A STUDY OF THE RELATIONSHIP BETWEEN THE GEOGRAPHIC LOCATIONS OF THE USER AND PARTICIPATION IN TWITTER DURING DIFFERENT TYPES OF NEWS EVENTS

by

Ghada Amoudi

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

at

Dalhousie University
Halifax, Nova Scotia
December 2016

Dedication

To my parent, and to my husband, their love and support guided my way to finish this work.

Table of Contents

List of Tables

List of Tables

List of Tables

List of Tables

List of Tables

List of Figures

List of Figures

List of Figures

Abstract

Twitter is one of the most active social networks in news sharing. People report local events sometimes faster than news agencies. During major events, such as earthquakes and presidential elections, people share tweets, retweets, images and links related to the event, creating an overwhelming number of posts. The amount of data generated by social media provides a powerful tool for knowledge discovery for individuals, organizations and governments. However, the large stream of tweets makes tracking interesting posts a challenging task. Understanding user behavior during different events provides substantial knowledge to get the most out of the social media power. A twitter post encapsulates several fields about the tweet and the users, such as posting date, users locations, time zones, and geographic coordinates. The main aim of this work is to investigate the relationship between user participation, news type and geographic locations of users in Twitter. To achieve this goal, posts related to a certain event were retrieved by keyword search, and geographic information was obtained from users' profiles, then posts were analyzed by news type and geographic location. The results showed that finance news tweets had distinct user behavior, finance tweets have more original tweets, more links and less hashtags, while politics tweets had the largest number of hashtags, compared to disaster and finance tweets. The investigation of the relationship between users' participation and news type would provide guidelines for social media management and advertisement, similarly the analysis of the relationship between the participation and users' locations provides insight for information diffusion research and location-aware news recommendation systems.

List of Abbreviations and Symbols Used

API	Application Program Interface
CSV	Comma Separated Value
DF	Degrees of Freedom
DNLP	Dalhousie Natural Language Processing
ETL	Extract Transfer Load
IE	Information Extraction
JSON	JavaScript Object Notation
Logit	Log Odds
NB	Naïve Bayes
p-value	A statistical indicator for accepting or rejecting the null hypothesis
SE	Standars Error
SNA	Social Network Analysis
SQL	Structured Query Language
SVM	Support Vector Machine
URL	Uniform Resource Locator
UTC	Coordinated Universal Time
Zipf	A statistical distribution found by George Kingsley Zipf

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Carolyn Watters for her countinouse support, help and pateince. I would like to thank my supervisory committee, Prof. Michael Shepherd, Prof. Vlado Keselj, and Dr. Bonnie Mackay for their insightful comments and encouragement. My sincere thanks also goes to Prof. Bruce Smith for his help in statistics.

Chapter 1 Introduction

News sharing and reporting is one of the most popular social media activities. Twitter, with 340 million users generating 6000 tweets per second on average (Oliveira, 2016). Kwak, Lee, Park and Moon (2010) found that over 85% of topics shared in Twitter are headline news or comments about news. Users who are distributing and creating news are known as citizen journalists (Bruns, 2010). People report news sometimes before official news channels; for example, the US Airways plane crash that took place in 2009 was reported in Twitter by pictures before any local news agency appeared in the scene (Choi & Kim, 2013). Additionally, the retweet feature provided by Twitter, i.e., the process of reposting messages of other users increases traffic considerably. Kwak et al. (2010) found that on average messages get 1000 retweets, even when the number of followers of the original message poster is low.

Choi and Kim (2013) analyzed Twitter data for tracking breaking news and showed that Twitter can be used for tracing news over the world. Additionally, due to the large number of users accessing social networks from their smart phones, information related to the location of the users is gaining increased importance in social networks (Lima & Musolesi 2012).

1.1 Twitter Fields and Characteristics

A tweet, is a message that consists of up to 140 characters written by a registered user of the Twitter service. Retweeting is the act of reposting a message, so that this message will be posted in the retweeter's timeline and to her followers. Retweets are one of the main propagation mechanisms in Twitter (Wu & Chen, 2015). A user can retweet

any message even if it is not from followers or friends, all of her followers will then receive the retweeted message and they can retweet it again creating a path of retweets.

The limited length of messages causes people to adopt different strategies. People often share links to full articles, websites, pictures, and videos, and use abbreviations and incomplete sentences. Users can also share hashtags in their posts, where a hashtag is text preceded by the symbol '#', and used to indicate a topic. For instance, when the hashtag #London is included in a message it is assumed that the message contents are about London or something related to London. Hashtags are used to organize topics, create discussion about an event and to promote ideas or products.

1.2 News

News has been defined as "information or a report about something that had happened recently in a newspaper, magazine, or television news program" (Merriam Webster Online Dictionary, 2016). With the advances of digital media, most well-known news agencies are now available online on the form of Websites, Twitter accounts and Facebook pages. BBC, CNN, Reuters, and New York Times are examples of news agencies, which publish news online and on Twitter (Bhattacharya & Ram, 2012). Other sources of news online include Yahoo news, and Wikipedia current events, which are news portals that list events, trends and developments with links to full articles. News also now includes events not yet reported formally, such as the US Airways plane crash example mentioned earlier.

1.3 Problem Definition

Social media sites, such as Twitter, experience heavy traffic during major events, e.g., earthquakes and presidential elections, as people tweet, retweet and reply to each

other. This makes tracking important posts from around the world a challenging task (Stanger, Thomee, Popescu, Pennacchiotti & Jaimes, 2013). Additionally, an individual may have many social media accounts, for instance an account in Twitter, Facebook, Snapchat, or Pinterest. The vast amounts of data generated by these applications continually, forces the user to randomly pick which to read, thus, important information could be overlooked.

While access to news is now global, the influence of local differences remains and may introduce bias to the reporting. Social media allows people to follow like-minded friends from around the globe, however, the desire to learn about local events is still important in some cases (Takhteyev, Gruzd & Wellman, 2011). Twitter includes several fields and information about the poster, such as user location, time zone, and geographic coordinate, which are not directly accessible by the ordinary user. Twitter users can also identify their location by specifying their user location, geotagging the tweet location, or attaching an event place name to a tweet.

People also use tweets to share links to sites, articles, videos, and images (Nizam, Watters & Gruzd, 2014). Cao and Caverlee, (2014) stated that 25% – 29% of tweets posted daily contain links. By analyzing URL sharing patterns it may be possible to identify the links shared in a certain geographical region and which may be of interest to other users in that region. Additionally, link analysis may provide potential applications in recommendation, advertising and detecting a spam (Cao & Caverlee, 2014).

Hashtags are one of the most used features in Twitter, it is estimated that 25% of tweets contain hashtags (Shi, Ifrim & Hurley, 2016). Hashtags are used to connect people, organize topics, propagate ideas and promote specific topics (Maity, Saraf & Mukherjee,

2016). News organizations use hashtags to promote new content and to engage people (Shi et al., 2016). Including hashtags in a post increases its chance to be seen by communities of interested users. Analyzing hashtags provides potential value in modeling user behavior (Bogdanov, Busch, Moehlis, Singh & Szymanski, 2014), and generating recommendations (Shi et al., 2016).

The abundance of user generated contents provided by Twitter poses a challenge to find quality news and information. Employing geographic features available in Twitter for developing location-aware news services provides two possible advantages. First, allowing users to view news-related tweets by location, and second, reducing the amounts of tweets, thus providing a filtering mechanism. Yang, Ghoting, Ruan and Parthasarathy (2012) indicated that interactive querying and analytics could be developed that would enhance the reading experience of Twitter posts. That is, given the scale and the immediacy of Twitter in news reporting it has become necessary to investigate different ways for viewing and tracking tweets related to major world events. This investigation could facilitate creating location-aware applications, predict news popularity, and provide real-time filtering.

1.4 Objectives

The main objective of this thesis is to understand the relationship between geographic location of users and the characteristics of their participation in Twitter for news related tweets. We examine this over three different types of news, specifically finance, disaster and world politics. The objectives can be summarized as follows:

1. To examine the following behavioral characteristics of users tweeting about these news types: generation of original tweets, retweets, hashtags and links

2. To develop an overall model by country, which could be used by users and developers to profile the expected geographic patterns of twitter behavior for different types of news.

1.5 Research Questions

The main research question is how the behavior of participants using Twitter for news reporting and commenting is related to geographic location and the type of news. Three types of news are considered in this research, finance, disaster and politics. This question can be divided into three sub question as follows:

RQ1: Is there a relationship between country and the type of user participation in Twitter?

1- Is there a relationship between country and the generation of original tweets or retweets?

2- Is there a relationship between country and the use of hashtags?

3- Is there a relationship between country and the use of links?

RQ2: Is there a relationship between news type and user participation in Twitter?

1- Is there a relationship between news type and the generation of original tweets to retweets?

2- Is there a relationship between news type and the use of hashtags?

3- Is there a relationship between news type and the use of links?

RQ3: Is there a relationship between <country, news type> and user participation in Twitter? In other words, Is there an interaction between country and news type? And if interaction exists, how does it influence the user participation in Twitter?

1.6 Outline of Thesis

The thesis is organized as follows, chapter 2 presents background and related research in the areas of news sharing in Twitter, social network analysis and modeling user behavior. Next, chapter 3 presents an exploratory study conducted to gain general knowledge about Twitter and sports news. The aim of this study was to analyze the basic characteristics of two sports datasets, then these characteristics are compared and articulated, along with the results and motivation for the next phase. Chapter 4 explains the research methodology as well as the details of data collection and analysis. Next, chapter 5 introduces the study results and findings. Chapter 6 discusses the results and explains the limitations. Lastly, chapter 7 summarizes the research, and presents contributions, implications of research findings and future work.

Chapter 2 Background

Twitter research covers a broad spectrum of topics, ranging from popular areas such as event discovery and network analysis to novel ones such as mental depression detection (Wang et al., 2013). As background, we investigated three major research areas, that are relevant to this thesis, news sharing in Twitter, social network analysis, and modeling user behavior. Research concerning news sharing provides answers to questions such as how fast and how far a news story travels. Geographic characteristics of social media users have been extensively researched due to the wide use of smart phones, which allows sharing location information. User behavior research aims at modeling the behavior of users in order to predict influence, recommend friends and build user profiles. In the following subsection we present the literature related to these research areas.

2.1 News Sharing in Twitter

News is among the most common content shared in Twitter. Twitter is ranked second in news discussions after Facebook (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016). According to Subasic and Berendt (2011) news sharing is among the most important reasons for using Twitter, and due to the widely-adopted retweet feature, and the use of hashtags, news spreads in Twitter in a very fast rate. Subasic and Berendt (2011) found that users' roles in Twitter focus on reporting news and commenting on news. While the types of news posted in Twitter vary from serious news events to movie and celebrities' news, in this research we focus on finance, disaster and politics news.

2.1.1 Financial news. The increasing activity of news reporting in social media has influenced the financial investment community. Researchers have investigated the relationship between social media activity and market prices to improve investment

strategies (Yang & Mo, 2016). Karkulahti, Pivovarova, Du, Kangasharju, and Yang Arber (2016) studied the relationship between news, social media and stock prices. They measured social media presence by counting Wikipedia page hits. The study used online business news sources and applied Information Extraction (IE) to online news articles, and then used the extracted entities to compose queries to fetch information from Twitter using the product or company name. The study presented a visual analysis of correspondence between Wikipedia views, news and stock prices. The results showed a correlation between the mentions of the company in the online news source and the views of the company's Wikipedia page and a less obvious correlation was found with stock prices. Ruan, Alfantoukh and Durresi (2015) used Twitter data to monitor stock price change and trade volume. They created a trust network of Twitter users, using the followers of the Financial Times Twitter account. The trust network was built by using the interaction between these users and the sentiment in their messages. The study found that trade volume was more correlated than the daily price change with tweets, and the trust network approach improved predicting trade volume.

2.1.2 Disaster news. Disaster news shared in social media gained special attention from researchers due to the potential of social media in assisting affected people during these disastrous situations. The real-time nature of these events in addition to the wide use of Twitter has contributed to the fast spreading of news events. The advances of smart phones and the availability of location services, allow eye witnesses to share pictures from the physical locations of these events, and to report the locations of affected people. Stollberg and Groeve (2012) presented a Twitter parser to capture and analyze tweets to support the work of the Global Disaster Alert and Coordination System

(GDACS). Information gathered from Twitter about disasters has been shown to aid in identifying side effects such as collapsing buildings. Additionally, the fast response of Twitter users affected by the disaster, which sometimes come before news media, allows emergency responders to act promptly, according to Stollberg and Groeve (2012). Their study analyzed Twitter data containing the word "earthquake" from the end of October 2011 until the middle of January 2012. They found three peaks of heavy activity coinciding with three earthquakes that happened in that period. Toriumi and Baba (2016) devised a real-time tweet clustering method to find and classify tweets quickly in disaster situations. The method was based on retweets using network clustering methods, and the results were clustered by time and by keywords. The dataset used for the study consists of approximately 30 million tweets posted from March 5th to 24th of 2011. This period included the Great Eastern Japan Earthquake that took place on March 11th, 2011. The results showed that the system could cluster tweets by victims' requirements, such as shelters, rescues and train schedules.

2.1.3 Politics news. Soon after Twitter's launch in 2006, political activists started using Twitter to support or criticize politics around the world. For example, Obama's presidential campaign in 2008, the Iranian presidential election in 2009 and during the Arab Spring in 2010 and 2011 (Jungherr, 2015). Some research in the political domain has been focused on the idea of using Twitter data to investigate political debates, using frameworks for sentiment analysis and friend recommendations. Lai, Bosco, Patti and Virone (2015) used Twitter data to analyze debates about political reform in France, where public opinion was used to lead change and affect decisions. The research goal was to investigate the different aspects of communication in Twitter, which would be used for

sentiment analysis for French tweets. To collect the data, they used the main hashtag/slogan of the political campaign from December 2010 to July 2013. They collected data from three different sources, Twitter, a French newspaper and parliamentary debates. The study analyzed the lifeline of the hashtag, i.e., where was the hashtag first created, how did the community accept it, and how did it spread through Twitter. They also compared public opinion extracted from Twitter to what formal media announced. They found that the hashtag frequency in tweets was influenced by the news announced and shared in other media like newspaper, and parliamentary debates.

Cuesta, Barrero and R-Moreno (2013) analyzed Spanish tweets related to the Boston terror attack gathered during one week, from April 16 to April 23, 2013. The goal of the study was to characterize the tweeting activity generated by the Spanish-speaking community. The study found that during this event the retweet frequency was high about 44%, compared to replies of 5%. From these results, they concluded that users did not use Twitter as a communication tool, but rather to broadcast messages or breaking news.

Elghazaly, Mohmoud and Hefny (2016) used Twitter data to compare the performance of two classic classifiers, Support Vector Machine (SVM), and Naïve Bayes (NB) in performing political sentiment analysis for Arabic tweets. The tweets analyzed were collected during the presidential elections in Egypt from March 1st to June 24th, 2012. The study employed supervised learning approach. They found that NB classifier performed better for the political data analyzed than SVM.

2.1.4 Sports news. Sport events are widely shared in Twitter. Fans share their comments and reactions to live games broadcast on television. Ineson and Anderson (2016) investigated the motives behind using social media in sports, and found the

motivations are: fandom, entertainment and to obtain information. The study also found that people used social media when criticizing clubs and athletes.

Seron, Zorzal, Quiles, Basgalupp and Breve (2015) conducted a descriptive statistics analysis to 8 million tweets about World Cup 2014. The analysis included the most common languages, countries, users and hashtags. The study illustrated the number of tweets generated by each country in a map as shown in Figure 2-1. The study found that English was the most common language, even though the event was in Brazil. The study reported that in the day before the opening ceremony a high activity in twitter was witnessed, where almost all countries around the world took part in the tweeting activity. The study showed that the distribution of tweets per user followed a power law distribution, i.e., most of the tweets were generated by a small number of users, while most users generated few tweets.

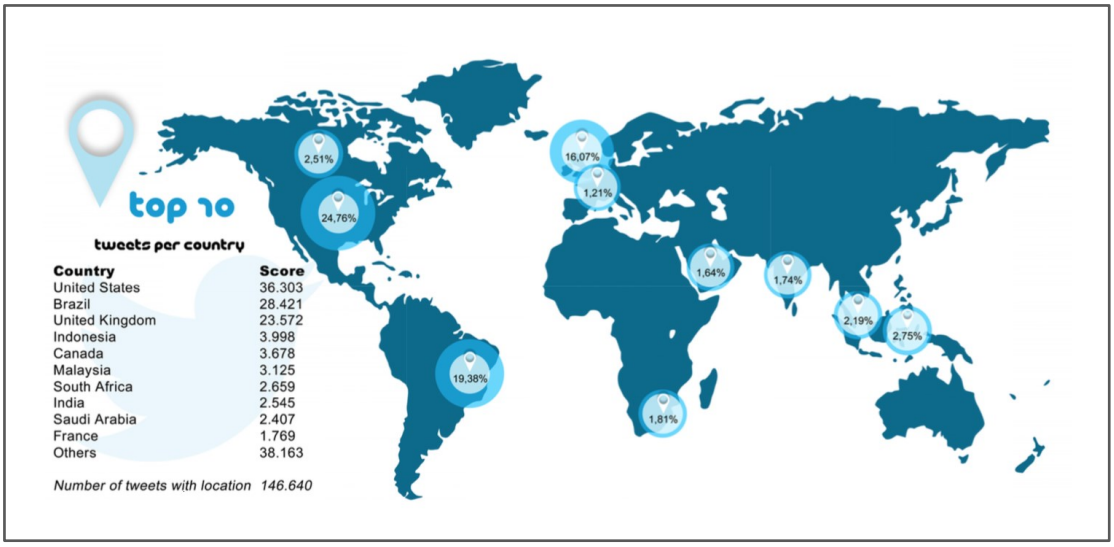

Figure 2-1 Top 10 countries tweeted about World Cup 2014 (Seron et al., 2015)

2.2 Information Propagation

2.2.1 Social network analysis. A social network such as Twitter can be modeled as a directed graph $G = (V, E)$, such that V is the set of vertices and E is the set of edges, and $E \subseteq V \times V$. A vertex or a node u represents an individual in the social network, and an arc (u, v) indicates a relationship between u and v. Information propagation or information diffusion is the process of spreading a piece of information from a user to another in a social network. Information propagation is associated with discrete time steps $t = 1, 2, \ldots n$. A node v in a social graph is either active or inactive, active indicates that the user adopted the new information, and inactive indicates that the user did not adopt the new information (Chen, Lakshmanan & Castillo, 2013). Kwak et al., (2010) investigated the topological characteristics and information diffusion in Twitter. They crawled the whole Twitter website in the period of June 6th to June 31st in 2009. They analyzed Twitter social network including, following patterns, reciprocity, users rank, homophily, and the impact of retweets on the spread of news. They found that the follower-following topology did not follow a power-law distribution, and had low reciprocity. Regarding homophily, they found that the median time zone difference between users and friends increased as the number of friends increased, i.e., users with 50 friends or less had their friends within 1.07 hours difference in time zone, while users with 5000 friends or more, have their friends within 6 hours difference.

Louni and Sabbalakshmi (2014) divided models of information propagation into three models; the contagion model, the social influence model and the social learning model. The contagion model treats information propagation in a manner similar to contagious diseases spreading among a population. The social influence model is similar

to the contagion model with a slight difference; the user receiving the information will adopt it only if the number of users in the network adopting that piece of information are above a certain threshold. In the social learning model users decide whether to adopt a piece of information based on the outcomes of prior adopters.

The spread of information in social networks has been considered also as a type of diffusion of innovation (Gruhl, Guha, Liben-Nowell and Tomkins, 2004), which is defined as the spread of new ideas, practices, and technologies among users in a network. The theory behind diffusion of innovation is based on sociology research to establish basic methods to study the spread of information within a social group, as well as the factors that facilitate or hinder the flow of information (Easley & Kleinberg, 2010). Information cascade is a model of diffusion of innovation (Chen et al., 2013). Easley and Kleinberg, (2010) stated, "information cascade has the potential to occur when people make decisions sequentially, with later people watching the actions of earlier people and from these actions inferring something about what the earlier people know". Gruhl et al., (2004) employed an information cascade model to study topic propagation in weblogs, i.e., the transmission of a topic from blog to blog based on the text of the post rather than links shared in the post. The study used a dataset of 401,021 posts, which were analyzed to find topics shared in the blog space and their paths through individuals. The topics were divided into three categories, chatter (daily normal topics exchanged between bloggers), spike, (a topic of global interest like world news), and a mix of chatter and spike. The study showed how topics in these different categories are distributed in the period of one year. The study also found that chatter and spike topics exhibited different patterns, and most spike topics lasted from 5-10 days.

2.2.2 News propagation in Twitter. Researchers have also studied how news spreads in social media and in Twitter in particular, to answer questions such as, which nodes are most influential (Gabielkov, 2016), and how messages travel in the network. Ye and Wu (2010) analyzed the propagation patterns in tweets during breaking news. They collected more than 58,000,000 tweets about Michel Jackson's death. They measured how many hops a message was propagated, how fast, and how long it lasted by constructing propagation trees where each tree consisted of a message and its replies. Each tree was considered to be a message flow, and the collection of all trees represented all the different paths the message took. They found that in normal days, 37% of the messages propagated more than three hops away from the first tweeter, 75% of replies happened within 16 minutes, and 75% of message flows lasted less than an hour. They also found a large spike on the day of the memorial service and that tweets continued steadily between day 16 and day 60, which means that breaking news sometimes lasts longer and does not vanish quickly. Trung, Jung, Lee and Kim (2013) created TweetScope to collect and monitor Twitter data in order to capture propagation patterns. The study's goal was to understand propagation patterns in order to optimize business strategies. They collected data from Twitter accounts of three major telecommunication companies in Korea over a period of 4 months, from March 16th to October 30th, 2012. One of the study's finding indicated that the most retweeted posts were for promotions of goods and services. Zhou, Bandari, Kong, Qian, and Roychowdhury (2010) studied message propagation on Twitter by analyzing message sharing related to the Iranian election in 2009. They gathered and analyzed more than three million tweets in the period of two months. They found that 99% of message flows trees had depth less than 3, and

63.7% of retweeted messages were from users the retweeters followed directly. They also found that the retweet probability depended heavily on tweet's contents. Bhattacharya and Ram (2012) studied the propagation patterns from twelve recognized news sources in Twitter, such as the BBC and the New York Times. They employed network analysis and visualization techniques to compare the twelve news sources in terms of the extent of articles spread, the rate of spread, and the life span of tweets related to articles. The study illustrated the spread using graphs, with the number of nodes, edges, diameter and edge/node ratio for each of the twelve sources. The study found that BBC had the maximum spread and longest lifespan of the twelve news sources.

2.2.3 Geographic information propagation in Twitter. The geographic properties of users in a social network have been found to be important in the discussion of a news events (Agrawal, Budak, El Abbadi, Georgiou, & Yan, 2014). The Internet and social networks allow a user to connect to other users regardless of the physical distance between them. However, distance, language, country and the number of airplane flights between the different locations have been shown to have an influence on the formation of Twitter ties (Takhteyev et al., 2011). Consequently, information propagation in Twitter is influenced by the proximity of user locations. Geographic information propagation is an emerging research area, due to the growing use of mobile devices and location aware services. Geographic information propagation researches ask questions such as "to what extent does the geographic distribution of friendship networks affect where the information will be potentially propagated?" (Lima & Musolesi, 2012).

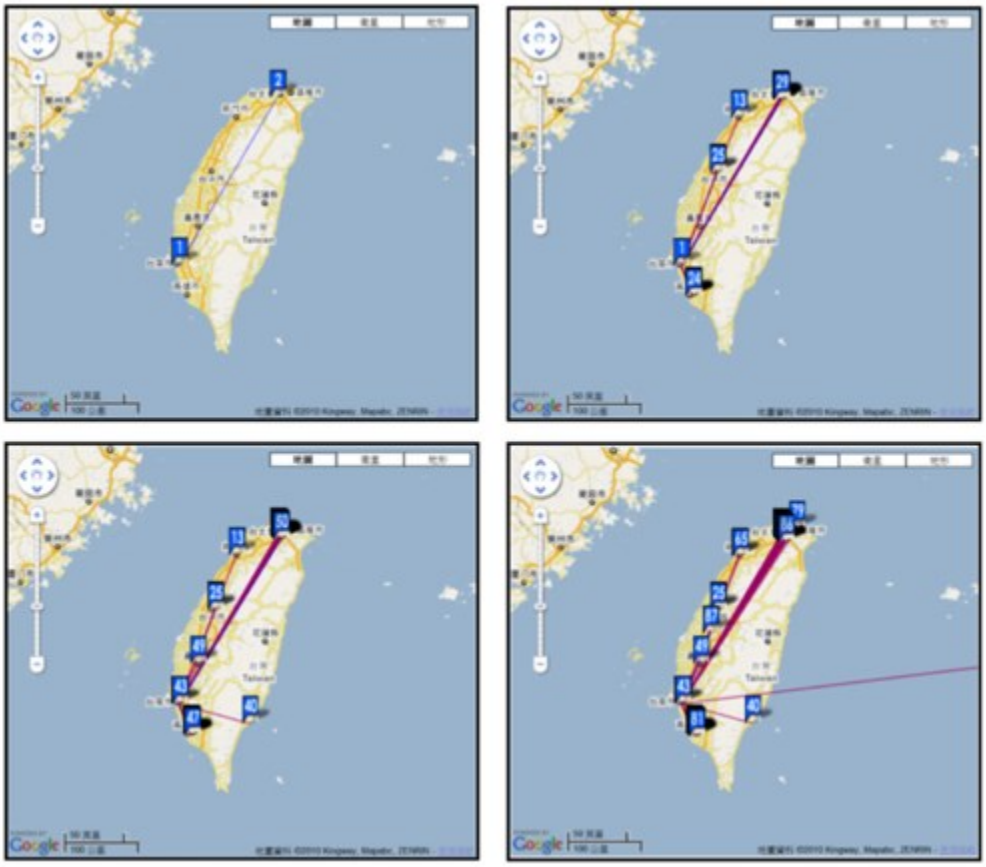

Figure 2-2 Message propagation pattern (Li et al., 2013)

Li et al. (2013) investigated and evaluated information propagation in Plurk, a microblogging service. The study considered the number of people influenced, the speed of propagation and the geographic distance of the propagation. The study illustrated how the propagation of messages were monitored and presented on a world map, Figure 2-2 presents an example, where the marks on the map appear in the order they were posted. The method used in the study depended on constructing propagation trees, with the original poster as the root of each tree and the re-posters and repliers as children nodes. They also created a framework to evaluate information diffusion models, predict propagation routes, and find influential users. Lima and Musolesi (2012) presented

information diffusion metrics to capture and measure geographic importance and centrality of users in specific geographic and social network. They evaluated their metrics using data from Twitter and Foursquare. The method used combined geographic information to traditional centrality measures, i.e., indegree, outdegree, closeness and betweenness, to develop spatial closeness centrality. These metrics were used in the study to depict spatial closeness centrality of Twitter users in London and San Francisco. Vosecky, Jiang, and Ng (2013) developed an interactive system for modeling geographic users' interests. The proposed system employed content analysis to extract geographical information and topics of interest from a Twitter user by examining the user's tweets. The system differentiated between the user's mentioned locations and visited locations, and displayed both locations on an interactive map. Finally, the system recommended three news lists for the user, the first list included top news stories from the user's region, the second list included topics of interest to the user from areas around the user's region, and the third list included news some related to the user location and some related to her interests.

Research methodologies have also been applied to investigate geographic features in social media using geographic measures and content analysis. Van Liere, (2010) defined the geographic distance a message traveled in a social network as the distance between the sender and the receiver in kilometers. He also defined geographic information diffusion pattern as "the distribution of geographic distances between the sender and the receiver". Li et al. (2013) defined geographic distance as the distance from the root (sender) to the users both directly and indirectly. To calculate the distance between two users in twitter, geographic locations were obtained from their profiles, and

their location information was geocoded, i.e., the longitude and latitude of the location (Li et al., 2013). Van Liere (2010) studied information propagation in Twitter and defined three possible propagation patterns, random, local and information brokerage. He used retweets as a measure for propagation, and showed that the random pattern had a uniform distribution, the local pattern was left skewed, i.e., the geographic distances between sender and receiver are short. The last pattern, information brokerage followed users with similar interests and this pattern was right skewed. The study concluded that a retweeter (information broker) transferred information and acted as bridge between the other two distinct groups.

2.3 Modeling User Behavior

Twitter data provides a wealth of information about user behavior during various types of events. Twitter has a unique network structure, unlike regular social networks where links are reciprocal in most of the cases. Twitter includes a large number of unidirectional ties. For instance, news media outlets and celebrities accounts are followed by thousands, but they do not necessarily follow back their followers. Thus, understanding the influence of users in such a network cannot rely solely on network structure, and the behavior of users must be considered and analyzed (Bogdanov et al., 2014).

User behavior modeling in social networks has been studied by a number of different methods. Analytic approaches dealing with user behavior modeling have included trend analysis and opinion mining among other methods. Trend analysis involves identifying how important a topic is, typically using accumulated interest in the topic, for instance counting YouTube video views (Agrawal et al., 2014). Time, location,

demographics and opinions are elements that that have been considered along with trend analysis to understand other characteristics of user behavior (Agrawal et al., 2014). User behavior modeling has been used in areas such as discovering influential users, providing recommendation (Yeung & Iwata, 2011), building users profiles (Macskassy & Matthew, 2011), and predicting new information spread (Bogdanov et al., 2014).

Combining geographic and news type dimensions with user behavior is an area that has not been adequately addressed in the literature. In this thesis, we study and model user behavior in Twitter in terms of three common behaviors, posting original tweets vs retweets, using hashtags and using links.

2.3.1 Retweet. Retweet behavior has been shown to be a prevalent message spreading mechanism. When a user retweets a post, it becomes visible to all her followers, who can in turn retweet the post and thus, people see posts of users they do not follow directly. According to Ota, Maruyama and Terada (2012) users receiving a retweet from a friend were likely to be interested in the contents of the tweet, so retweets could be considered as a "barometer" of interests. Additionally, interesting tweets get many retweets, so a retweeted post can be considered a filtered post. Retweet behavior has been extensively studied by researchers to improve friend recommendation, define propagation paths and filter contents.

Retweets have been widely adopted in disaster situations. For instance, communication between the government of Japan and people during the of nuclear radiation leakage disaster that took place after the Fukushima earthquake in 2011, was conducted largely through Twitter and by means of retweets in particular (Li, Vishwanath & Rao, 2014). Their study collected retweets using keyword searches in the month

following the earthquake, and conducted content analyses to categorize tweets into one of the four categories; alarm, assurance, doubt and government. Results showed that the most retweeted messages were in the alarm category. The study also found that retweeted messages differed dramatically from what newspapers and television broadcasts, where the latter contained more reassuring messages, while Twitter messages contained more alarm. Toriumi and Baba (2016) used retweets as a filtering mechanism for real-time tweet retrieval in disaster events. They analyzed 34,000 tweets that were retweeted more than 100 times each. Results showed that the retweet-based methods were well suited for quick tweet retrieval in disaster situations. Kogan, Palin and Anderson (2015) investigated the creation of new relationships between users during disaster events, and how the behavior of people affected by the disaster differ from other users. They applied the analysis to Twitter data collected during Hurricane Sandy. The study found that people form dense interconnected retweet networks during disaster events, and a high proportion of the retweets were retweeted from 10 to 80 times.

Kwak et al. (2010) found that people got information in Twitter not only from the people they followed, but also by retweets. They counted the number of additional recipients of a tweet who are not direct followers of the original tweet poster. They found that the average number of additional recipients was a 1000, and this number was not affected by the number of the followers of the original tweet poster. They constructed retweet trees in order to discover how far and deep retweets move in Twitter, and they found that most retweet trees have a height smaller than 6 and up to a maximum of 11. Additionally, the study investigated how fast retweets started and how long they lasted, and found that 50% of retweets occur in the first hour, and 75% within a day.

2.3.2 Hashtags use in Twitter. Social media users create hashtags in order to form group discussions about a certain topic, event or product. Hashtags allow users to find trending topics, join discussions, and spread related hashtag information to their followers (Lai et al., 2015). Hashtag related research has been widely conducted due to their large use not only in Twitter but in many other social media platforms such as Facebook and Google+ (Maity et al., 2016). Researchers have studied hashtags to analyze adoption rate, predict popularity (Maity et al., 2016), analyze propagation patterns (Wang & Zheng, 2014), provide recommendation (Shi et al., 2016) and to model user behavior (Bogdanov et al., 2014).

Bogdanov et al. (2014) studied topic-specific user behavior, and introduced a model named genotype, which represents a summary of user interests, activity and susceptibility to adopt new ideas. The study analyzed 467 million tweets and 42 million users, and categorized the tweets by hashtags into 5 different categories; business, celebrities, politics, science/technology, and sports. The study showed that user behavior remains constant within the same topic. The study also found that politics has many more hashtags than the other categories.

Wang and Zheng (2014) analyzed the basic properties of hashtag diffusion in Twitter, including tweet spreading speed, retweet ratio, duration of tweets containing the hashtag and temporal patterns. Temporal patterns included single spike and fluctuation patterns. Single spike patterns occur when a hashtag appears suddenly and is used heavily in a short period of time. Fluctuation patterns occur when a hashtag is used moderately for a long period of time. The study found that both pattern have the same spreading speed, however single spike pattern has a larger retweet ratio, but lower duration.

Xu, Chiu, Chen and Mukherjee (2014) examined healthcare topics shared in Twitter using hashtags. They conducted an extensive social network analysis on a dataset of tweets using 14 common healthcare hashtags over the period of two months, where the selection of the hashtags was guided by the Healthcare Hashtag Project (The Healthcare Hashtag Project, 2014). The study found that the most common theme of healthcare conversation was knowledge sharing, and the second common theme was action, which involves conversation about activism, advocacy and promotion. The study also presented a visualization of the whole conversation network, the different roles people took in the conversation and the number of participants associated with each role. Bogdanov et al. (2014) examined the consistency of Twitter user interests, by examining hashtags shared in 5 different news types. Then, they used this model of user behavior to predict influencers and early adaptors of new topics.

2.3.3 Link sharing in Twitter. Sharing links is among the most common behaviors of Twitter users. Links allow users to complement Twitter short messages with extra contents including websites, images and videos. Link related researches investigate link usage properties (Antoniades et al., 2011), spam and security issues (Cao & Caverlee, 2014).

Nizam, Watters and Gruzd (2014) studied the characteristics of links shared in Twitter in order to improve website navigation. They analyzed 264,647 tweets for four events that have official websites, two of the events were related to sports, and the other two were related to entertainment. Among the characteristics analyzed in the study, were the percentages of tweets containing links, uniqueness of links, top 10 links, depth of links from the official website, and the type of contents shared in those links. The study

found that 25-47% of tweets contained links, and the number of links shared was affected by the type of the event and the characteristics of the official website. Christodoulou, Georgiou and Pallis (2012) tracked the diffusion of YouTube videos in Twitter. They measured the likelihood of a user retweeting a video by analyzing a million tweets containing YouTube video links. They concluded that the retweet functionality has an influence on the diffusion of YouTube videos. They also found that social cascading has an impact on tweeters navigation behavior. For instance, people were more influenced by friends who also followed them, i.e., they share mutual friendship. Another study (Hu et al., 2012) examined links shared in tweets during Osama Bin Laden death, and found that 9% of the tweets contained links, and of the links shared, 65% came from official news accounts such as CNN, NY Times and Reuters, while 35% of links were user generated contents such as YouTube and Tumblr. They also concluded that the event of Osama Bin Laden death was reported in Twitter 21 minutes before major TV channels announced it.

In this chapter, related areas to this thesis were introduced. These areas can be summarized as follows, investigating different types of news-related tweets, and modeling user behavior by analyzing originals, retweets, hashtags and links used in the different types of news tweets. Some of the findings presented in the literature overlap with some of our findings, which will be discussed in chapters 5 and 6.

Chapter 3 Preliminary Study

Social media are a significant reflection of real world events (Shuai, Lui, Xia, Wu & Guo, 2014), from political to sports to entertainment events. Sports events are one of the most engaging events shared in social media. In the initial part of the research study, two sports datasets were analyzed as an exploratory step to gain a better understanding of tweeting behavior during such events. This chapter outlines this study and concludes with recommendations and motivations for the next part of the study.

3.1 Objective

The main goal of this thesis is to compare tweeting activity around the world among different types of news events. As a preliminary step, two sports datasets were studied and modeled. The objective of this preliminary study was to better understand the characteristics of tweets related to sports events. Modeling the datasets started with profiling the tweets in the two datasets, examining retweets, geographic features, links and hashtags. This study provided the basis for the second larger phase of the research, by providing insight on the various aspects of processing and analyzing news-related tweets.

3.2 Research Question

The research question for this part of the study was:

RQ: Does user participation around the world during two similar sports event have similar characteristics? In order to answer this question a series of sub questions were posed:

1. What is the number of users, locations, retweets, and density in each dataset?
2. How do users use links during these sports event?
3. How do users use hashtags during the events?

3.3 Datasets

The datasets used in this study were a set of tweets collected during the 2013 World Junior Ice Hockey Championship, and curling events in 2014 winter Olympics created by Nizam and Watters (2014). The dataset was created originally for a link analysis research study. For this study, only some of twitter fields are used; tweet text, time stamp, user name, id, location and time zone. Location information was missing from the datasets and therefore further processing was needed to obtain users locations and time zones from Twitter. An example of tweets after obtaining the location fields is shown in Figure 3-2.

Geocoding was applied to the location field, which meant obtaining longitude and latitude of a location via a mapping service, such as MapQuest or Google Maps. In tweets where the user had no location or the name of the location was not recognized by the geocoder, the tweet was eliminated from the dataset. The geocoding was achieved using Google geocoding API. Examples of final tweets after adding longitude and latitude are presented in Figure 3-1. The total counts of tweets in each dataset before and after geocoding are presented in Table 3-1.

	Ice Hockey	**Curling**
No. tweets before geocoding	1809	14047
No. tweets After geocoding	1277 (71%)	9535 (68%)

Table 3-1 Number of tweets before and after geocoding

Created at	Tweet ID	Tweet text	User name	User ID	User location	Time zone	URL
2012-12-11 18:45	278571275243372544	So one of the lines Canada is trying out is Jonathan	chrisnpeters	60671176	Iowa, USA	Eastern Time (US & Canada)	http://t.co/cG52xzrcNq
2012-12-11 18:46	278571433179893760	But as @coreypronman told me.... IF it doesn't wor	chrisnpeters	60671176	Iowa, USA	Eastern Time (US & Canada)	http://t.co/cG52xzrcNq
2012-12-11 19:00	278574895129042944	Looking for info from this morning's @HockeyCana	NHLFlames	27487343	Calgary, AB	Mountain Time (US & Canada)	http://t.co/X4qqrf7qt7
2012-12-11 19:00	278575061827473409	RT @NHLFlames: Looking for info from this mornin	knighter1990	8.76E+08	London Canada	Atlantic Time (Canada)	
2012-12-11 19:30	278582431865511936	The @HockeyCanada #2013WJC Selection Camp Re	NHLFlames	27487343	Calgary, AB	Mountain Time (US & Canada)	http://t.co/X4qqrf7qt7

Figure 3- Tweets after adding location fields

Created at	Tweet ID	Tweet text	User name	User ID	User location	country	state	city	longitude	latitude	Time zone	URL
2012-12-11 18:45	278571275243372544	So one of the lines Canada is trying out is Jonathan	chrisnpeters	60671176	Iowa, USA	United States	Iowa		-93.097702	41.8780025	Eastern Time (US & Canada)	http://t.co/cG52xzrcNq
2012-12-11 18:46	278571433179893760	But as @coreypronman told me.... IF it doesn't wor	chrisnpeters	60671176	Iowa, USA	United States	Iowa		-93.097702	41.8780025	Eastern Time (US & Canada)	http://t.co/cG52xzrcNq
2012-12-11 19:00	278574895129042944	Looking for info from this morning's @HockeyCana	NHLFlames	27487343	Calgary, AB	Canada	Alberta	Calgary	-114.07085	51.0486151	Mountain Time (US & Canada)	http://t.co/X4qqrf7qt7
2012-12-11 19:00	278575061827473409	RT @NHLFlames: Looking for info from this mornin	knighter1990	876078565	London Canada	Canada	Ontario	London	-81.243177	42.9869502	Atlantic Time (Canada)	
2012-12-11 19:30	278582431865511936	The @HockeyCanada #2013WJC Selection Camp Re	NHLFlames	27487343	Calgary, AB	Canada	Alberta	Calgary	-114.07085	51.0486151	Mountain Time (US & Canada)	http://t.co/X4qqrf7qt7

Figure 31 Tweets after geocoding

3.4 Methodology

The research methodology used in this study was quantitative, in particular, tweets from both datasets were profiled to examine their spread between people and across locations. For each dataset, to model its spread through Twitter, we examined the following:

1. The number of different users involved, the higher the number of users the more spread and vice versa.
2. The number of different locations involved, defined by the location information.
3. The retweet ratio; which is the percentage of posts that were retweets. The retweet ratio was used as an indicator of spread.
4. The distance of the news spread, which is the distance between the geographic location the story took place and the farthest point it reached. The user location field provided a name of a city or country; location information was then used to find the longitude and latitude. The distance was calculated by applying the Haversine formula, which is a mathematical formula that calculates the distance between two points given their longitudes and latitudes (Van Liere, 2010).
5. The density of the spread, which is the number of tweets in a predefined radius of origin, for instance 10, 100, or 1000 kilometers, divided by the total number of tweets. For our purpose, a density of 0.8 was high, and a density 0.3 was considered low. For each news story and for each post the distance between the user location and the origin was obtained, then the number of posts within the predefined distance or less was counted, and divided by the total number of posts.
6. The number of posts with links.

7. The number of unique links included.
8. The number of retweets with links.
9. The top 10 hashtags used with the event.
10. Percentage of posts containing one or more of the top 10 hashtags.
11. Top 10 locations with posts that used the top 10 hashtags.

3.5 Results

The result of profiling the tweets to obtain general characteristics from both datasets is presented in Table 3-2. The use of retweets varied between the datasets; hockey (46%) and curling (37%). There were more unique users with curling event (85%) perhaps indicating wider population participating in the tweeting activity for an Olympic event. Similarly, number of unique locations is greater with the curling event (35%) vs hockey (29%). Maximum distance in both datasets were close. However, density exhibited different patterns. The maximum percentage of tweets (85%) occurred within 9,000 KM radius with the hockey event, this distance included mainly North American locations. In the curling event, the maximum number of tweets occurred (30%) within 12,000 KM radius, again this distance included USA and Canada, however, within 9,000 KM, and 6,000 KM a little less ratios were reported (28%) and (26%), these distances included UK and other European countries. This finding indicates that the curling event had a wider interest than the hockey event.

Characteristic	IIHF – Ufa, Russia	Curling – Sochi, Russia
% Retweet	46%	37%
Unique users/ total	703/ 1,277 (55%)	8,066/ 9,535 (85%)
Unique locations/ total locations	330/ 1,154 (29%)	2,634/7,519 (35%)
Maximum distance km.	West Coast, New Zealand 15,260	Chile 16,696
Density the number of tweets in ranges (km) 3000, 6000, 9000, 12000	2%, 4%, %85, 8%	10%, 26%, 28%, 30%

Table 3-2 Basic characteristics of the two sports datasets

Table 3-3 presents the results of link analysis for the sports datasets. Similar ratios of unique links shared in both datasets, hockey (47%) and curling (48%). Curling had larger ratios of posts with links and retweets with links. This may be due to the difference of scale between the two events.

Characteristic	IIHF – Ufa, Russia	Curling – Sochi, Russia
Posts with Links	388 (30%)	3201 (34%)
% Retweets with Links	51%	66%
Unique Links	183 (47%)	1551 (48%)

Table 3-3 URL characteristics of the two sports datasets

IIHF – Ufa, Russia	Count	Curling – Sochi, Russia	Count
#2013WJC *	1162	#Sochi2014*	8039
#2013wjc *	77	#curling	2402
#Flames	60	#sochi2014 *	1031
#TeamCanada	59	#Curling	883

#TSN	55	#lovecurling	830
#TSNs	50	#TeamGB'	250
#TeamUSA	37	#Olympics *	240
#Habs	24	#WeAreWinter	225
#Canada	21	#Biathlon	175
#OHL	21	#IceHockey	174

Table 3-4 presents the top 10 hashtags used in each event, and the frequencies of hashtags in the tweets. The hashtags marked with an asterisk were used in the keyword search. The top 10 hashtags used in both events were the most related, where in the bottom of the list of all hashtags many irrelevant hashtags appeared. This may indicate the importance on focusing on the top 10 hashtags when analyzing user behavior by using hashtags.

IIHF – Ufa, Russia	**Count**	**Curling – Sochi, Russia**	**Count**
#2013WJC *	1162	#Sochi2014*	8039
#2013wjc *	77	#curling	2402
#Flames	60	#sochi2014 *	1031
#TeamCanada	59	#Curling	883
#TSN	55	#lovecurling	830
#TSNs	50	#TeamGB'	250
#TeamUSA	37	#Olympics *	240
#Habs	24	#WeAreWinter	225
#Canada	21	#Biathlon	175
#OHL	21	#IceHockey	174

Table 3-4 Hashtags characteristics of the two sports datasets

IIHF – Ufa, Russia	**Count**	**Curling – Sochi, Russia**	**Count**
Calgary, AB	150	London, UK	407

Toronto, ON	124	Toronto, ON	323
North Liberty, IA	81	Canada	216
London, ON	56	United Kingdom	171
Canada	53	New York, NY	163
Rochester, NY	40	Winnipeg, MB	159
New Jersey	32	Edinburgh, UK	133
Ottawa, ON	22	Zaragoza, Spain	105
Vancouver, BC	20	Ottawa, ON	100
Massachusetts	19	Scotland, UK	91

Table 3-5 presents the top 10 countries that used any of the top 10 hashtags in each sport event. In the hockey event Calgary was the location with the largest hashtag use. In curling event UK was the location with the most tweet with hashtags. The hashtags analysis exhibited a reflection of users' activity during the events, where the analysis revealed which country participated more. This finding would guide further investigation about the reasons and factor behind such behavior. For instance, the hockey event was held in Russia, and Russia team participated in the games, however Russia did not appear in the top 10 countries. Investigating such case, would provide information about social media adoption in Russia.

IIHF – Ufa, Russia	**Count**	**Curling – Sochi, Russia**	**Count**
Calgary, AB	150	London, UK	407
Toronto, ON	124	Toronto, ON	323
North Liberty, IA	81	Canada	216
London, ON	56	United Kingdom	171
Canada	53	New York, NY	163

Rochester, NY	40	Winnipeg, MB	159
New Jersey	32	Edinburgh, UK	133
Ottawa, ON	22	Zaragoza, Spain	105
Vancouver, BC	20	Ottawa, ON	100
Massachusetts	19	Scotland, UK	91

Table 3-5 Top 10 countries in hashtags use

3.6 Limitations

The two datasets varied in size considerably, which might affect the accuracy of the result. The hashtags were used in the keyword search to create the datasets, which influenced the result of the number of popular hashtags.

3.7 Conclusion and Recommendations

While the size and global perspective of the data sets are different, the profiles are quite similar except for the number of unique users, percent of retweets, and the density (tweets by distance). The pattern of location in Tweets with hashtags exhibits Zipf characteristics while the use of locations is more linear in both sets. Sports is considered a type of entertainment news, in the next phase of the study we would explore these characteristics in other large social media datasets from more traditional types of news, specifically natural events and top news stories leading to the development of profiles that can be used to inform guidelines. The characteristic we found interesting to investigate in the next phase are retweets, links, hashtags and the distribution of the usage of these features among countries and news types.

Chapter 4 Research Methodology

According to Creswell (2009) quantitative approaches are best for problems that calls for identification of factors that influence or predict an outcome. Quantitative research examines the relationships among variables. These variables can be measured computationally, for instance by SQL queries, then the results can be statistically analyzed. By taking into consideration bias prevention, findings can be generalized and replicated (Creswell, 2009). A quantitative analysis is typically retrospective and starts with datasets that have been collected in advance of the study (Imran, Castillo, Diaz, & Vieweg, 2015). In this phase of the research study three types of news stories were examined.

News events have different levels of geographical interest; as some events attract the attention of the whole world, while other concern a small population. The participation of users in Twitter in response to an event is influenced by their geographic proximity and whether they are directly affected by the event. For example, the Malaysian airplane crash took place in 2014 was an event that all world news media reported, however the crash has more direct impact on Malaysia and its people. In order to compare geographic features in news-related tweets, the news stories must bear a global interest as far as possible, therefore the types selected for our study are world politics, disaster, and finance. In this study, politics category refers to social disaster stories that have a political impact. In the following sections data collection, analysis, assumptions and problems faced are presented.

4.1 Datasets

Sampling Twitter can be achieved by using two basic methods, the search API and the streaming API. The maximum number of tweets returned by any of these services does not exceed 1% of all tweets (Twitter.com, 2015). However, a comparative study (Wang, Callan & Zheng, 2015) comparing the streaming API sample to a complete sample showed that the streaming API sample maintained enough information for research concerning general Twitter statistics, sentiment analysis, and user activity.

Data collection and preparation process for this research was done by applying the ETL (Extract-Transform-Load) approach (Ji et.al, 2015). The extract step started by collecting the data from Twitter Streaming API in JSON format (JavaScript Object Notation), which is a format readable by human and easily parsed by machine. The transform step was achieved by parsing fields of interest into a CSV (Comma Separated Value) file. Lastly, the load step was accomplished by parsing the CSV file into a relational database for querying and analysis. The overall dataset preparation process is shown in Figure 4-1, and in the following subsections the details of this process are explained.

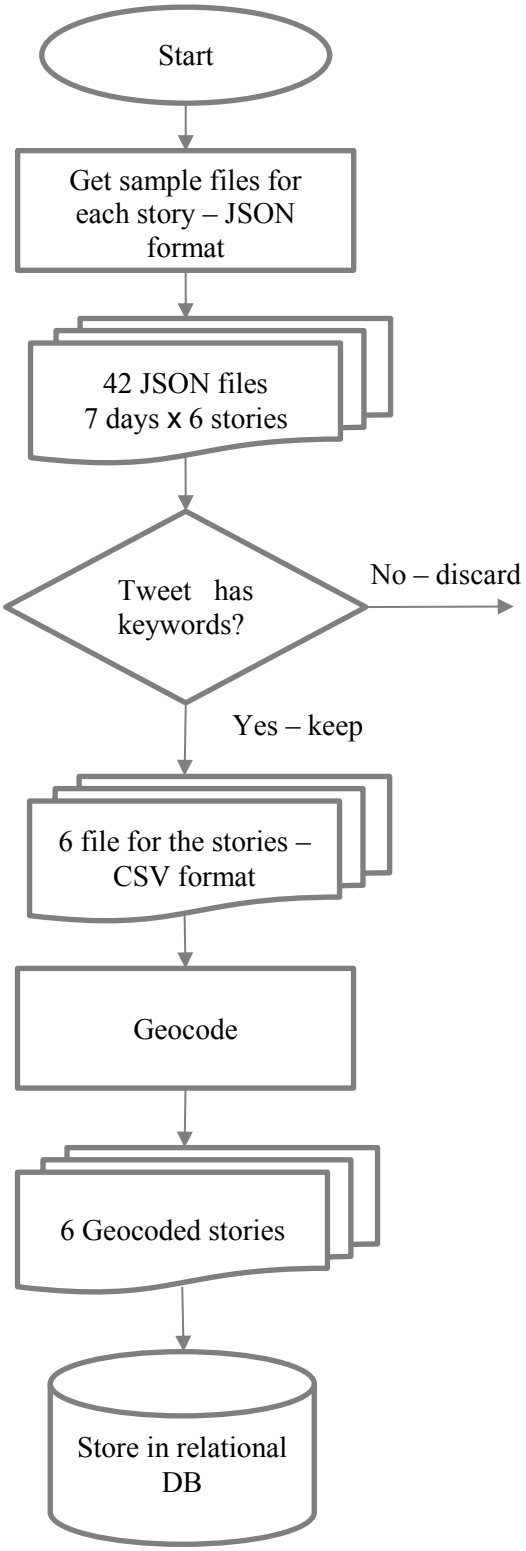

Figure 4-1 Dataset preparation process

4.1.1 News stories selection. News sources are not consistent in their categorization of news, where categories differ from one media outlet to the other. For instance, CNN categorize news by country and by news type. The first category (country) includes USA and Canada, Africa, Asia, Europe, Latin America, Middle East, and the second category (news type) includes Business, Entertainment, and Technology. BBC uses news, sport, weather, shop, earth, travel, with sub-categories, world, UK, business, tech, science, entertainment and art. The New York Times uses world, U.S. politics, N.Y., business, opinion, tech, science, health, sports, arts, style, food, travel, and real estate. A study of news trends diffusion (Lima & Musolesi 2012) used the categories art, disaster, science, sports and technology.

News channels broadcast a large number of different stories from different news types daily, however not all news stories motivate social media users to post and comment. During some events the tweeting and retweeting activity between social media users increase significantly (Imran et al., 2014). Tweeting may be initiated for many reasons, which may be endogenous or exogenous. Endogenous trends occur when popular ideas spread widely by viral contagion or information cascade. Exogenous trends are associated with real-world events, such as emergencies and earthquakes.

For purposes of this research, the selection of stories was limited to the ones that were associated with exogenous causes. Additionally, in order to find enough tweets in the 1% sample of tweets available to the study, the event had to be important and significant enough to trigger social media users to tweet and retweet generating huge amounts of tweets. The types of news that fit this profile and chosen for this research were, finance, disaster and politics. We searched for tweets about stories from these news

categories that took place in 2015, then we selected stories that return enough posts (> 10,000).

The two stories chosen for political news were the shooting in Charlie Hebdo magazine office in Paris, and the story of the teenage boy who was arrested in a Texas high school for inventing a clock that his teacher mistakenly thought it was a bomb. For financial news the stories chosen were the Chinese stock market fall causing disturbance in shares market around the world, and the story of the automaker Volkswagen, which was accused for manipulating pollution control systems so that it emits low levels of pollutant during emission tests. The stories selected for disaster were the Germanwings airplane crash, an airbus crashed in French Alps killing 150 people, and Nepal earthquake of 7.8 magnitude that killed over 9,000 people. Stories as appeared in official news websites are presented in Figure 4-2.

Figure 4-2 News stories selected

4.1.2 Sampling. The sample used was collected from the streaming API by DNLP, Dalhousie Natural Language Processing research group which consisted of a collection of tweets collected during 2014 and 2015, and organized by days and categories, such as finance, auto, retail and sample (general) among other categories, as shown in Figure 4-3.

Figure 4-3: Raw dataset files organized by days and by categories

All the six stories selected took place in 2015, and the categories utilized were sample, finance and auto. For each news story, tweets were collected for a 7-day period, starting at the day the event took place. We chose 1-week period because according to observation stated by Kwak et al. (2010), tweeting activity for most trending topics lasts for a week or less. A total of 42 compressed files were downloaded and extracted, seven

for each dataset. Next keyword searches were applied to find related tweets for each news story. Table 4-1, Table 4-2 and Table 4-3 show the details of each news type dataset, and Table 4-4 presents a summary of the dataset collection process. The total count of tweets for the six datasets before processing was 371,379 tweets, with the size of approximately 800 GB of memory.

Event	Search Keywords	Date	Tweets in sample	Tweets returned
Chinese Shares drop	'chinese shares', 'chinese stock'	24/08/2015	1,394,841	3,851
		25/08/2015	1,430,283	1,801
		26/08/2015	1,329,505	2,275
		27/08/2015	1,327,922	901
		28/08/2015	1,442,474	1,787
		29/08/2015	1,148,202	69
		30/08/2015	1,230,819	66
Volkswagen Scandal	All: ‘emissions’ ‘volkswagen’	19/09/2015	2,064,797	2,920
		20/09/2015	1,949,912	1,799
		21/09/2015	2,291,496	18,974
		22/09/2015	2,543,039	33,700
		23/09/2015	2,572,296	29,139
		24/09/2015	2,614,319	18,395
		25/09/2015	2,385,107	13,635

Table 4-1 Finance news stories data collection details

Event	Search Keywords	Date	Tweets in sample	Tweets returned
Germanwings airplane crash	'Germanwings', 'flight 9525' ‘Lufthansa’	24/03/2015	4,462,240	7,299
		25/03/2015	4,650,410	3,800
		26/03/2015	4,318,080	4,379
		27/03/2015	4,731,955	3,392
		28/03/2015	4,808,650	1,522
		29/03/2015	4,729,602	1,104
		30/03/2015	4,601,930	1,149
Nepal Earthquake	All:‘nepal’, ‘earthquake’	25/04/2015	4,187,583	7,671
		26/04/2015	4,230,153	7,824
		27/04/2015	4,229,411	6,273
		28/04/2015	4,241,342	3,936
		29/04/2015	1,013,868	540
		02/05/2015	3,221,767	1,033
		03/05/2015	4,568,149	1,148

Table 4-2 Disaster news stories data collection details

Event	Search Keywords	Date	Tweets in sample	Tweets returned
Charlie Hebdo Shooting	‘Hebdo’, ‘JeSuisCharlie’	07/01/2015	4,504,568	63,794
		08/01/2015	4,380,187	41,707
		09/01/2015	4,297,932	25,284
		10/01/2015	4,314,595	11,359
		11/01/2015	4,302,956	13,753
		12/01/2015	4,158,101	7,715
		13/01/2015	4,420,819	7,391
Ahmed Mohammed Clock	‘IStandWithAhmed’, OR All:‘arrested’ ‘clock’	16/09/2015	3,673,312	11,700
		17/09/2015	3,696,966	6,564
		18/09/2015	3,579,579	1,550
		19/09/2015	3,692,748	615
		20/09/2015	3,756,820	411
		21/09/2015	3,699,213	171
		22/09/2015	3,729,180	201

Table 4-3 Politics news stories data collection details

News Story	Tweets Returned	Story Date
Charlie Hebdo newspaper shooting	170,992	January 7
The “Clock Boy”	21,164	September 16
Germanwings airplane crash	22,644	March 24
Nepal earthquake	28,324	April 25
Chines stock market collapse	10,649	August 24
Volkswagen emission test cheating	117,606	September 19

Table 4-4 Dataset collection summary

Each tweet returned from Twitter API contained a number of different fields, the fields of interest to this research are presented in Table 4-5 and Table 4-6 (Twitter Developers, 2014). The fields used were the following; created_at, tweet identification number, tweet text or status, user name, user identification number, time zone, user location and URL if available. Hashtags were parsed from tweet text in database creation stage.

Field	Description
Created_at	Coordinated Universal Time (UTC) when the tweet was posted.
Retweeted	Boolean value indicating whether or not the tweet was retweeted.
Tweet id	Integer represents the tweet identification number
User id	integer representation of a unique user identification number.
User name	name of the user.
User screen_name	unique user name.
Status	the last user’s tweet or retweet.
URL	Link extracted from status (tweet text) if exist

Table 4-5 Tweet fields used in the study

Field	Details
User location	A user-defined field, may contain a real "correct" location, and may not.
Time zone	Time zone of the user, selected from a dropdown list.
geo_enabled	Boolean indicating whether or not the user's tweet will be geotagged.
Coordinate	The geographic location of a tweet, represented by longitude and latitude, included in the tweet only if geo_enabled is true
Place	Indicate that a tweet is associated with a place, and not necessarily tweeted from that place.

Table 4-6 User location fields in a tweet

4.1.3 User locations. The geographic information available in a tweet falls into one of the following; user location, time zone, coordinate and place. User location is typed by the user, therefore it may contain any word and sometimes typos, time zone is selected by the user from a dropdown list, coordinates represents the longitude and latitude as detected by location services if the location detection feature was enabled in the user's smart device, and place is associated by Twitter with a place, however it does not necessarily mean that the tweet was posted from that place. The details of the location fields in a tweet are summarized in Table 4-6. User location and time zone fields in the tweet were used to identify user location. These two feature were used due to their availability in more than 70% of the tweets, while the other two existed in only 2% of the posts.

Using the user location and time zone fields, an algorithm was applied to fetch longitude, latitude and the correct country name. This geocoding, i.e., assigning coordinate to a location name, was performed using MapQuest geocoding API. The algorithm checked if the user had a location, and if a location was found, the algorithm

attempted to geocode the location. If the user location was not found after geocoding, the time zone was used.

From our initial skimming of the results returned by the geocoding service, a number of mistakenly geocoded locations were found. This was because the location typed by the user was either incorrect (not a location) or had a spelling mistake. However, the geocoder attempted to find a matching location, and in some case it found an incorrect match. Table 4-7 demonstrates some examples of incorrect geocoding attempts.

Original Location	Geocoded Location
Earth	United States
My Home	United States
Paris – Hamburg – In the Air	United States
Pisa, Italy - Pescia, Italy	United States

Table 4-7 Examples of geocoding errors

To find the geocoding error rate, manual screening was applied to 1000 randomly selected sample of geocoded user locations. The result of the manual error check shows that 72% of the users' locations were correctly geocoded, i.e., the percentage of error is 28%. Table 4-8 presents the total counts of tweets, geocoded tweets and the percent of geocoded tweets to total tweets in each news story dataset.

Dataset	Total Tweets	Geocoded Tweets	Percent
Chinese Shares	10,649	6,784	64%
Volkswagen	117,606	86,364	73%
Germanwings	22,644	17,351	77%
Nepal Earthquake	28,324	20,796	73%
Charlie Hebdo	170,992	133,409	78%
Clock Boy	21,164	15,732	74%

Table 4-8 Total counts of tweets, geocoded tweets and the percent of geocoded to total tweets in each news story dataset

Next, distance from the origin of the story is calculated by applying the Haversine formula (Van Liere, 2010). The result of this step is a CSV file with the same fields mentioned above plus longitude, latitude and country name. A snap shot of a file with the fields used is illustrated in Figure 4-4, and an example of the result of the geocoding process is shown in Figure 4-5.

Created at	Tweet ID	Tweet text	User name	User ID	User location	User location	Time zone
2015-09-19 0:05	645025793429622784	RT Volkswagen, Audi accused of using software to ch	Joe_The_Wizard	351798304	Miami, FL		Eastern Time (US & Canada)
2015-09-19 0:05	645025870680223744	Volkswagen Accused Of Cheating EmissionsŒæRules	iGetMoneyOrg	943474375			Casablanca
2015-09-19 0:05	645025912501616641	Volkswagen, Audi accused of using software to cheat	BerkleyBearNew	787546010	Doghouse		Eastern Time (US & Canada)
2015-09-19 0:05	645025933821276160	RT Vox Sentences: Volkswagen made cars smart enou	amrokadri	125565507	Egypt		Cairo
2015-09-19 0:05	645025947931013120	ams6110 comments on "Volkswagen Is Ordered to Re	ExplodingAds	1155188726	Lincoln, NE		Eastern Time (US & Canada)

Figu ple CSV file with the fields use

Figur ple geocoded CSV file

Created at	Tweet ID	Tweet text	User name	User ID	User location	country	longitude	latitude	Time zone	URL
2015-09-19 0:05	645025793429622784	RT Volkswag	Joe_The_Wizard	351798304	Miami, FL	United States	-80.194702	25.775084	Eastern Time (US & Canada)	
2015-09-19 0:05	645025870680223744	Volkswagen	iGetMoneyOrg	943474375		Morocco	-7.5898434	33.5731104	Casablanca	
2015-09-19 0:05	645025912501616641	Volkswagen,	BerkleyBearNews	787546010	Doghouse	United Kingdom	0.638927	51.881035	Eastern Time (US & Canada)	
2015-09-19 0:05	645025933821276160	RT Vox Sente	amrokadri	125565507	Egypt	Egypt	29.0548	26.317301	Cairo	
2015-09-19 0:05	645025947931013120	ams6110 cor	ExplodingAds	1155188726	Lincoln, NE	United States	-96.70261	40.813599	Eastern Time (US & Canada)	http://ExplodingAds.com

4.1.4 Database creation and querying. The CSV file containing all the parsed fields from the tweet, plus the location information obtained by geocoding (Figure 4-5), was then transformed into a relational database. Hashtags were extracted from tweet text and stored in a separate table. The final database schema consisted of the tables, locations, users, tweets, hashtags and links. Database was created and queried using Python version 2.7 and SQLite3.

4.1.5 Data collection summary. The decision was made regarding which data to use, live Twitter stream or archived tweets. If a live stream of tweets was the choice then Twitter streaming API would be used, however, for this research archived data was used that had previously been collected from the live stream.

4.2 Analysis

The aim of this research is to examine the relationships between user participation, geographic locations, and news types. Using both quantitative and statistical analysis. This section outlines the processing applied. The analysis started by examining the general characteristics of the whole dataset, i.e., the 6 stories combined. Next, analyses were applied by: country, news type and news story. Finally, statistical analysis was applied to model the relationships between country, news type and user behavior. The following subsections explain the details of these analyses.

4.2.1 General Descriptive Statistics for All 6 Stories. The data used for this research consisted of 6 datasets of tweets for 6 major news events that took place in 2015 in 5 countries. The tweets were generated by users from 235 countries around the world. The first step in analyzing this data was to combine the 6 datasets into one dataset, resulting in a total of 363,720 tweets, of these 363,720 tweets 280,436 were correctly

geocoded tweets, and were stored in the database. Hence, the term database is used to refer to the geocoded and saved tweets used for the analysis in this study. The general characteristics used to describe the database include the counts of total geocoded tweets, original tweets, retweets, tweets with hashtags and tweet with links. We also obtained the number of originals with links and the number of originals with hashtags, and the same for retweets. This analysis provides an overall description of all the tweets used in the study.

4.2.2 Characteristics by country. Using the whole dataset (all 6 stories), the top 10 countries were identified, and analyzed by the same characteristics for each country as in the previous step. Counts of originals and retweets, counts of tweets with hashtags, tweets with links, the original/retweet distribution for tweets with hashtags and tweets with links were obtained for each of the top 10 countries. Only three countries appeared in each individual story, the tweets for these countries were identified and analyzed in the same manner.

4.2.3 Characteristics by news type. News stories of the same type were then collapsed into one dataset, resulting in three datasets by news type: finance, disaster and politics. The three datasets were analyzed by the same characteristics used for the whole database. Top 10 countries for each news type were identified and analyzed.

4.2.4 Characteristics by news story. Each of these news story exhibited some unique characteristics, depending on the nature of the story and the population it affected, among other factors. The characteristics of each individual story were analyzed. Additionally, we analyzed the data by the top 10 countries, the top 10 hashtags, and the top 10 links for each story.

4.2.5 Statistical analysis. Binary logistic regression was applied to provide two basic outcomes. First, to find the significance of the co-relations of the variables country and news type. Second, to model the relationship between the variables country, news type and behavior. From this model we calculated probabilities of the occurrences of the different combinations of variable, for instance, the probability of country A in news type X to generate hashtags. Also, odds ratios were obtained from this model, which allowed performing comparisons among countries and among news types.

4.3 Assumptions

Users locations considered for this research were obtained from the users' profiles. For the purpose of this study we assumed that for a certain user, the location obtained was the location where the user lived, but not necessarily where the tweet came from. We also assumed that the user stayed in the original location during the period of data collection, so if the user appeared twice in the dataset only the first location was fetched and saved in the database.

4.4 Problems Faced

Quantitative research methods require examining the relationships among variables statistically, thus, a statistical procedure suitable for our data was needed. However, due to the nature of the data used, and its lack of normal distribution, finding the right tool was one of the difficulties faced in this study. Another problem faced was a technical one, which was the rate limit of the geocoding service. This rate limit allowed limited access of the mapping service per day, so to geocode a large number of tweets, the process took longer than expected, it took several days in some cases.

Chapter 5 Results

The results of applying the analyses described in chapter 4 are presented in this chapter. First, the results describing the whole dataset are presented, the results by country, news type and story are reported next. Finally, we explain the results of the statistical tests applied for analyzing the relationships between original/retweet, hashtags, links, and both country and news type.

5.1 General Descriptive Statistics for All 6 Stories

The general characteristics for the whole dataset provides an overall description of all the tweet used in the study, these results are presented in Table 5-1. The first column presents the characteristics examined, the second and third columns are counts and percentages of total tweets, and the last column presents the percentage of the characteristic to total retweets or originals.

The original to retweet ratio was 41% to 59%, indicating that most of the tweets were retweets. 70% of retweets contained hashtags, and 77% contained links. This high link use ratio did not agree with the finding reported in (Cao & Caverlee, 2014), which were between 25% – 29%. This could be due to the different nature of the tweets used in our research as opposed to a stream of tweets with no specific event. 46% of originals contained hashtags and 78% contained links. Retweets containing both hashtags and links represented 31% of all tweets, while originals containing both hashtags and links were only 12% of all tweets.

Characteristics	Count	% Total Tweets	%Original /Retweet
Total tweets	280,436		
Unique users	203,363		
Average tweets per user	1.38		
Unique countries	235		
Originals	114,695	41%	
Average originals per user	0.57		
Retweets	165,741	59%	
Average retweets per user	0.81		
Tweets with Hashtags	169,517	60%	
Tweets with Links	214,697	77%	
Hashtags & Links	122,325	44%	
Originals & Hashtags	52,749	19%	46%
Originals & Links	89,166	32%	78%
Originals Hashtags & Links	34,278	12%	30%
Retweets & Hashtags	116,768	42%	70%
Retweets & Links	125,531	45%	76%
Retweets Hashtags & Links	88,047	31%	53%

Table 5-1 General characteristics of tweets in the database of the 6 stories combined

5.2 Characteristics by Country

To analyze the distribution of tweets by country, ideally, the distributions of originals, retweets, tweets with hashtags and links for the 235 countries must be analyzed and compared. A closer look at the distribution of tweets by country, however, revealed that 73% of tweets originated from 10 countries, and three countries were in the top 10 producers of tweets in each of the 6 stories. Therefore, two subsets of the main database, the first with 10 countries and the second with 3 countries were created and analyzed. The

two subsets shown in Table 5-2 exhibited a pattern similar to the whole database in terms of percentages of originals, retweets, tweets with hashtags and tweets with links to total tweets in each one of them.

	Dataset	Total Tweets	Originals	Retweets	Tweets with Hashtags	Tweets with links
Counts	**All countries**	280,436	114,695	165,741	169,517	214,697
	Top 10	205,296	83,977	121,319	125,318	156,272
	Common 3	132,240	59,409	72,831	72,409	103,729
Percent	**All countries**		41%	59%	60%	77%
	Top 10		41%	59%	61%	76%
	Common 3		45%	55%	55%	78%

Table 5-2 Counts of originals, retweets, tweets with hashtags and tweets with links in the whole database and in each subset and percentages to total tweets in each of them

In the following subsections, the analyses of the characteristics of originals, retweets, tweets with hashtags and links for the top 10 countries and the common 3 countries are presented.

5.2.1 Top 10 countries. The following 10 countries produced 73% of the dataset: USA, France, UK, India, Netherland, Canada, Spain, Greece, Indonesia and Germany. These countries along with their ranks are presented in Table 5-3. The common three countries USA, UK and India were the first, second and fourth ranks respectively. France occupied the second place due to the large number of tweets generated by Charlie Hebdo story, which represented a large portion of the dataset.

Rank	Country	Tweets	Original	% Total	Retweets	% Total
1	USA	99,295	43992	44%	55303	56%
2	France	32,463	8130	25%	24333	75%
3	UK	21,316	9776	46%	11540	54%
4	India	11,629	5641	49%	5988	51%
5	Netherlands	8,604	2800	33%	5804	67%
6	Canada	7,854	3547	45%	4307	55%
7	Spain	7,813	2397	31%	5416	69%
8	Greece	6,872	1991	29%	4881	71%
9	Indonesia	4,841	3674	76%	1167	24%
10	Germany	4,609	2029	44%	2580	56%
Totals/ % of Total		205,296	83,977	41%	121,319	59%

Table 5-3 Raw counts of tweets, originals and retweets in the top 10 countries and percentages of originals and retweets to total tweets in the top 10 countries dataset

5.2.1.1 Original and retweets. Table 5-3 presents the raw counts of tweets in each of the top 10 countries, originals, retweets, and percentages of originals and retweets to total tweets in each country. The number of retweets was more than originals in all countries except for Indonesia.

USA generated 99,295 tweets, which represented 35% of all tweets in the database, next France (32,463), UK, (21,316) and India, (11,629), the rest of the countries generated less than 10,000.

Figure 5-1 shows the distribution of the number of tweets, originals and retweets in the top 10 countries. USA generated the largest number of tweets, France generated around 1/3 of USA, UK 2/3 of France, and India approximately 1/2 of UK. This pattern is

similar to a power law distribution, with the top few countries generating most of the tweets, and many countries generating few tweets.

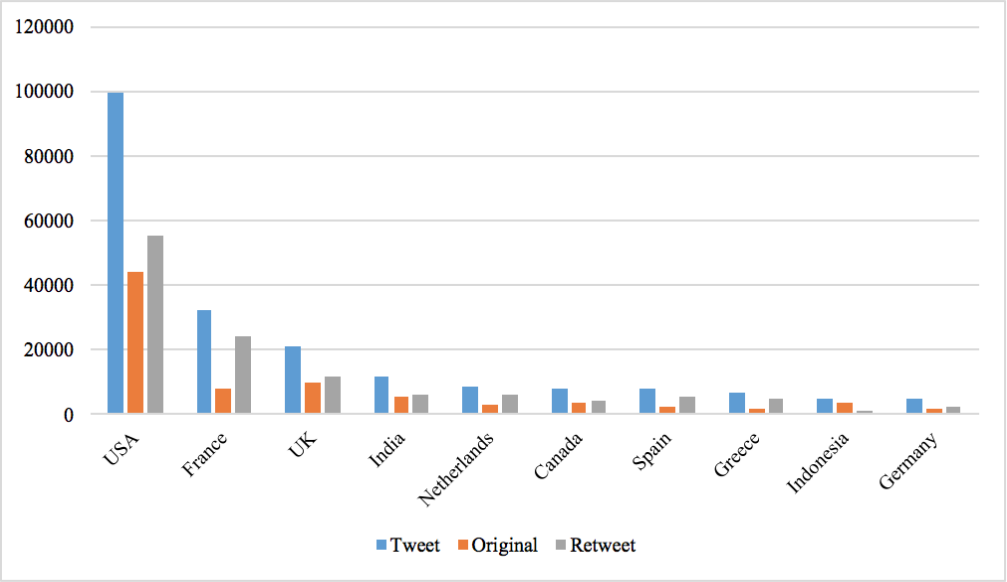

Figure 5-1 Tweets, originals and retweets counts in the top 10 countries

5.2.1.2 Hashtags. Hashtag usage in the top 10 countries is presented in Table 5-4. The table presents the raw counts of tweets with hashtags, percentages of tweets with hashtags to total tweets in each country, counts of originals and retweets with hashtags, and the percent of originals and retweets with hashtags to total tweets with hashtags in each country. A high percentage of tweets (51% – 83%) contained hashtags in all the top 10 countries, except Indonesia (41%).

Originals and retweets

All countries in the top 10 had higher percentages of retweets with hashtags than originals with hashtags (69%) except Indonesia (40%). The hashtags occurrences in both

originals and retweets (30/70) were similar to the overall pattern of the original to retweet ratio (40/60).

Country	Tweets with Hashtags	%Total Tweets	Originals with Hashtags	% Tweets with Hashtags	Retweets with Hashtags	% Tweets with Hashtags
USA	54,118	55%	18,354	34%	35,764	66%
France	26,946	83%	6,193	23%	20,753	77%
UK	11,870	56%	4,125	35%	7,745	65%
India	6,421	55%	2,415	38%	4,006	62%
Netherlands	6,182	72%	1,647	27%	4,535	73%
Canada	4,005	51%	1,303	33%	2,702	67%
Spain	5,432	70%	1,404	26%	4,028	74%
Greece	5,182	75%	1,192	23%	3,990	77%
Indonesia	1,973	41%	1,188	60%	785	40%
Germany	3,189	69%	1,296	41%	1,893	59%
Total/%Total	125,318	61%	39,117	31%	86,201	69%

Table 5-4 Raw counts of tweets, originals and retweets with hashtags in the top 10 countries and percentages of originals and retweets to total tweets in each country

Figure 5-2 illustrates the distribution of the usage of hashtags in the top 10 countries. Hashtags were more frequent in retweets than in originals. This might indicate that retweets were a factor in spreading hashtags.

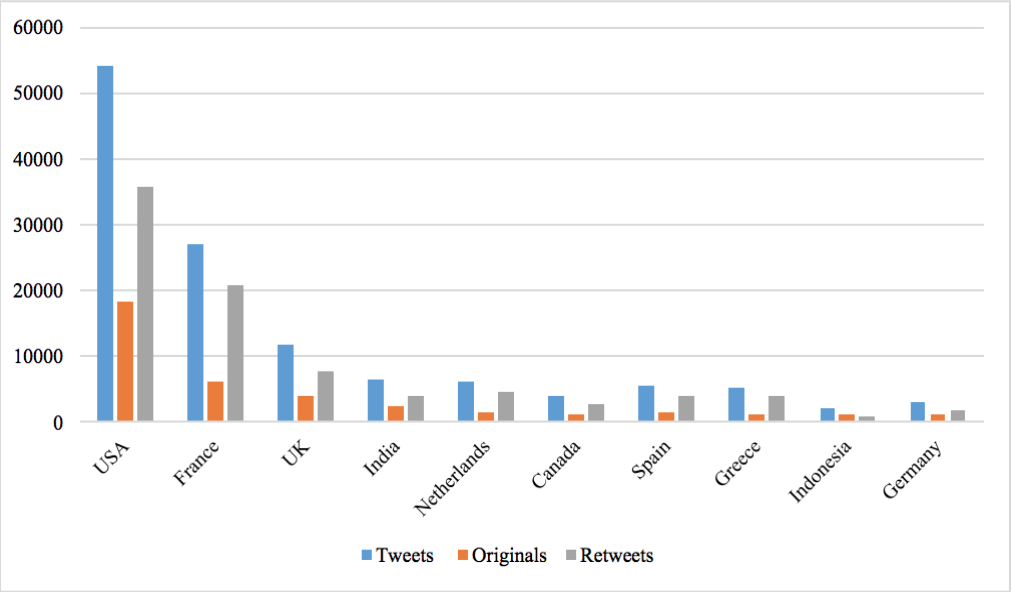

Figure 5-2 Raw counts of tweets, originals and retweets with hashtags in the top 10 countries

5.2.1.3 Links. In Table 5-5 link usage in the top 10 countries is presented. The second column presents the total number of tweets with links in the corresponding country, and the third column presents the percentage of tweet with links to total tweets in each country. The last four columns show the number of originals with links, retweets with links and their percentage to total tweets with links in each country. All the top 10 countries had high percentages of tweets with links, the lowest in Greece (67%) and the highest in Indonesia (90%).

Originals and retweets

All countries had higher or equal percentages of retweets with links than originals with links, except for Indonesia. Interestingly Indonesia had a higher ratio of originals overall.

Country	Tweets with Links	%Total Tweets	Originals with Links	%Tweets with Links	Retweets with Links	%Tweets with Links
USA	78336	79%	35,443	45%	42,893	55%
France	21840	67%	4,480	21%	17,360	79%
UK	16943	79%	7,848	46%	9,095	54%
India	8450	73%	4,253	50%	4,197	50%
Canada	6340	81%	2,953	47%	3,387	53%
Netherlands	6060	70%	1,903	31%	4,157	69%
Spain	5809	74%	1,747	30%	4,062	70%
Greece	4633	67%	1,357	29%	3,276	71%
Indonesia	4362	90%	3,458	79%	904	21%
Germany	3499	76%	1,558	45%	1,941	55%
Total	156,272	76%	65,000	42%	91,272	58%

Table 5-5 Raw counts of tweets with links, originals and retweets in the top 10 countries and percentages of originals and retweets to total tweets in each country

Figure 5-3 illustrates the distribution of tweets, originals and retweets with links in the top countries. USA generated around 80,000 tweets with links, next France, 20,000, 1/4 of USA, UK a bit less than France, then the rest of the countries generated fewer than 5,000 tweets with links. Generally, the ratio of original to retweet with links was close to 40/60.

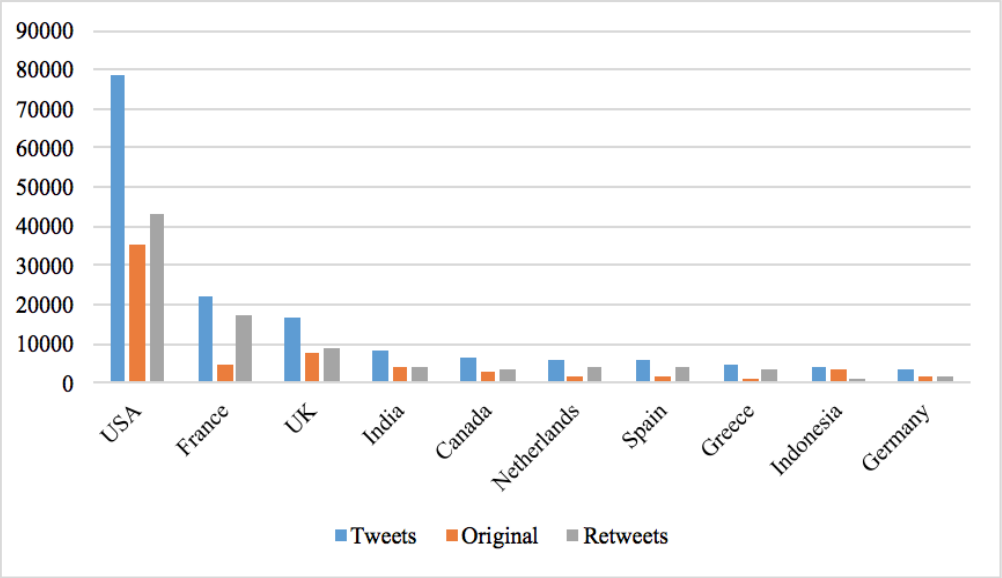

Figure 5-3 Counts of tweets, originals and retweets with links in the top 10 countries

5.2.2 Common three countries. Three countries were found in all the six stories used for this research study, USA, UK and India, and they represented 47% of all the tweets in the database. The analysis applied to the top 10 countries was repeated for the 3 common countries and the results are presented in the following subsections. Table 5-6 presents the top 10 countries, rank, and raw counts of tweets in each country.

5.2.2.1 Originals and retweets. Table 5-6 shows the tweets, originals, retweets, counts and ratios of originals and retweets to total tweets in each country. The 3 countries were not quite as similar to the overall ratio of original/retweet, 40/60, as the top 10 countries. Originals total in the 3 countries (59,409) is 52% of the total originals in the database, whereas retweets (72,831) is 44% of total retweets in the database. That may indicate that the top countries contributed more in creating originals than other countries.

Rank	Country	Tweets	Original	% Total	Retweets	% Total
1	USA	99,295	43,992	44%	55,303	56%
3	UK	21,316	9,776	46%	11,540	54%
4	India	11,629	5,641	49%	5,988	51%
Totals		132,240	59,409	45%	72,831	55%

Table 5-6 Counts of tweets, originals and retweets in the common 3 countries and percentages of originals and retweets to total tweets in the common 3 countries dataset

Figure 5-4 illustrate the raw counts of tweets, originals and retweets in the top 3 countries. USA generated about four times the number of tweets as UK, and UK generated twice as India.

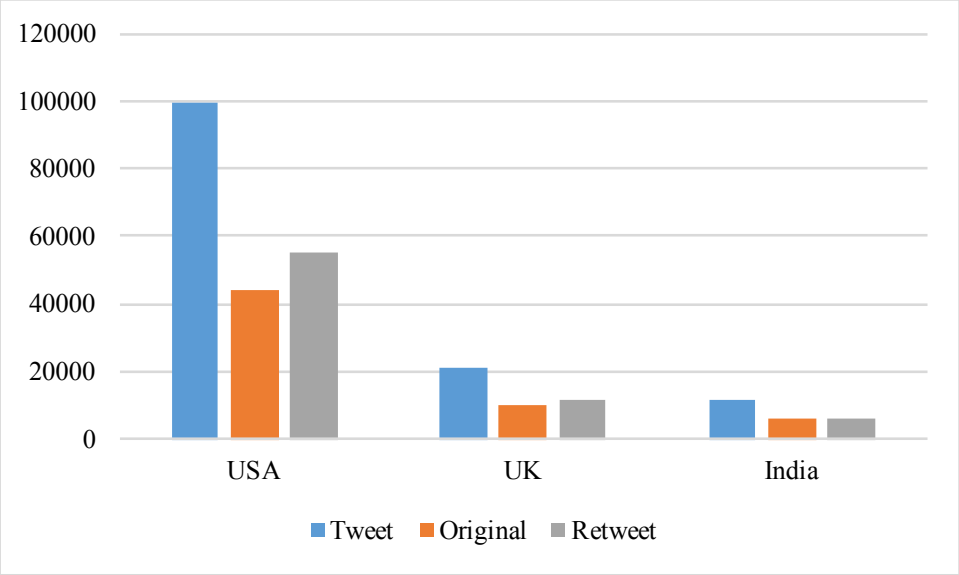

Figure 5-4 Tweets, originals and retweets counts in the common 3 countries

The distributions of the counts of tweets, originals, and retweets had somewhat similar patterns to the top 10 countries presented earlier. The common 3 countries generated around half of tweets, retweets and originals of the whole in the database. This

confirms the finding presented earlier, that few countries generated high proportion of tweets both originals and retweets.

5.2.2.2 Hashtags. Table 5-7 presents the raw counts of tweets with hashtags, percentages of tweets with hashtags to total tweets in each country, counts of originals and retweets with hashtags, and the percent of originals and retweets with hashtags to total tweets with hashtags in each country. The common 3 countries had similar ratios of tweets with hashtags, i.e., 55% of all the tweets of the common 3 countries contained hashtags.

Country	Tweets With Hashtags	%Total Tweets	Originals with Hashtags	% Tweets with Hashtags	Retweets with Hashtags	% Tweets with Hashtags
USA	54,118	55%	18,354	34%	35,764	66%
UK	11,870	56%	4,125	35%	7,745	65%
India	6,421	55%	2,415	38%	4,006	62%
Total	72,409	55%	24,894	34%	47,515	66%

Table 5-7 Counts of tweets, originals and retweets with hashtags and percentages of originals and retweets with hashtags to total tweets with hashtags in the common 3 dataset

Originals and retweets

The percentages of originals with hashtags among the 3 countries was 34%, i.e., a little more than one third of the tweets with hashtags were originals, and two thirds were retweets. The overall original to retweet ratio was 31% to 69%, in the common 3 countries the ratio was 34% to 66%.

Figure 5-5 illustrates the counts of tweets, originals and retweets with hashtags in the common countries. The figure highlights that hashtags occurred more often in

retweets than in originals, which is similar to the distribution in whole database, and in the top 10 countries.

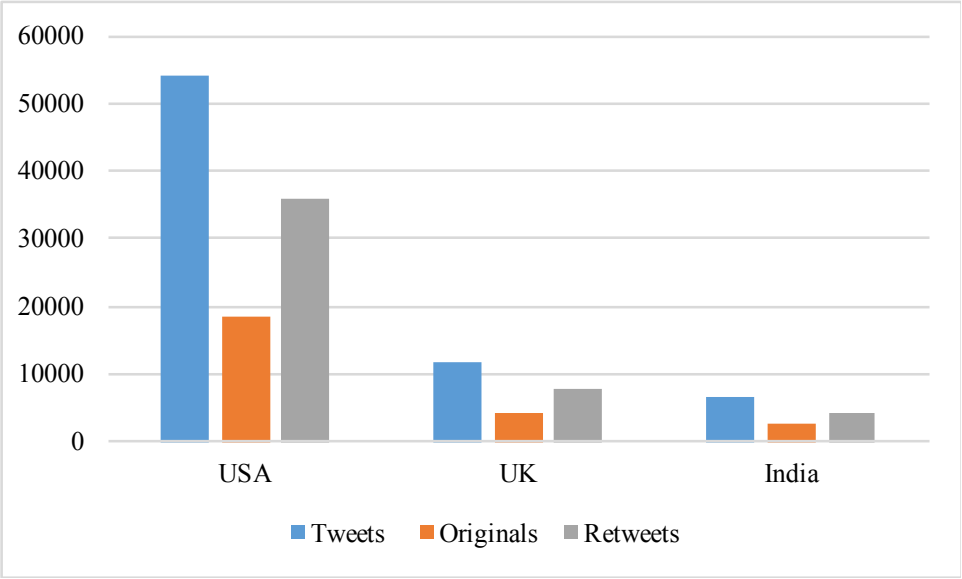

Figure 5-5 Counts of tweets, originals and retweets with hashtags in the common 3 countries

The USA generated 35% of all tweets with hashtags in the database, where the 3 countries combined generated 43% of all tweets with hashtags in the database.

5.2.2.3 Links. The use of links in the common 3 countries is presented in Table 5-8. The second column presents the total number of tweets with links in each country, and the third column presents the percentage of tweet with links to total tweets in each country. The last four columns show the number of originals with links, retweets with links and their percentage to total tweets with links in each country. USA generated 40% of tweets with links in the whole database, the 3 countries combined generated 48% of all tweets with links in the whole database.

Country	Tweets With Links	%Total Tweets	Originals with Links	% Tweets with Links	Retweets with Links	% Tweets with Links
USA	78,336	79%	35,443	45%	42,893	55%
UK	16,943	79%	7,848	46%	9,095	54%
India	8,450	73%	4,253	50%	4,197	50%
Total	103,729	78%	47,544	46%	56,185	54%

Table 5-8 Counts of tweets, originals and retweets with links and percentages of originals and retweets with links to total tweets with links in the common 3 dataset

Originals and retweets

The percentages of originals with links in the 3 countries was 46%, i.e., a little less than half of the tweets with links in each country were originals, and more than half of tweets with links were retweets.

Figure 5-6 shows the counts of tweets, originals and retweets with links in the common countries. The figure reflects that retweets with links were more frequent than originals with links within each country.

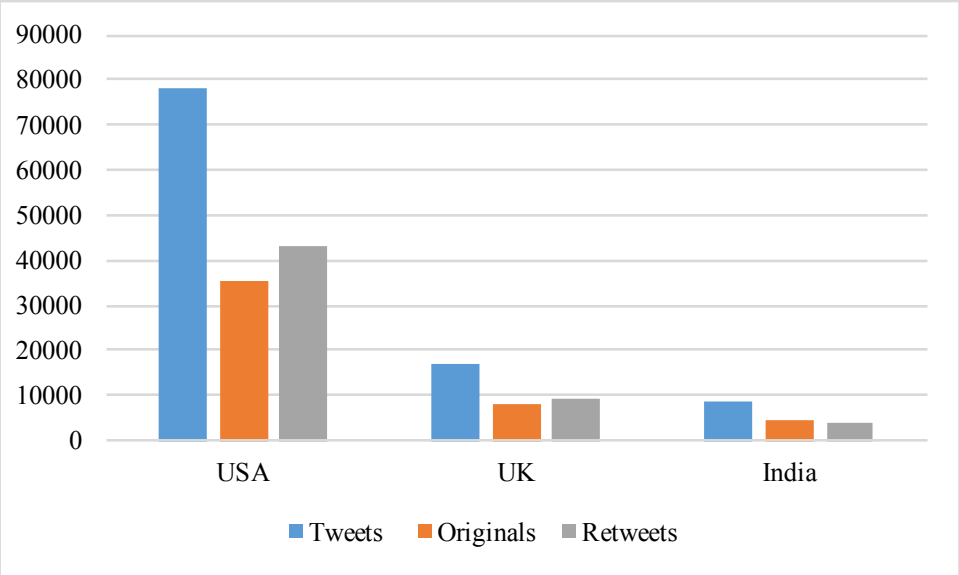

Figure 5-6 Counts of tweets, originals and retweets with links in the common 3 countries

From all the tweets with links in the database, the common countries showed higher ratios of originals with links (46%) than the whole database (42%). This may indicate that the common countries contributed more original content than retweets of existing posts. Also, link sharing activity in the common 3 was higher than hashtag sharing during the events used in this study.

5.2.3 Summary of results by country. The analyses above revealed that 73% of all tweets came from the top 10 countries and about half of tweets (47%) originated from the common three countries, USA, UK and India. The log log scale of the distributions of raw counts of all tweets from all countries in the database, tweets with hashtags, and tweets with links followed a power law distribution as presented in Figure 5-7. Figure 5-8 illustrates the log log scale of the distribution of raw counts of the top 10 countries, which also followed a power law distribution with a smaller slope compared to the chart with all countries.

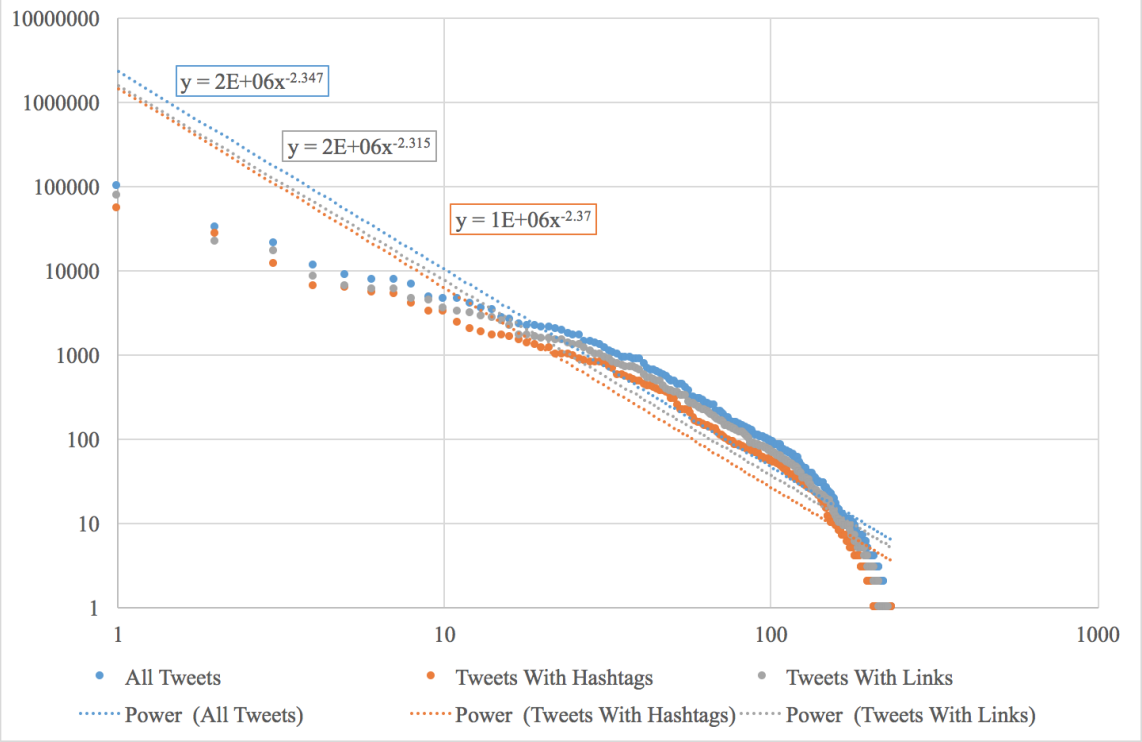

Figure 5-7 Log log scale of the distribution of raw counts of tweets, tweets with hashtags and tweets with links in all countries

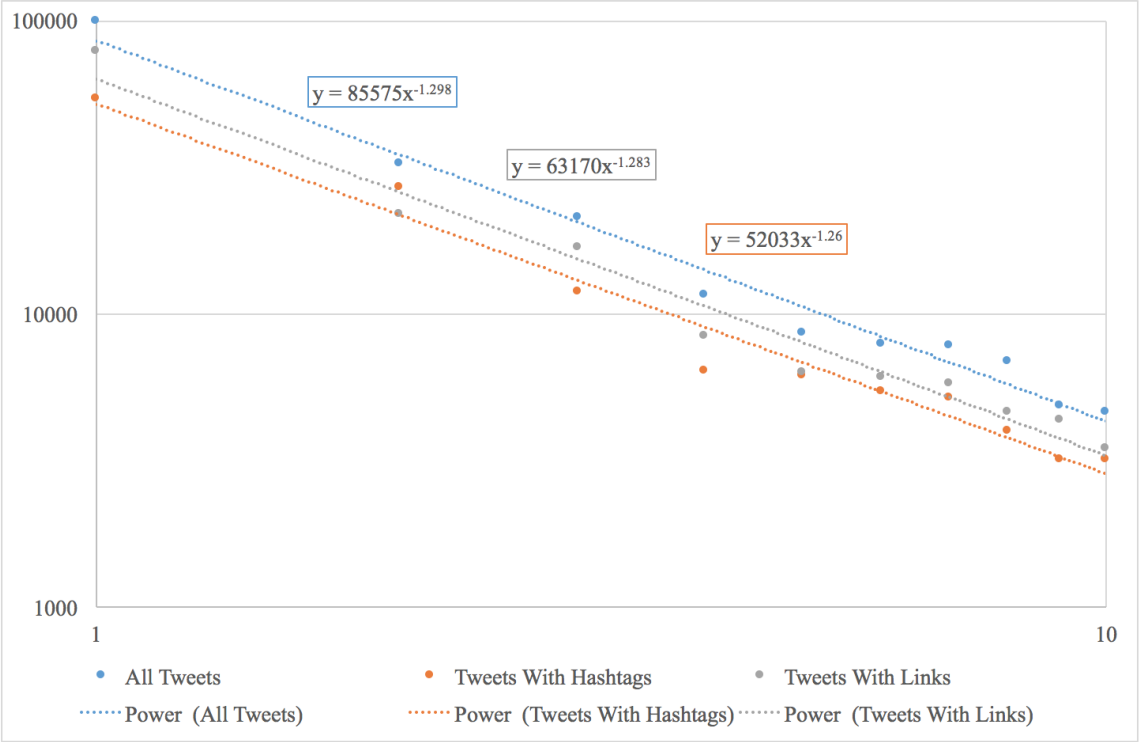

Figure 5-8 Log log scale of the distribution of raw counts of tweets, tweets with hashtags and tweets with links in the top 10 countries

The overall ratio of original to retweets tweets was 41% vs. 59%. A high retweet ratio has been associated with news with negative contents (Hansen, Arvidsson, Nielsen, Colleoni, & Etter, 2011), where people tend to forward the tweet containing the news to their followers, generating large amounts of tweets. Table 5-9 presents the summary of the results by country.

Dataset (# of Tweets)	Originals/ Retweet Ratio	Tweets with Hashtags	Tweets with links
All countries (280,436)	41/59	60%	77%
Originals/ Retweet Ratio		31/69	42/58
Top 10 (205,296)	41/59	61%	76%
Originals/ Retweet Ratio		31/69	42/58
Common 3 (132,240)	45/55	55%	78%
Originals/ Retweet Ratio		34/66	46/54

Table 5-9 percent of originals, retweets, hashtags and links within each dataset

The top 10 countries had similar percentages to the whole database, however USA, UK and India had somewhat higher ratios of originals to retweets than the overall database.

5.3 Characteristics by News Type

To examine characteristics of tweets by news type, and to a develop a sense of what differentiates the news types, news stories of the same type were collapsed into one dataset, resulting in three datasets by news type: finance, disaster and politics. In the following subsection the general descriptive statistics are presented for the three datasets. Following that, the characteristics of original, retweet, hashtag and link use of the top 10 countries are examined.

5.3.1 General characteristics by news type. We used the same characteristics for examining the results for each of the combined datasets as we did with whole test dataset presented in section 5.1. Table 5-10 shows the general characteristics of the combined datasets for each news type, the total count of tweets in all types was 280,436, i.e., the whole database. The column labeled '% Total' presents the percent of the characteristic to the total tweets in each news type, while the column labeled '% Original or Retweet' present the percent of the characteristic to total originals or retweet counts in each news type.

Characteristics	Finance			Disaster			Politics		
	Tweets Count	% Total	% Originals or Retweets	Tweets Count	% Total	% Originals or Retweets	Tweets Count	% Total	% Originals or Retweets
Total tweets	93,148			38,147			149,141		
Originals	59,179	64%		14,861	39%		40,655	27%	
Retweets	33,969	36%		23,286	61%		108,486	73%	
Tweets with Hashtags	28,635	31%		21,697	57%		119,185	80%	
Tweets with Links	84,439	91%		26,693	70%		103,565	69%	
Hashtags & Links	25,287	27%		15,290	40%		81,748	55%	
Originals & Hashtags	17,557	19%	30%	6,671	17%	45%	28,521	19%	70%
Originals & Links	54,234	58%	92%	11,248	29%	76%	23,684	16%	58%
Originals Hashtags & Links	15,595	17%	26%	4,326	11%	29%	14,357	10%	35%
Retweets & Hashtags	11,078	12%	33%	15,026	39%	65%	90,664	61%	84%
Retweets & Links	30,205	32%	89%	15,445	40%	66%	79,881	54%	74%
Retweets Hashtags & Links	9,692	10%	29%	10,964	29%	47%	67,391	45%	62%

Table 5-10 General characteristics of the combined datasets of each news type

Table 5-10 shows that the highest proportion of original tweets was in finance stories (64%) and 92% of these original tweets contained links. The lowest proportion of original tweets was in political stories (27%) and (39%) in disaster stories. The retweet ratio was highest in political stories (73%), with 84% of them containing hashtags. The patterns of behavior in the different news type were different from the patterns of the whole dataset.

The top 10 countries for each news type were identified and presented in Table 5-11. These countries are not the same countries used in the previous section which are the most frequent countries in the whole database.

Rank	Finance	Disaster	Politics
1	USA	USA	USA
2	UK	India	France
3	India	UK	UK
4	Canada	Spain	Netherlands
5	Nigeria	Indonesia	Greece
6	Indonesia	Germany	Spain
7	Australia	Mexico	Italy
8	Netherlands	Canada	Canada
9	Germany	France	India
10	France	Netherlands	Germany

Table 5-11 Top 10 countries in each news type

5.3.2 Finance stories. The first three columns of Table 5-12 display rank, common countries, and the number of tweets for each country for finance stories. USA

generated 42,226 tweets in finance dataset, which represented 45% of total tweets in finance, next UK (11%), and the rest of the countries made 5% or less each.

5.3.2.1 Original and retweets. Original tweets were more frequent than retweets in finance, 64% versus 36%, Table 5-12 presents the originals and retweets counts and percentages to the total tweets in each country in finance stories. Most countries had similar percentages of originals and retweets to the whole finance dataset except for Indonesia, which had higher percentage of originals (93%).

Rank	Country	Tweets	Originals	% Total	Retweets	% Total
1	**USA**	42,226	27,044	64%	15,182	36%
2	**UK**	10,061	6,538	65%	3,523	35%
3	**India**	4,976	3,417	69%	1,559	31%
4	**Canada**	4,081	2,463	60%	1,618	40%
5	**Nigeria**	1,897	1,321	70%	576	30%
6	**Indonesia**	1,845	1,717	93%	128	7%
7	**Australia**	1,748	996	57%	752	43%
8	**Netherlands**	1,598	1,017	64%	581	36%
9	**Germany**	1,465	893	61%	572	39%
10	**France**	1,318	704	53%	614	47%
Total		71,215	46,110	65%	25,105	35%

Table 5-12 Tweets, originals and retweets counts in the top 10 countries and percentages to total tweets in each country in finance

Figure 5-9 illustrates the distribution of tweets, originals and retweets in the top 10 countries in finance. USA generated around 42,000 tweets, UK 10,000 tweets, which around 1/4 the count of USA, India generated 5,000 tweets, which is around 1/8 the count of USA and 1/2 the count of UK.

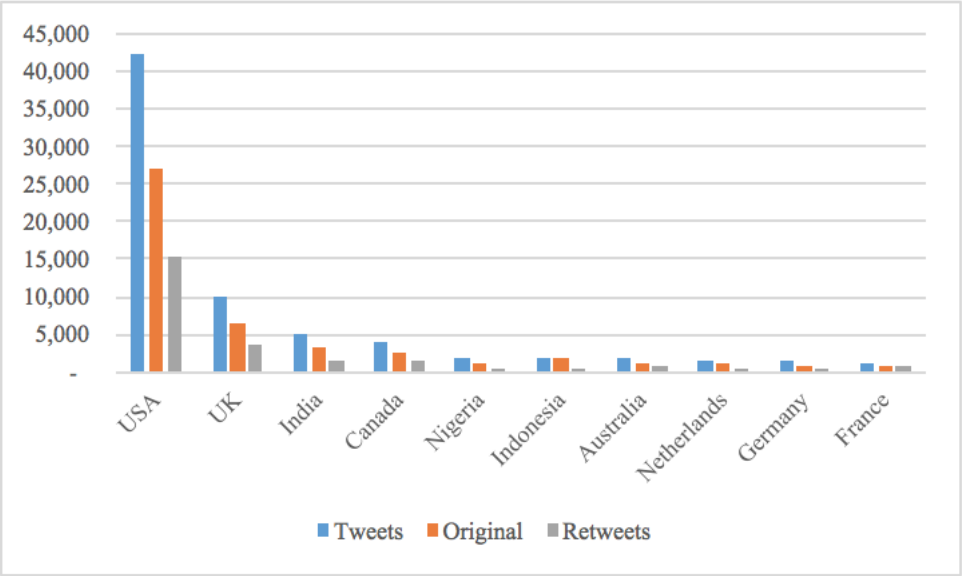

Figure 5-9 Tweets, originals and retweets counts in the top 10 countries in finance dataset

5.3.2.2 Hashtags. Table 5-10 shows that 31% of finance tweets contained hashtags. Table 5-13 shows the counts of tweets with hashtags, originals with hashtags and retweets with hashtags and the percentages to total tweets with hashtags in each country. The use of hashtags in most of the top 10 countries was close to 31%, the overall percent of the finance dataset, the percentages ranged from 29% - 47%.

Originals and retweets

61% of finance tweets with hashtags were originals and 39% were retweets. The countries followed a similar distribution of originals with hashtags and retweets with hashtags as the whole finance dataset, except Indonesia with (95%) of originals with hashtags.

Country	Tweets with Hashtags	%Total Tweets	Originals with Hashtags	% Tweets with Hashtags	Retweets with Hashtags	% Tweets with Hashtags
USA	12,037	29%	7,674	64%	4,363	36%
UK	3,305	33%	1,903	58%	1,402	42%
India	1,485	30%	1,063	72%	422	28%
Canada	1,170	29%	616	53%	554	47%
Nigeria	567	30%	254	45%	313	55%
Indonesia	704	38%	667	95%	37	5%
Australia	770	44%	436	57%	334	43%
Netherlands	534	33%	338	63%	196	37%
Germany	688	47%	464	67%	224	33%
France	528	40%	291	55%	237	45%
Total	21,788	31%	13,706	63%	8,082	37%

Table 5-13 Counts of tweets, originals and retweets with hashtags in the top 10 countries and percentages to total tweets in each country in the finance dataset

Figure 5-10 illustrates the distribution of tweets, originals and retweets with hashtags in top 10 countries. USA generated around 12,000 tweets with hashtags, next UK with less than 4,000. All countries had higher counts of originals with hashtags except for Nigeria, where retweets with hashtags were more than originals. That may be due to the Chinese shares drop being negative news for Nigeria, so high retweet ratio is expected (Hansen et al., 2011).

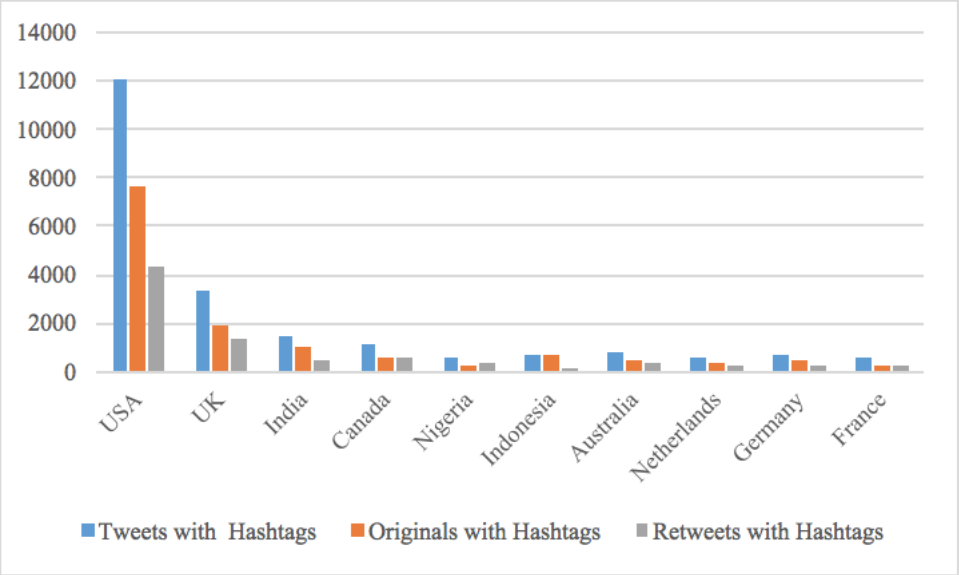

Figure 5-10 Counts of tweets, originals and retweets with hashtags in the top 10 countries in the finance dataset

The top 10 countries made 76% of the total finance tweets with hashtags. USA represented 42% of all finance tweets with hashtags, 44% of finance originals with hashtags, and 39% of retweets with hashtags, next UK with around 12%, 11% and 13%, and the rest of the countries made around 20%, 21% and 19% in total. USA, India, and Indonesia had higher ratio of originals with hashtags than the other countries.

5.3.2.3 Links. Table 5-14 displays the counts of tweets with links, percent to total finance tweets in each country, originals, retweets with links and the percentages to total tweets with links in each country. 90% of the tweets of the top 10 countries contained links.

Country	Tweets with Links	%Total Tweets	Originals with Links	%Tweets with links	Retweets with Links	% Tweet with links
USA	38,278	91%	24,902	65%	13,376	35%
UK	8,945	89%	5,918	66%	3,027	34%
India	4,298	86%	2,903	68%	1,395	32%
Canada	3,615	89%	2,280	63%	1,335	37%
Nigeria	1,855	98%	1,293	70%	562	30%
Indonesia	1,793	97%	1,671	93%	122	7%
Australia	1,580	90%	901	57%	679	43%
Netherlands	1,443	90%	935	65%	508	35%
Germany	1,354	92%	829	61%	525	39%
France	1,153	87%	578	50%	575	50%
Total	64,314	90%	42,210	66%	22,104	34%

Table 5-14 Counts of tweets, originals and retweets with links in the top 10 countries and percentages to total tweets in each country in the finance dataset

Originals and retweets

In finance dataset 91% of tweets contained links. 64% of finance tweets with links were originals and 36% were retweets. Similar to hashtags distribution, most countries followed the same distribution of originals with links and retweets with links as the whole finance dataset, except Indonesia (93%).

Figure 5-11 shows the distribution of tweets, originals and retweets with links in the top 10 countries. The counts of originals with links were more than retweets with links across all the top 10 countries in finance.

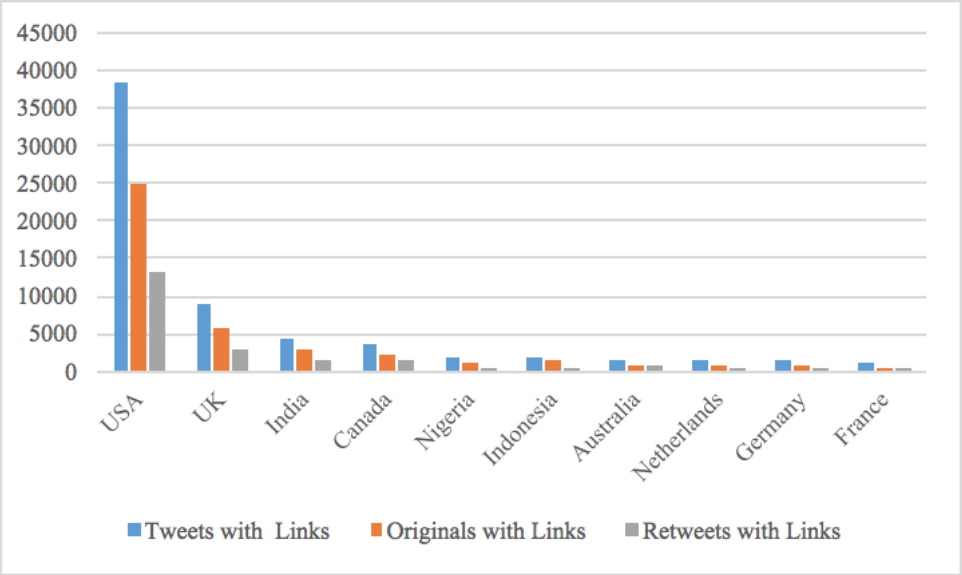

Figure 5-11 Counts of tweets, originals and retweets with links in the top 10 countries in the finance dataset

All countries showed consistent percentages of tweets, originals and retweets, except for Indonesia, which had a higher ratio of originals compared to retweets. In general, all the top 10 countries for finance stories had more originals, more originals with hashtags and more originals with links than retweets. Within each country, the ratio of originals was high, with Indonesia the highest ratio of originals (93%).

5.3.3 Disaster stories. The first 3 columns of Table 5-15 presents the top 10 countries in the disaster dataset. Four countries were common among the two stories, USA, India, UK and Indonesia.

5.3.3.1 Originals and retweets. Table 5-15 presents the counts of originals and retweets and percentages to total tweets in each country in the disaster dataset. The percentages of originals and retweet in the disaster dataset were 39% and 61% as

presented in Table 5-10. The top 10 countries followed this pattern except for Indonesia (73/27) and Mexico (61/39).

Rank	Country	Tweets	Original	% Total	Retweets	% Total
1	**USA**	12,847	4,690	37%	8,157	63%
2	**India**	3,990	1,363	34%	2,627	66%
3	**UK**	2,177	815	37%	1,362	63%
4	**Spain**	1,830	609	33%	1,221	67%
5	**Indonesia**	1,367	1,004	73%	363	27%
6	**Germany**	1,001	462	46%	539	54%
7	**Mexico**	971	592	61%	379	39%
8	**Canada**	783	303	39%	480	61%
9	**France**	775	277	36%	498	64%
10	**Netherlands**	723	285	39%	438	61%
Totals		26,464	10,400	39%	16,064	61%

Table 5-15 Counts of Tweets, originals and retweets and percentages to total tweets in each country in the disaster dataset

Figure 5-12 presents the raw counts of tweet, originals and retweets in the top 10 counties in the disaster dataset. In the disaster dataset, only Indonesia and Mexico had higher original to retweet ratios.

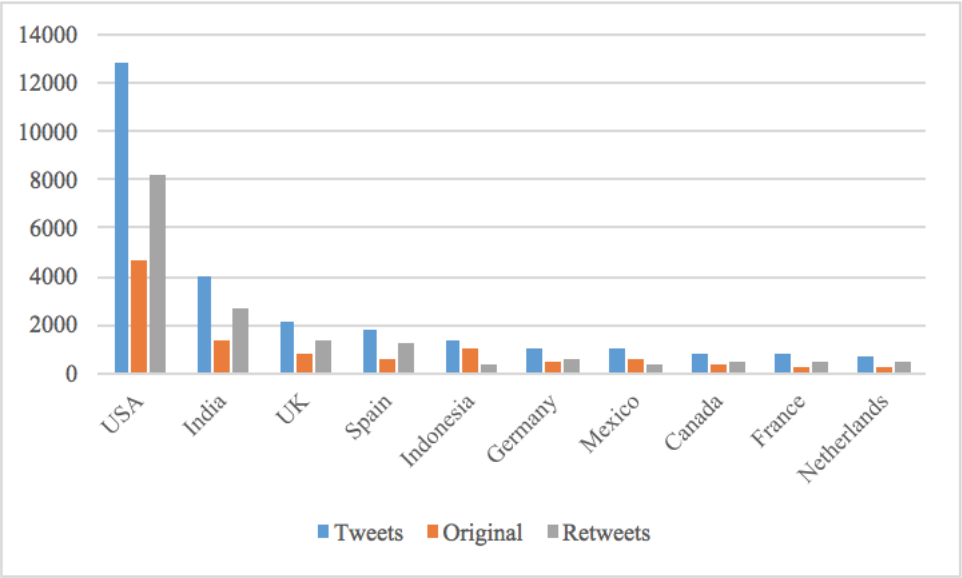

Figure 5-12 Raw counts of tweets, originals and retweets in the top 10 countries in the disaster dataset

USA generated 1/3 of disaster tweets, retweets and originals, next India, generated 10% and UK 5%.

5.3.3.2 Hashtags. Table 5-16 presents the number of tweets, originals, retweets with hashtags, and percentages to total tweets in the disaster dataset. More than 50% of the top 10 countries tweets contained hashtags, except Indonesia (31%) and Mexico (36%). USA presented around 33% of total disaster tweets with hashtags, India 13% and UK 6%.

Originals and retweets

According to Table 5-10, 57% of disaster dataset tweets contained hashtags, 31% of these were originals and 69% were retweets. The top 10 countries in the disaster dataset had the same pattern, except for Indonesia which had 49% originals and 51% retweets, and Germany 43% original and 57% retweets as presented in Table 5-16.

Country	Tweets with Hashtags	%Total Tweets	Originals with Hashtags	% Tweets with Hashtags	Retweets with Hashtags	% Tweets with Hashtags
USA	7,137	56%	2,128	30%	5,009	70%
India	2,855	72%	762	27%	2,093	73%
UK	1,207	55%	381	32%	826	68%
Spain	1,322	72%	383	29%	939	71%
Indonesia	427	31%	209	49%	218	51%
Germany	681	68%	292	43%	389	57%
Mexico	346	36%	125	36%	221	64%
Canada	442	56%	139	31%	303	69%
France	505	65%	175	35%	330	65%
Netherlands	416	58%	151	36%	265	64%
Total	15,338	58%	4,745	31%	10,593	69%

Table 5-16 Counts of tweets, originals and retweets with hashtags in the top 10 countries and percentages to total tweets in each country in disaster dataset

Figure 5-13 illustrates the distribution of tweets, originals and retweets with hashtags in the top 10 countries, USA generated around 7,000 tweets which is roughly third the disaster dataset, next in rank India, around 3,000, Spain around 1,300, UK 1,200, and the rest of the countries generated around 500 or fewer tweets with hashtags. The chart also shows that hashtags occurred more frequently in retweets than in originals.

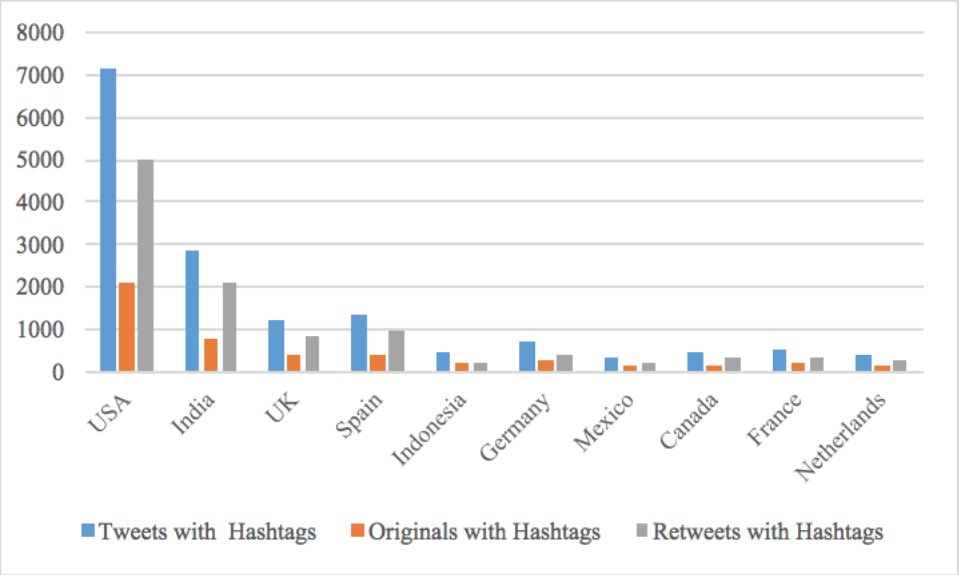

Figure 5-13 Counts of tweets, originals and retweets with hashtags in the top 10 countries in disaster dataset

5.3.3.3 Links. Table 5-14 displays the counts of tweets with links, percent to total disaster tweets in each country, originals, retweets with links and the percentages to total tweets with links in each country. More than 60% of disaster tweets contained links in all the top 10 countries.

Originals and retweets

Links were in 70% of disaster dataset tweets, 42% were originals and 58% were retweets. The top 10 countries followed a similar pattern, except for Indonesia and Mexico. This difference could be because the counts of originals in these countries were higher, therefore the pattern of original\retweet with hashtags and links followed the pattern original/retweet distribution.

Country	Tweets with Links	%Total Tweets	Originals with Links	%Tweets with Links	Retweets with Links	%Tweets with Links
USA	9,113	71%	3,556	39%	5,557	61%
India	2,428	61%	829	34%	1,599	66%
UK	1,487	68%	584	39%	903	61%
Spain	1,277	70%	411	32%	866	68%
Indonesia	1,229	90%	955	78%	274	22%
Germany	618	62%	298	48%	320	52%
Mexico	751	77%	501	67%	250	33%
Canada	580	74%	234	40%	346	60%
France	498	64%	182	37%	316	63%
Netherlands	491	68%	211	43%	280	57%
Total	18,472	70%	7,761	42%	10,711	58%

Table 5-17 Counts of tweets, originals and retweets with links in the top 10 countries and percentages to total tweets in each country in the disaster dataset

Figure 5-14 illustrates the counts of tweets, originals and retweets with links in the top 10 countries, the distribution was similar to the distribution of the top 10 countries in the whole database.

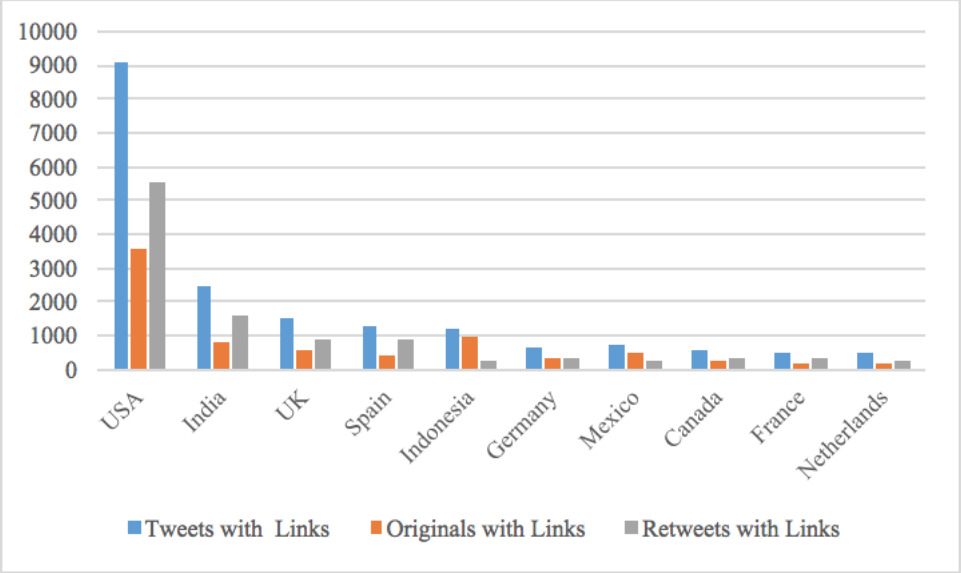

Figure 5-14 Counts of tweets, originals and retweets with links in the top 10 countries in disaster dataset

USA accounted for 34% of total disaster tweets with links, India 9% and UK 6%. Most countries had higher retweet with links ratio, except for Indonesia, which had a higher ratio of originals with links compared to retweets. This may indicate that during disaster events, the top countries were retweeting more than creating new tweets compared with all the other countries in the disaster dataset.

Generally, in the disaster dataset, all countries followed the overall original/retweet ratio except for Indonesia and Mexico, which had higher original to retweet ratios. These ratios were reflected on original/retweet ratios in tweets with hashtags and tweets with links in these two countries.

5.3.4 Politics stories. Table 5-18 presents the top 10 countries in politics dataset and the number of tweets in each country. Five countries from the top ten in each politics

story were common among the two stories, USA, France, UK, Canada and India. USA presented 30%, next France 20% and UK 6% of all politics dataset.

5.3.4.1 Originals and retweets. Table 5-18 presents the counts of originals, retweets and their percentages to total tweets in each country. Most of politics tweets were retweets, 73% retweets versus 27% originals. The top 10 countries followed the same pattern as the whole politics dataset.

Rank	Country	Tweets	Original	% Total	Retweets	% Total
1	**USA**	44,222	12,258	28%	31,964	72%
2	**France**	30,370	7,149	24%	23,221	76%
3	**UK**	9,078	2,423	27%	6,655	73%
4	**Netherlands**	6,283	1,498	24%	4,785	76%
5	**Greece**	5,281	1,095	21%	4,186	79%
6	**Spain**	4,925	1,223	25%	3,702	75%
7	**Italy**	3,056	1,066	35%	1,990	65%
8	**Canada**	2,990	781	26%	2,209	74%
9	**India**	2,663	861	32%	1,802	68%
10	**Germany**	2,143	674	31%	1,469	69%
Totals		111,011	29,028	26%	81,983	74%

Table 5-18 Counts of tweets, originals and retweets in the top 10 countries and percentages to total tweets in each country in the politics stories

Figure 5-15 presents counts of tweets, originals and retweets in the top 10 countries in the politics dataset. The counts of retweets were higher in all the top 10 countries in the politics dataset.

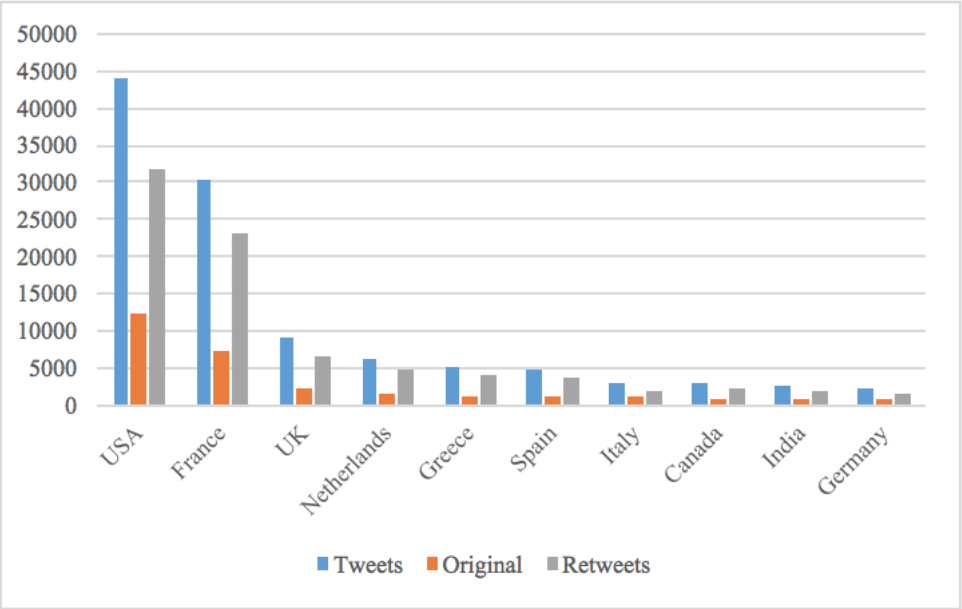

Figure 5-15 Tweets, originals and retweets counts in the common countries in politics stories

5.3.4.1 Hashtags. Table 5-19 presents tweets with hashtags, percent of these to total tweets in each country, originals with hashtags, retweets with hashtags and percentages to total tweets with hashtags in each of the top 10 countries in the politics dataset. USA presented around 29% of total politics tweets with hashtags, France 22% and UK 6%.

Originals and retweets

80% of tweets in politics dataset contained hashtags, 76% of tweets with hashtags were retweets, and 24% were originals. The top 10 countries were similar to the overall proportions.

Country	Tweets with Hashtags	%Total Tweets	Originals with Hashtags	%Tweets with Hashtags	Retweets with Hashtags	%Tweets with Hashtags
USA	34,944	79%	8,552	24%	26,392	76%
France	25,913	85%	5,727	22%	20,186	78%
UK	7,358	81%	1,841	25%	5,517	75%
Netherlands	5,232	83%	1,158	22%	4,074	78%
Greece	4,438	84%	808	18%	3,630	82%
Spain	3,787	77%	861	23%	2,926	77%
Italy	2,557	84%	847	33%	1,710	67%
Canada	2,393	80%	548	23%	1,845	77%
India	2,081	78%	590	28%	1,491	72%
Germany	1,820	85%	540	30%	1,280	70%
Total	90,523	82%	21,472	24%	69,051	76%

Table 5-19 Counts of tweets, originals and retweets with hashtags in the top 10 countries and percentages to total tweets in each country in politics dataset

Figure 5-16 displays the distribution of tweets, originals and retweets with hashtags in the common countries in politics. USA generated 35,000 tweets, and France generated a little more than 25,000. All countries had higher counts of retweets with hashtags than originals with hashtags.

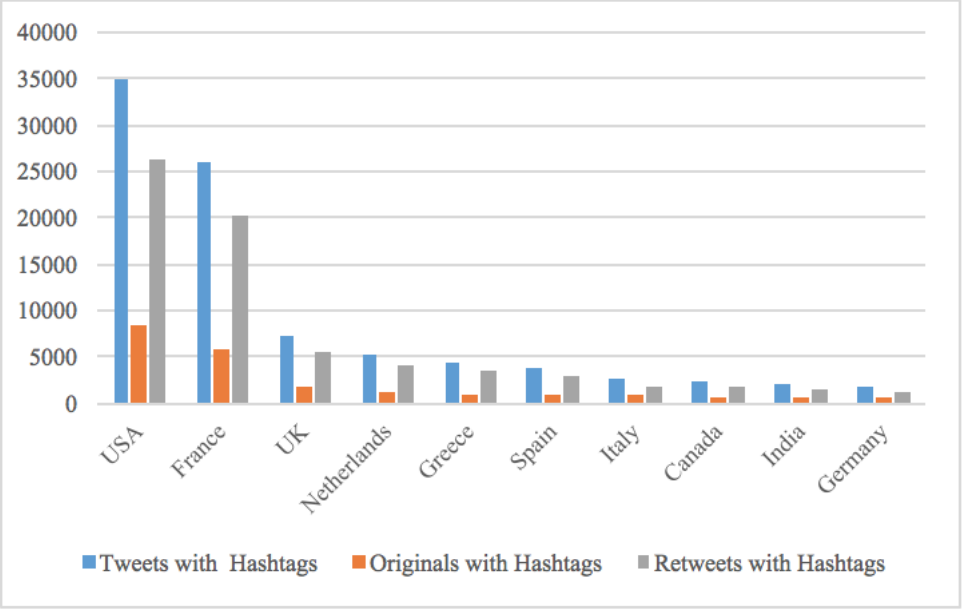

Figure 5-16 Counts of tweets, originals and retweets with hashtags in the top 10 countries in politics dataset

In politics, the analysis showed that all the top 10 countries had more retweets with hashtags than originals with hashtags. Similar behavior was observed in the disaster dataset.

5.3.4.2 Links. Table 5-20 presents tweets with links, percent of these to total tweets in each country, originals with links, retweets with links and percentages to total tweets with links in each of the top 10 countries in the politics dataset. USA presented 30% of total politics tweets with links, France 19% and UK 6%.

Originals and retweets

69% of politics tweets contained links, 23% of them were original, and 77% were retweets. All top countries had lower percentages of originals with links, which is similar to the whole politics dataset with few variations.

Country	Tweets with Links	%Total Tweets	Originals with Links	%Tweets with Links	Retweets with Links	%Tweets with Links
USA	30,945	70%	6,985	23%	23,960	77%
France	20,189	66%	3,720	18%	16,469	82%
UK	6,511	72%	1,346	21%	5,165	79%
Netherlands	4,126	66%	757	18%	3,369	82%
Greece	3,313	63%	531	16%	2,782	84%
Spain	3,549	72%	801	23%	2,748	77%
Italy	2,057	67%	629	31%	1,428	69%
Canada	2,145	72%	439	20%	1,706	80%
India	1,724	65%	521	30%	1,203	70%
Germany	1,527	71%	431	28%	1,096	72%
Total	76,086	69%	16,160	21%	59,926	79%

Table 5-20 Counts of tweets, originals and retweets with links in the top 10 countries and percentages to total tweets in each country in politics dataset

Figure 5-17 displays the distribution of tweets, originals and retweets with links in the top 10 countries in politics dataset. The distribution of original/retweets with links followed the overall original/retweet pattern of the politics dataset.

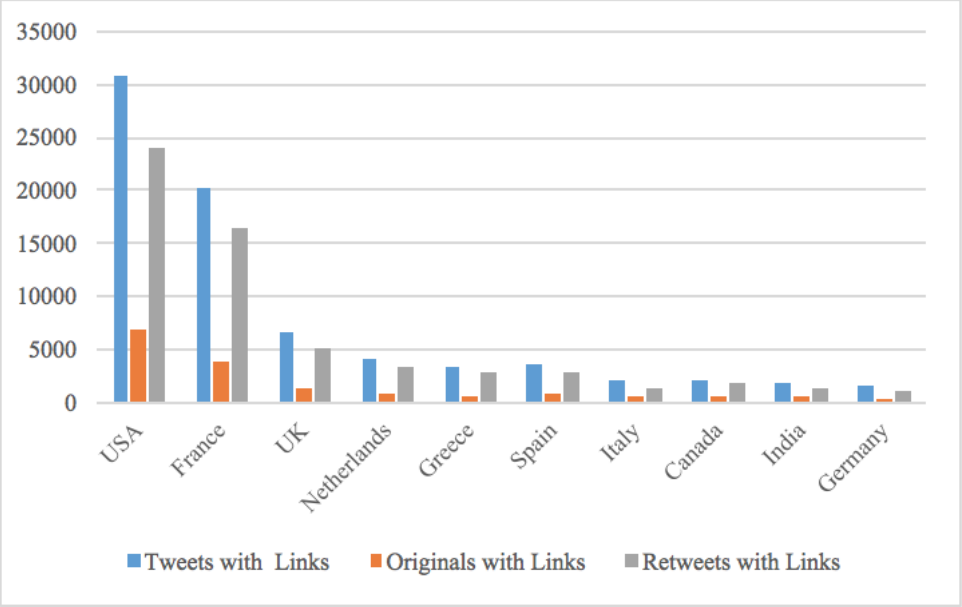

Figure 5-17 Counts of tweets, originals and retweets with links in the top 10 countries in politics dataset

USA presented 30% of tweets, originals and retweets of the whole politics dataset. France presented more than 20% of retweets and 15% of originals, that indicates that France contributed proportionately more retweets to the politics dataset than originals.

5.3.5 Summary of results by news type. The characteristics of tweets by news type exhibited different distributions of originals vs retweets, tweets with hashtags, and links across the different news types. Table 5-21 summarizes the results by news type. Finance had the highest originals ratio (64%), while politics had the highest retweets ratio (73%). Finance tweets had the most links (91%), and political tweets had the most hashtags (80%). USA represented 45% of all finance tweets, UK 11%, i.e., more than half of all finance tweets originated from USA and UK alone. In disaster stories, the common three countries, USA, UK and India contributed around half of total tweets. France was in

the second place in the top 10 countries in politics tweets because of the Charlie Hebdo story, which represented 20% of total political tweets. The top 10 countries followed the overall pattern of original/retweet ratios in tweets with hashtags and links in each dataset. This is likely because these countries represented such a high proportion of each dataset. Indonesia, however, showed different behavior in many scenarios, but since the total counts of Indonesia's tweets is relatively small, the different pattern Indonesia exhibited did not affect the overall results. This finding agrees with the finding of (Poblete, Garcia, Mendoza & Jaimes, 2011) to some extent. They studied the top 10 countries they found active in Twitter, and reported that Indonesia ranked first in tweets per user, and had the fewest retweets, compared to the other countries.

	Finance	Disaster	Politics	Overall
Originals	64%	39%	27%	41%
Hashtags	31%	57%	80%	60%
Links	91%	70%	69%	77%

Table 5-21 Summary of results by type

Figure 5-18 illustrates the distribution of tweets in finance, disaster and politics across all participating countries in a log log scale. Figure 5-19 shows the distribution of tweets in the different news types across the top 10 countries in each news type. The distributions followed a power law, although the top 10 charts have less steep slopes, 1.38, 1.47 and 1.23 for finance, disaster and politics respectively. The distributions of tweets with hashtags and links also followed a power law distribution.

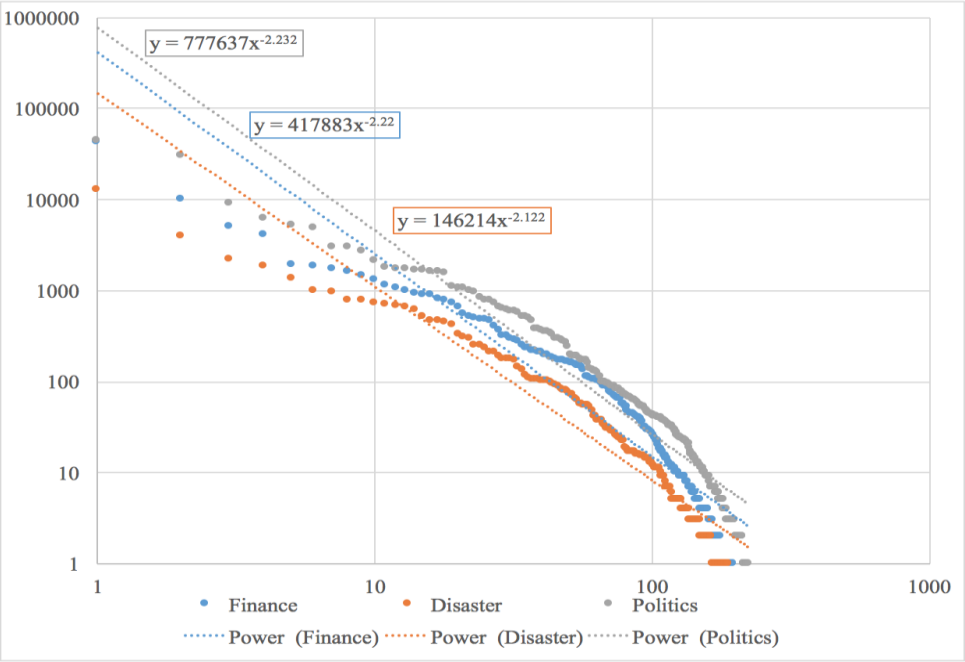

Figure 5-18 Log log scale of the distribution of tweets in finance, disaster and politics across all participating countries

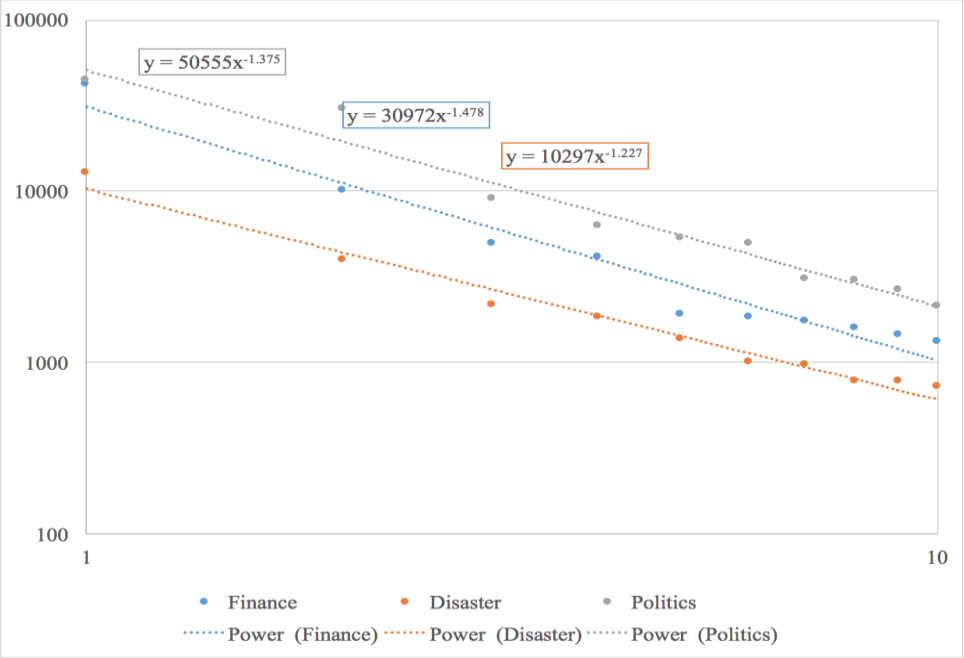

Figure 5-19 Log log scale of the distribution of tweets in finance, disaster and politics across the top 10 countries

5.4 Characteristics by News Story

In the previous section, we examined the characteristics for each news type in general, in this section we analyze each individual story. The stories representing each news type were introduce in section 4.1.1. The stories within each news type had different sizes, for instance in finance, Chinese shares story had 6,784 tweets and Volkswagen story had 86,364 tweet. However, the analysis revealed similar ratios among the stories of the same type in most of the characteristics examined, as presented in Table 5-22, Table 5-23 and Table 5-24, especially in the ratios of tweets with hashtags and tweets with links.

In the following subsections, the general characteristics are introduced by news type and by story. Similar to general characteristics introduced in previous sections, the general characteristics include the number of tweets, originals, retweets, number of countries, number of tweets with hashtags and links. The first column of Table 5-22 shows the full list of the general characteristics examined.

5.4.1 Finance stories. The two finance stories were the Chinese shares market fall and the Volkswagen scandal. Table 5-22 displays the general characteristics for each finance story. Although the two datasets differed in size, 6,700 versus 46,000, they shared similar characteristics. They both had higher original to retweet ratios (81%, 62%). Links were found in 92% of tweets in Chinese share story, and in 91% in Volkswagen story. Hashtags were found in 19% of Chinese shares and 38% of Volkswagen tweets.

	Chinese Shares			Volkswagen		
	Count	%Total	%Original/ Retweets	Count	%Total	%Original/ Retweets
Total Tweets	6,784			86,364		
Unique Users	4,014			46,219		
Avg. Tweets/ User	1.69			1.87		
Unique Countries	114			196		
Originals	5,485	81%		53,694	62%	
Avg. Originals/ User	1.37			1.16		
Retweets	1,299	19%		32,670	38%	
Avg. Retweet/ User	0.32			0.71		
Tweets with Hashtags	1,566	23%		27,069	31%	
Tweets with Links	6,273	92%		78,166	91%	
Hashtags & Links	1,483	22%		23,804	28%	
Unique Hashtags	480			3,585		
Unique Links	4,712			46,883		
Originals & Hashtags	1,279	19%	23%	16,278	19%	30%
Originals & Links	5,067	75%	92%	49,167	57%	92%
Originals Hashtags & Links	1,213	18%	22%	14,382	17%	27%
Retweets & Hashtags	287	4%	22%	10,791	12%	33%
Retweets & Links	1,206	18%	93%	28,999	34%	89%
Retweets Hashtags & Links	270	4%	21%	9,422	11%	29%

Table 5-22 General characteristics of the two finance stories

These results are consistent with our earlier finding that finance stories can be characterized by having high original/retweet ratio, and contain high proportion of links.

5.4.2 Disaster stories. Disaster stories were the Germanwings airlines plane crash and the Nepal earthquake. The general characteristics of the two stories are shown in Table 5-23. Originals ratio was 52% in the Germanwings, and 28% in the Nepal earthquake story. Both stories had a little more than 50% of tweets with hashtags, roughly 70% of tweets with links, and nearly 75% of originals in both stories contained links.

	Germanwings			Nepal Earthquake		
	Count	%Total	%Original/ Retweets	Count	%Total	%Original/ Retweets
Total Tweets	17,351			20796		
Unique Users	15,282			19060		
Avg. Tweets/ User	1.14			1.09		
Unique Countries	160			177		
Originals	9,064	52%		5797	28%	
Avg. Originals/ User	0.59			0.3		
Retweets	8,287	48%		14999	72%	
Avg. Retweet/ User	0.54			0.79		
Tweets with Hashtags	9,194	53%		12503	60%	
Tweets with Links	13,086	75%		13607	65%	
Hashtags & Links	6,456	37%		8834	42%	
Unique Hashtags	2,125			2041		
Unique Links	10,522			9123		
Originals & Hashtags	3,513	20%	39%	3158	15%	54%
Originals & Links	7,188	41%	79%	4060	20%	70%
Originals Hashtags & Links	2,343	14%	26%	1983	10%	34%
Retweets & Hashtags	5,681	33%	69%	9345	45%	62%
Retweets & Links	5,898	34%	71%	9547	46%	64%
Retweets Hashtags & Links	4,113	24%	50%	6851	33%	46%

Table 5-23 General characteristics of the two disaster stories

The two disaster stories in this dataset did not have same original/retweet ratio. This may be related to the fact that the Germanwings story had different countries in its top 10 list, and we have seen in the previous sections that countries have different behaviors in general. Additionally, this story was not as purely a natural disaster story as the Nepal earthquake story. Soon after the crash took place, headlines started announcing information about the co-pilot's medical history, and legal issues associated with incident. So, we anticipate that the discussion over Twitter took other directions than a normal disaster story might take. Although the Nepal earthquake had 6% more hashtags, and 10% fewer links, these percentages are similar to some extent

5.4.3 Politics stories. The two politics stories were the Charlie Hebdo shooting and the story of the boy who was arrested for inventing a clock that his teacher thought was a bomb. The result of the general characteristics analysis is presented in Table 5-24. The two datasets varied in size, 170,000 versus 21,000, nevertheless, similarities between their characteristics can be observed. Both stories had high retweets percentages around 75%, and high percentages of tweets with hashtags, (79% in Charlie Hebdo story, and 87% in clock boy story). More than 80% of retweets contained hashtags in both stories. A high number of tweets contained links in both stories, (69% and 76%).

	Charlie Hebdo			**Clock Boy**		
	Count	%Total	%Original/ Retweets	Count	%Total	%Original/ Retweets
Total Tweets	133409			15732		
Unique Users	114159			15168		
Avg. Tweets/ User	1.17			1.04		
Unique Countries	223			164		
Originals	37211	28%		3444	22%	
Avg. Originals/ User	0.33			0.23		
Retweets	96198	72%		12288	78%	
Avg. Retweet/ User	0.84			0.81		
Tweets With Hashtags	105525	79%		13660	87%	
Tweets With Links	91665	69%		11900	76%	
Hashtags & Links	71382	54%		10366	66%	
Unique Hashtags	10575	8%		820	5%	
Unique Links	52046	39%		4161	26%	
Originals & Hashtags	25748	19%	69%	2773	18%	81%
Originals & Links	21973	16%	59%	1711	11%	50%
Originals Hashtags & Links	13090	10%	35%	1267	8%	37%
Retweets & Hashtags	79777	60%	83%	10887	69%	89%
Retweets & Links	69692	52%	72%	10189	65%	83%
Retweets Hashtags & Links	58292	44%	61%	9099	58%	74%

Table 5-24 General characteristics of the two politics stories

The politics stories were somewhat consistent in the original/retweet, tweets with hashtags, tweets with links ratios.

5.4.4 Top ten countries. Tweets generated for each news story came from different parts of the globe. However, each story had a different geographic distribution of tweets. Table 5-25 presents the top 10 countries for each news story. For these countries, we extracted the number of retweets, originals, tweets with hashtags and tweets with links. These results are presented in Appendix A.

USA, UK and India appeared in all the six stories, Indonesia in five stories, France, Netherlands, Canada and Australia in four stories, Germany in three stories, Nigeria, Greece and Spain in two, and the rest of the countries appeared in one story only, including Mexico, Venezuela, Nepal, Saudi Arabia, Pakistan, Malaysia, Thailand, Philippines and Italy.

No.	Chines Shares	Volkswagen	Germanwings	Nepal Earthquake	Charlie Hebdo	Clock Boy
1	USA	USA	USA	USA	USA	USA
2	UK	UK	Spain	India	France	UK
3	India	India	Indonesia	UK	UK	Malaysia
4	Nigeria	Canada	UK	Canada	Netherlands	Canada
5	Indonesia	Indonesia	Germany	Nepal	Greece	India
6	Australia	Australia	Mexico	Australia	Spain	Saudi Arabia
7	Greece	Nigeria	Venezuela	Thailand	Italy	Indonesia
8	Netherlands	Netherlands	France	Brazil	Canada	France
9	Canada	Germany	Netherlands	Indonesia	India	Pakistan
10	France	France	India	Philippines	Germany	Australia

Table 5-25 Top 10 countries for each news story

5.4.5 Top ten hashtags analysis. For each news story the tweets were processed to extract hashtags, calculate the frequency of each hashtag and calculate the distribution

of hashtags. After examining the list of hashtags sorted by frequency, it was noticed that hashtags in the top of the list were more meaningful and relevant to the event than the ones in the bottom of the list. Therefore, more processing was applied to the top ten hashtags to find the counts and ratios of tweets, retweets and originals that include any of the top ten hashtags. The result of hashtags analysis shows that the top 10 hashtags are used heavily in all the 6 stories compared to the rest of the hashtags. For instance, 63% of hashtags were from the top 10 hashtags in Volkswagen story, 84% in Germanwings story and 95% in Charlie Hebdo. The full results of hashtag analysis are presented by news type and by story in Appendix B.

5.4.6 Links analysis. Links analysis was applied to each news story to extract the links and to calculate the frequency of each link. Also the distribution of the counts of links and the number of links in tweets, retweets, and originals were analyzed. Similar to hashtags analysis, more analysis was applied to the top 10 links, including analyzing the counts and ratios of tweet, retweets and originals including any of the top 10 links. Link unshortening and parsing services were employed to find the domain names of website shared, these results are provided in Appendix C. Additionally, the domain names of the top 100 links of each story were examined to get an idea of what kind of information people shared during the events used in the study. These results are presented below by news type and by news story.

5.4.6.1 Finance. Among the links shared in both finance stories originated from BBC, Economist, CNN Money, and Bloomberg, which is a news site that delivers business and market news. Other links included other Twitter posts for different accounts.

Table 5-26 presents the 3 most common domain names in each finance story among the top 100 links.

	Chinese Shares	Count	Percent	Volkswagen	Count	Percent
1	bbc.in	331	27%	bloom.bg	2,742	23%
2	goo.gl	187	15%	bbc.in	1,780	15%
3	twitter.com	138	11%	twitter.com	1,241	10%
Total		656	53%		5,763	48%

Table 5-26 The 3 most used links in the top 100 links in finance

5.4.6.2 Disaster. Table 5-27 presents the 3 most common domain names in each disaster story in the top 100 links.

	Germanwings	Count	Percent	Nepal Earthquake	Count	Percent
1	bbc.in	184	21%	twitter.com	414	28%
2	twitter.com	176	20%	bbc.in	284	19%
3	cnn.it	131	15%	cnn.it	226	15%
Total		491	56%		924	62%

Table 5-27 The 3 most used links in the top 100 links in disaster datasets

The 3 common domain names in both stories were the same, but with slightly different ranks. The distribution of link use showed that few links made most links, while the majority of links appeared few times.

5.4.6.3 Politics. In politics, most links pointed to other twitter accounts, in addition to BBC and NBC. Table 5-28 presents the 3 most common domain names in each politics story among the top 100 links.

	Charlie Hebdo	Count	Percent	Clock Boy	Count	Percent
1	twitter.com	4,945	69%	twitter.com	4,011	85%
2	ysoheil.com	1,060	15%	amp.twimg.com	246	5%
3	bbc.in	547	8%	nbcnews.to	170	4%
Total		6,552	91%		4,427	94%

Table 5-28 The 3 most used links in the top 100 links in politics datasets

5.5 Statistical Analysis

The descriptive analysis provided a deeper basis for understanding the data. To answer the research questions, the relationships between tweeting activity and news type and country were statistically analyzed. The goal of the statistical analyses is to determine whether the variations between countries and between news types are significant. Additionally, these analyses provide the basis for a model that integrates the different variables.

Statistical tests were applied on the test dataset that combined the tweets for all six stories from the common three countries USA, UK and India. The reason for using this subset, is due to the difficulty of applying regression on all 235 countries, and because of the consistency found in the common 3 countries with the whole database with all the 6 stories in terms of the percentages of originals, retweets, tweets with hashtags and tweets with links as presented in Table 5-2.

Binary logistic regression was used for analyzing the relationships among the study variables. Logistic regression is a statistical procedure for creating a prediction model for an outcome variable that is binary, i.e., takes two values 0 or 1, where 0 indicates nonoccurrence, and 1 indicates occurrence (Verma, 2013). In this study the outcome variables are original (1) or retweet (0), having a hashtag (1) or not (0) and

having a link (1) or not (0). The predictive variables used in this study are the categorical variables: country, news type and the interaction between country and news type. Each of the predictive variables (country and news type) have three levels. Country has the levels USA, UK and India, and news type has the levels finance, disaster, and politics. The interaction includes all the different combinations between country and news type. Thus, logistic regression was applied three times, i.e., creating three prediction models, one for each of the outcome variables (originals, hashtags and links) using the two predictive variables (country, news type) and the interaction terms.

In logistic regression, the dependent variable is in log odd, also called logit, which is the probability that the dependent variable = 1. The regression equation is written as shown in Equation 5-1, where B_0 is the intercept, and B_1, $B_2 \ldots B_n$ are the coefficients of the variables X_1, $X_2 \ldots X_n$ respectively. These variables are the different levels of country and news type.

$$logit = \ln\left(\frac{p}{1-p}\right) = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n$$

Equation 5-1 Regression equation

To interpret the results of the regression analysis, regression coefficients are converted into odds ratios. The odds ratio of an event can be defined as the ratio of the probability of success to the probability of failure. In logistic regression, the odds ratio is calculated by finding the exponent of the coefficient B. For example, if the regression coefficient B is equal to 1.5 then the odds ratio is exponent of (1.5), which is equal to 4.48 (Verma, 2013).

SPSS was used for the statistical analysis conducted for this thesis. SPSS starts the regression analysis process by creating dependent variables parameter coding or "dummy" variables, as shown in Table 5-29. Since the degrees of freedom for NewsType is 2, SPSS creates two dummy variables, Finance and Disaster using Politics as the reference category, i.e., when both Finance and Disaster are equal to 0. Similarly, country has two degrees of freedom, so two dummy variables, UK and India were created and USA was used as a reference category, when both UK and India are equal to 0.

		Frequency	Parameter coding	
			(1)	(2)
NewsType	Finance	57240	1	0
	Disaster	19004	0	1
	Politics	55931	0	0
Country	USA	99257	0	0
	UK	21291	1	0
	India	11627	0	1

Table 5-29 Dependent variables coding

The logistic regression analysis starts by creating a "beginning block" or the null model with only the constant or intercept in the equation, in this case the constant is the coefficient of USA in Politics. This null model normally has a prediction percentage similar to random guessing. Next, the regression proceeds with the next block or model, which includes all the dependent variables and the interactions terms. The total percentage of the prediction of the full model determines how well the model correctly classifies the cases (Starkweather & Herrington, 2016). In the following subsections, we explain the results of each logistic regression test.

5.5.1 Analyzing the relationship between <Country, NewsType> and originals. The process of the analysis started by "Block 0" or the null model which had an overall percentage of 55%, whereas the full model percentage is 67.7% as shown in the lower right corner of Table 5-30, which means that the prediction of the model had improved after adding the predictive variables.

Table 5-30

Observed			Predicted		
			Original		Percentage Correct
			Retweet	Original	
Step 1	Original	Retweet	52519	20237	72.2
		Original	22416	37003	62.3
	Overall Percentage				67.7
a. The cut value is .500					

Classification table of originals regression analysis

Table 5-31 presents the chi-square value, 16620.446, and the significance (p-value < 0.05), this result indicates that the full model is significantly different from the null model, and there is a significant effect for including the predictive variables (country, news type) and the interaction terms on the outcome variable, which is the originals count.

		Chi-square	df	Sig.
Step 1	Step	34.102	2	.000
	Block	16620.446	8	.000
	Model	16620.446	8	.000

Table 5-31 Omnibus tests of model coefficients

Table 5-32 shows the value of the -2 Log likelihood and Cox & Snell and Nagelkerke R square values (Verma, 2013). The -2 Log likelihood was equal to 165264.96. This value represents the unexplained variance or deviation in the dependent variable, the smaller the value of the variance the better the fit. Cox & Snell and Nagelkerke R square values are measures that explain the ratio of the variability of the dependent variable by the independent variables in the model. Nagelkerke R square measure is more reliable than the first (Verma, 2013). The Nagelkerke R square value in the model was 0.158, which means that roughly 16% of the variation of the value of originals can be explained by the model.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	165264.960[a]	.118	.158
a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.			

Table 5-32 Model summary

In Table 5-33, B represents the logistic coefficient for each predictive variable, which is the expected amount of change in the outcome variable (original) when the predictive variable is equal to one. The smaller the value of the coefficient the less effect it has in predicting the outcome variable. Table 5-33 also includes S.E. (Standard Error) associated with the coefficients, which measures the precision of the coefficient, the smaller the standard error the more precise the estimate. Wald test, which is the ratio of the square of the coefficient to the square of the standard error, df (degrees of freedom), and significance, which indicates whether or not the logistic coefficient is different than

zero. The last column Exp(B), the exponent of B, which represents the odds ratio for each predictive variable.

	B	S.E.	Wald	df	Sig.	Exp(B)
Country			34.724	2	**.000**	
Country(1) **(UK)**	-.067	.026	6.616	1	**.010**	.935
Country(2) **(India)**	.215	.043	25.152	1	**.000**	1.240
NewsType			11314.269	2	**.000**	
NewsType(1) **(Finance)**	1.533	.015	10902.718	1	**.000**	4.633
NewsType(2) **(Disaster)**	.401	.021	357.334	1	**.000**	1.493
Country * NewsType			57.974	4	**.000**	
Country(1) by NewsType(1)	.108	.035	9.614	1	**.002**	1.114
Country(1) by NewsType(2)	.104	.055	3.628	1	.057	1.110
Country(2) by NewsType(1)	-.009	.054	.026	1	.872	.991
Country(2) by NewsType(2)	-.313	.057	29.805	1	**.000**	.731
Constant	-.954	.011	8081.058	1	**.000**	.385

Table 5-33 Logistic model for interaction between <Country, NewsType> and originals

Thus, applying Equation 5-1 using the beta values from Table 5-33 and the levels of the variables country and news type yields the regression model for the interaction between <Country, NewsType> and originals as presented in Equation 5-2.

$$logit$$
$$= -.954 \pm .067 * UK + .215 * India + 1.533 * Finance + .401 * Disaster + .108$$
$$* UKFinance + .104 * UKDisaster + -.009 * IndiaFinance + -.313$$
$$* IndiaDisaster$$

Equation 5-2 Regression equation example

To interpret the results, the main effects were examined first, which are the effects of country and news type, and the interaction was examined next. The effect of country on originals generation has to be analyzed while keeping news type constant, similarly the effect of news type must be analyzed while keeping country constant.

Table 5-29, presented earlier, shows the corresponding Country and NewsType encoding, i.e., Country (1) is UK, Country (2) is India, NewsType (1) is Finance, and NewsType (2) is Disaster. The overall statistics of Country was significant (p-value < .05, Wald 34.724 with 2 degrees of freedom). The comparison was based on USA in politics as the reference categories, so there was a statistically significant difference in originals generation between UK and USA in politics, (p-value .01 and odds ratio of .935), i.e., UK was 0.94 times (6.5%) less likely to generate originals in politics than USA, and S.E. of .026, which is relatively small. There was a statistically significant difference in original tweets generation between India and USA in politics (p-value < .05), India was 1.24 times more likely to generate originals in politics than USA.

In NewsType the overall statistics was significant, (p < .05, Wald 11314.27 with 2 degrees of freedom), USA was 4.6 times more likely to generate originals in finance than in politics, and 1.49 times (49%) more likely to generate originals in disaster than in politics.

The odds ratios discussed above were directly obtained from the model shown in Table 5-33, however the odds ratio to compare the rest of the variables are calculated by the method explained in (Vittinghoff, Glidden, Shiboski & McCulloch, 2005). For instance, to find the odd ratio for UK vs. USA in finance, first the coefficient of (UK in Finance) is calculated by adding the beta values for the following: Constant, UK, Finance

and the interaction term between UK and Finance, similarly the coefficient of (USA in Finance) is obtained, and then finding the exponent of the difference between the two values. If the p-value of a variable or an interaction term is insignificant, then the beta value for this variable is omitted from the calculation. These odds ratios are presented in Table 5-34 and Table 5-35.

UK was almost the same as USA in generating originals in all news types, the differences were between 4% and 6.5%. India was 1.24 times (24%) more likely to generate originals than USA in both finance and politics respectively, and 0.907 times (9%) less likely to generate originals than USA in disaster. India was 1.190 times (19%) and 1.326 (33%) more likely to generate originals than UK in both finance and politics respectively, and 0.969 (3%) less likely to generate originals than UK in disaster.

	Finance	Disaster	Politics
UK vs USA	1.042	0.935	0.935
India vs USA	1.240	0.907	1.240
India vs UK	1.190	0.969	1.326

Table 5-34 Odds ratios of comparing countries in generating original tweets

Table 5-35 presents the odds ratios for comparing news types. Comparing finance with both disaster and politics, all countries were 3 – 4 times more likely to generate original tweets in finance than in disaster, and 4.5 – 5 times more likely to generate original tweets in finance than in politics.

	USA	UK	India
Finance vs Disaster	3.102	3.456	4.242
Finance vs Politics	4.632	5.160	4.632
Disaster vs Politics	1.493	1.092	1.092

Table 5-35 Odds ratios of comparing news types in generating original tweets

The difference between disaster and politics across countries was not as high as the differences between finance and both politics and disaster. Disaster was 1.493 times (50%) more likely to generate original tweets than politics in USA, and 1.092 times (9%) in UK and India.

Odds ratios are used to compare two countries or news types to one another as shown above. However, to find the probabilities of generating originals for each individual country in each news types Equation 5-3 was applied to calculate these probabilities.

$$P = \frac{e^{B_0+B_1X_1+\cdots}}{1 + e^{B_0+B_1X_1+\cdots}}$$

Equation 5-3 Probability equation

The coefficients of each country and news type combination was obtained, as illustrated in Equation 5-2, then the exponent of this coefficient was calculated ($e^{B_0+B_1X_1+\cdots}$) and divided by 1+ ($e^{B_0+B_1X_1+\cdots}$).

Table 5-36 presents the probabilities of generating original tweets in each country and new type. The probabilities were higher in finance (0.64 – 0.69) than both disaster and politics in the 3 countries. Disaster had higher probabilities of generating originals than politics in the 3 countries as well.

	Finance	Disaster	Politics
USA	0.64	0.37	0.28
UK	0.65	0.35	0.26
India	0.69	0.34	0.32

Table 5-36 Probabilities of generating originals in all news types and countries

5.5.2 Analyzing the relationship between <Country, NewsType> and hashtags. The result of applying logistic regression to analyzing hashtags generated a null model with an overall percentage of 54.5%, and a full model with a percentage of 72.4%. This indicates that the prediction of the model has improved after adding the predictive variables. The chi-square value was 29891.399, with a p-value of less than 0.05, i.e., the full model was significantly different from the null model, and there was a significant effect for including the predictive variables, country, news type and the interaction terms on hashtags count. The -2 Log likelihood was equal to 152,264.71, and Nagelkerke R square value in the model was 0.271, which means that roughly 27% of the variation of the value of hashtags can be explained by the model. Tables showing these results are provided in Appendix D.

Table 5-37 presents the logistic model for analyzing hashtags, the overall statistics of Country was significant (p-value < .05, Wald 26.477 with 2 degrees of freedom). Similar to originals analysis, the comparison was based on USA and politics as the reference categories.

The result showed a statistically significant difference in the use of hashtags between UK and USA in politics (p-value < .05), UK was 1.15 times more likely to

generate hashtags in politics than USA. The difference between India and USA in politics was not statistically significant, (p-value .218 > .05).

NewsType had an overall significant difference (p-value < .05, Wald 19715.718 with 2 degrees of freedom). Finance was 0.108 times (89%) less likely to include hashtags in tweets than politics in USA, and disaster was 0.340 times (66%) less likely to include hashtags in tweets than politics in USA.

	B	S.E.	Wald	df	Sig.	Exp(B)
Country			26.477	2	**.000**	
Country(1) **(UK)**	.141	.029	23.560	1	**.000**	1.152
Country(2) **(India)**	-.059	.048	1.520	1	.218	.943
NewsType			19715.718	2	**.000**	
NewsType(1) **(Finance)**	-2.221	.016	19683.423	1	**.000**	.108
NewsType(2) **(Disaster)**	-1.079	.021	2587.057	1	**.000**	.340
Country * NewsType			234.912	4	**.000**	
Country(1) by NewsType(1)	.064	.038	2.872	1	.090	1.066
Country(1) by NewsType(2)	-.144	.055	6.847	1	**.009**	.866
Country(2) by NewsType(1)	.127	.058	4.815	1	**.028**	1.136
Country(2) by NewsType(2)	.747	.062	145.603	1	**.000**	2.111
Constant	1.297	.012	12542.735	1	**.000**	3.660

Table 5-37 Logistic model for interaction between <Country, NewsType> and hashtags

The rest of the odds ratios were calculated and presented in Table 5-38 and Table 5-39. UK was 1.151 times (15%) more likely to generate hashtags than USA in finance, and had about the same odds ratio as USA in disaster. Comparing India and USA, India was nearly twice as likely to generate tweets with hashtags in disaster, 14% more in finance. India versus UK, India was 2.117 times more likely to generate tweets with hashtags in disaster, 0.872 times (13%) less in finance, and 0.819 times (18%) less in politics.

	Finance	Disaster	Politics
UK vs USA	1.151	0.997	1.151
India vs USA	1.135	2.111	1.000
India vs UK	0.986	2.117	0.868

Table 5-38 Odds ratios of comparing countries in generating tweets with hashtags

The odds ratios of India versus both USA and UK in generating tweets with hashtags in disaster were 2.11 and 2.12, which were largest odds ratios in the model for comparing countries.

Table 5-39 presents odds ratios for comparing news types while keeping country constant. Finance was 0.319 times (68%), 0.369 times (63%) and 0.172 times (83%) less likely to include hashtags in tweets than disaster in USA, UK and India respectively. Comparing finance and politics, finance was 0.109 times (89%) less likely to include hashtags in tweets than politics in USA, UK, and 0.123 times (88%) less likely to include hashtags in India. Lastly, disaster tweets were 0.717 times (28%) less likely to include hashtags in both UK and India compared to political tweets.

	USA	UK	India
Finance vs Disaster	0.319	0.369	0.172
Finance vs Politics	0.109	0.109	0.123
Disaster vs Politics	0.340	0.717	0.717

Table 5-39 Odds ratios of comparing news types in generating tweets with hashtags

These results show that finance exhibited different behavior in generating tweets with hashtags within each of the 3 countries. Finance was the lowest in terms of tweets

with hashtags, and the largest differences are between finance and politics. Additionally, these result were similar within each country.

Table 5-40 presents the probabilities of generating tweets with hashtags in each country and news type. The probabilities were higher in politics (0.78 – 0.81) in the 3 countries than both disaster and finance. Disaster had higher probabilities of generating hashtags than finance in the 3 countries.

	Finance	Disaster	Politics
USA	0.28	0.55	0.79
UK	0.33	0.55	0.81
India	0.30	0.71	0.78

Table 5-40 Probabilities of generating hashtags in all news types and countries

5.5.3 Analyzing the relationship between <Country, NewsType> and links.

The result of applying logistic regression to analyze the relationship between country, news type and links had a null model and full model with the same overall percentage, 78.4%. The reason for this similarity was because links were in more than 50% in all countries across all news types as the cross tabulation in Table 5-41 reveals. That explains the reason why adding the dependent variables to the model did not add to the value of the prediction of the model.

		NewsType			Total
		Finance	Disaster	Politics	
NoLink	USA	3,955	3,728	13,283	20,966
	UK	1,116	691	2,568	4,375
	India	678	1,562	938	3,178
Link	USA	38,267	9,096	30,928	78,291
	UK	8,931	1,485	6,500	16,916
	India	4,293	2,442	1,714	8,449

Table 5-41 Counts of links in the 3 countries across the 3 news types

The -2 Log likelihood was equal to 129,063.67, and Nagelkerke R square value in the model was 0.099, which means that roughly 10% of the variation of the value of hashtags can be explained by the model. Tables showing these results are provided in Appendix D.

Table 5-42 presents the logistic model for analyzing links. The overall statistics of Country is significant (p-value < .05, Wald 48.433 with 2 degrees of freedom). Similar to originals and hashtags analysis, the comparison was based on USA and politics as the reference categories. There was a statistically significant difference among all countries and news types as indicated in the Sig. column. UK was 1.087 times (9%) more likely to generate tweets with links than USA in politics, while India was about 0.785 times (22%) less likely than USA to generate tweets with links in politics. USA was about 4 time more likely to generate tweets with links in finance than in politics, and only 1.048 times (5%) more in disaster.

	B	S.E.	Wald	df	Sig.	Exp(B)
Country			48.433	2	**.000**	
Country(1) **(UK)**	.083	.026	10.712	1	**.001**	1.087
Country(2) **(India)**	-.242	.042	33.423	1	**.000**	.785
NewsType			5516.737	2	**.000**	
NewsType(1) **(Finance)**	1.424	.020	5248.405	1	**.000**	4.155
NewsType(2) **(Disaster)**	.047	.022	4.506	1	**.034**	1.048
Country * NewsType			52.991	4	**.000**	
Country(1) by NewsType(1)	-.273	.044	38.552	1	**.000**	.761
Country(1) by NewsType(2)	-.210	.056	14.059	1	**.000**	.810
Country(2) by NewsType(1)	-.182	.061	8.815	1	**.003**	.834
Country(2) by NewsType(2)	-.203	.056	12.909	1	**.000**	.816
Constant			48.433	2	**.000**	2.328

Table 5-42 Logistic model for interaction between <Country, NewsType> and links

The rest of the odd ratio for comparing countries while keeping news type constant are presented in Table 5-43. UK was 0.827 times (17%) less and 0.881 times (12%) less likely to generate tweets with links than USA in both finance and in disaster respectively. India was also less than USA by 0.654 times (35%) and 0.641times (36%) in finance and disaster respectively. India was less than UK by 0.791 times (21%), 0.728 times (27%) and 0.723 (28%) in finance, disaster and politics respectively.

	Finance	Disaster	Politics
UK vs USA	0.827	0.881	1.087
India vs USA	0.654	0.641	0.785
India vs UK	0.791	0.728	0.723

Table 5-43 Odds ratios of comparing countries in generating tweets with links

Table 5-44 presents the odd ratios of comparing news types while keeping country constant. Finance odd ratios across all countries ranged from 3 to 4 times, i.e., finance was more likely to include links in tweets than disaster and politics. While disaster was

about 1.048 times (5%) more likely to have links than politics in USA, disaster also is 0.856 times (14%) less likely to have links than politics in UK and India.

	USA	UK	India
Finance vs Disaster	3.963	3.721	4.047
Finance vs Politics	4.154	3.161	3.463
Disaster vs Politics	1.048	0.856	0.856

Table 5-44 Odds ratios of comparing news types in generating tweets with links

These results showed that finance tweets had different behavior in generating tweets with links than both politics and disaster in all the 3 countries. Finance had higher odds ratios for generating tweets with links, and these result were similar within each country to some extent.

Table 5-45 presents the probabilities of generating tweets with links in each country and new type. The probabilities were generally high across all news types and countries, but higher in finance (0.86 – 0.91) than both disaster and politics.

	Finance	Disaster	Politics
USA	0.91	0.71	0.70
UK	0.89	0.68	0.72
India	0.86	0.61	0.65

Table 5-45 Probabilities of generating links in all news types and countries

5.5.4 Statistical tests summary. The summary of the results of the statistical tests are presented in Table 5-46. In general, all the tests had high chi-square values, high significant levels (.000) and high -2 log likelihood, these high values are expected

because of the large sample size (Moyé, 2006). Also the S.E. in all the results was relatively small, indicating that the coefficients estimate was fairly precise.

	Chi-square	df	Sig.	-2 Log likelihood	Nagelkerke R Square
Originals	16620.446	8	.000	165264.960	.158
Hashtags	29891.399	8	.000	152,264.71	0.271
Links	8794.306	8	.000	129063.67	.099

Table 5-46 Summary of statistical tests results

The logistic models for interaction between <Country, NewsType> and originals, hashtags and links, were statistically significant. However, the variation among news types in general was greater than the variation among countries. For instance, the maximum difference among countries in original tweets generation was between India and USA, with odd ratio of 1.33. In hashtags, the largest difference was also between India and USA, and India and UK, with odds ratio of around 2. However, regarding news type, in finance all countries were 3 to 5 times more likely to generate originals and links than politics and disaster. Finance was 61% to 83% less likely to generate hashtags than both politics and disaster, these findings are summarized in Table 5-47.

User Behavior	New Types	USA	UK	India
Originals	Finance vs Disaster	3.102	3.456	4.242
	Finance vs Politics	4.632	5.160	4.632
Hashtags	Finance vs Disaster	0.319	0.369	0.172
	Finance vs Politics	0.109	0.109	0.123
Links	Finance vs Disaster	3.963	3.721	4.047
	Finance vs Politics	4.154	3.161	3.463

Table 5-47 Summary of the large odds ratios scored in the different regression models

These results showed that news type was more influential than country on user behavior. These results are consistent with our descriptive analysis results presented in section 5.3. Finance exhibited different behaviors from the other two news types in generating originals, retweets, tweets with links and tweets with hashtags; finance had larger original/retweet ratio, more tweets with links and fewer tweets with hashtags than disaster and politics.

Chapter 6 Discussion

In the previous chapter the detailed results of each analysis conducted were presented, by country, news type and news story. Statistical analyses were conducted to evaluate the relationships between country, news type and user participation. The main focus of this chapter is to examine how these results could expand the knowledge and understanding of the different user behaviors during the types of news events used in the study. This understanding would provide the basis for developing methods that could improve the user experience in similar contexts. To achieve this understating, the first step is to answer the research questions. Next a summary of all the results is provided.

6.1 RQ1: Is there a relationship between country and the type of user behavior in Twitter?

The descriptive analysis conducted to analyze the relationship between country and the different types of user participation, the use of originals, retweets, hashtags and links, showed that overall user behavior for all countries, top 10 and common 3 was different.

The logistic regression analyses conducted to model the relationships between the common 3 countries and user behavior, proved that relationships exist, and country was a significant factor in all the different participation for these countries. Table 6-1 summarizes the odds ratios of the regression analysis results, the second column in the table display the countries compared, and asterisks are placed near the larger values (> 0.20).

	Countries	Finance	Disaster	Politics
Originals	UK vs USA	1.042	0.935	0.935
	India vs USA	1.240 *	0.907	1.240 *
	India vs UK	1.190	0.969	1.326 *
Hashtags	UK vs USA	1.151	0.997	1.151
	India vs USA	1.135	2.111 *	1.000
	India vs UK	0.986	2.117 *	0.868
Links	UK vs USA	0.827	0.881	1.087
	India vs USA	0.654 *	0.641 *	0.785 *
	India vs UK	0.791 *	0.728 *	0.723 *

Table 6-1 Odds ratios of comparing countries in generating original tweets, tweets with hashtags and tweets with links

The results also showed that distributions of tweets among all the participating countries and among the top 10 countries followed a power-law distribution in all the different behaviors, with an exponent of -2.4, whereas the top 10 countries followed a power law distribution but with exponent of -1.3. This finding is consistent with the finding of Cao and Caverlee (2014), where the study reported that a few links have been used up to 100,000 times, whereas the majority of links have been used once or twice. The exponent value explains the relationship between the counts of tweets and the countries, i.e., the higher exponent for all countries reflects that the number of countries with high counts are much fewer than the number of countries with low counts. The lower exponent for the top 10 countries reflects the same fact, but with less variability among countries than the case with all the countries. For instance, in the case of the higher exponent, if we assume that the number of countries with higher counts are the first 10, then the number of countries with the lower counts is 225, whereas with the case of the lower exponent (the top 10 countries), if we assume that the number of countries with higher counts is the first 3 countries, then the number of countries with fewer tweets

would be 7, the difference is 4, hence, less variability than the case with the larger exponent. In the following sections, a broader discussion is provided for the descriptive analysis and the regression model for each participation type.

6.1.1 Is there a relationship between country and the generation of original tweets or retweets? By referring to the results by country presented in section 5.2, it can be observed that for all countries the participations in terms of the number of originals and retweets varied considerably. The USA generates around 35% of all the tweets, the next in the rank was France, which generated around 12% of all tweets, and that may be due to Charlie Hebdo story, which had 61% of all the tweets in the database.

By examining the results by news story (section 5.4), we found that the distribution of participating countries differed with news event change. For instance, in the Germanwings airplane crash story, the plane departed from Barcelona, Spain, to Düsseldorf, Germany, and crashed in France, the countries Spain, Germany and France appeared in the top 10 list. While close proximity of countries to the event might be one of the influential factors, other factors emerged after analyzing other news types. For instance, Nigeria appeared in the top 10 list of Chinese Shares story, and had a relatively high retweet ratio. Among the reasons for Nigeria high participation, is that China is a major trading partner for most African countries (BBC, 2015). Moreover, Nigeria and China established diplomatic relations in 1971. Thus, Nigeria became a major source of oil for China, nearly 90% of Nigeria's export is oil (Geopolitical Futures, 2016), and in return China supported Nigeria economically and politically (Wikipedia, 2016). When the Chinese market collapsed, it affected many countries around the world including Nigeria,

which recorded losses of around 228 billion Nigerian Naira on the day of the event, which is equivalent to around 722 million USD (Ventures, 2015).

The model of the relationship between <Country, NewsType> and originals presented in section 5.5.1, showed that country was a significant factor in the generation of originals. The variation between countries in original tweets generation varied from 3% to 33% among the different news types. The smallest differences were between USA and UK, while larger odds ratios were found between India and both UK and USA, where India was 24% and 33% more likely to generate originals than the USA and the UK in political news, and India also 24% more likely to generate originals than USA in finance.

In some cases, the difference between countries' behaviors can be explained by factors such as language, population, economy level, technology use, and closeness to origin of the news event. Other variations did not have such direct explanations, and may need further investigation. For instance, India generated fewer original posts than the UK and the USA in disaster, i.e., India generated more retweets. This could be due to the closeness of India to the Nepal earthquake, as people tend to retweet more in such disastrous situations (Hansen et al., 2011). In politics, India was more likely to generate originals than both UK and USA, this case is an example of a scenario that need further research. The differences between the USA and the UK was relatively smaller, for instance UK was 4.2% more likely to generate originals than USA in finance, and 6.5% less likely to generate original tweets in disaster and politics.

Table 6-2 summarizes the descriptive and statistical analysis results for the relationship between country and originals. As the table shows, the descriptive results were consistent with the statistical results. The differences between USA and UK were

relatively small, whereas the differences between India and both USA and UK were larger.

Comparison	Descriptive – % Total Within Country			Statistical		
	Finance	Disaster	Politics	Finance	Disaster	Politics
UK/ USA	65% vs 64%	37% vs 37%	27% vs 28%	1.042	0.935	0.935
India/ USA	69% vs 64%	34% vs 37%	32% vs 28%	1.240 *	0.907	1.240 *
India/ UK	69% vs 65%	34% vs 37%	32% vs 27%	1.190	0.969	1.326 *

Table 6-2 Descriptive and statistical analysis results summary for the relationship between country and originals

The discussion above shows how the relationship between countries and the generation of original tweets differ. However, in the analysis of the top 10 countries provided in section 5.2, the 3 common countries were among the countries with higher originals ratios compared to the rest of the top 10 countries. For instance, USA, UK and India generated higher ratios of originals compared to the other countries in the top 10 list, and together they make more than 50% of originals of the whole database. Therefore, the variations among all countries are expected to be larger than among the 3 countries, and the power-law distribution of countries in original generation across all news types supports this finding.

6.1.2 Is there a relationship between country and the use of hashtags?

Hashtags use across countries varied as Table 5-4 shows, for instance 83% of France tweets contained hashtags, 72% in Netherlands, and 70% in Spain. Referring to Table 5-3, which presents retweet ratios in the top 10 countries, we found that high hashtags ratios were associated with high rewets ratios. For instance, 83% of France tweets contained hashtags, and 75% of France tweets were retweets.

The distributions of tweets, originals and retweets with hashtags of the top 10 countries presented in Figure 5-2, showed that all countries had higher counts of retweets containing hashtags than originals overall, except for Indonesia. However, on the story level, most tweets in Chinese share story were originals within each of the top 10 countries, except for Nigeria, which had higher retweet ratio than originals, and represented more than 50% of all retweets with hashtags in this story. The distribution of tweets with hashtags in the whole database among all countries followed a power-law distribution.

The common 3 countries make roughly half of the total originals with hashtags (47%), and had higher ratios of originals with hashtags of the total original with hashtags in the whole database than retweets. This result might indicate that these English-speaking countries were creating more hashtags than retweeting existing ones.

As presented in section 5.5.2, the logistic model for analyzing hashtags had a significant overall statistic for Country, i.e., the hashtags usage was different among countries. Odds ratios for comparing countries are presented in the Hashtags row in Table 6-1. The largest differences were between India and each of UK and USA in disaster news, as India was twice as likely to generate tweets with hashtag than USA and UK.

That could be due to the Nepal earthquake event, in which India was in the second country in the top 10 list for that story. The other differences between countries in tweeting with hashtags ranges from 13% to 15%. For instance, UK was 15% more likely to generate tweets with hashtags than USA in finance.

Table 6-3 presents the summary of descriptive and statistical analyses for the relationship between country and hashtags. The descriptive results were consistent with statistical results, for instance, the statistical result indicates that India was 2 times more likely than both UK and USA to generate tweets with hashtags in disaster, and the descriptive result shows that 72% of India disaster tweets contained hashtags, whereas hashtags were in 56% and 55% of USA and UK disaster tweets respectively.

Comparison	Descriptive – % Total Within Country			Statistical		
	Finance	Disaster	Politics	Finance	Disaster	Politics
UK/ USA	33% vs 29%	55% vs 56%	81% vs 79%	1.151	0.997	1.151
India/ USA	30% vs 29%	72% vs 56%	78% vs 79%	1.135	2.111 *	1.000
India/ UK	30% vs 33%	72% vs 55%	78% vs 81%	0.986	2.117 *	0.868

Table 6-3 Descriptive and statistical analysis results summary for the relationship between country and hashtags

Countries behavior in generating tweets with hashtags differ significantly as the regression analysis show. In some cases, we were able to justify the difference, such as the case where India was more likely to generate hashtags in disaster stories. In other cases, no clear reason for the different behavior, for example, UK was more likely to generate tweets with hashtags than US and India in finance, and in politics.

6.1.3 Is there a relationship between country and the use of links? The descriptive statistics showed that there was a variation among countries in generating

tweets with links. USA generates a little more than third of the tweets, originals and retweets with links. As presented earlier in Table 5-5, Indonesia had the largest ratio of tweets with links, Canada come next, and USA and UK were in the third rank.

Additionally, the top common countries, USA, UK and India make more than 52% of originals with links of all originals with links in whole database. That might indicate that these countries had the larger portion of users creating the tweets with links, more than the rest of the countries.

There was a statistically significant difference among countries in generating tweets with links as indicated in section 5.5.3. UK was about 9% more likely to generate tweets with links than USA in politics, while India was about 22% less than USA in politics. The larger differences were between India and USA, India was about 35% less likely to generate tweets with links than USA in both finance and disaster news.

Table 6-4 presents the summary of descriptive and statistical analyses for the relationship between country and links. The descriptive and the statistical analyses reflect the same results, for instance, the descriptive result shows that India had links in 86% of tweets in finance, and USA 91%, and the statistical result indicates that India was 35% less likely to generate tweets with links in finance.

Comparison	Descriptive – % Total Within Country			Statistical		
	Finance	Disaster	Politics	Finance	Disaster	Politics
UK/ USA	89% vs 91%	68% vs 71%	72% vs 70%	0.827	0.881	1.087
India/ USA	86% vs 91%	61% vs 71%	65% vs 70%	0.654 *	0.641 *	0.785 *
India/ UK	86% vs 89%	61% vs 68%	65% vs 72%	0.791 *	0.728 *	0.723 *

Table 6-4 Descriptive and statistical analysis results summary for the relationship between country and links

Link sharing activity in the 3 countries vary, the comparisons between UK and USA showed less variations, than the comparisons involved India. India was less likely to generate tweets with links than both USA and UK, with greater variations with USA.

6.2 RQ2: Is there a relationship between news type and the type of user behavior in Twitter?

The descriptive statistics by news types showed that behavior of users change as news type change. Finance had the highest originals ratio, 64% of total finance tweet, politics had the highest retweets ratio, 73%. Finance tweets had links in 91% of them, and political tweets contained hashtags in 80%. Disaster had 40% originals, links in 70%, and hashtags in 57% of the total disaster tweets.

The logistic regression models showed that news type was significant in the relationship between <Country, NewsType> and the different user behaviors including generating originals versus retweet, using hashtags and including links during the events used for this study. Table 6-5 summarizes these results, the second column in the table display the news types compared, and asterisks are placed near the larger values (> 0.20).

		USA	UK	India
Originals	Finance vs Disaster	3.102*	3.456*	4.242*
	Finance vs Politics	4.632*	5.160*	4.632*
	Disaster vs Politics	1.493 *	1.092	1.092
Hashtags	Finance vs Disaster	0.319 *	0.369 *	0.172 *
	Finance vs Politics	0.109 *	0.109 *	0.123 *
	Disaster vs Politics	0.340 *	0.717 *	0.717 *
Links	Finance vs Disaster	3.963 *	3.721 *	4.047 *
	Finance vs Politics	4.154 *	3.161 *	3.463 *
	Disaster vs Politics	1.048	0.856	0.856

Table 6-5 Odds ratios of comparing news types containing original tweets, tweets with hashtags and tweets with links among the 3 common countries

From the results presented in the table, originals and links odds ratios were higher in finance compared to the other news types among all countries, and hashtags odds ratios were lower in finance than both disaster, and politics. Next, a detailed discussion is provided for the relationship between news type and each of the different behaviors.

6.2.1 Is there a relationship between news type and the generation of original tweets or retweets? The descriptive analysis presented in section 5.3 showed that originals ratios were different among news types, 64%, 39% and 27% in finance, disaster and politics respectively.

The distribution of originals and retweets in all news types followed a power-law distribution. Figure 6-1 presents the log log scale of the distribution of originals counts across all countries, and Figure 6-2 presents the log log scale of the distribution of originals counts across top 10 countries.

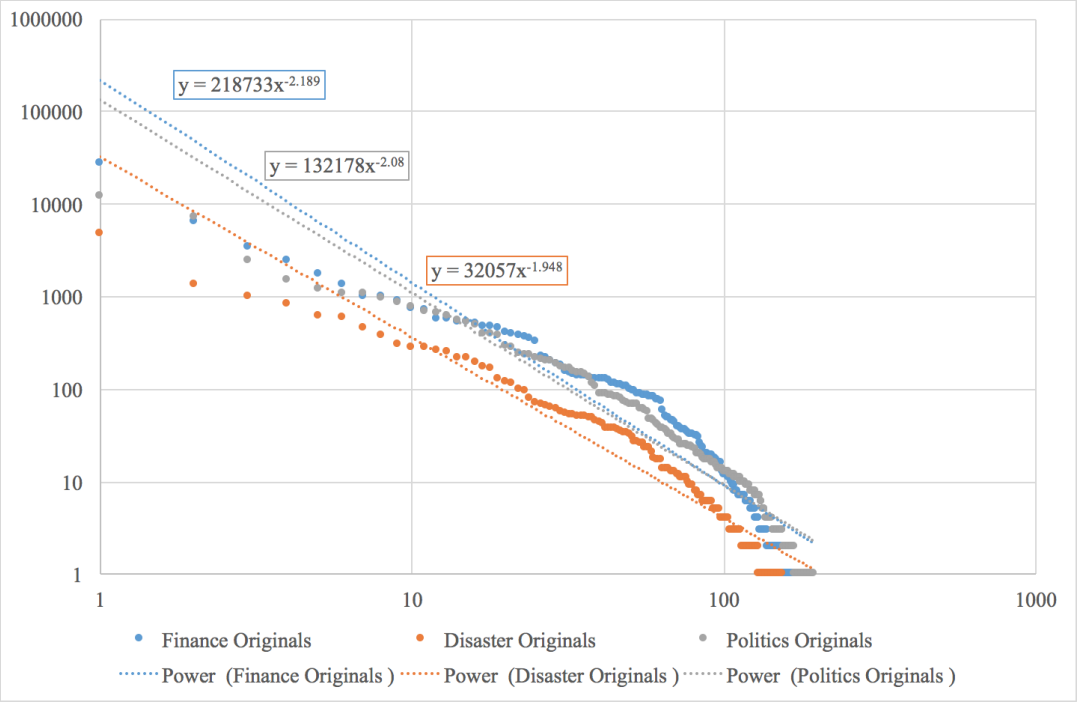

Figure 6-1 Log log scale of the distribution of originals counts across all countries in the three news types

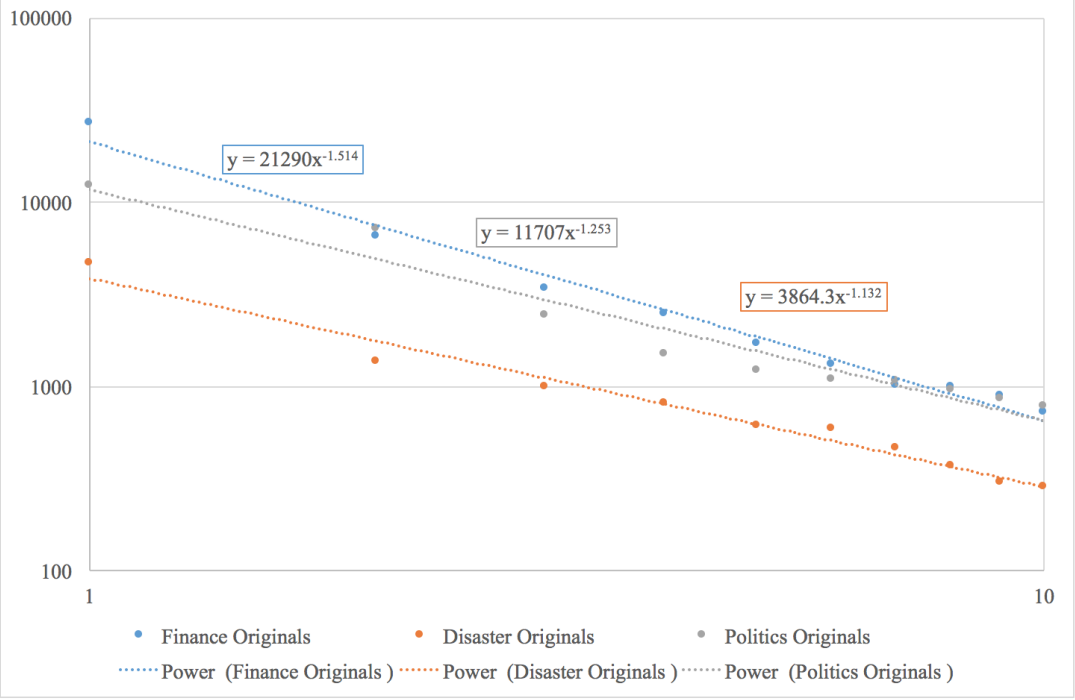

Figure 6-2 Log log scale of the distribution of originals counts across top 10 countries

The distributions among all countries had exponents values of around -2.2 in finance and -2 in politics and disaster. In the top 10 countries, the exponents of the slope of the line had the values -1.5, -1.3 and -1.1 in finance, politics and disaster respectively.

In the model of analyzing <Country, NewsType> and originals, the overall statistics was significant. Finance was 3 to 5 times more likely to generate original tweets than disaster and politics across the three countries. The differences between disaster and politics in original tweets creation was not as high as the case of finance. Disaster tweets were 50% more likely to be original than politics in USA, and only 10% more in UK and India. The characteristics by news type presented in section 5.3 support these results, where original tweets ratios were 64% in finance, 39% in disaster and 27% in politics. Table 6-6 summarizes the descriptive and statistical results.

Comparison	Descriptive – % Total Within Country			Statistical		
	USA	UK	India	USA	UK	India
Finance/Disaster	64% vs 37%	65% vs 37%	69% vs 34%	3.102*	3.456*	4.242*
Finance/Politics	64% vs 28%	65% vs 27%	69% vs 32%	4.632*	5.160*	4.632*
Disaster/Politics	37% vs 28%	37% vs 27%	34% vs 32%	1.493*	1.092	1.092

Table 6-6 Descriptive and statistical analysis results summary for the relationship between news type and originals

The table show that the odds ratio is aligned with the parentages generated from the descriptive analysis. Originals in finance had the largest odd ratios compared to politics in UK, USA and India respectively, and finance had also larger odds ratio compared to disaster in India, UK and USA respectively.

6.2.2 Is there a relationship between news type and the use of hashtags? The characteristics by news type results presented in section 5.3 showed that tweets

percentages with hashtags were 80%, 57% and 31% in politics, disaster and finance tweets respectively. These ratios were higher than ratios reported by Shi et al. (2016), where the study stated that tweets with hashtags represented 25%. This would indicate that users tend to use more hashtags during specific events as opposed to chatting about every day activity.

The ratios of original/retweet in tweets with hashtags were also different in each news type. In finance the original/retweet ratio in tweets with hashtags was around 60/40, 30/70 in disaster and 25/75 in politics. Similar to originals distributions, the distribution of hashtag counts across countries in all news types followed a power-law distribution.

The regression analysis results between <Country, NewsType> and hashtags had an overall significant p-value < .05. Similar to originals, the differences in tweets with hashtags was greater between finance and both disaster and politics as presented in the Hashtags row in Table 6-5. Finance was 68%, 61% and 83% less likely to have hashtags in tweets than disaster in USA, UK and India respectively. In finance versus politics, finance was 90% less likely to have hashtags in tweets in USA, and 88% less likely to have hashtags in tweets in both UK and India. Lastly, disaster tweets were 66% less likely to include hashtags than political tweets in USA, and 28% less likely to include hashtags in both UK and India.

To summarize these results, tweets with hashtags were less in finance tweets than in politics and disaster tweets across the three countries, and that might reflect the nature of finance posts, which are normally written by financial experts and concerned users, and include information and analytics about a current financial situation, rather than spreading some kind of breaking news like in politics and in disaster, where people use

hashtags to guarantee that their post would reach the maximum number of interested audience. Table 6-7 summarizes the results of the descriptive and the statistical analyses.

Comparison	Descriptive – % Total Within Country			Statistical		
	USA	UK	India	USA	UK	India
Finance/Disaster	29% vs 56%	33% vs 55%	30% vs 72%	0.319*	0.369*	0.172*
Finance/Politics	29% vs 79%	33% vs 81%	30% vs 78%	0.109*	0.109*	0.123*
Disaster/Politics	56% vs 79%	55% vs 81%	72% vs 78%	0.340*	0.717*	0.717*

Table 6-7 Descriptive and statistical analysis results summary for the relationship between news type and hashtags

The results of the descriptive analysis were consistent with the statistical analysis. Hashtags were more likely to be in politics tweets more than both disaster and finance tweets, and the difference was greater between politics and finance, than the difference between politics and disaster. This result agrees with the findings of Bogdanov et al. (2014) presented in the literature, in their work, they found that the politics topic had many more hashtags compared to the other categories they studied including business, celebrities, sports, science and technology.

6.2.3 Is there a relationship between news type and the use of links? The characteristics by news type results presented in section 5.3 showed that tweets with links percentages were 91%, 70% and 69% in finance, disaster and politics tweets respectively. The ratios of original/retweet in tweets with links were different in each news type. In finance the original/retweet ratio in tweets with links was around 60/40, 40/60 in disaster and 25/75 in politics. Similar to originals and hashtags distributions, the distribution of links counts across countries in all news types followed a power-law distribution.

Table 5-44 presents the odd ratios of comparing news types in generating tweets with links while keeping country constant. Finance odd ratios across all countries ranges from 3 to 4 times as likely to have links in tweets than in disaster and politics. While disaster was about 5% more likely to have links in than politics in USA, and 14% less likely to have links than politics in UK and India. From these results, we conclude that the use of links was more in finance news, that may be because when there was a financial event, such as stock price drop, people need to understand the influences and the consequences of such an event, and this elaboration might need more than 140 characters, so users tend to share links to articles that have more information. The examination of the top 10 links shared in both finance stories supports this assumption, where links to BBC, CNN Money, Economist, and Bloomberg were in that top 10 list, the full list of links in the two finance stories is available in Appendix C.

Table 6-8 summarizes the results of descriptive and statistical analyses. The table shows that the odds ratio is consistent with the parentages generated from the descriptive analysis.

Comparison	Descriptive – % Total Within Country			Statistical		
	USA	UK	India	USA	UK	India
Finance/Disaster	91% vs 71%	89% vs 68%	86% vs 61%	3.963*	3.721*	4.047*
Finance/Politics	91% vs 70%	89% vs 72%	86% vs 65%	4.154*	3.161*	3.463*
Disaster/Politics	71% vs 70%	68% vs 72%	61% vs 65%	1.048	0.856	0.856

Table 6-8 Descriptive and statistical analysis results summary for the relationship between news type and links

Links in finance had the largest odd ratios compared to politics in USA, India and UK respectively. Finance was more likely to have links than disaster in India, USA and UK respectively. Comparing disaster and politics, disaster was less likely to have links in UK and India, but slightly more in USA.

6.3 RQ3: Is there a relationship between <Country, NewsType> and user participation in Twitter?

This question investigates whether an interaction between country and news type exists. By applying the interaction model of regression analysis, we found that there was an interaction between county and news type in the different user behaviors, Table 6-9 summarizes these results.

Country * NewsType	Wald	df	Sig.
Originals	57.974	4	.000
Hashtags	234.912	4	.000
Links	52.991	4	.000

Table 6-9 Interaction between country and news type Wald and significance

This interaction was an integral part of the analysis and the answers to questions 1 and 2 involved the interaction in the calculations of odd ratios and probabilities.

However, if the p-value of an interaction was insignificant the corresponding beta value was not added to the regression equation, hence the insignificant value was not considered in the calculation of odds ratios or probabilities.

6.4 Summary of findings

The discussion above provides detailed answers the research questions, in this section a summary of all the research results is provided. Table 6-10 summarizes the answers to the research questions.

		RQ1: Country	RQ2: News Type			RQ3: Interaction
			Finance	Disaster	Politics	
Originals	Answer	Yes	Yes			Yes
	Overall	41%	64%	39%	27%	Significant except: UK + D India + F
	Top 10	25%-76%	53%-93%	33%-73%	21%-35%	
	Common 3	44%-49%	64%-69%	34%-37%	27%-32%	
	Regression	C1: .28-.64 C2: .26-.65 C3: .32-.69	.64 -.69	.34 -.37	.26-.32	
Hashtags	Answer	Yes	Yes			Yes
	Overall	60%	31%	57%	80%	Significant except: UK + F
	Top 10	41%-83%	29%-47%	31%-72%	77%-85%	
	Common 3	55%-56%	29%-33%	55%-72%	78%-81%	
	Regression	C1: .28-.79 C2: .33-.81 C3: .30-.78	.28-.33	.55-.71	.78-.81	
Links	Answer	Yes	Yes			Yes
	Overall	77%	91%	70%	69%	Significant
	Top 10	67%-90%	86%-98%	61%-90%	63%-72%	
	Common 3	73%-79%	86%-91%	61%-71%	65%-72%	
	Regression	C1: .70-.91 C2: .68-.89 C3: .61-.86	.86-.91	.61-.71	.65-.72	

Table 6-10 Summary of findings

The three columns present the three research questions, 'RQ1' column shows the overall percentages and the range of variability across countries. Overall, top 10 and common 3 rows represent the descriptive part of the analysis. The descriptive results showed that news type had more effect on participation than country, and the statistical part proves this finding. Country participation varied based on some known factors, such as closeness to the event, economic relations and language, other variations are associated with unknown factors.

Beside answering the research questions, the analyses conducted revealed some additional findings:

1. By country

a. The distributions of tweets by countries followed a power-law distribution, similar to a Zipf curve, and that includes retweets, original, tweets with hashtags and tweets with links.

b. The Top 10 countries generated nearly three quarter of the total tweets in the whole database, and the common 3 generated around 50% of all tweets.

c. The top 10 and the common 3 countries had similar patterns of the use of originals, retweets, tweets with hashtags and tweets with links to total tweets in each of them. The common 3 countries created higher ratio of new content, rather than retweeting.

2. By news type

a. Finance news tweets had distinct user behavior, finance tweets had more original tweets, more links and less hashtags.

b. Politics tweets had the largest number of hashtags, compared to disaster and finance tweets.

c. The retweet mechanism is an important factor for spreading hashtags and links in disaster and politics tweets. Hashtags occur more in retweets, than in originals, in both disaster and politics tweets.

d. Retweets were used more in disaster and breaking news, e.g. Nepal earthquake, Charlie Hebdo and clock boy stories. This result is consistent with the finding of (Wang & Zheng, 2014), presented in the literature. The study stated that the single-spike pattern, pattern that happen as result of an event, had a larger retweet ratio.

e. The distribution of links used also followed the power-law distribution, i.e., few links were shared many times, and large number of links were shared few times. The same applies for hashtags.

f. BBC appeared in the top 3 most shared links in 5 stories, and since the stories used in the study had global interest, this may indicate that this news agency has more worldwide recognition. The results found by Bhattacharya and Ram (2012) presented in the literature, supports this finding, where they found that BBC had the maximum article spread and longest lifespan than the other news sources they investigated.

6.5 Limitations

6.5.1 Keyword search. To create the datasets for the news stories selected, keyword search was used to retrieve relevant tweets from DNLP (Dalhousie Natural Language Processing) research group dataset, which is a collection of tweets gathered from the Twitter stream. The returned tweets from the keyword search normally contain a fraction of irrelevant posts, which may affect the analysis results. The other drawback of using keyword search, to find enough tweets for a story, we had to use the hashtags used

in the story by the tweeters. This influenced the number of tweets with hashtags retrieved by the search, which might yield to a slight inaccuracy in the analysis.

6.5.2 User location. The location of the tweeter can be obtained from user profile; however, these geographic locations do not necessarily indicate the real physical location or nationality of the user. Additionally, this field is a user defined field, therefore, it could be inaccurate or incorrect. Another limitation with locations, was the errors in geocoding. User location may contain incorrect location name, such as 'Heaven', however, the geocoder mapped it to some real location. Finding geographic locations of users accurately is an open research area, and may require other methods such as the analysis of friendship networks (Ren, Zhang, & Lin, 2012), and contents analysis (Cheng, Caverlee & Lee, 2010).

6.5.3 Dataset size. The number of stories used in the study was limited to six stories. This number might be insufficient to draw general conclusions. Due to the time and resources constraints we kept the data limited to this size. Also, the sizes of the stories datasets vary. Although it is preferable in research involving more than one sample to keep samples sizes equal or near equal, in our case, the large variation between datasets sizes could not be avoided. Since we were dealing with real events, and real data, these variations were likely to happen, and they reflect the size of the event. The last limitation with size involve the use of the 1% sample of twitter stream, this may have influenced the rate of tweets per user, which might be larger if the full Twitter stream was analyzed.

Chapter 7 Conclusion and Future Work

The discussion provided in the previous chapter highlighted the findings of the study of the relationships between countries, news types and users' behavior in Twitter. In this chapter, we provide a summary of the research, contribution and future work.

7.1 Summary

News sharing is one of the most popular activities in Twitter, over 85% of topics are headline news or comments about news. In this study we aim to investigate the relationship between news types, geographic locations and some of the elements that characterize users' behavior or participation type, namely, posting an original tweet vs retweet, the use of hashtags and the use of links. We formulated three research questions, and proposed the methodology that will enable answering these questions. The research questions are:

RQ1: Is there a relationship between country and the type of user participation in Twitter?

RQ2: Is there a relationship between news type and user participation in Twitter?

RQ3: Is there an interaction between country and news type? And if interaction exists, how does it influence the user participation in Twitter?

We found that user behavior is related to country, news type and the interaction between country and news type. Users from different geographic locations reacted differently to the different news events. Countries generated varying number of tweets, for instance, the top 10 countries generated 73% of all the tweets, and the 3 common countries, USA, UK and India generated 47% of all tweet. Additionally, the common countries contributed more in creating original contents compared to the rest of the 223

countries in the dataset. The investigation also revealed that the behavior of a country may be explained partially by factors such as, the close proximity to the origin of the event e.g. India in Nepal earthquake story, whether there are economic or diplomatic relations between the country and the country of the event origin e.g. Nigeria and China, and whether the country is involved in the event, e.g. Spain in Germanwings plane crash. These examples are associated with the news stories under investigation, different scenarios may have different explanations. Some countries' behaviors could not be explained within the scope of this research, and further inquiry of countries' social, economic, political situations might be needed to obtain knowledgeable explanation of their behaviors.

The descriptive analysis for the three news types datasets showed that users' behavior varied considerably by the type of news. Finance had the highest proportion of originals (64%), politics had the highest proportion of retweets (73%). Finance tweets had the most links (91%), and political tweets contained the most hashtags (80%).

Statistical analyses were conducted to create three logistic regression interaction models, one for each dependent variable (originals, hashtags and links) and the independent variables country, news type, and the interaction term country*news type. Country has three levels, USA, UK and India, and news type has three levels, finance, disaster and politics. The results of the regression analysis supported the results from the descriptive analysis. Odds ratios for comparing countries showed variations ranging from 4.5% to 33%, whereas odd ratios for comparing news types ranged from -89% to 5 times. News types variations were greater than countries variations. Finance stood out with high

odds ratio in generating both original tweets and tweets with links, and low odds ratios in generating tweets with hashtags compared to both politics and disaster news.

The distribution of tweets among all countries in the whole database followed power-law distribution, where high counts of tweets were generated by few countries, and high number of countries generated few tweet. The distribution of retweets, originals, tweets with links and tweets with hashtags all followed the same pattern, in the whole database and in the individual news types datasets. The same applies for links and hashtags shared in each story, few hashtag and links appeared large number of times, while the majority of the other hashtags and links appeared few times.

A framework for applying the methodology used in this research include the following steps:

1. Defining objectives and research questions to identify what the research is trying to find out and why.
2. Collecting data using live stream or pre-collected sets of tweets.
3. Preprocessing the tweets including both cleaning and formatting. This step depends on the objectives of the study, for instance if sentiment analysis is required the tweet text should be cleaned by eliminating symbols, and misspelled words.
4. Collecting additional data, such as geographic locations, as appropriate for the research question.
5. Storing all the collected data in an accessible form for quick and efficient retrieval and to allow the extraction of other descriptive data, such as counts and averages.
6. Understanding the data, through descriptive analysis, including summarizing data to identify trends in the data.

7. Perform appropriate statistical analysis on the data to test the hypothesis posed by the research questions and to find relationships among study variables.
8. Using the results of the statistical procedure, each hypothesis is tested and answers to research questions clearly articulated.

7.2 Research Contributions

The main research contributions of this research study are:

1- We found that there are relationships between countries, news types and user behavior for news related tweets.

2- Applying logistic regression to analyze the relationships between the variables of the study is an appropriate approach for modeling Twitter data, countries and news types.

3- Geographic locations can be used to analyze the distribution of tweets in addition to traditional information propagation methods.

4- News type is strongly correlated to users' behavior.

5- News type has a stronger relationship to users' behavior than does geographic location.

7.3 Implications of Research Findings

The findings of the research study provide insight on how the different user behaviors (hashtags, links and retweets) were used across news type and countries for some news events. This knowledge could be employed in various social media applications, such as news recommendation, hashtag recommendation, filtering, map-based social media applications and disaster management.

Based on the finding of the study, we provide the following recommendations:

1- The behaviors of users by country varied depending on the type of events among other factors. Understanding outliers or shifts from expectations, for instance Indonesia or Nigeria may be useful in the design of real-time filtering methods and location-aware news recommendations.

2- Tweets with hashtags and tweets with links showed different patterns among countries. Combining this knowledge with an investigation of the hashtags and links used by each country, would be useful in recommending hashtags and links and popularity prediction for users based on their location.

3- The top 10 hashtags were used heavily in most of the stories examined, 47% - 99% of the hashtags used were from the top 10. However, the top 10 links were not frequently used in the tweets. This finding could be utilized when designing news recommendation using hashtags and links.

4- Investigating user behavior in finance related tweets could assist in both price prediction research and marketing applications.

5- Studying changes in user participation behavior by location during disasters could help in disaster management and awareness, by recognizing on-site posts vs comments from further away.

6- The findings of this research would aid in the design and implementation of map-based browsing applications. Finding hashtags, links and sentiment by country or by city, could also be incorporated in such application.

7- Twitter is a rich source of social data, considering collaboration between computer science and social science would provide more insight for further research.

7.4 Future Work

Twitter usage is an active research area, nevertheless research involving countries and associated behaviors are limited. This research study provided some insight on users' behaviors in global political, financial and disaster events. Other genre of news, or perhaps other stories such as sports or entertainment could be investigated to fill the gaps in understanding of other aspects of user behavior.

The popularity of smart phones and location aware devices, in addition to the availability of geocoding and mapping services, encourages us to develop a prototype of a map-based browsing application for news-related tweets. Employing the findings and the knowledge gained from this study, would give the opportunity to create a visual framework for finding tweets along with their geographic distribution. Finding hashtags, links and sentiment by country or by city, could also be incorporated in the framework.

The use of hashtags is a widely adopted feature in social media and in Twitter in particular. Research involving hashtag recommendation is now emerging as a result of this wide use of this feature. Using social network analysis as well as user location would allow the suggestion of hashtags based on the type of news and geographic location. Additionally, information retrieval or machine learning methods can be employed to explore the area of hashtags ranking and recommendation.

In the early stages of the study, sentiment analysis was applied to tweets and stored in the database, but not used. In the future we intend to analyze sentiment in conjunction with news types and countries. Applying sentiment analysis to news-related tweets provide an insight on public opinions, which journalists and news curators seek, especially during political events.

Social media applications are advancing in a rapid rate, however, the human factor still has a substantial impact on shaping the way information shared in social media. In this study, the behavior of social media users was analyzed considering two dimensions, geographic location and news type. The results indicated that user behavior is greatly affected by the type of news, while geographic location had less influence on the behavior of users. For instance, people tend to write more original tweets in finance than politics and disaster, and include more hashtags in politics that finance and disaster. The findings of this study could assist researches and developers in areas concerning modeling user behavior, detecting anomalies, filtering, ranking and recommendation systems.

References

About Twitter, Inc. | About. (n.d.). Retrieved November 16, 2014, from https://about.twitter.com/company

Agrawal, D., Budak, C., El Abbadi, A. E. A., Georgiou, T., & Yan, X. (2014). Big Data in Online Social Networks: User Interaction Analysis to Model User Behavior in Social Networks. In *Databases in Networked Information Systems* (p. pp 1–16). Springer International Publishing.

Angeli, E., Wagner, J., Lawrick, E., Moore, K., Anderson, M., Soderlund, L., & Brizee, A. (2010, May 5). General format. Retrieved from http://owl.english.purdue.edu/owl/resource/560/01/

Antoniades, D., Polakis, I., Kontaxis, G., Athanasopoulos, E., Ioannidis, S., Markatos, E. P., & Karagiannis, T. (2011). we.b: the web of short urls. *Proceedings of the 20th international conference on World wide web - WWW '11*. doi:10.1145/1963405.1963505

Bhattacharya, & Ram. (2012). Sharing news articles using 140 characters: A diffusion analysis on Twitter. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* doi:10.1109/asonam.2012.170

Bogdanov, P., Busch, M., Moehlis, J., Singh, A. K., & Szymanski, B. K. (2014, June). *Modeling individual topic-specific behavior and influence backbone networks in social media*. doi:10.1007/s13278-014-0204-6

Cao, C., & Caverlee, J. (2104, July). *Behavioral Detection of Spam URL Sharing: Posting Patterns versus Click Patterns*.

Chen, W., Lakshmanan, L. V. S., & Castillo, C. (2013). Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, *5*(4), 1–177. doi:10.2200/s00527ed1v01y201308dtm037

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*. doi:10.1145/1871437.1871535

Choi, D., & Kim, P. (2013). Sentiment analysis for tracking breaking events: A case study on Twitter. In A. Selamat et al. (Eds.) *Intelligent Information and Database Systems,* 7803, 285-294

Christodoulou, G., Georgiou, C., & Pallis, G. (November 28-30, 2012). The role of Twitter in YouTube videos diffusion. *In Web Information System Engineering: 13th International Conference*, Paphos, Cyprus. doi: 10.1007/978-3-642-35063-4_31

Cuesta, Á., Barrero, D. F., & R-Moreno, M. D. (2013). A descriptive analysis of Twitter activity in Spanish around Boston terror attacks. *Computational Collective Intelligence. Technologies and Applications*. doi:10.1007/978-3-642-40495-5_63

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. United Kingdom: Cambridge University Press.

Elghazaly, T., Mahmoud, A., & Hefny, H. A. (2016). Political sentiment analysis using Twitter data. Proceedings of the International Conference on Internet of things and Cloud Computing - ICC '16. doi:10.1145/2896387.2896396

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016, June). *Social Clicks: What and Who Gets Read on Twitter?*

Gruhl, D., Guha, Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. *Proceedings of the 13th Conference on World Wide Web - WWW '04*. doi:10.1145/988672.988739

Hansen, L. K., Arvidsson, A., Nielsen, F. A., Colleoni, E., & Etter, M. (2011). Good friends, bad news - affect and Virality in Twitter. In *Communications in Computer and Information Science* (pp. 34–43). doi:10.1007/978-3-642-22309-9_5

Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., & Ma, K. Breaking news on Twitter in *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM,* 2751-2754

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency. ACM Computing Surveys, 47(4), 1–38. doi:10.1145/2771588

Ineson, D., & Anderson, M. (2016). Understanding sporting fandom in social media: A UK perspective of professional Rugby league clubs. 2016 SAI Computing Conference (SAI). doi:10.1109/sai.2016.7556137

Ji, X., Chun, S. A., Wei, Z., & Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, *5*(1). doi:10.1007/s13278-015-0253-5

Jungherr, A. (2015). Twitter, usage and research. *Analyzing Political Communication with Digital Trace Data*. doi:10.1007/978-3-319-20319-5_2

Karkulahti, O., Pivovarova, L., Du, M., Kangasharju, J., & Yangarber, R. (2016). Tracking interactions across business news, social media, and stock fluctuations. doi:10.1007/978-3-319-30671-1_61

Kogan, M., Palen, L., & Anderson, K. M. (2015, February). *Think local, Retweet global*. doi:10.1145/2675133.2675218

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proc. WWW '10*, *ACM* 591–600.

Lai, M., Bosco, C., Patti, V., & Virone, D. (2015). Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. doi:10.1109/dsaa.2015.7344884

Li, C., Kuo, T., Ho, C., Hong, S., Lin, W., Lin, S. (2013). Modeling and evaluating information propagation in a microblogging social network. *Social Network Analysis and Mining*, 3(3), 341-357. doi: 10.1007/s13278-012-0082-8

Li, J., Vishwanath, A., & Rao, H. R. (2014). Retweeting the Fukushima nuclear radiation disaster. *Communications of the ACM*, *57*(1), 78–85. doi:10.1145/2500881

Lima, A., & Musolesi, M. (2012). Spatial dissemination metrics for location-based social networks. Proceedings from UbiComp '12: *The 2012 ACM Conference on Ubiquitous Computing*

Louni, A., & Subbalakshmi, K. (2014). Diffusion of information in social networks. *Intelligent Systems Reference Library*, 1–22. doi:10.1007/978-3-319-05164-2_1

Maity, S. K., Saraf, R., & Mukherjee, A. (2016, February). *#Bieber + #Blast = #BieberBlast: Early Prediction of Popular Hashtag Compounds*.

Merriam-Webster (n.d.). Retrieved November 16, 2014, from http://www.merriam-webster.com/

Moyé, L. A. (2006). P-values, power, and efficacy. In *Statistical Reasoning in Medicine* (pp. 137–156). doi:10.1007/978-0-387-46212-7_7

Nizam, N. Watters, C., & Gruzd, A. (2013). Link sharing on Twitter during popular events: Implications for social navigation on websites. *In the 47th Hawaii International Conference on System Science,* HI.

Ota, Y., Maruyama, K., & Terada, M. (2011). *Discovery of interesting users in Twitter by overlapping propagation paths of retweets*.

Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org

RefME | Free Reference Generator. (n.d.). Retrieved December 16, 2014, from http://www.refme.com

Ren, K., Zhang, S., & Lin, H. (2012). Where are you settling down: Geo-locating Twitter users based on tweets and social networks. *Lecture Notes in Computer Science*, 150–161. doi:10.1007/978-3-642-35341-3_13

Ruan, Y., Alfantoukh, L., & Durresi, A. (2015). Exploring stock market using Twitter trust network. 2015 IEEE 29th International Conference on Advanced Information Networking and Applications. doi:10.1109/aina.2015.217

Seron, W., Zorzal, E., Quiles, M. G., Basgalupp, M. P., & Breve, F. A. (2015). #Worldcup2014 on Twitter. doi:10.1007/978-3-319-21404-7_33

Shi, B., Ifrim, G., & Hurley, N. (2016, April). *Learning-to-Rank for Real-Time High-Precision Hashtag Recommendation for Streaming News*.

Starkweather, J. S., & Herrington, R. H. RSS SPSS short course. Retrieved November 1, 2016, from http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/SPSS_SC/

Stollberg, B., & de Groeve, T. (2012). The use of social media within the Global Disaster Alert and Coordination System (GDACS). *Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion*. doi:10.1145/2187980.2188185

Subašić, I., & Berendt, B. (2011). Peddling or creating? Investigating the role of Twitter in news reporting. *Advances in Information Retrieval*, 207–213. doi:10.1007/978-3-64220161-5_21

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*(1), 73–81. doi: 10.1016/j.socnet.2011.05.006

The Healthcare Hashtag Project. (n.d.). Retrieved December 16, 2014, from http://www.symplur.com/healthcare-hashtags

Toriumi, F. T. (2016, April). Real-time Tweet Classification in Disaster Situation. In S. B. Baba (Ed.), *WWW'16 Companion*. . doi:10.1145/2872518.2889365

Trung, D. N., Jung, J., Lee, N., & Kim, J. (2013). Thematic analysis by discovering diffusion patterns in social media: An exploratory study with TweetScope. *Lecture Notes in Computer Science*, 266–274. doi:10.1007/978-3-642-36543-0_28

Twitter Developers. (n.d.). Retrieved December 16, 2014, from https://dev.twitter.com/

Van Liere, D. (2010). How far does a tweet travel? *Proceedings of the International Workshop on Modeling Social Media - MSM '10*. doi:10.1145/1835980.1835986

Verma, J. P. (2012). Logistic regression: Developing a model for risk analysis. In *Data Analysis in Management with SPSS Software* (pp. 413–442). doi:10.1007/978-81-322-0786-3_13

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. M. (2005). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models* (1st ed.). New York, NY: Springer-Verlag New York.

Vosecky, J., Jiang, D., & Ng, W. (2013). Limosa: A system for geographic user interest analysis in Twitter. *Proceedings of the 16th International Conference on Extending Database Technology - EDBT '13*. doi:10.1145/2452376.2452460

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013). A depression detection model based on sentiment analysis in micro-blog social network. *Trends and Applications in Knowledge Discovery and Data Mining*, 201–213. doi:10.1007/978-3-642-40319-4_18

Wang, Y., & Zheng, B. (2014). On macro and micro exploration of hashtag diffusion in Twitter. *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. doi:10.1109/asonam.2014.6921598

Wu, B., & Shen, H. (2015). Corrigendum to "Analyzing and predicting news popularity on Twitter" [Int. J. Inf. Manage. 35 (6) (2015) 702–711]. *International Journal of Information Management*. doi: 10.1016/j.ijinfomgt.2015.09.004

Xu, W. W., Chiu, I.-H., Chen, Y., & Mukherjee, T. (2014). Twitter hashtags for health: applying network and content analyses to understand the health knowledge sharing in a Twitter-based community of practice. *Quality & Quantity*. doi:10.1007/s11135-014-0051-6

Yang, S. Y., & Mo, S. Y. K. (2016). Social media and news sentiment analysis for advanced investment strategies. Sentiment Analysis and Ontology Engineering. doi:10.1007/978-3-319-30319-2_11

Yang, X., Ghoting, A., Ruan, Y., & Parthasarathy, S. (2012). A Framework for summarizing and analyzing Twitter feeds. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12*. doi:10.1145/2339530.2339591

Ye, S., & Wu, F. (2010). Measuring message propagation and social influence on Twitter.com. *Lecture Notes in Computer Science*, 216–231. doi:10.1007/978-3-642-16567-2_16

Yeung, C. A. Y., & Iwata, T. (Eds.). (2011). Modelling User Behaviour and Interactions: Augmented Cognition on the Social Web. In *Foundation of Augmented Cognition* (pp. 277–287). Japan.

Zhou, Z., Bandari, R., Kong, J., Qian, H., & Roychowdhury, V. (2010). Information resonance on Twitter: Watching Iran. *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*. doi:10.1145/1964858.1964875

Appendix A – Characteristics of the Top 10 Countries by News Story

The top ten countries for each news story were presented in section 5.4, below we show the analyses of the use of originals, hashtags and links in the top ten countries in each news story.

Original and retweets

Tweet, original and retweets raw counts in top countries in each story are shown in Figure A-1. Generally, the distributions followed a power law distribution, with some variations among the stories. Both finance stories showed higher counts of originals, while disaster and politics stories had higher retweets counts.

Ratios of tweets, original and retweets to total tweets, original and retweets in each story across the top 10 countries are presented in Figure A-2. In Chinese shares story USA accounted for around 35%, UK and Nigeria 15% each, of total retweets. In disaster stories, and in Germanwings story in particular, around 40% of all retweets were originated from USA and Spain, whereas in Nepal earthquake story around 50% of all tweets were originated from USA and India, in retweets and originals the ratios were almost the same. In Charlie Hebdo around 50% of total retweets originated from USA and France, while in clock boy story 50% of retweets were originated from USA.

Figure A-3 illustrates the ratios of originals and retweets within each of the top 10 countries. As introduced earlier the ratios of originals were higher in finance stories as the orange bars of the figure indicate. Contrarily, in politics, stories showed higher retweets ratio. Disaster stories showed inconsistent ratios, high originals ratio in Germanwings story and high retweet ratio in Nepal Earthquake story.

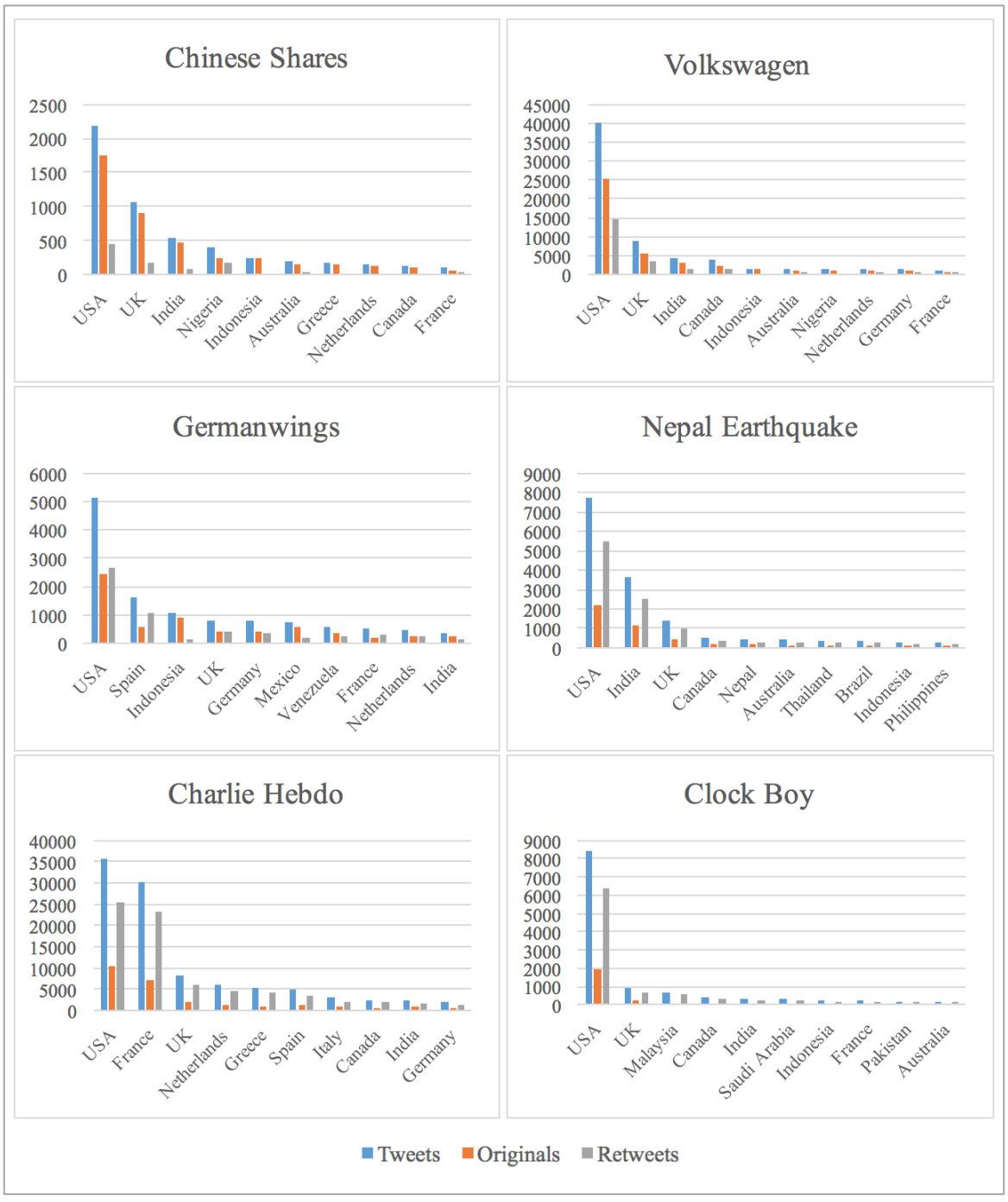

Figure A-1 Counts of tweet, original and retweets in the top 10 countries

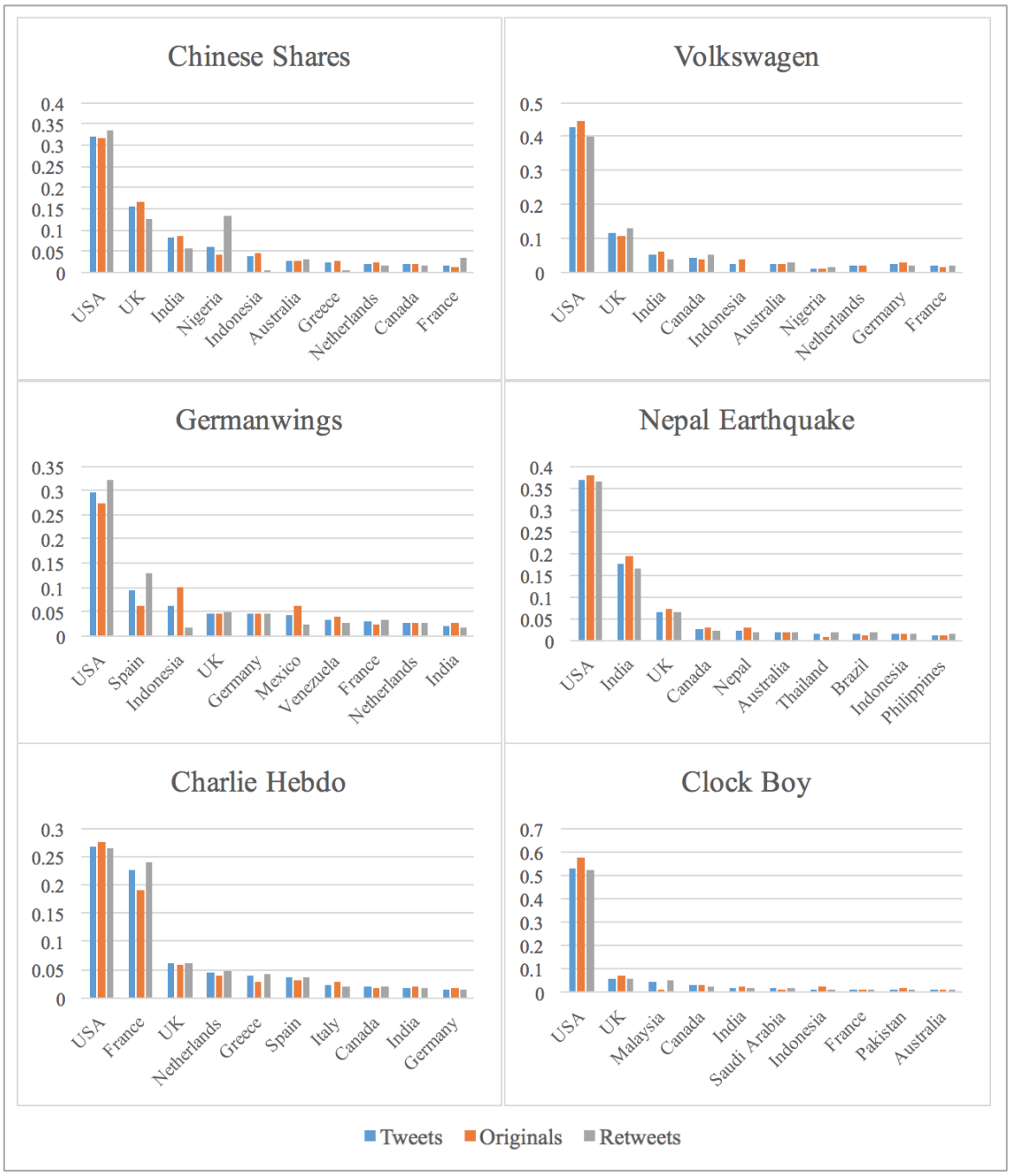

Figure A-2 Ratios of tweets, original and retweets to total tweets, original and retweets in each story across the top 10 countries

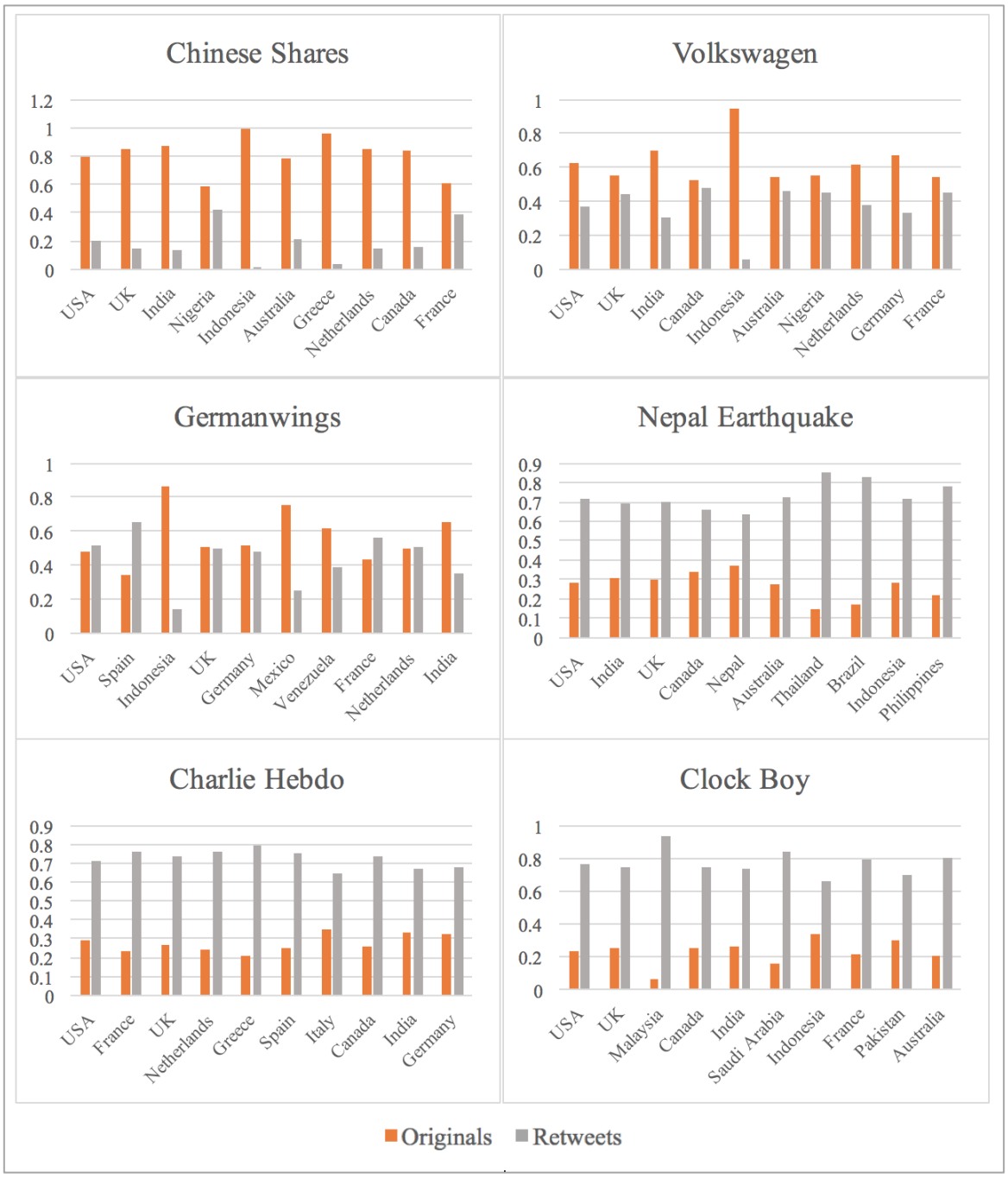

Figure A-3 Originals and retweets ratios within each of the top 10 countries in each story

Hashtags

Figure A-4 shows the hashtags distribution among the top 10 countries in each news story. In finance stories originals count including hashtags were higher than retweets, except for Nigeria in Chinese shares story. In disaster stories retweets with hashtags counts were higher except for Indonesia and Venezuela in Germanwings story. In both politics stories retweets with hashtags were more than originals.

Figure A-5 shows the ratios of tweets, original and retweets to total tweets, original and retweets containing hashtags in each story across the top 10 countries. In finance more than 50% of retweets with hashtags were originated from Nigeria, in Volkswagen story around 50% of originals originated from USA and UK, and the same for retweets. In Germanwings story around 45% of retweets, and 40% of originals with hashtags came from USA and Spain, while in Nepal earthquake story more than 55% of originals and around 55% of retweets with hashtags originated from USA and India. Lastly in Charlie Hebdo story 50% of retweets with hashtags originated from USA and France, and about the same for originals, and in clock boy story USA makes 50% of retweets with hashtags, and originals as well.

Figure A-6 presents originals and retweets with hashtags ratio within each of the top 10 countries in each story. In both finance stories, originals with hashtags ratio were higher than retweets, except for Nigeria in chines shares story which showed the opposite. In most of the countries in disaster stories the retweets with hashtags ratio were higher, however, Indonesia, India and Venezuela had different ratios, were the originals with hashtags ratios were higher. In both politics stories the retweets with hashtags ratios were higher than originals in all of the top countries.

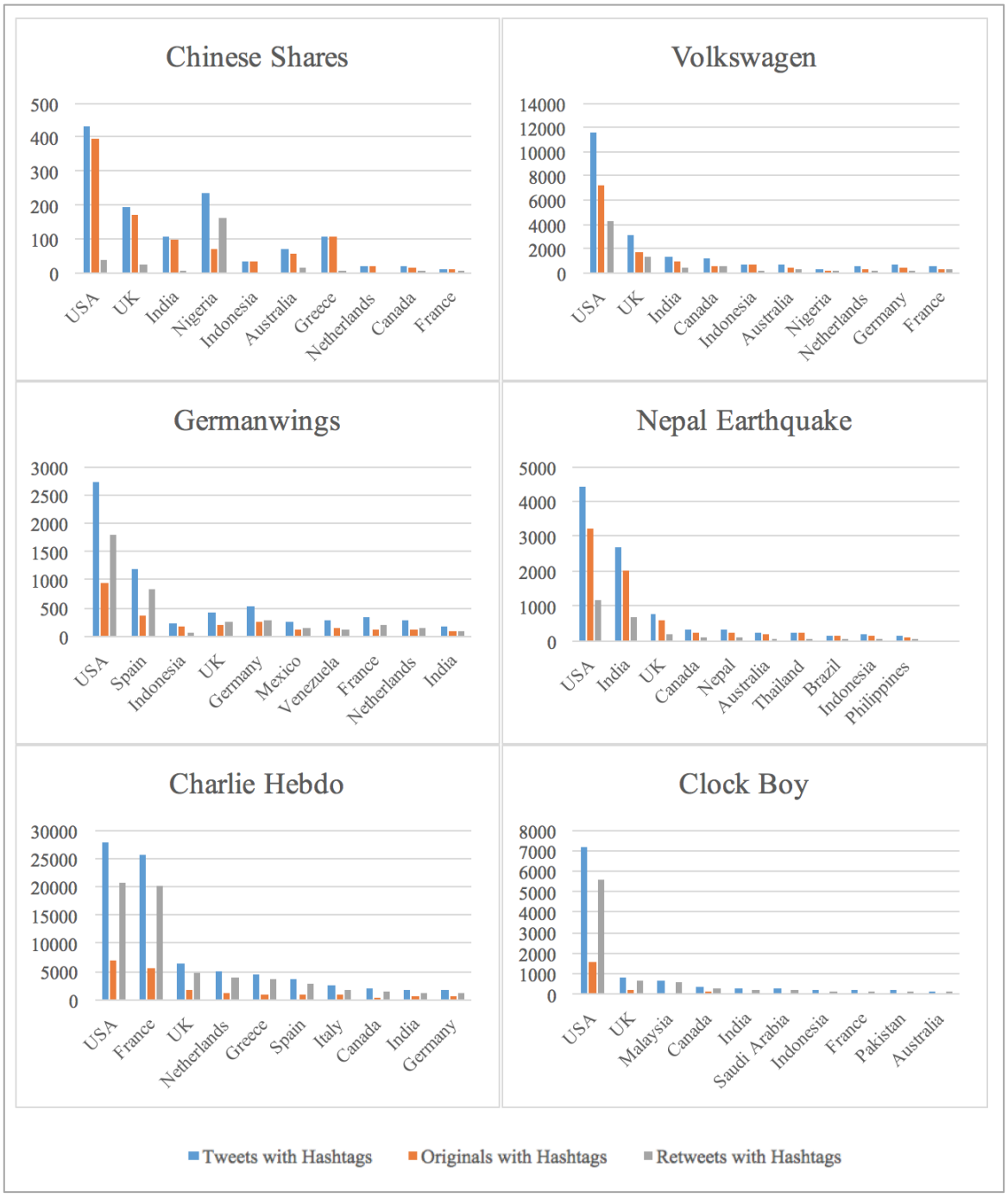

Figure A-4 Counts of tweet, original and retweets containing hashtags in the top 10 countries

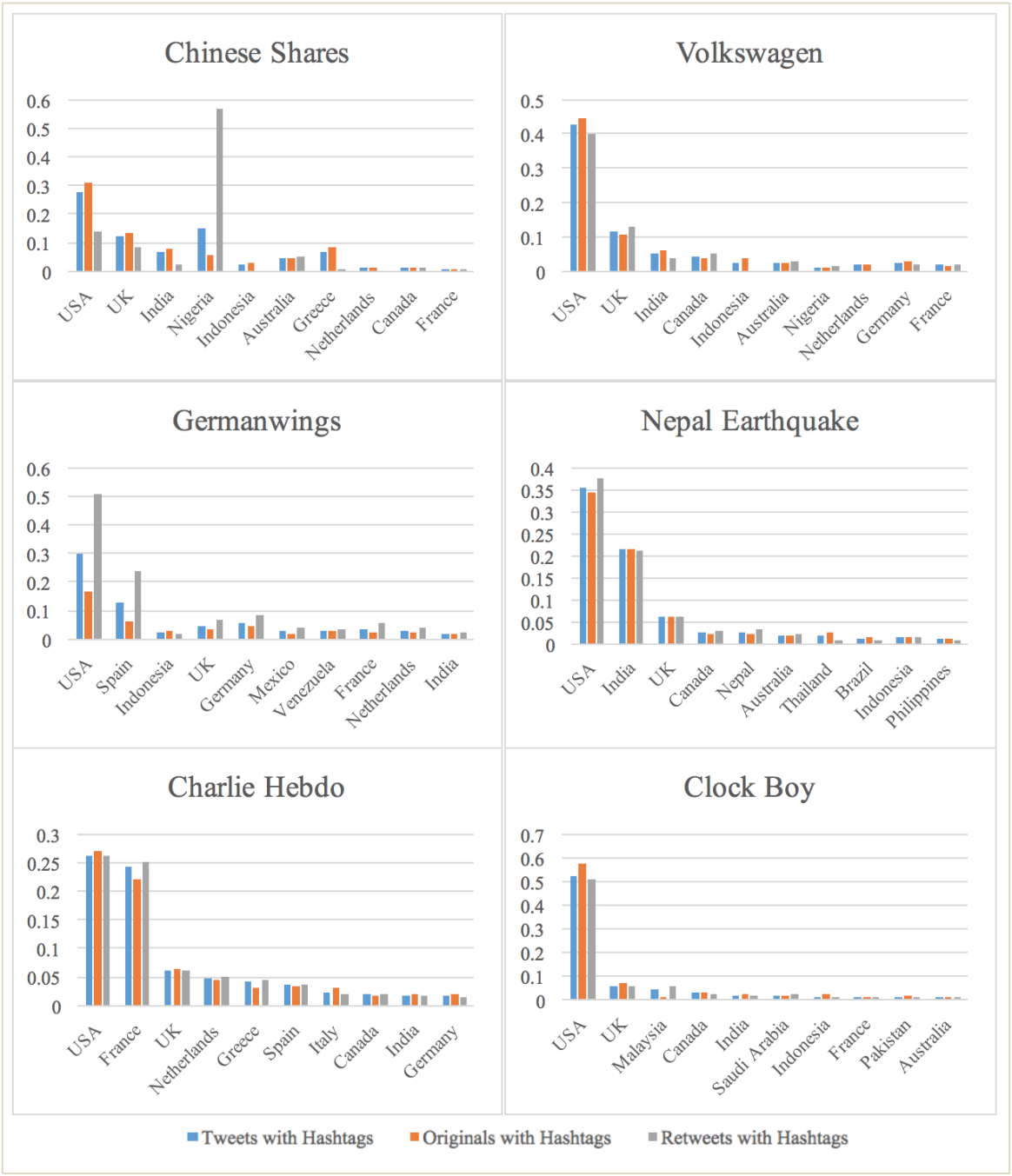

Figure A-5 Ratios of tweets, original and retweets to total tweets, original and retweets containing hashtags in each story across the top 10 countries

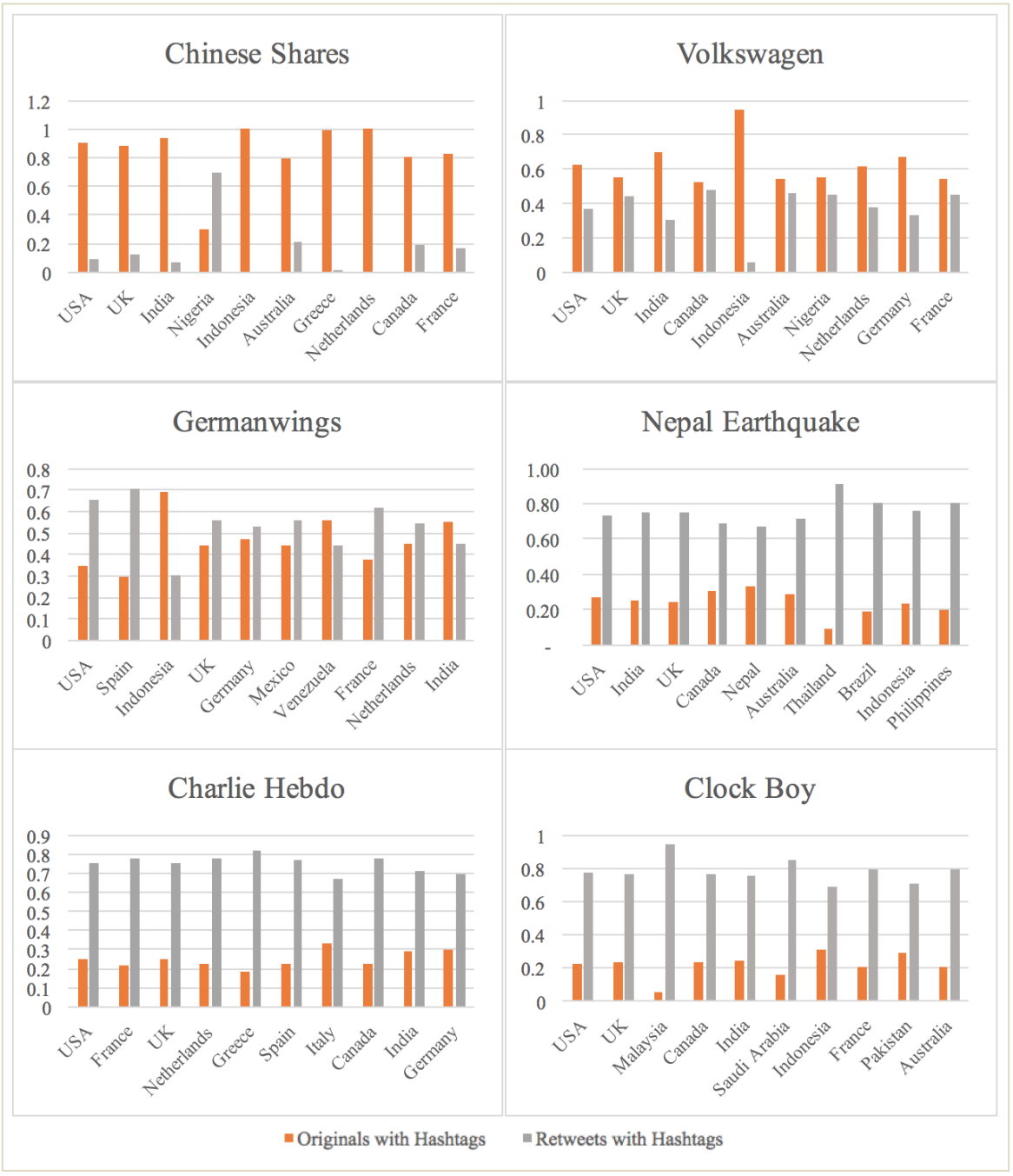

Figure A-6 Originals and retweets with hashtags ratios within each of the top 10 countries in each story

Links

Figure A-7 displays the counts of tweet, original and retweets containing links in the top 10 countries. In finance, similar to the distributions of originals and hashtags, originals with links counts were more than retweets. In disaster stories, and in Germanwings story in particular, originals with links counts were more than retweets, while in Nepal earthquake story retweets with links were more than originals. In politics, retweets with links counts were higher than originals with links across all the countries in both stories.

Figure A-8 presents the ratios of tweets, original and retweets to total tweets, original and retweets containing links in each story across the top 10 countries. USA accounted for 25% to 50% of all tweets with links, in all stories. In finance, and in Chinese shares story, USA, Nigeria and UK presented approximately 60% of all retweets with links, and around 50% of originals. In Volkswagen story more than 50% of tweets with links came from USA and UK, with similar distributions for originals and retweets. In Germanwings story more than 40% of retweets with links originated from USA and Spain, and more than 45% of originals with links came from USA, Spain, Indonesia and Mexico. In Nepal earthquake story more than 60% of tweets originated from USA, India and UK. Lastly, in politics and in Charlie Hebdo story, more than 50% of retweets with links originated from USA and France, and in clock boy story more than 50% of retweets with links originated from USA.

Figure A-9 illustrates the ratios of originals and retweets with links within each of the top ten countries in each story. In finance, both stories have higher link sharing ratio in originals than retweets, except for France in Chinese shares story, where the ratios of

retweets and originals were equal. In disaster stories, the distributions were not consistent, in Germanwings originals with links ratios were higher than retweets, except for France and Spain. In Nepal earthquake story, retweets with links have higher ratios in all countries. In both politics stories the retweets with links ratios were higher than originals in all countries.

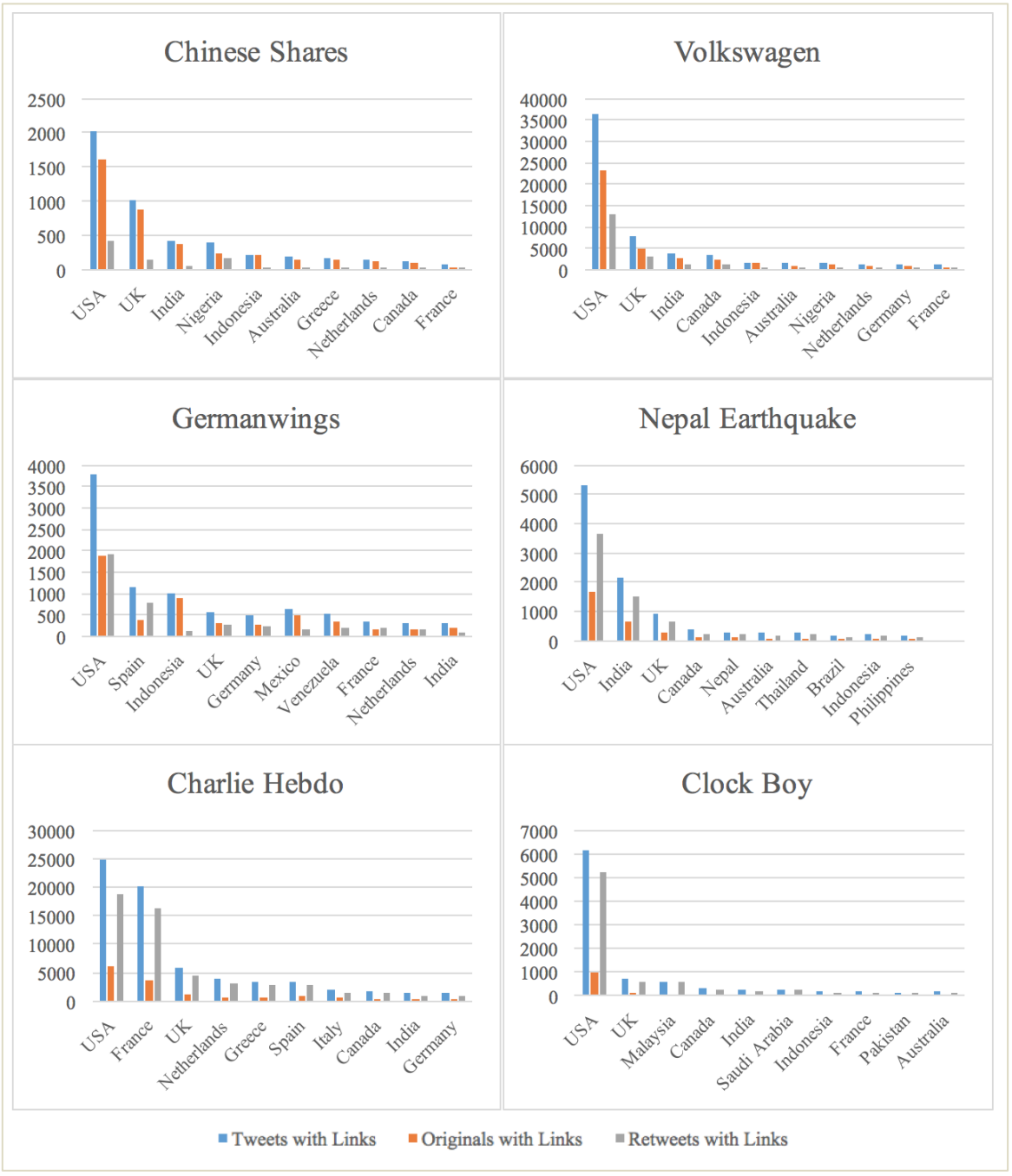

Figure A-7 Counts of tweet, original and retweets containing links in the top 10 countries

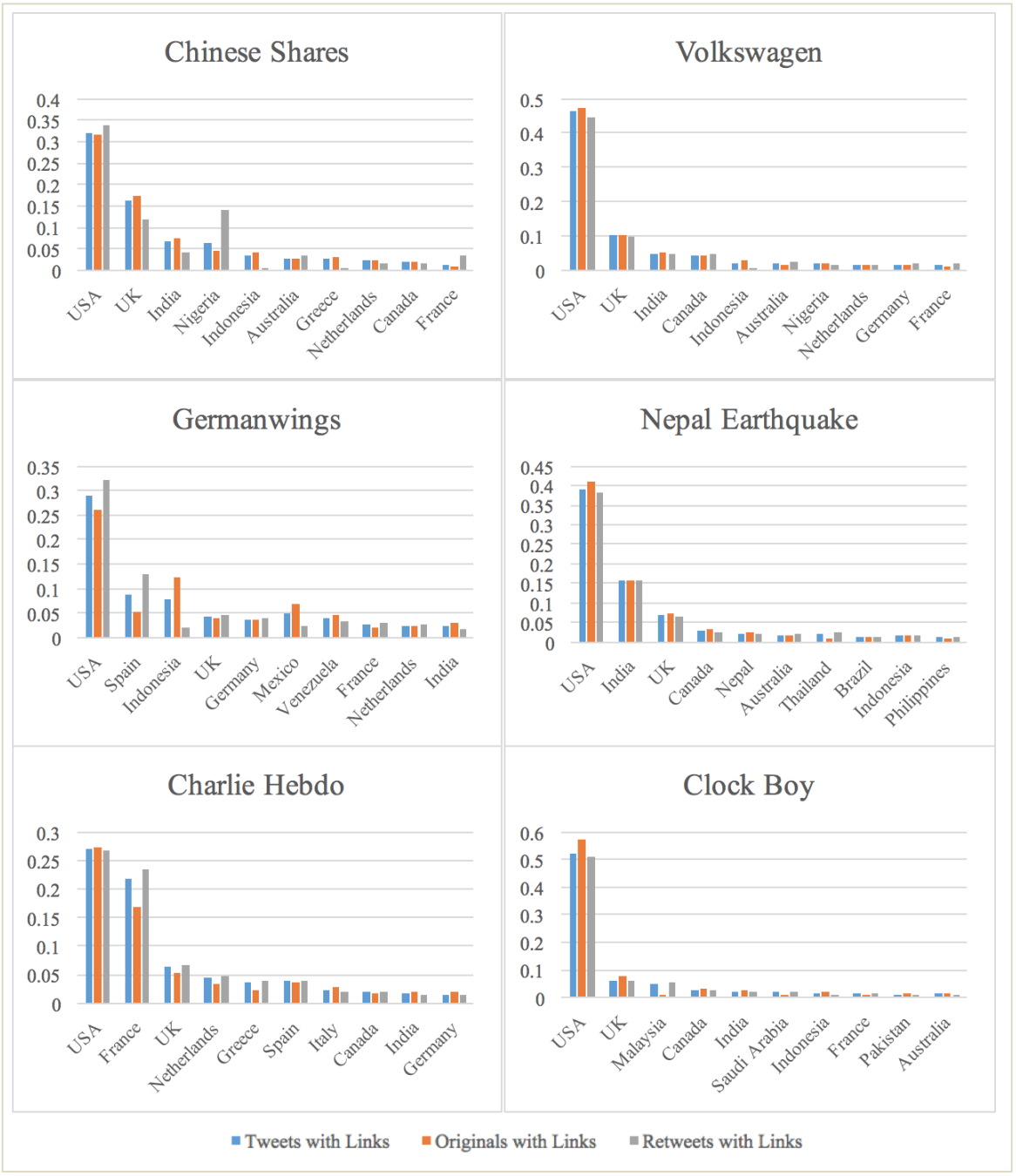

Figure A-8 Ratios of tweets, original and retweets with links to total tweets, original and retweets containing links in each story across the top 10 countries

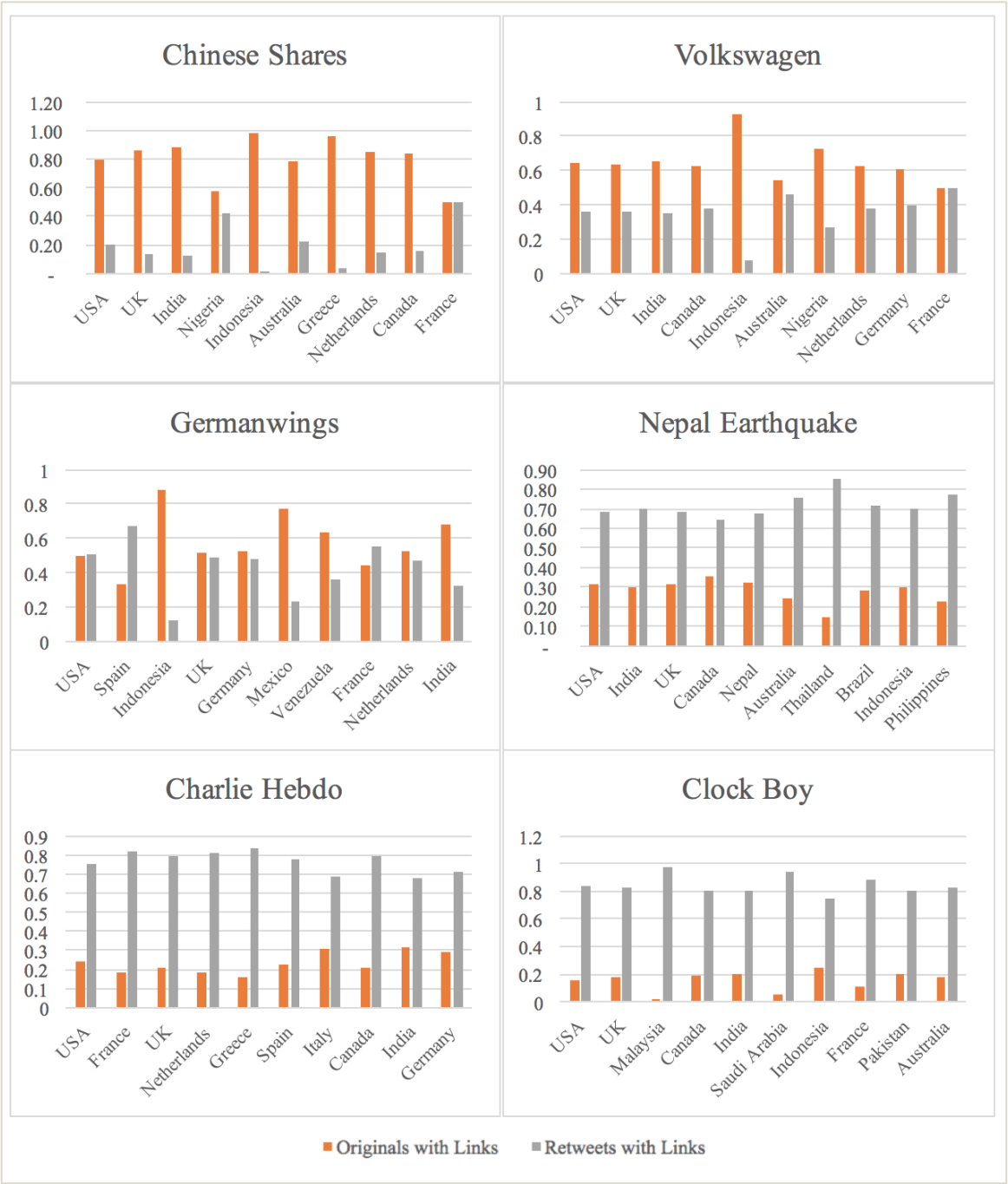

Figure A-9 Ratios of originals and retweets with links within each of the top 10 countries in each story

Appendix B – Top 10 Hashtags Analysis

Finance stories

The top 10 hashtags in each finance story are presented in Table B-1. Hashtags in Chinese shares story included words such as news, worldnews, china, aisa, world, bbc, and business. Two hashtags were unrelated, these are focusonmeiscoming and sharehumanity. In Volkswagen story hashtags included the words volkswagen, auto, scandal, and martinwinterkorn, which is the name of the chairman of the board of directors. The hashtags news and business appeared in both stories.

No.	Chinese Shares	Count	Volkswagen	Count
1	news	329	volkswagen	12,870
2	business	176	emissions	2,102
3	bbc	82	news	1,623
4	china	65	vw	1,246
5	world	51	business	990
6	focusonmeiscoming	43	scandal	636
7	chinese	41	tech	545
8	asia	37	autos	481
9	sharehumanity	35	breaking	451
10	worldnews	35	martinwinterkorn	429
Total		894		21,373

Table B-1 Top 10 hashtags in finance stories

In Chinese shares story hashtags were in 23% of all tweets, while the top ten hashtags appeared in 11% of all tweets, and tweets with one of the top ten hashtags made 47% of tweets with any hashtag, as shown in Table B-2. Originals and retweets with hashtags have similar percentages to total tweets, however, the top 10 hashtags appeared in only 6% of retweets. In Volkswagen story hashtags were in 31% of all tweets, the top ten hashtags were in 20% of all tweets, and 63% of all tweets with hashtags contained one

of the top 10 hashtags, i.e., most of the hashtags used in Volkswagen story were from the top ten. Originals and retweets with hashtags had similar percentages to total tweets with hashtags, however, the top 10 hashtags appeared in 15% and 27% of originals and retweets respectively.

Characteristic	Chinese Shares		Volkswagen	
	Count	%	Count	%
Tweets with Hashtags	1566	23%	27069	31%
Tweets with one or more of top 10 Hashtags	742	11%	17075	20%
% Tweets of Top 10 to Tweets with Any Hashtag		47%		63%
Originals with Hashtags	1279	23%	16278	30%
Originals with One or more of Top 10 Hashtags	664	12%	8310	15%
% Originals of top 10 to originals with hashtag		52%		51%
Retweets with Hashtags	287	22%	10791	33%
Retweets with One or more of Top 10 Hashtags	78	6%	8765	27%
% Retweets of top 10 to Retweets with hashtag		27%		81%

Table B-2 Top 10 hashtags analysis for finance stories

Figure B-1 illustrates hashtags counts distribution in finance stories, and Figure B-2 shows the log log scale of hashtags counts distribution in finance stories.

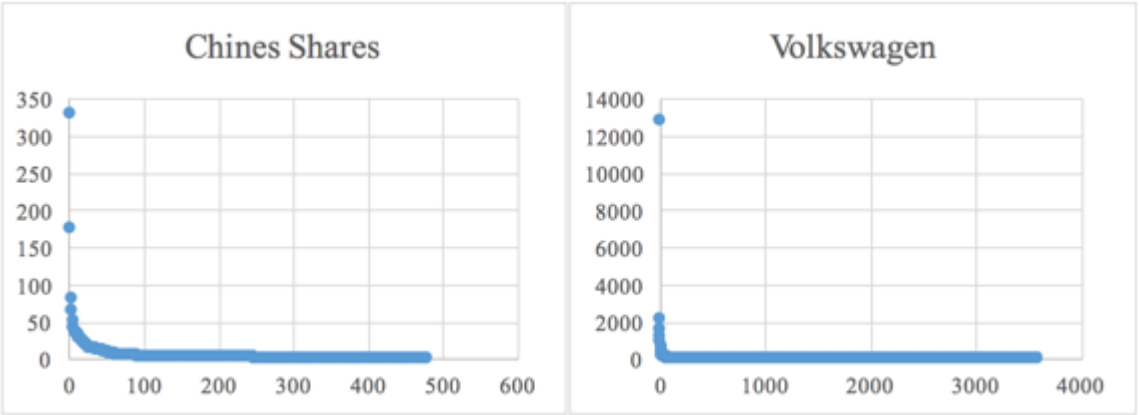

Figure B-1 Hashtags counts distribution in finance stories

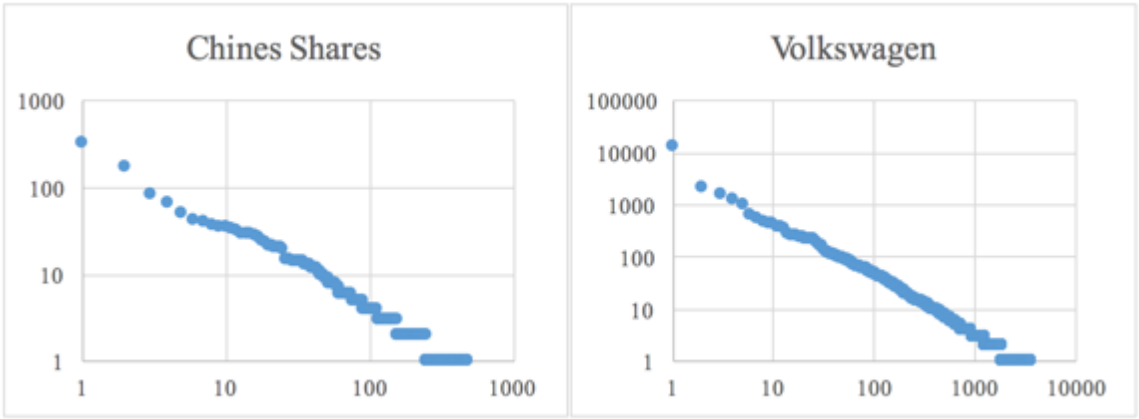

Figure B-2 Log log scale of hashtags counts distribution in finance stories

Disaster stories

The top 10 hashtags in each disaster story were presented in Table B-3. In Germanwings story, the first hashtag was used in the search for related tweets, other hashtags include germanwings, airbus, lufthanza, andreaslubitz, which is the name of the pilot, and crasha320. In Nepal earthquake story hashtags included nepalearthquake, nepal, earthquake and kathmandu which is the name of the city where the earthquake happened. The hashtag news was the only one that appeared in both stories.

No.	Germanwings	Count	Nepal Earthquake	Count
1	germanwings	6,951	nepalearthquake	7,919
2	4u9525	1,085	nepal	2,855
3	a320	277	earthquake	2,292
4	lufthansa	192	nepalquake	600
5	news	186	msghelpearthquakevictims	505
6	germanwingscrash	171	prayfornepal	469
7	andreaslubitz	169	kathmandu	296
8	crasha320	168	nepalquakerelief	194
9	accidenteaereo	108	india	184
10	airbus	88	news	130
Total		9,395		15,444

Table B-3 Top 10 hashtags in disaster stories

In Germanwings story tweets with hashtags made 53% of all tweets, and the top ten hashtags were in 44% of all tweets as shown in Table B-4. Hashtags and the top ten hashtags were more in retweets than originals, 69% and 63% in retweets versus 39% and 27% in originals. Tweets with one of the top 10 hashtags made 84% of tweets with any hashtag, 70% in originals, and 93% in retweets. In Nepal earthquake story tweets with hashtags made 60% of all tweets, and the top ten hashtags were in 56% of all tweets. Hashtags and the top ten hashtags ratio were slightly higher in retweets than originals, 62% and 59% in retweets versus 54% and 46% in originals. Tweets with one of the top 10 hashtags made 92% of tweets with any hashtag, 84% in originals, and 95% in retweets.

Characteristic	Germanwings		Nepal Earthquake	
	Count	%	Count	%
Tweets with Hashtags	9194	53%	12503	60%
Tweets with one or more of top 10 Hashtags	7711	44%	11559	56%
% Tweets of Top 10 to Tweets with Any Hashtag		84%		92%
Originals with Hashtags	3513	39%	3158	54%
Originals with One or more of Top 10 Hashtags	2449	27%	2651	46%
% Originals of top 10 to originals with hashtag		70%		84%
Retweets with Hashtags	5681	69%	9345	62%
Retweets with One or more of Top 10 Hashtags	5262	63%	8908	59%
% Retweets of top 10 to Retweets with hashtag		93%		95%

Table B-4 Top 10 hashtags analysis for disaster stories

Figure B-3 presents hashtags counts distribution in disaster stories, and Figure B-4 presents the log log scale of hashtags counts distribution in disaster stories.

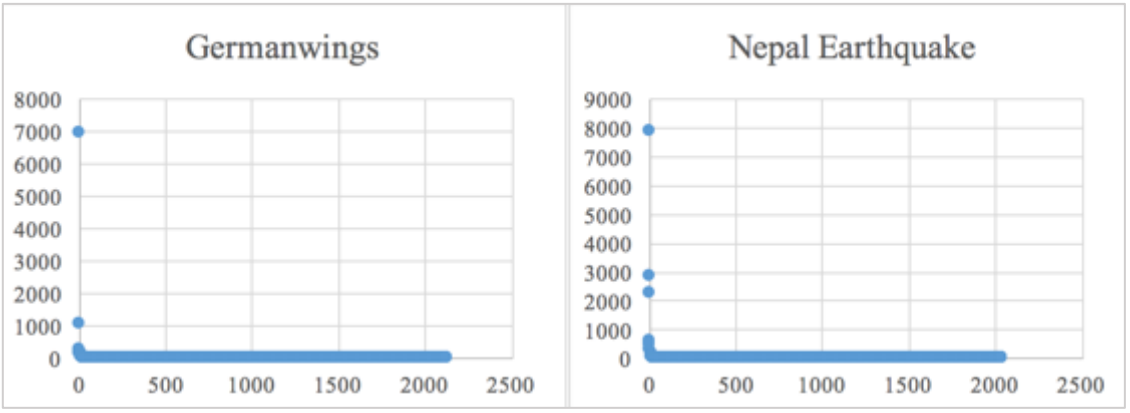

Figure B-3 Hashtags counts distribution in disaster stories

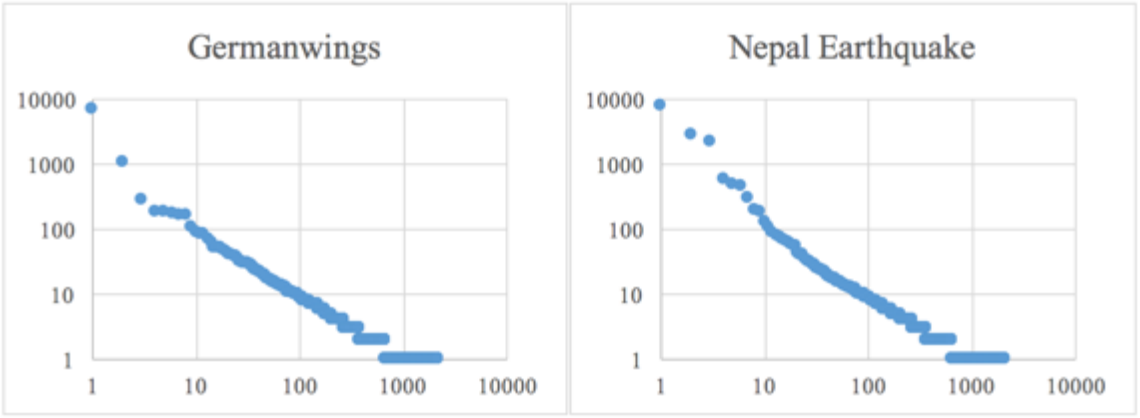

Figure B-4 Log log scale of hashtags counts distribution in disaster stories

Politics stories

The top 10 hashtags in each politics story are presented in Table B-5. In Charlie Hebdo story, the top two hashtags were used in the search for related tweets, and that may explains the large counts. Other hashtags included paris, france, parisshooting, jesuisahmed, which is a name of one of the victims, and afp which is a name of a French news agency. Similarly, in clock boy story, the first hashtag was used in the keyword search to find relevant tweets. The rest of the hashtags included words such as notabomb, media, blacklivesmatter and ahmedmohamed.

No.	Charlie Hebdo	Count	Clock Boy	Count
1	charliehebdo	63,366	istandwithahmed	13,459
2	jesuischarlie	46,144	thankyouforstandingwithme	153
3	paris	1,625	ahmedmohamed	122
4	chaliehebdo	1,556	target	83
5	parisshooting	1,377	notabomb	72
6	afp	1,064	islamophobia	57
7	france	1,041	blacklivesmatter	52
8	noussommescharlie	826	inners	47
9	marcherepublicaine	808	solidarity	40
10	jesuisahmed	670	media	37
Total		118,477		14,122

Table B-5 Top 10 hashtags in politics stories

In Charlie Hebdo 79% of all tweets contained hashtags, and 75% of all tweets contained one of the top ten hashtags as presented in Table B-6. Hashtags and the top 10 hashtags were more in retweets than originals, 83% and 80% in retweets versus 69% and 62% in originals. Tweets with one of the top 10 hashtags made 95% of tweets with any hashtag, 90% in originals, and 96% in retweets. In clock boy story 87% of all tweets contained hashtags, 86% of all tweets contained hashtags. Hashtags and the top 10 hashtags had higher ratios in retweets than originals, 89% and 88% in retweets versus 81% and 77% in originals. Tweets with one of the top 10 hashtags made 99% of tweets with any hashtag, 96% in originals, and 99% in retweets.

	Charlie Hebdo		**Clock Boy**	
	Count	%	Count	%
Tweets with Hashtags	105525	79%	13660	87%
Tweets with one or more of top 10 Hashtags	99794	75%	13476	86%
% Tweets of Top 10 to Tweets with Any Hashtag		95%		99%
Originals with Hashtags	25748	69%	2773	81%
Originals with One or more of Top 10 Hashtags	23155	62%	2668	77%
% Originals of top 10 to originals with hashtag		90%		96%
Retweets with Hashtags	79777	83%	10887	89%
Retweets with One or more of Top 10 Hashtags	76639	80%	10808	88%
% Retweets of top 10 to Retweets with hashtag		96%		99%

Table B-6 Top 10 hashtags analysis for politics stories

Figure B-5 presents hashtags counts distribution in politics stories, and Figure B-6 presents the log log scale of hashtags distribution in both politics stories.

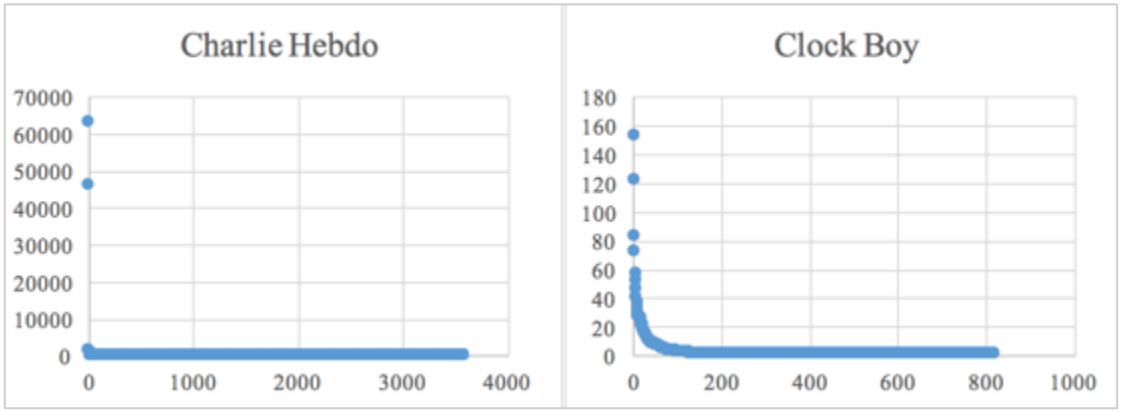

Figure B-5 Hashtags counts distribution in politics stories

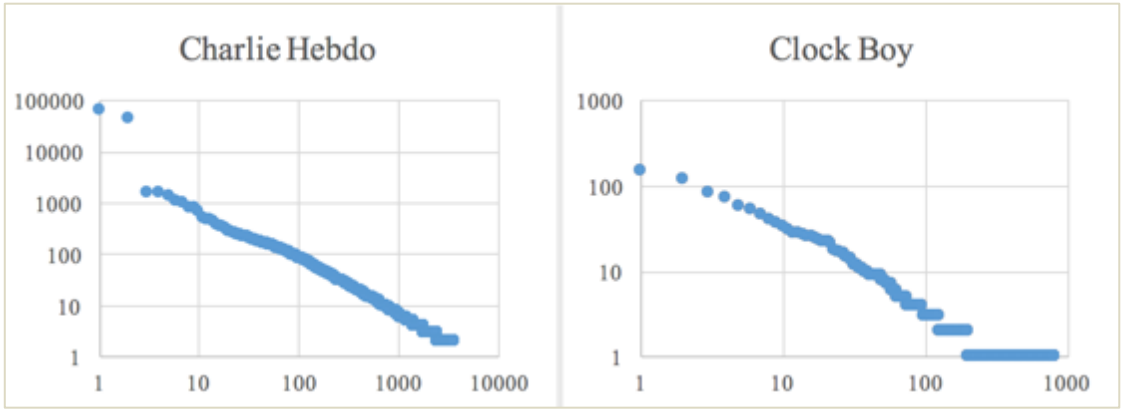

Figure B-6 Log log scale of hashtags counts distribution in politics stories

Appendix C – Top ten links analysis

Links analysis was applied to each news story dataset to extract links and to calculate the frequency of each link. Also, the distribution of the counts of links and the number of links in tweets, retweets, and originals were analyzed. Similar to hashtags analysis, more analysis was applied to the top ten links, and link parsing services were employed to examine the links sources. Top ten links analysis included analyzing the counts and ratios of tweet, retweets and originals including any of the top ten links. In the following subsections, links analysis is presented by news type and by story.

Finance stories

The top 10 links in each finance story are presented in Table C-1. In Chinese shares story links shared were from BBC, Economist, and Bloomberg, which is a news site that delivers business and market news. Other links include other Twitter posts for different accounts. The same websites appeared in Volkswagen story, in addition to CNN Money.

No.	Chinese Shares	Count	Volkswagen	Count
	URL/ Unshorten domain		URL/ Unshorten domain	
1	http://t.co/kr4KTivLAK	127	http://t.co/5ZD5WGl9PJ	1336
	twitter.com		bloom.bg	
2	http://t.co/aMIFThEYf7	84	http://t.co/ULiwm9Eojc	727
	bbc.in		bbc.in	
3	http://t.co/GbutZOSnR7	83	http://t.co/j6ZbCmGAiz	406
	bbc.in		cnnmon.ie	
4	http://t.co/LlqQAEsIyu	70	http://t.co/ezJvzgMnpa	375
	bbc.in		bloom.bg	
5	http://t.co/EyrtbV5oLT	64	http://t.co/Zj7bzLdBNh	296
	econ.st		twitter.com	
6	http://t.co/d4GAJfZvGG	50	http://t.co/QiOooe2Lvo	280
	bloom.bg		bloom.bg	
7	http://t.co/D3tDpHCof8	46	http://t.co/wl2PvNcERu	191
	twitter.com		cnnmon.ie	
8	http://t.co/mHFnQrWQYc	42	http://bit.ly/1lxXC	190
	xhne.ws		www.scriptlance.com	
9	http://t.co/gvLojgefcR	42	http://t.co/178X4LWOMy	187
	bbc.in		www.latimes.com	
10	http://t.co/wuWUEd62qZ	35	http://t.co/DOjcDXCDbt	180
	twitter.com		nyti.ms	
Total		643		4168

Table C-1 Top 10 links in finance stories

Generally, link sharing activity in finance news stories was high according to the results introduced in section 5.3, however, the top 10 links appeared in less than 10% of all tweets, which indicates that links shared in finance events include more variety of links than hashtags. In Chinese shares 92% of all tweets contained links, and 9% of all tweets contained one of the top ten links, as presented in Table C-2. Links and the top 10 links ratios were slightly higher in retweets than originals, 93% and 45% in retweets versus 92% and 12% in originals. Tweets with one of the top 10 links made 10% of tweets with any link, 1% in originals, and 48% in retweets. In Volkswagen story 91% of all tweets contained links, and 5% of all tweets contained one of the top ten links. 92% and 89% of originals and retweets respectively contained links, and 12% and 45% of

originals and retweets respectively contained one of the top ten links. Tweets with one of the top ten links made only 5% of tweets with any link, 0% in originals, and 14% in retweets, these results are shown in Table C-2.

Characteristic	Chinese Shares		Volkswagen	
	Count	%	Count	%
Tweets with Links	6273	92%	78166	91%
Tweets with a Link from top 10 Links	643	9%	4168	5%
% Tweets of Top 10 to Tweets with Any Link		10%		5%
Originals with Links	5067	92%	49167	92%
Originals with a Link from Top 10 Links	62	12%	205	8%
% Originals of top 10 to Originals with Any Link		1%		0%
Retweets with Links	1206	93%	28999	89%
Retweets with a Link from Top 10 Links	581	45%	3963	12%
% Retweets of top 10 to Retweets with Any Link		48%		14%

Table C-2 Top 10 links analysis for finance stories

Figure C-1 presents the links counts distribution in finance stories, and Figure C-2 presents the log log scale of links count distribution in finance stories.

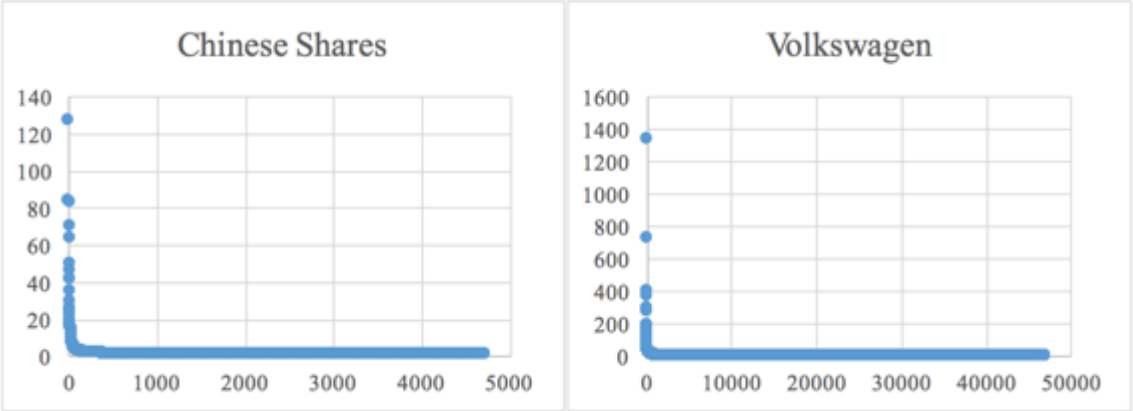

Figure C-1 Links counts distribution in finance stories

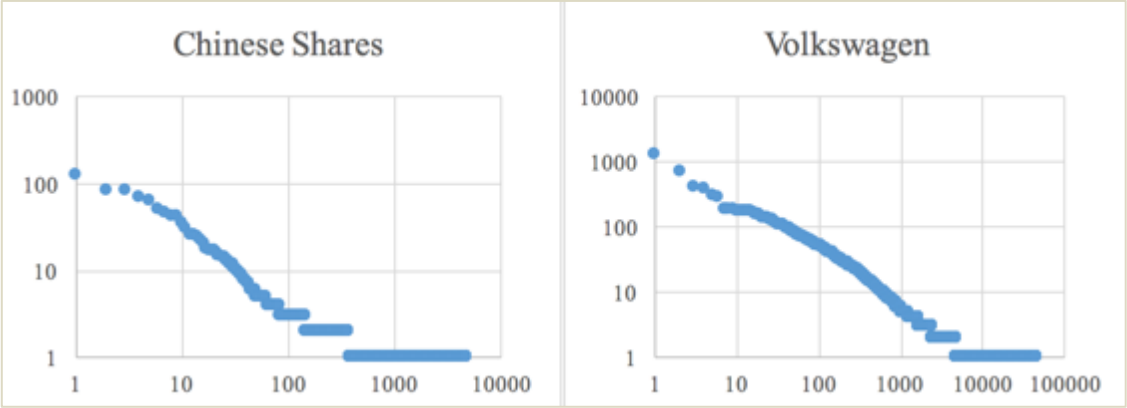

Figure C-2 Log log scale of links counts distribution in finance stories

Disaster stories

The top 10 links in each disaster story are presented in Table C-3. In Germanwings story links shared included some news sites, like BBC, RT and CNN. Other sites included flights news websites, such FlightRadar24 and AIRLIVE. Similarly, in Nepal earthquake story links shared included news website such as CNN, BBC and ABC.

No.	Germanwings URL/ Unshorten domain	Count	Nepal Earthquake URL/ Unshorten domain	Count
1	http://t.co/FHoX6q0GHt www.flightradar24.com	48	http://t.co/E8Fh03tnSi cnn.it	87
2	http://t.co/yecmaFK8JC twitter.com	36	http://t.co/ivv02aGAY3 www.supportunicef.org	59
3	http://t.co/yNlWbNJmYI bbc.in	31	http://t.co/hTcl4Kv0M5 bbc.in	56
4	http://t.co/zU6hn03xzU on.rt.com	29	http://t.co/8ix4HGFurO bbc.in	39
5	http://t.co/iw1QLzJQKW bbc.in	29	http://t.co/3BTo9l1QZ4 bbc.in	35
6	http://t.co/LtGqg9BUqF cort.as	27	http://t.co/ilMVRmI3AS tw.appstore.com	34
7	http://t.co/HAhU3MmiMf cnn.it	24	http://t.co/ZetrMJxtrC bbc.in	32
8	http://t.co/wFg8KTSve2 www.airlive.net	22	http://t.co/yXKpIsBixf abcn.ws	29
9	http://t.co/VxurYtdujU lpce.co	22	http://t.co/hCyjO7YyS7 cnn.it	29
10	http://t.co/ZH3wXGxyUn es.rt.com	18	http://t.co/R0XUNuoxj4äó□ goo.gl	26
Total		286		426

Table C-3 Top 10 links in disaster stories

As introduced in section 5.3, links were in 70% of all tweets in disaster stories combined. Similar to finance stories, the top ten links were not highly shared in tweets. In Germanwings story 75% of all tweets contained links, and 4% of all tweets contained one of the top ten links, as presented in Table C-4. 79% of originals contained links, but non from the top ten appeared in any original tweet. In retweets, 71% contained links, and 9% contained one of the top ten links. Tweets with one of the top ten links made 6% of tweets with any link, 0% in originals, and 13% in retweets. In Nepal earthquake story 65% of all tweets contained links, and 5% of all tweets contained one of the top ten links. 70% of originals contained links, and no link from the top 10 appeared in any original tweet. 64% of retweets contained links, and 6% contained one of the top 10 links. Tweets with one of

the top 10 links made only 7% of tweets with any link, 0% in originals, and 10% in retweets, these results are shown in Table C-4.

Characteristic	Germanwings		Nepal Earthquake	
	Count	%	Count	%
Tweets With Links	13,086	75%	13,607	65%
Tweets With a Link From top 10 Links	765	4%	958	5%
% Tweets of Top 10 to Tweets with Any Link		6%		7%
Originals With Links	7,188	79%	4,060	70%
Originals With a Link From Top 10 Links	0	0%	0	0%
% Originals of top 10 to Originals with Any Link		0%		0%
Retweets with Links	5,898	71%	9,547	64%
Retweets With a Link From Top 10 Links	765	9%	958	6%
% Retweets of top 10 to Retweets with Any Link		13%		10%

Table C-4 Top 10 links analysis for disaster stories

Figure C-3 presents links counts distribution in disaster stories, and Figure C-4 presents the log log scale of links counts distribution in disaster stories.

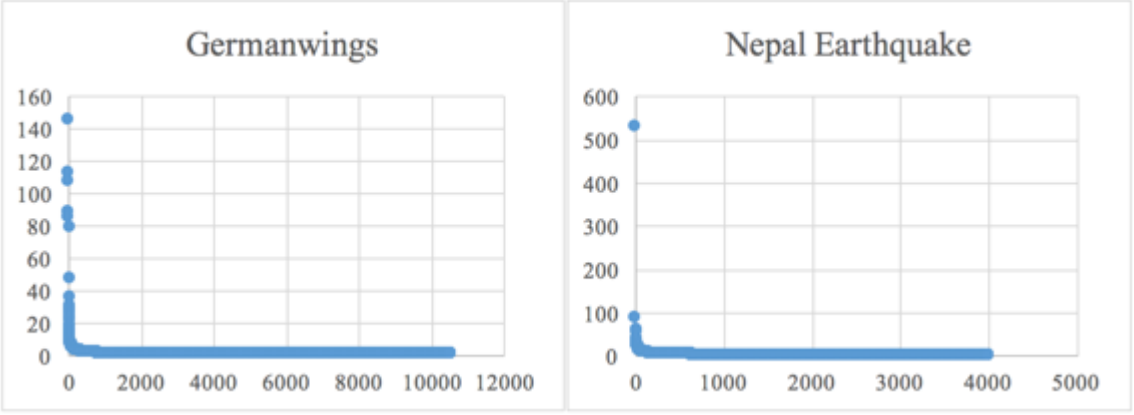

Figure C-3 Links counts distribution in disaster stories

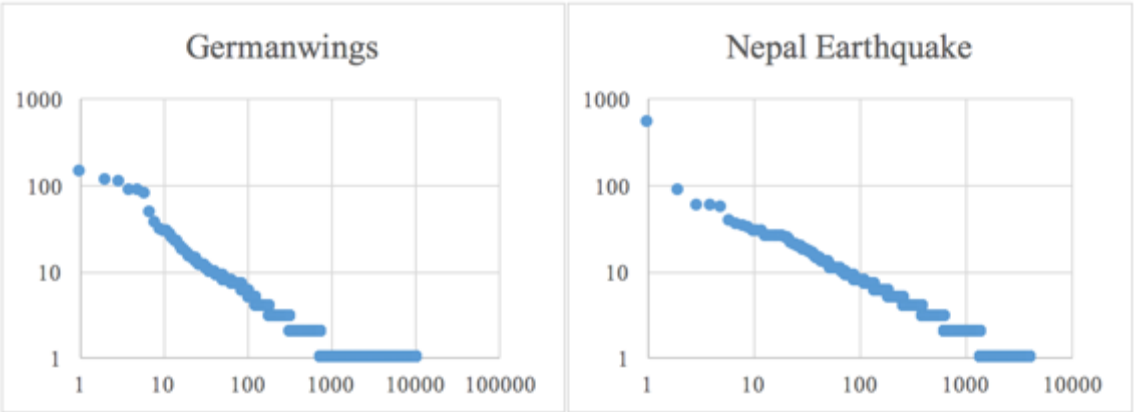

Figure C-4 Log log scale of links counts distribution in disaster stories

Politics stories

In politics most links pointed to other twitter accounts, one link only in clock boy story pointed to NBC News website, these result are presented in Table C-5.

No.	Charlie Hebdo URL/ Unshorten domain	Count	Clock Boy URL/ Unshorten domain	Count
1	http://t.co/_Ê_ ysoheil.com	395	https://t.co/tZGJFBWYrs amp.twimg.com	214
2	http://t.co/a2JOhqJZJM twitter.com	324	http://t.co/osgVM53g0f twitter.com	194
3	http://t.co/_ _ ysoheil.com	290	http://t.co/YCxOOeOz3Z twitter.com	160
4	http://t.co/_†_ ysoheil.com	260	http://t.co/NQCZ9rW2lw twitter.com	142
5	http://t.co/4T0uI4sF2h twitter.com	247	http://t.co/fBlmckoafU twitter.com	140
6	http://t.co/OP6h1YZUWs twitter.com	241	http://t.co/GyHlounDJh twitter.com	116
7	http://t.co/8KwTipn3Wp twitter.com	232	http://t.co/2sm0McBJH1 nbcnews.to	113
8	http://t.co/15O4YC2KWg twitter.com	208	http://t.co/J7yyyKixOC twitter.com	109
9	http://t.co/9sCF1EN5DH twitter.com	201	http://t.co/mIyMbhg2Mp twitter.com	94
10	http://t.co/Ksbl89WLsE twitter.com	185	http://t.co/KPLVCvVVDi twitter.com	88
Total		2,583		1,370

Table C-5 Top 10 links in politics stories

Table C-6 presents the top 10 links analysis in politics stories. The top ten links were not highly shared in politics stories tweets, In Charlie Hebdo story, 69% of all tweets contained links, and 3% of all tweets contained one of the top ten links. 59% of originals contained links, and no link from the top ten appeared in any original tweet. 72% of retweets contained links, and 4% contained one of the top ten links. Tweets with one of the top ten links made 4% of tweets with any link, 0% in originals, and 6% in retweets. In clock boy story 76% of all tweets contained links, and 9% of all tweets contained one of the top ten links. 50% of originals contained links, and no link from the top ten appeared in any original tweet. 83% of retweets contained links, and 11% contained one of the top 10 links. Tweets with one of the top ten links made only 12% of tweets with any link, 0% in originals, and 13% in retweets.

Characteristic	**Charlie Hebdo**		**Clock Boy**	
	Count	%	Count	%
Tweets With Links	91,665	69%	11,900	76%
Tweets With a Link From top 10 Links	3,958	3%	1,370	9%
% Tweets of Top 10 to Tweets with Any Link		4%		12%
Originals With Links	21,973	59%	1,711	50%
Originals With a Link From Top 10 Links	0	0%	0	0%
% Originals of top 10 to Originals with Any Link		0%		0%
Retweets with Links	69,692	72%	10,189	83%
Retweets With a Link From Top 10 Links	3,958	4%	1,370	11%
% Retweets of top 10 to Retweets with Any Link		6%		13%

Table C-6 Top 10 links analysis for politics stories

Figure C-5 presents links counts distribution in politics stories, and Figure C-6 presents log log scale of links counts distribution in politics stories.

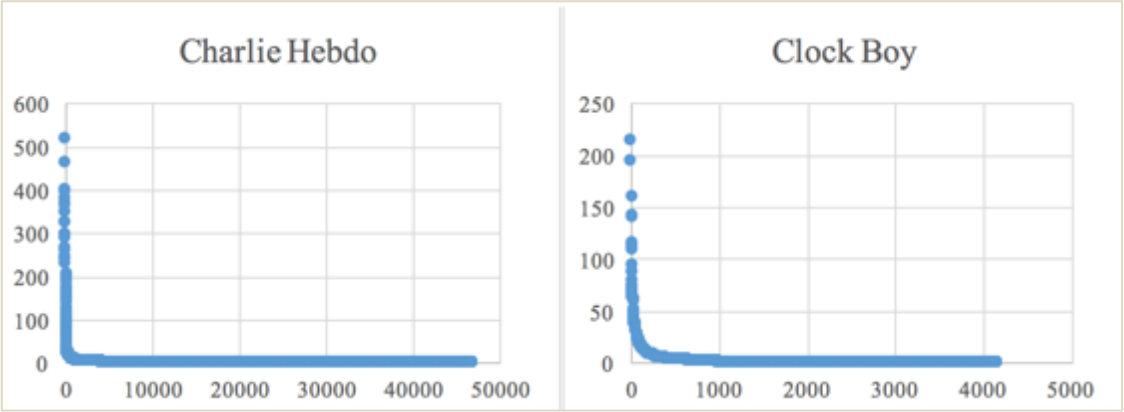

Figure C-5 Links counts distribution in politics stories

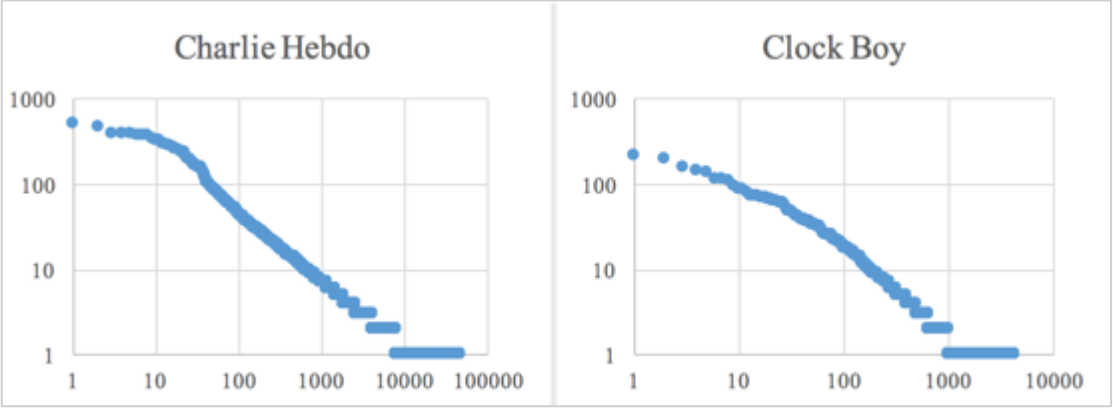

Figure C-6 Log log scale of links counts distribution in politics stories

Appendix D – Logistic Regression Results for Hashtags and Links Analyses

Hashtags

Table D-1 shows the Chi-square values and significance of the hashtag analysis regression model. Table D-2 presents the -2 log likelihood and the Nagelkerke R square values of the model. Table D-3 presents the classification table, which shows the overall ratio of the prediction of the model.

		Chi-square	df	Sig.
Step 1	Step	29891.399	8	.000
	Block	29891.399	8	.000
	Model	29891.399	8	.000

Table D-1 Omnibus Tests of Model Coefficients

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	152264.705	.202	.271

Table D-2 Model Summary

	Observed		Predicted		
			Hashtags		Percentage Correct
			NoHashtag	Hashtag	
Step 1	Hashtags	NoHashtag	40465	19660	67.3
		Hashtag	16775	55275	76.7
	Overall Percentage				72.4

Table D-3 Classification Table

Links

Table D-4 presents the Chi-square values and significance of the link analysis regression model. Table D-5 presents the -2 log likelihood and the Nagelkerke R square values of the model. Table D-6 presents the classification table, which shows the overall ratio of the prediction of the model.

		Chi-square	df	Sig.
Step 1	Step	8794.306	8	.000
	Block	8794.306	8	.000
	Model	8794.306	8	.000

Table D-4 Omnibus Tests of Model Coefficients

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	129063.667[a]	.064	.099

Table D-5 Model Summary

	Observed		Predicted		
			Link		Percentage Correct
			NoLink	Link	
Step 1	Link	NoLink	0	28519	.0
		Link	0	103656	100.0
	Overall Percentage				78.4

Table D-6 Classification Table