29023	0.32728	0.32116	0.32435	0.33016	0.34487	0.31974	0.34641	0.33425	0.29449	0.27971	0.31705	0.31563	0.32420
6	0.29252	0.29747	0.32884	0.31250	0.31472	0.32014	0.34778	0.31982	0.34554	0.34256	0.29186	0.28186	0.32153	0.31512	0.31185
7	0.30328	0.29765	0.33518	0.31534	0.31407	0.30812	0.34876	0.31633	0.33189	0.34676	0.29939	0.28288	0.32616	0.30904	0.32593
8	0.29726	0.28522	0.32974	0.31293	0.31766	0.31810	0.33031	0.31472	0.33756	0.32059	0.29523	0.29697	0.31886	0.31846	0.32797
9	0.29661	0.28899	0.33067	0.31925	0.32965	0.31772	0.34079	0.31793	0.32215	0.32780	0.29325	0.29587	0.31835	0.31174	0.32186
10	0.28992	0.29264	0.32796	0.30932	0.31583	0.31537	0.34211	0.31814	0.34100	0.33127	0.29868	0.28746	0.32509	0.31796	0.32597
11	0.29496	0.29215	0.33037	0.30960	0.31086	0.31976	0.33993	0.31606	0.34330	0.32916	0.29387	0.27632	0.31841	0.31545	0.31951
12	0.29319	0.29518	0.32941	0.31502	0.31900	0.32097	0.34167	0.31889	0.32995	0.32343	0.29277	0.27539	0.31867	0.31135	0.31788
13	0.29808	0.28904	0.33244	0.31743	0.31759	0.31620	0.34353	0.31899	0.33426	0.32268	0.29452	0.28114	0.31885	0.30771	0.32994
14	0.29537	0.29020	0.33059	0.31568	0.31446	0.31345	0.34767	0.32184	0.33963	0.32919	0.29718	0.27940	0.31750	0.31644	0.32185
15	0.30534	0.29271	0.32839	0.31836	0.32059	0.31345	0.34357	0.32435	0.33320	0.32570	0.29494	0.27518	0.32466	0.31183	0.31359

Appendix I

Features Removed for Project 7 - Experiment 1

I.1 Textual features removed prior to Latent Dirichlet Allocation

Canada

(1) *BMO*: BMO, BMO Harris Private Banking, BMO Financial Group, Canadian, BMO Life Assurance, BMO Financial Corp, Commercial Banking Canada, BMO Global Asset Management, BMO Life Insurance, BMO InvestorLine, BMO Capital Markets, BMO Financial Group, 119 rue Saint-Jacques, Montreal, Quebec, Québec, H2Y 1K6, Phone: 1-800-340-5021, Canada, U.S. 514-982-7800, Fax: 416-263-9394.

Total words removed: 9,259 (57,482 bytes)

(2) *CIBC*: CIBC, Canadian, Imperial, Bank, Commerce, Canadian-based, Canada, (800) 465-2422, Commercial, North America, CXNPS, Client Experience, Net Promoter Score Index, FirstCaribbean, Commercial Banking, Wealth Management, 81 Bay Street, Toronto, Ontario, M5L 1A2.

Total words removed: 9,166 (54,467 bytes)

(3) *National Bank*: National Bank, Credigy, ABA Bank, Canada, Canadian, The Bank, The Bank's, Bank's, PCAF, United Nations (UN), Principles for Responsible Banking, Specialty Finance and International, (USSF&I), 800 Rue Saint-Jacques, Street, Montreal, Quebec, Québec, Square-Victoria-OACI or Bonaventure.

Total words removed: 6,570 (37,941 bytes)

(4) *Royal Bank of Canada*: Royal Bank of Canada, RBC, Canada, Canadian, (RY on TSX and NYSE), Global Technology and Operations, Abacus Financial Services Group, Thomson Financial, Bloomberg, Financial Post, rbc.com, GTO, Global Functions, Dexia Investor Services, RBC Dexia IS, Carlin Financial Group, Toronto, 200 Bay Street, North Tower, P.O. Box 75, Ontario, M5J 2Z5.

Total words removed: 18,477 (109,142 bytes)

(5) *Scotiabank*: Scotia, Scotiabank, Canada, Bank, Canadian, Scotiabankers, Private Investment Counsel Firms, Structural Cost Transformation, SCT, Smart Automation, Digital Retail Sales, Toronto, Ontario, 44 King St. West, M5H 1H1, 40 Temperance Street, M5H 0B4, (416) 866-6161.

Total words removed: 13,477 (95,142 bytes)

Europe

(1) *Barclays*: Barclays, PLC, Syndications, Private Banking, UK, United Kingdom, London, 1 Churchill Place, England, E14 5HP, Phone, Fax, +44 (0) 20 7116 1000, +44 (0) 20 7116 7665, Johnson Smirke Building, 4 Royal Mint Court, EC3 N4H, British.

Total words removed: 34,274 (186,693 bytes)

(2) *Deutsche Bank*: The Group, Group, DBK GR, Asset & Wealth Management Preserving, Supervisory Board, Deutsche Bank AG, Group Executive Committee, GEC, Fax, +49 69 910-34 225, German, Wealth, & Wealth, DeAWM, Clients-Corporate Banking & Securities, CCB&S, Frankfurt, AG 1 Columbus Circle New York, 10019-8735, Taunusanlage 12 60325, am Main, Tel., +49 69 910-00.

Total words removed: 28,827 (190,341 bytes)

(3) *Erste Group AG*: Erste, Am 1 Belvedere, 1100 Vienna, Austria, E-Mail:, investor.relations@erstegroup.com, internet, www.erstegroup.com/en/investoren, Investor Relations & Accounting Teams, Foundation, European Central Bank, ECB, European.

Total words removed: 2,291 (18,952 bytes)

(4) *Rabo Bank*: Rabo, Rabobank, Group's, Group, Global Reporting Initiative, GRI, Communications & Corporate Affairs, Croeselaan 18 Postbus 17100 3500 HG, Utrecht, Nederland, Phone, +31 (0) 30 216 0000, E-Mail, AGRI3 Fund, Foundation, Security Threat Elimination Program, STEP, Council of European Development Bank, CEB, Fair Finance Guide International, FFGI, Europe, European, Netherlands.

Total words removed: 10,351 (87,283 bytes)

(5) *Sberbank*: Sberbank, SBER, London, Moscow, Russia, Russian, Sberbank PSC, ShreBank, Bank 2 Eherbank OJSC, Europe, JSC, LLC, DeliveryClub, Local Kitchen, Samokat, Citimobil, YouDrive, YourMoney LLC, SberDevices, SberHealth, DublGIS, Sberbank-Telecom, Centre of Speech Technologies, Sberbank-Automated, SberBusiness AP, JPSC Sberbank, 19 Valilova St, 117997, Phone, +7 (495) 974-66-77, Fax, +7 (495) 957-57-31, www.sberbank.com.

Total words removed: 2,354 (15,953 bytes)

US

(1) *Bank of New York Mellon*: BNY, Mellon, Bank, New York, U.S., United States, Brookfield, Place, Manhattan, Seventh Circuit, One Wall Street, CIBC Mellon, Federal Reserve, Investment Management, EMEA, Global Collateral Service, N.A., Ancillary Regulatory Requirements, Securities Exchange Act of 1934, Exchange Act, 240 Greenwich Street, 10286, 1735 Market Street 2nd Floor, AIM 193-0261.

Total words removed: 44,451 (214,005 bytes)

(2) *Citi Bank*: Citi, Holdings, Citigroup, Inc., U.S., Bike®), bike, Mobile Challenge, Mobile®, Kinder Morgan, CitiDirect, ®), United States, UNITED STATES, SECURITIES AND EXCHANGE COMMISSION, Washington, D.C., New York, 388 Greenwich Street, 10012, U.S.A., 153 East 53rd Street, 16/F, Zone 19, 10022.

Total words removed: 9,428 (75,354 bytes)

(3) *JPMorgan Chase & Co*: JPMORGAN CHASE & Co., NYSE: JPM, Bank One, Bank One/JPMorgan Chase & Co, N.A., Chase MyHome, Chase Mobile, Corporate & Investment Bank, CIB, Chase, Midtown, Manhattan, J.P. Morgan, Morgan, United States, U.S., 270 Park Avenue, New York, 1-800-935-8835, www.jpmorganchase.com, 383 Madison Avenue.

Total words removed: 9,766 (49,091 bytes)

(4) *US Bank*: U.S, Bancorp, BANCORP, Bank, United States, USA, American, FlexControl, www.usbank.com, 1-800-872-2657, 800 Nicolette Mall Ste 1500, Minneapolis, MN, Minnesota, 55402-7014, Floor Two.

Total words removed: 3,646 (26,556 bytes)

(5) *Wells Fargo & Company*: Wells Fargo, & Company, First Security, U.S., United States, USA, Community Banking, Financial, Norwest Financial, American, San Francisco, California, 420 Montgomery St., Street, 94104.

Total words removed: 11,243 (71,068 bytes)

I.2 Topic features for Latent Dirichlet Allocation evaluation

Top Topic Features Removed

The number of topics generated for each country were different, ranging from 21 to 33 between both “named” and “unnamed”. For standardization where dimensions x (known reports) and y (unknown reports) had to be the same shape, we used the lowest number for all — 21 topics. Therefore, features were removed (top or bottom) to achieve 21 topics.

Named - Known

Canada: credit, financial, risk, asset, value

Europe: financial, total, credit, risk, value, include, loan, year asset, loss, loan

US: loan, loss, asset, include, credit, value, risk, financial, total

Named - Unknown

Canada: loan, asset, risk, credit, value, loss

Europe: interest, value, total, year, asset, include, risk, financial report, management, loss, credit, loan
US: asset, risk, value, loss, interest, credit, financial, include

Unnamed - Known

Canada: risk, asset

Europe: asset, financial, include, business, year, credit, interest, risk, total, value, market, loss

US: asset, loss, credit, market, financial, value, include, business, year

Named - Unknown

Canada: include, asset, value, risk

Europe: business, financial, total, include, interest, credit, asset, year, loss, market
US: interest, financial, market, total, include, business, asset, credit, loss, year

Bottom Features Removed

Named - Known

Canada: capital, amount, security, fair, share

Europe: rate, capital, fair, amount, income, eur, service, corporate, employee, sberbank, client, rabo

US: business, market, management, find, capital, fair, change, citis, derivative

Named - Unknown

Canada: statement, annual, service, capital, share, fair, security Europe: business, share, capital, rate, fair, net, income, amount, result, annual, service, employee, corporate, sberbank

US: management, fund, citis, rate, change, fair, year, mortgage

Unnamed - Known

Canada: share, capital

Europe: share, rate, fair, market, net, eur, amount, company, client, annual, service, corporate

US: fund, market, result, business, change, rate, report, fair, mortgage

Named - Unknown

Canada: instrument, fair, capital, report

Europe: rate, fair, eur, amount, result, bank, statement, annual, service, corporate

US: citis, fair, income, report, management, company, commercial, percent, mortgage, change, rate

Appendix J

Topic Scores

TOPIC SCORES - CANADA - NAMED																									
include	credit	financial	risk	total	group	year	asset	income	loss	interest	value	loan	management	net	market	capital	business	rate	instrument	bank	amount	security	fair	share	
bmo																									
bmo	0.009	0.009	0.013	0.009	0.008	0.010	0.009	0.011		0.010	0.010	0.008	0.008												
bmo	0.010	0.012	0.017	0.009		0.010	0.010	0.010		0.009	0.010	0.010		0.011											
bmo	0.011	0.013	0.014				0.010	0.012	0.009	0.009	0.009	0.009		0.010											
bmo	0.015	0.012	0.016				0.010	0.013	0.009	0.009	0.008	0.009		0.010											
bmo	0.010	0.010	0.015				0.009	0.010	0.010		0.008	0.009		0.010											
bmo	0.010	0.014	0.015				0.012	0.010	0.009	0.008	0.009	0.009		0.008				0.008							
cbc	0.008	0.016	0.011				0.010	0.012	0.009	0.009	0.009			0.008											
cbc	0.010	0.012	0.018				0.010	0.010	0.010		0.010	0.009		0.008											
cbc	0.012	0.013	0.022				0.013	0.010	0.010		0.009	0.009		0.009			0.008								
cbc	0.009	0.014	0.009				0.008	0.011	0.010		0.009						0.009								
cbc	0.009	0.013	0.021				0.012	0.009	0.011	0.009	0.008						0.011								
cbc	0.009	0.013	0.021				0.009	0.010	0.008	0.009	0.010														
cbc	0.009	0.014	0.019				0.012	0.011	0.008	0.009	0.010			0.010				0.009							
cbc	0.009	0.014	0.017				0.009	0.009	0.011	0.009	0.008			0.009											
nbc	0.014	0.018	0.020				0.013	0.012	0.009		0.011	0.009									0.011				
nbc	0.012	0.018	0.021				0.015	0.014	0.012	0.011	0.009										0.027	0.008	0.008		
nbc	0.013	0.025	0.018				0.008	0.010	0.013	0.013	0.011										0.021	0.008			
nbc	0.011	0.022	0.017				0.013	0.014	0.009	0.011	0.010			0.010							0.015				
nbc	0.013	0.022	0.021				0.011	0.011	0.009	0.011	0.010			0.010							0.019				
nbc	0.013	0.020	0.018				0.010	0.013	0.012	0.009	0.010			0.009							0.020				
nbc	0.011	0.019	0.019				0.012	0.014	0.012	0.011	0.010			0.009							0.009				
nbc	0.013	0.020	0.017				0.013	0.010	0.010	0.009	0.011			0.009							0.024				
nbc	0.010	0.013	0.016				0.014	0.012	0.016	0.009	0.009			0.009							0.018				
rbc	0.010	0.011	0.016				0.011	0.009	0.008		0.013	0.009			0.009										
rbc	0.010	0.011	0.021				0.011	0.009	0.008	0.008	0.008			0.009											
rbc	0.010	0.013	0.013				0.013	0.010			0.008			0.009		0.008	0.008								
rbc	0.009	0.010	0.014				0.010	0.009	0.010		0.011			0.009			0.010								0.009
rbc	0.010	0.012	0.011				0.010	0.009	0.010		0.011			0.009			0.012								
rbc	0.011	0.010	0.013				0.012	0.009	0.009		0.012	0.009		0.011											
rbc	0.015	0.011	0.019				0.009	0.011	0.010		0.010	0.010									0.018				
scotia	0.012	0.014	0.018				0.010	0.012	0.015	0.010	0.011	0.012									0.017				
scotia	0.010	0.009	0.025				0.009	0.012	0.010	0.009	0.009	0.010									0.018				0.009
scotia	0.013	0.012	0.015				0.010	0.013	0.011	0.009	0.009	0.009									0.017				
scotia	0.009	0.015	0.020				0.011	0.012	0.013	0.009	0.009			0.012							0.018				
scotia	0.013	0.012	0.016				0.010	0.012	0.014	0.009	0.009			0.012							0.020				
scotia	0.013	0.012	0.015				0.010	0.015	0.016	0.009	0.009			0.010							0.020				
scotia	0.018	0.012	0.016				0.009	0.012	0.012	0.009	0.009			0.010							0.021	0.009			

Figure J.1: Topic scores for each feature for Canada “Named”

TOPIC SCORES - EUROPE - NAMED																																		
financial	total	credit	market	risk	rate	business	share	interest	value	include	loan	year	group	report	capital	asset	fair	loss	management	amount	income	customer	eur	rabobank	bank	statement	service	corporate	employee	sberbank	client			
barclays	0.011				0.013	0.007	0.007	0.007	0.008	0.007	0.007	0.007	0.012	0.008	0.011																			
barclays	0.011				0.019	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.012	0.01	0.01																			
barclays	0.010				0.021	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.015	0.007	0.008																			
barclays	0.010				0.02	0.009	0.008	0.008	0.008	0.007	0.008	0.008	0.013	0.008	0.01																			
barclays	0.012				0.017				0.007	0.007	0.007	0.008	0.014	0.009	0.01																			
barclays	0.008	0.009			0.019	0.009	0.007	0.008	0.008	0.009	0.009	0.009	0.013	0.01	0.01																			
barclays	0.011				0.019	0.009	0.009	0.009	0.007	0.011	0.008	0.011	0.014	0.009	0.01																			
barclays	0.011	0.01	0.007		0.017	0.009	0.008	0.007	0.008	0.008	0.008	0.014	0.014	0.007	0.012																			
barclays	0.012	0.008			0.021	0.007	0.009	0.009	0.009	0.009	0.009	0.013	0.009	0.009	0.009																			
barclays	0.013	0.008			0.018	0.009	0.009	0.009	0.008	0.007	0.008	0.013	0.008	0.012	0.012																			
db	0.018		0.01	0.02				0.009	0.009	0.009	0.009	0.011	0.008	0.008	0.015	0.010	0.008	0.008	0.008	0.008	0.009	0.009	0.011											
db	0.020		0.01	0.009	0.02			0.012	0.009	0.009	0.009	0.008	0.009	0.008	0.015	0.009	0.008	0.008	0.008	0.009	0.008	0.009	0.011											
db	0.017	0.008	0.01	0.017	0.017			0.009	0.009	0.009	0.009	0.008	0.009	0.009	0.014	0.010	0.010	0.009	0.009	0.010	0.010	0.009	0.014											
db	0.017		0.01	0.019				0.011	0.009	0.011	0.009	0.007	0.01	0.009	0.012	0.012	0.009	0.012	0.009	0.012	0.009	0.012	0.008											
db	0.017	0.008	0.01	0.02				0.011	0.011	0.009	0.011	0.009	0.01	0.008	0.013	0.008	0.008	0.008	0.008	0.008	0.008	0.013	0.014	0.014	0.014	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	
db	0.020	0.008	0.01	0.023				0.01	0.011	0.009	0.008	0.008	0.01	0.012	0.013	0.013	0.013	0.013	0.008	0.008	0.009	0.009	0.011											
erste	0.015		0.01	0.02				0.009	0.009	0.009	0.009	0.009	0.013	0.014	0.011	0.011	0.011	0.011	0.008	0.008	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	
erste	0.016	0.008	0.01	0.008	0.014			0.009	0.009	0.009	0.009	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.008	0.008	0.008	0.008	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	
erste	0.017		0.01	0.015				0.009	0.009	0.009	0.009	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.016		0.01	0.008	0.013			0.009	0.009	0.009	0.008	0.011	0.012	0.012	0.012	0.012	0.012	0.012	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.016		0.01	0.014				0.008	0.008	0.008	0.008	0.008	0.013	0.013	0.012	0.012	0.012	0.012	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.017		0.01	0.016				0.008	0.008	0.008	0.008	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.018	0.007		0.014				0.008	0.008	0.008	0.008	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.018	0.007		0.014				0.008	0.008	0.008	0.008	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.018	0.006	0.006	0.013				0.006	0.008	0.008	0.008	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.019		0.01	0.015				0.007	0.007	0.007	0.007	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.019		0.01	0.015				0.007	0.007	0.007	0.007	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.022		0.01	0.015				0.008	0.008	0.008	0.008	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021		0.01	0.015				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
erste	0.021	0.007		0.016				0.007	0.007	0.007	0.006	0.011	0.011	0.01																				

		TOPIC SCORES - US - NAMED																														
business	risk	investment	market	management	fund	capital	interest	total	security	loan	net	less	financial	asset	value	include	credit	fair	change	citis	year	firm	derivative	rate	income	company	percent	mortgage				
bny								0.009	0.013			0.010	0.012	0.014	0.009	0.013	0.009															
bny	0.009	0.014			0.008		0.009	0.010	0.013			0.009	0.011	0.013	0.009	0.008																
bny	0.009	0.012	0.009				0.010	0.010	0.010		0.009	0.010	0.011	0.013	0.009	0.008																
bny	0.009	0.009		0.009			0.011	0.012	0.011		0.008	0.010	0.012	0.015	0.012	0.014																
bny	0.008	0.013					0.011	0.011	0.011	0.008	0.008	0.010	0.013	0.015	0.010	0.010																
bny	0.010	0.010	0.010				0.009	0.011			0.009	0.009	0.009	0.015	0.009	0.013																
bny	0.010	0.013			0.009		0.010	0.010			0.009	0.010	0.012	0.017	0.009	0.009																
bny	0.009	0.011	0.009				0.009	0.010	0.010		0.009	0.009	0.013	0.012	0.010	0.010																
bny	0.010	0.014					0.010	0.011	0.011		0.009	0.011	0.012	0.011	0.010	0.01																
bny	0.010	0.011			0.008		0.011	0.011	0.011		0.010	0.010	0.013	0.012	0.011	0.010	0.011															
citi							0.009	0.010	0.010	0.013	0.011	0.010	0.010	0.012	0.011	0.011	0.011															
citi	0.010						0.009	0.010			0.018	0.009	0.010	0.013	0.010	0.010	0.010			0.009												
citi							0.010	0.010	0.009	0.009	0.009	0.011		0.011	0.014	0.010	0.010															
citi	0.009						0.009	0.010	0.015	0.01	0.011		0.010	0.010	0.010	0.010	0.010			0.009												
citi	0.010						0.008	0.010	0.014	0.008	0.013		0.013	0.009	0.009	0.012																
citi	0.011						0.009	0.011	0.011	0.009	0.011	0.009	0.012	0.013	0.012	0.011	0.011															
citi							0.008	0.010	0.009	0.01	0.011	0.013	0.009	0.010	0.012	0.012	0.012															
citi							0.009	0.010	0.011	0.017	0.01	0.012	0.011	0.011	0.011	0.014																
citi							0.011	0.010	0.008	0.013	0.009	0.009	0.010	0.012	0.012	0.012	0.012															
citi	0.010						0.010	0.010	0.013	0.011	0.009	0.011	0.010	0.011	0.011	0.013				0.008												
jpm	0.011						0.009	0.009	0.018	0.008	0.008	0.008	0.009	0.008	0.008	0.011						0.019										
jpm	0.010						0.009	0.009	0.019	0.009	0.011		0.010	0.012	0.010	0.013						0.009	0.017									
jpm	0.011						0.009	0.009	0.016	0.009	0.009	0.009	0.010	0.012	0.010	0.014						0.009	0.025									
jpm	0.011						0.009	0.022	0.009	0.009	0.022	0.009	0.011	0.010	0.010	0.016						0.017										
jpm	0.013						0.016	0.016	0.016	0.016	0.011	0.010	0.012	0.010	0.011	0.014						0.019										
jpm	0.015						0.014	0.014	0.014	0.014	0.010	0.010	0.008	0.008	0.010	0.015						0.018	0.008									
jpm	0.014						0.008	0.008	0.02	0.009	0.009	0.009	0.010	0.009	0.010	0.012						0.008	0.026									
usb							0.008	0.008	0.010	0.018	0.008	0.008	0.008	0.008	0.008	0.010	0.015															
usb							0.009	0.010	0.010	0.02	0.009	0.009	0.010	0.009	0.010	0.012																
usb							0.010	0.010	0.010	0.021	0.009	0.012	0.010	0.011	0.010	0.013																
usb							0.01	0.01	0.017	0.017	0.009	0.009	0.010	0.010	0.010	0.010	0.010															
usb							0.013	0.01	0.02	0.01	0.012	0.012	0.010	0.011	0.010	0.015																
wf							0.009	0.009	0.011	0.022	0.01	0.012	0.012	0.012	0.012	0.010	0.009															
wf	0.010						0.009	0.009	0.01	0.022	0.009	0.009	0.011	0.011	0.016	0.009																
wf							0.010	0.009	0.012	0.022	0.009	0.009	0.011	0.012	0.012	0.012	0.010															
wf							0.011	0.01	0.011	0.02	0.012	0.011	0.010	0.010	0.012	0.009																
wf							0.010	0.009	0.01	0.022	0.010	0.010	0.010	0.014	0.011	0.011	0.011															
wf	0.011						0.01	0.024	0.01	0.024	0.011	0.011	0.010	0.009	0.010	0.010	0.011															
wf							0.011	0.010	0.012	0.018	0.011	0.010	0.010	0.013	0.009	0.013	0.011															
wf							0.010	0.010	0.010	0.02	0.011	0.010	0.012	0.012	0.012	0.011	0.011															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.010	0.010	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012	0.010	0.026	0.010	0.012	0.014	0.012	0.012	0.012	0.012															
wf							0.012	0.012																								

TOPIC SCORES - CANADA - UNNAMED																							
	total	income	interest	asset	year	risk	loss	security	management	value	include	rate	loan	net	credit	financial	share	market	business	capital	bank	instrument	
bmo		0.009	0.009	0.012	0.009	0.017	0.008			0.010				0.010	0.010	0.010							
bmo	0.009	0.011		0.011	0.010	0.016	0.008	0.009			0.010	0.01			0.013	0.009							
bmo		0.009		0.011	0.010	0.016	0.008	0.008					0.009	0.011	0.012	0.010							
bmo		0.011	0.009	0.009		0.012				0.010		0.01		0.013	0.011	0.011							
bmo	0.009	0.011	0.009	0.010	0.013	0.017			0.009		0.008		0.008	0.008	0.011								
bmo		0.010	0.009	0.010	0.009	0.017	0.012						0.009	0.010	0.009	0.009							
bmo		0.011	0.011	0.011		0.012	0.010			0.009		0.01			0.013	0.009							
cibc		0.012		0.009		0.016	0.010			0.012	0.011	0.01		0.011	0.013	0.01							
cibc	0.009	0.010	0.010	0.010	0.008	0.019		0.008		0.010	0.009			0.015	0.014								
cibc		0.010		0.011	0.010	0.024				0.009	0.010				0.013	0.011		0.008	0.009				
cibc	0.009	0.009		0.015		0.022	0.009			0.011	0.010	0.01		0.014	0.013	0.01							
cibc	0.009	0.009		0.011	0.009	0.019				0.009	0.011		0.009	0.012	0.009								
cibc	0.009	0.008	0.011		0.020	0.012				0.009			0.009	0.016	0.011				0.011				
cibc					0.011	0.019		0.009		0.009	0.011	0.01		0.009	0.015	0.011			0.010				
cibc	0.009			0.011	0.019	0.019	0.009			0.009		0.010		0.009	0.011	0.011		0.009					
cibc	0.010	0.009	0.009	0.011	0.009	0.016	0.009			0.010	0.008			0.013	0.015								
nbc	0.016			0.011	0.009	0.020	0.011			0.012					0.021	0.021		0.010					
nbc	0.012			0.011		0.016	0.011			0.011			0.010	0.012	0.019								
nbc	0.012			0.015		0.020			0.009	0.010			0.009	0.010	0.012	0.024							
nbc	0.009	0.009	0.009	0.013		0.022	0.011	0.008		0.010				0.010	0.019								
nbc	0.012			0.012		0.016	0.010			0.012				0.009	0.015	0.022							
nbc	0.016			0.014		0.021	0.012	0.009		0.008				0.008	0.010	0.016							
nbc	0.011			0.013		0.016	0.010			0.009			0.011	0.014	0.019			0.009					
rbc				0.014		0.019	0.010			0.009	0.010			0.010	0.013	0.010		0.009	0.011				
rbc	0.009	0.011		0.009		0.013	0.009		0.009	0.012				0.009	0.014				0.009				
rbc	0.010			0.010		0.013				0.010	0.010	0.01			0.010	0.010		0.008	0.011				
rbc	0.009	0.009	0.013			0.016	0.008			0.011		0.01		0.010	0.015			0.008	0.008				
rbc	0.009			0.013		0.015				0.013	0.011	0.01	0.008		0.011	0.010			0.008				
rbc	0.009	0.009	0.016	0.016	0.010	0.016	0.008			0.010				0.008	0.014	0.013			0.011				
rbc		0.011		0.010		0.016				0.009	0.009		0.009	0.010	0.011	0.012		0.009					
scotia	0.010	0.014		0.010	0.011	0.018	0.011			0.009			0.008	0.014	0.011								
scotia		0.012		0.010	0.010	0.016	0.010			0.010		0.01		0.013	0.011	0.012							
scotia	0.017			0.014	0.010	0.015	0.009	0.009					0.009	0.010	0.014	0.013		0.009					
scotia		0.015	0.010	0.014		0.016				0.008	0.010			0.012	0.013	0.012							
scotia		0.013		0.014	0.012	0.022	0.010			0.012			0.011	0.009	0.014	0.011							
scotia	0.013	0.009	0.011		0.017	0.009	0.009		0.009	0.010			0.011	0.010	0.014	0.014							
scotia	0.014			0.014	0.010	0.018	0.009			0.010			0.011	0.010	0.013	0.012							
scotia	0.013			0.015	0.009	0.020	0.011			0.009			0.011	0.015	0.014	0.014							
scotia	0.013	0.013	0.009	0.015	0.009	0.020	0.011			0.009			0.011	0.015	0.015	0.014	0.01						
scotia	0.010	0.013	0.009	0.012	0.011	0.020	0.010			0.009				0.009	0.012	0.013							

Figure J.4: Topic scores for each feature for Canada “Unnamed”

TOPIC SCORES - EUROPE - UNNAMED																																	
	asset	financial	share	include	report	rate	group	business	year	credit	loan	interest	risk	income	total	value	management	fair	market	net	loss	capital	eur	customer	amount	statement	bank	company	client	annual	service	corporate	
barclays	0.009	0.012		0.010	0.009	0.007	0.013	0.007	0.007		0.008	0.016				0.008																	
barclays	0.012	0.012	0.009	0.009	0.008	0.008	0.014	0.009	0.007			0.019																					
barclays	0.009	0.012	0.009	0.009	0.008		0.014	0.008	0.009			0.019				0.008																	
barclays	0.012	0.009		0.009	0.007		0.012	0.007	0.008	0.007		0.019		0.007																			
barclays	0.009	0.012	0.009	0.008	0.009		0.013	0.010	0.009			0.019		0.007																			
barclays	0.011	0.009	0.008		0.008		0.012	0.011	0.007			0.007		0.007																			
barclays	0.011	0.009	0.008	0.007	0.008	0.007	0.015	0.008	0.008			0.018		0.007																			
db	0.015	0.016		0.007	0.011					0.009		0.024				0.012	0.008	0.009			0.009												
db	0.015	0.018			0.011					0.010		0.020		0.009		0.010	0.009			0.012	0.008												
db	0.012	0.016			0.010			0.010	0.009			0.022				0.012	0.008			0.008	0.008												
db	0.014	0.018						0.020		0.008		0.015		0.008		0.012	0.008	0.008	0.008	0.013	0.009												
db	0.014	0.017			0.010			0.010	0.009	0.007		0.015		0.007		0.012	0.009	0.009	0.009	0.01	0.009												
db	0.013	0.017			0.010			0.020		0.008		0.019		0.008		0.012	0.007	0.008	0.008	0.009	0.008												
db	0.012	0.020		0.008	0.011					0.008		0.020		0.008		0.01	0.009			0.008	0.008												
db	0.013	0.019		0.008	0.009			0.008	0.008	0.008		0.020		0.008		0.011			0.009	0.011	0.009												
db	0.014	0.021			0.009			0.009	0.009	0.009		0.022				0.010	0.008		0.008	0.009	0.011	0.009											
db	0.013	0.018			0.010			0.010		0.010		0.018				0.011	0.008	0.008	0.008	0.010	0.010												
db	0.013	0.019			0.009			0.009		0.008		0.020				0.009	0.009	0.009	0.009	0.009	0.008												
db	0.016	0.016			0.010					0.008		0.015		0.008		0.008	0.009		0.008	0.009	0.008		0.010										
erste	0.015	0.015					0.011	0.009	0.009			0.017		0.008					0.008	0.009	0.008		0.009										
erste	0.010	0.016	0.008				0.013	0.008	0.008			0.012							0.008	0.009	0.008		0.009										
erste	0.011	0.016					0.017	0.008	0.009			0.015							0.008	0.009	0.008		0.009										
erste	0.012	0.016					0.016	0.008	0.008			0.012							0.008	0.009	0.008		0.009										
erste	0.010	0.016					0.018	0.008	0.008			0.015							0.008	0.009	0.008		0.011										
rabobank	0.009	0.023			0.015	0.006						0.018		0.008		0.008	0.008		0.008	0.008													
rabobank	0.008	0.021		0.006	0.012							0.016		0.006		0.008	0.010		0.010	0.011													
rabobank	0.008	0.018			0.011							0.007		0.007		0.008	0.007		0.007	0.011													
rabobank	0.008	0.023			0.013							0.007		0.007		0.008	0.007		0.007	0.010													
rabobank	0.009	0.021			0.011							0.008		0.006		0.009	0.008		0.008	0.010													
rabobank	0.012	0.019			0.012							0.008		0.008		0.009	0.008		0.008	0.015													
rabobank	0.012	0.018		0.007	0.011							0.007		0.007		0.009	0.008		0.007	0.012													
sberbank	0.007	0.010			0.009			0.007	0.007			0.011				0.007	0.008		0.008	0.011													
sberbank	0.011				0.010				0.007			0.011				0.008	0.008		0.008	0.009	0.008		0.009										
sberbank	0.006	0.008			0.006			0.006				0.010				0.007	0.007		0.007	0.008													
sberbank	0.008	0.010			0.008			0.008		0.007		0.010				0.007	0.007		0.007	0.008													
sberbank	0.007	0.009			0.008			0.007	0.007			0.015				0.006	0.006		0.006	0.015													
sberbank	0.011				0.008			0.007	0.007			0.010				0.008	0.008		0.008	0.010													
sberbank	0.009				0.007			0.008				0.015				0.006	0.006		0.006	0.010													
sberbank		0.009			0.007			0.008				0.015				0.006	0.006		0.006	0.010													
sberbank		0.009			0.007			0.008				0.015				0.006	0.006		0.006	0.010													

Figure J.5: Topic scores for each feature for Europe “Unnamed”

TOPIC SCORES - US - UNNAMED																												
asset	loss	credit	fund	market	total	financial	value	net	result	security	investment	service	capital	include	risk	interest	business	loan	change	rate	report	firm	year	income	fair	company	percent	mortgage
bny	0.015	0.010			0.010	0.013		0.009		0.010	0.014		0.009	0.010	0.011													
bny	0.016	0.010	0.008		0.014	0.011		0.009		0.009	0.010		0.009	0.011	0.009													
bny	0.013	0.008			0.009	0.009	0.010	0.010		0.012	0.013		0.009	0.010	0.010													
bny	0.011	0.009	0.010		0.010	0.009	0.009	0.009		0.011	0.011		0.009	0.010	0.012													
bny	0.012	0.009		0.009	0.012	0.011	0.013	0.009	0.008	0.013	0.010		0.009	0.011	0.011													
bny	0.014				0.010	0.011	0.009			0.009	0.012			0.010	0.010	0.009												
bny	0.015	0.012	0.008		0.012	0.010		0.010		0.010	0.011			0.010	0.010													
bny	0.016	0.011			0.009	0.012	0.010			0.014	0.011			0.011	0.009	0.010												
bny	0.013	0.010			0.009	0.010	0.010	0.011		0.009	0.011			0.012	0.009	0.010												
bny	0.013	0.009			0.009	0.012	0.013	0.010		0.011	0.013			0.010	0.009	0.009												
bny	0.014	0.009		0.008	0.009	0.009	0.010			0.011	0.009			0.013	0.011	0.009												
cit	0.013	0.011			0.011		0.010	0.010		0.010				0.012	0.009													
cit	0.011	0.010	0.011		0.009		0.012	0.011		0.009				0.009	0.010													
cit	0.010	0.010	0.014		0.010	0.010	0.011	0.011		0.010	0.010			0.012	0.010													
cit	0.011	0.011	0.010		0.010	0.010	0.010	0.010		0.010				0.011	0.010													
cit	0.012	0.010	0.012		0.009		0.011	0.009		0.008				0.010	0.009	0.009												
jpm	0.012	0.017			0.009		0.011	0.010		0.008				0.008	0.012						0.008		0.019					
jpm	0.012	0.009	0.013		0.009	0.009	0.011	0.009		0.008	0.010			0.011	0.010							0.020						
jpm	0.011	0.009	0.009		0.008		0.011			0.008				0.012	0.011							0.017						
jpm	0.012	0.014			0.008	0.009	0.008			0.009				0.011	0.011							0.022						
jpm	0.009	0.013			0.009		0.008	0.008		0.008				0.011	0.011							0.024						
jpm	0.010	0.009	0.015		0.009	0.009	0.011	0.009		0.009	0.011			0.011	0.012	0.009					0.008	0.024						
jpm	0.011	0.008	0.012		0.008		0.011			0.009	0.008			0.011	0.011							0.021						
jpm	0.011	0.009	0.014		0.009		0.011	0.008		0.008				0.009	0.008	0.008						0.023	0.008					
jpm	0.010	0.010	0.015		0.009	0.010	0.010	0.008		0.009	0.010			0.011	0.012	0.009						0.024						
jpm	0.010	0.010	0.010		0.009	0.010	0.013			0.015				0.010	0.013							0.018	0.010					
usb	0.010	0.010			0.009	0.010	0.010	0.008		0.009				0.011	0.014	0.008						0.020	0.010					
usb	0.010	0.010	0.010		0.009	0.009	0.009	0.008		0.015				0.012	0.013							0.016	0.010					
usb	0.010	0.010	0.010		0.009	0.009	0.009	0.009		0.020				0.012	0.013							0.015	0.010					
usb	0.014	0.012	0.011		0.009	0.009	0.011	0.009		0.019	0.009			0.009	0.009							0.019	0.009					
usb	0.011	0.010	0.012		0.009	0.009	0.009	0.009		0.025				0.011	0.011							0.025	0.009					
usb	0.011	0.010	0.013		0.010	0.009	0.009	0.010		0.019				0.010	0.010							0.019	0.009					
usb	0.009	0.009	0.014		0.010	0.009	0.009	0.010		0.018				0.012	0.012							0.018	0.010					
usb	0.010	0.011			0.010	0.009	0.011			0.015				0.011	0.011							0.015	0.010					
usb	0.009	0.012	0.013		0.010	0.009	0.011	0.009		0.022				0.009	0.010	0.009						0.015	0.010					
usb	0.013	0.010	0.012		0.009	0.009	0.009	0.009		0.016				0.010	0.009	0.009						0.016	0.009					
usb	0.010	0.013	0.012		0.010		0.010	0.009		0.017				0.010	0.011							0.017	0.009					
usb	0.009	0.009	0.011		0.010	0.009	0.010	0.009		0.017				0.009	0.009							0.017	0.009					
wf	0.012	0.010	0.011		0.012	0.013	0.009	0.009		0.022				0.009	0.010	0.009						0.019	0.010					
wf	0.010	0.010			0.010	0.010	0.014	0.010		0.011	0.011			0.010	0.010							0.023	0.010					
wf	0.012	0.011	0.011		0.011	0.010	0.011	0.010		0.011	0.011			0.013	0.013							0.023	0.010					
wf	0.010	0.010			0.009	0.012	0.012	0.012		0.013				0.010	0.009	0.010						0.023	0.010					
wf	0.011	0.009			0.009	0.012	0.012	0.012		0.012				0.012	0.012							0.010	0.010					
wf	0.013	0.011	0.011		0.011		0.010	0.010		0.011				0.011	0.011							0.022	0.011					

Figure J.6: Topic scores for each feature for the United States “Unnamed”