
Theses and Dissertations

Fall 2012

Discovering entities' behavior through mining Twitter

Hung Viet Tran
University of Iowa

Copyright 2012 Hung Viet Tran

This dissertation is available at Iowa Research Online: <https://ir.uiowa.edu/etd/3545>

Recommended Citation

Tran, Hung Viet. "Discovering entities' behavior through mining Twitter." PhD (Doctor of Philosophy) thesis, University of Iowa, 2012.
<https://ir.uiowa.edu/etd/3545>.

Follow this and additional works at: <https://ir.uiowa.edu/etd>



Part of the [Computer Sciences Commons](#)

**DISCOVERING ENTITIES' BEHAVIOR
THROUGH
MINING TWITTER**

by

Hung Viet Tran

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

December 2012

Thesis Supervisors: Professor Padmini Srinivasan
Assistant Professor Gautam Pant

ABSTRACT

The unprecedented amount of user generated content from emerging social media platforms like Facebook and Twitter make them invaluable sources of information for research. Twitter in particular has about 500 million registered accounts globally who are generating approximately 340 million messages daily containing personal updates, general life observations, opinions, moods, etc. Twitter's vast amount of data, which is generally available, offers an ideal source for mining entities' behaviors. This thesis explores two research streams involving mining Twitter data. In the first work, we seek to understand the Twitter-based stakeholder communication strategies of firms. We analyze tweets posted by firms to build a system that can automatically predict target stakeholder groups of a given tweet. We also examine and incorporate firm characteristics into the system for performance improvement. The result will potentially provide valuable business intelligence to market analysts who would like to discover social media strategies and behaviors of firms. In the second work, we investigate how readers from different parts of the world react to news headlines through their Twitter messages. We design a framework for data collection, statistical analysis, sentiment analysis, and language model comparison to understand the interests and reactions of Twitter users towards news headlines. The results from this work can possibly help news organizations have better understanding of their audience for better services. Though the two research directions may seem distinct, there are points of connection. In both cases, we are interested in the impact

of companies (firms and news organizations). Moreover the methods used are similar. Our results illustrate that just by gathering Twitter data stream and developing a framework to examine them, we are able to discover many interesting insights about news readers and firms.

Abstract Approved: _____
Thesis Supervisor

Title and Department

Date

Thesis Supervisor

Title and Department

Date

**DISCOVERING ENTITIES' BEHAVIOR
THROUGH
MINING TWITTER**

by

Hung Viet Tran

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

December 2012

Thesis Supervisors: Professor Padmini Srinivasan
Assistant Professor Gautam Pant

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Hung Viet Tran

has been approved by the Examining Committee for the
thesis requirement for the Doctor of Philosophy degree
in Computer Science at the December 2012 graduation.

Thesis Committee: _____
Padmini Srinivasan, Thesis Supervisor

Gautam Pant, Thesis Supervisor

Alberto Segre

Juan Pablo Hourcade

James Cremer

To my Family

ACKNOWLEDGEMENTS

Doing PhD is an exciting but challenging journey. One would not be able to reach the goals without helps and supports along the way from others. There are many people I would like to thank for helping me make this happen. First and foremost, I would like to thank my advisors, Professor Padmini Srinivasan (Department of Computer Science) and Professor Gautam Pant (Department of Management Science) for their tremendous advices and supports for my research works. Especially during the time I had to change my research focus from computer security to information retrieval and web mining, Professor Padmini gave me encouragement to move on.

I also would like to thank my thesis committee members, Professor Alberto Sergre, Professor James Cremer and, Professor Juan Pablo Hourcade for their valuable feedback and comments on my research.

I am thankful to have an opportunity to work with great colleagues from the Text Retrieval and Text Mining group, department of Computer Science. I would like to thank Chao Yang, Yelena Mejova, Sanmitra Bhattacharya, Viet Ha, and Christopher Harris for their helps and support, especially Chao and Sanmitra who helped me a lot with the *Discovering Public Reactions to News Headlines* project.

Outside of academic life, I would like to thank my great Iowan friends, Joe Tye and Thomas Hornbeck who helped me get familiar with U.S. culture and lifestyle which make it much easier for me to live and work in the U.S.

Last but not least, I would like to have special thanks to my family for their

love, encouragement, and supports. I would not be able to finish my journey without them.

ABSTRACT

The unprecedented amount of user generated content from emerging social media platforms like Facebook and Twitter make them invaluable sources of information for research. Twitter in particular has about 500 million registered accounts globally who are generating approximately 340 million messages daily containing personal updates, general life observations, opinions, moods, etc. Twitter's vast amount of data, which is generally available, offers an ideal source for mining entities' behaviors. This thesis explores two research streams involving mining Twitter data. In the first work, we seek to understand the Twitter-based stakeholder communication strategies of firms. We analyze tweets posted by firms to build a system that can automatically predict target stakeholder groups of a given tweet. We also examine and incorporate firm characteristics into the system for performance improvement. The result will potentially provide valuable business intelligence to market analysts who would like to discover social media strategies and behaviors of firms. In the second work, we investigate how readers from different parts of the world react to news headlines through their Twitter messages. We design a framework for data collection, statistical analysis, sentiment analysis, and language model comparison to understand the interests and reactions of Twitter users towards news headlines. The results from this work can possibly help news organizations have better understanding of their audience for better services. Though the two research directions may seem distinct, there are points of connection. In both cases, we are interested in the impact

of companies (firms and news organizations). Moreover the methods used are similar. Our results illustrate that just by gathering Twitter data stream and developing a framework to examine them, we are able to discover many interesting insights about news readers and firms.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	xi
CHAPTER	
1 INTRODUCTION	1
2 LITERATURE REVIEW	5
2.1 Twitter Use in Organizations	5
2.2 Surveys of News Readers	7
2.3 Twitter as a Communications Medium	9
2.4 Sentiment Analysis Using Twitter Data Stream	12
2.5 Related Works from TREC	14
3 DISCOVERING TARGET STAKEHOLDERS OF FIRM'S TWEETS	18
3.1 Background and Motivation	18
3.2 Research Questions	20
3.3 Methodology	20
3.4 Data Collection and Annotation	21
3.4.1 Data Collection	21
3.4.2 Data Annotation	22
3.4.2.1 Preliminary Content Analysis	22
3.4.2.2 Annotating Data	26
3.5 Experiments	30
3.5.1 Feature Description	30
3.5.2 Evaluation	33
3.5.3 Classifiers Using Tweet-based Features	35
3.5.4 Classifiers Using Firm-based Features	36
3.5.5 Classifiers Using both Tweet-based and Firm-based Features	37
3.6 Analysis	39
3.6.1 Relative Contribution of Features	39
3.6.2 Application on All Tweets	41
3.7 Summary	44
4 DISCOVERING PUBLIC REACTIONS TO NEWS HEADLINES	49

4.1	Background and Motivation	49
4.2	Research Questions	51
4.3	Methodology	51
4.4	Data Collection	53
	4.4.1 Headlines Collection	55
	4.4.2 Tweet Collection	57
	4.4.2.1 Query Generation	59
	4.4.2.2 Tweet Search and Retrieval	61
4.5	Analysis	66
	4.5.1 Within-Country Analysis	66
	4.5.1.1 Distribution of Tweets Over Headlines	66
	4.5.1.2 Statistical Analysis	66
	4.5.2 Cross-Country Analysis	78
	4.5.2.1 Cross-Country Tweet	78
	4.5.2.2 Sentiment Analysis	81
	4.5.2.3 Language Model Comparison	84
4.6	Summary	86
5	CONCLUSIONS	88
	APPENDIX	91
	A SUMMARY OF OTHER COMPLETED WORKS	91
	A.1 Spam Detection in Online Classified Advertisements	91
	A.2 Belief Surveillance with Twitter	93
	A.3 Discovering Health Beliefs in Twitter	95
	B TWEETS ANNOTATING GUIDELINES	98
	B.1 Introduction	98
	B.2 Annotating Tweets Online	98
	REFERENCES	102

LIST OF TABLES

Table		
3.1	Initial Companies	26
3.2	Percentage of Overlapping Labels among Judges	30
3.3	F-Measure of Base Classifiers	35
3.4	F-Measure of Heuristic Classifiers	36
3.5	F-Measure of Firm-based Classifiers	37
3.6	Tweet-based and Company-based Classifiers	39
3.7	Statistically Significant Firm-based Features	42
4.1	Geocode Information for the Countries in our Dataset	54
4.2	Example Headlines from Different Countries Collected on October 2, 2012	55
4.3	Differences in Top Ranked and Bottom Categories Ranked by Frequency	57
4.4	News Headline Corpus	60
4.5	Search Results with Different Values of k	60
4.6	Query Examples	61
4.7	Number of Headlines with Relevant Tweets in Cross-country Tweet Retrieval	64
4.8	Headlines with Largest Number of Tweets	65
4.9	Examples of Tweets with Additional Content	68
4.10	Most Frequent Words in Additional Content	69
4.11	Percentage of Tweets in each News Category	72

4.12 p-values for Comparison among News Categories (Non Significant Results are in Bold)	72
4.13 p-values for Comparison within each Countries (Significant Results are in Bold)	76
4.14 p-values for Comparison of Different Countries in the Same Category (Significant Results are in Bold)	80
4.15 Cross-Country Tweet Retrieval	81
4.16 Cross-country Tweet Examples	82
4.17 Selected Headlines for Sentiment Analysis	82
4.18 Sample Tweets with Sentiment Labels	83
4.19 Sentiment Annotation Results	83
4.20 Agreement Ratios	83
4.21 KL Divergence Score Averaged by Headline	85

LIST OF FIGURES

Figure	
3.1	Framework for Data Collection and Classifiers Training 21
3.2	Distribution of Tweets per Company 22
3.3	Target Stakeholders in Initial Companies 27
3.4	Online Labeling System 29
3.5	Label Ratio of each Stakeholder Group 31
3.6	Prediction and Actual Data 34
3.7	Splitting Data 39
3.8	Combining Tweet-based and Firm-based Features 40
3.9	Classifier Performance for Different Stakeholder Groups 41
3.10	Stakeholder Communication Strategy of Initial Companies 45
3.11	Stakeholder Communication Strategy of Different Industries 46
3.12	Stakeholder Communication Strategy of Initial Companies by Volume 47
3.13	Stakeholder Communication Strategy of Different Industries by Volume 48
4.1	Framework for Analyzing Twitter Response to News Headlines 52
4.2	Headline Distribution 56
4.3	Distributions of Headlines Collected from 09/30/12 to 10/14/12 58
4.4	Headlines and Tweets from 10/01/12 to 10/21/12 62
4.5	Headlines and Tweets from 10/01/12 to 10/14/12 63
4.6	Tweet Proportion: Distribution by Country of Origin 64

4.7	Within Country Tweet Retrieval: Distribution over Date and Country . .	65
4.8	Percentage of Headlines with at Least 5 Related Tweets	67
4.9	Tweet Content	70
4.10	Tweets with Additional Content	70
4.11	Percentage of Tweets in each News Category	72
4.12	Tweets in Different Countries	75
4.13	Tweets from Different Countries in the Same Category	79
B.1	Target Audience	99
B.2	Login Screen	100
B.3	Home Screen	100
B.4	Company Screen	100
B.5	Labeling Screen	101

CHAPTER 1 INTRODUCTION

Social media platforms like Facebook¹, Twitter², and Blogspot³ allow users to easily connect and share information. The unprecedented data generated by millions of users from all around the world make social media ideal places for mining trends and patterns of interests.

Among those social media platforms, Twitter stands out as a phenomenon in the research community. Twitter as a microblog site enables its users to post 140-character messages, or tweets, to the system and those tweets become available instantly to everyone. In Twitter the social relationship among users is *following* where one user subscribes to another user's tweet stream. The following user becomes a *follower* and will receive all the tweets from the followed user. Twitter also uses some special syntax like user mention (*@*), direct message (*D or DM*), where one user sends a tweet directly to another user, hashtag (*#*) to indicate a categorized topic, and retweet (*RT*) to relay tweets. As of July, 2012, Twitter has approximately 140 million active accounts over 500 million registered accounts who generate about 340 million tweets every single day⁴. Twitter is becoming an integral part of modern life. According to recent research by Pew Research Center [43, 44], about 15% of

¹<http://www.facebook.com>

²<http://www.twitter.com>

³<http://www.blogspot.com>

⁴<https://business.twitter.com/en/basics/what-is-twitter/>

online adults in the U.S. use Twitter to tweet various types of messages ranging from personal life updates to general life observations. In some special cases like the natural disasters or events that news reporters cannot approach, Twitter is the only medium for information diffusion [22, 55]. Given such advantages, Twitter has become a hot topic for research. Researchers are actively mining Twitter data stream for sentiment analysis [2, 3, 15], real time event detection [41, 41, 24] and news recommendation [35, 36, 1], tracking emergency situations [34, 7] and political campaigns [39].

This thesis explores two research streams (analyzing firms' behaviors and news readers' behaviors) involving social media, in particular involving Twitter. Though the directions are distinct there are points of connection primarily in the methods used. Both focus on examining entities' behaviors (firms or news readers) through their Twitter messages. Additionally understanding news readers' reactions to news headlines provides firms with valuable intelligence, while understanding firms' behaviors gives market analysts and other external observers insights about firms.

- *Discovering Target Stakeholders of Firm's Tweets:* We seek to uncover the twitter-based stakeholder communication strategy of firms. The proposed methodology involves the use of crowd sourcing and machine learning. In particular, we use Twitter messages posted by Fortune 100 companies to develop a system that can automatically predict target audience of a given tweet. We also try to understand if other publicly available information about firms' characteristics like industry, revenues, size, etc. in addition to textual information from Twitter message, when used in combination, helps increase the accuracy of our

prediction. Once we can predict the target audience of each Twitter messages using our system, we are be able to see how companies' focus on their stakeholders changes from time to time and how firms' characteristics like industry contribute to that change. The result will potentially provide valuable business intelligence to market analysts who would like to discover firms' social media strategies and behaviors.

- *Discovering Public Reactions to News Headlines:* We investigate how readers from different parts of the world react to news headlines through their Twitter messages. The first research question we ask is *Do the news readers actually discuss what they read in addition to sharing links?* Then we ask *How do the discussions vary across broad categories of news? How do the discussions vary among different categories of news within a geographical location? What is the difference among the discussions on the same news category across reader groups at different geographical locations?* In order to find answers to above research questions, we also ask several methodological questions: *How can we find the tweets mentioning/discussing a specific news article?* and *How do we analyze a user's reaction to news from her tweet messages?* We use headlines from Google News⁵ and relevant tweets in the Twitter data stream for our experiments. We explore the public interests and reactions to certain headlines or categories of news. We also want to see how these interests and reaction vary from one country to another. The results from this research can help news organizations

⁵<http://news.google.com>

have a better understanding of their audience leading to better strategies for providing news services. These can also help automatically aggregate relevant tweets as readers' comments for certain news.

Besides possible answers to the questions we raise, this thesis makes the following general contributions:

- We present our mining processes from Twitter data stream collection to crowd sourced data annotation to classification models training strategies to result analysis for understanding firms and news readers.
- We show some potential real world applications using our research results.

The rest of this thesis is organized as follows. Chapter 2 summarizes related works in mining Twitter data stream in literature. Chapter 3 presents details of our work on *Discovering Target Stakeholders of Firm's Tweets*. Chapter 4 describes our work on *Discovering Public Reactions to News Headlines*. Chapter 5 concludes the thesis and outlines some directions for future works.

CHAPTER 2 LITERATURE REVIEW

2.1 Twitter Use in Organizations

Researchers in Public Relations have conducted several works regarding how organizations leverage Twitter for stakeholder communication. Rybalko et al. [40] investigate tweets in a sample of 93 companies with an active Twitter account in Fortune 500 companies to understand the dialogic features of Twitter and the target public groups of those companies. They found the dialogic features when analyzing tweets, for example responses to users' posts (60.2%), posting newsworthy information about the company (58.1%), and posing questions (30.1%). They also found that a large portion of tweets are addressed to general audience, which is not explicitly identified (74.5%), while only a small number of tweets are targeted to customers, which are specific users with *@username* in the tweets (0.9%). Lovejoy et al. [26] examined 4655 tweets from 73 nonprofit organizations in the *Nonprofit Times 100*¹ list to see if those organizations fully make use of available communication tools in Twitter, e.g. following, hashtag, retweet, etc., to engage their stakeholders. The finding shows that most of the organizations only use Twitter as the one way communication tools which is opposite to the corporate world with 61% companies classified as dialogic.

In Computer Science, Twitter has become a hot research topic for years with a large number of publications. While many research works focus on topic and commu-

¹<http://www.thenonproffitimes.com/>

nity detection, sentiment analysis, and network structures of following and retweet behaviors, there are not many ones working on exploring Twitter users, especially business users. Perhaps, one of the earliest works on Twitter users is conducted by Java et al. [17]. They analyzed 1,348,543 posts from 76,177 users collected during two months in 2007 to understand *how* and *why* people tweet. The result from link analysis categorizes user intentions in four groups: *daily chatter*, *conversations*, *sharing information*, and *reporting news*. Also, Twitter users are categorized in three groups: *information source*, *friends*, and *information seeker*. Regarding business use of Twitter, Popescu et al. [38] conducted the initial work on the topic to understand how companies interact with customers in Twitter. They analyzed 1000 tweets from 5,245 business accounts to develop the business tweet taxonomy including five classes: *content + recommendations*, *engagement:specific*, *brand_awareness*, *announcements*, and *engagement:all*. They focused on classifying tweets in the *announcements* category into *deals* and *events*. The results are very promising: *deal* class with 95% precision and 93% recall, *events* class with 96% precision and 97% recall. Other works close to ours regarding tweets classification include [46] where Sriram et al. analyzed 5407 tweets from 684 users to extract 8 features including author type (personal and corporate) and 7 features from tweets' content. They use this feature set to train models to classify tweets in five different categories *news*, *opinions*, *deals*, *events*, and *private messages*. The experiment results show that the feature set provide significantly better accuracy (32.1%) than bag of word features.

Above studies explored how Twitter is used in business. However, none of

these studies were designed to understand stakeholder communication strategy, nor did they have a firm-level focus. Thus they are ad hoc and provide limited firm behavior information. Our study is the first systematic effort to look at social media efforts of firms through the lens of stakeholder theory.

2.2 Surveys of News Readers

News organizations and communications and behavioral scientists have been conducting research on audience's reaction on news media. Klein et al. [20] administered a survey among three different populations: middle/high school students (262), college students (332), and seniors (271) to investigate their behavior and reactions to the local television news. The survey includes 31 items in form of a questionnaire which each group will answer during the 6-year research period. The results show the comparison among groups in term of: Frequency and Motives for Viewing, News Effects, News Balance, News Reality, and News Contents. It's interesting to see that the majority agreed that the local news accurately portrays the world with violence, tragedy, and disaster, which make people feel unsafe. Also, the majority said that they want to see the good news which makes them feel happy but did not find much good one as there is no balance between good news and bad news on local TV. Among the groups, women are more adversely affected by news than men. College student men enjoy the violent content from the news. Senior tend to share what they learn from the news to others. In [53], Williams et al. conducted a survey with ten British newspapers to understand the newspaper reporting of crime and fear of crime. They

found that readers of newspapers reporting most crime have the highest level of fear of crime but the causal link is not clear. They also suggested to measure the impact on the readers if the newspapers report crime in more dispassionate, objective, and responsible fashion as different ways of reporting same crime may have different influence in readers' fear of crime. To help understanding the audience declines in local TV news, Southwell et al. [45] conducted a national survey of 2728 viewers across the U.S. for their attitudes toward local TV news. They found that viewers consider watching TV news for information rather than for entertainment which contradicts the belief of local TV news professionals. The results maybe helpful for local TV stations to have better broadcasting strategies to bolster viewership. In their research, Keinan et al. [18] surveyed 534 people in Israel about their attitudes and reactions to media coverage regarding terrorist attacks in Israel. They found that the audience would like to receive the detailed coverage of such events but if the coverage includes horrifying details the demand declines. Also when the audience was extensively exposed to that kind of coverage, they develop the symptoms similar to Post Traumatic Stress Disorder. In addition, they found that the attitudes and reactions are different among the demographic groups.

These researches provide some insights of the public opinion and reactions to news. However we can see that these researches required a lot of manual work which takes time to conduct the surveys even with a small sample of population. In this research, we are exploring a systematic approach to gather data and produce such results with much larger population and less time. For example, we would be able

to automatically collect tweets from millions of Twitter users mentioning about news across a wide range of topic categories. Also, the result from [18] on the differences in attitudes and reactions to news coverage among demographic groups motivated us to investigate the differences in reactions to news headlines among audience groups at different geographical locations.

2.3 Twitter as a Communications Medium

In recent years, Twitter data stream carrying hundreds of millions of messages with users' opinions a day becomes a hot topic among research communities. Researchers use Twitter data to automate the surveying processes. For example, Wakamiya et al. [51] monitor tweets from different locations regarding TV programs to estimate the TV viewing rates. Also, with the ability of spreading messages throughout the world in real time, Twitter is an excellent tool to diffuse news. Kwak et al. [22] analyzed Twitter in its early days to understand its topological characteristics and its power of information sharing. They crawled the whole site from June 6, 2009 to June 31, 2009 as well as collected the user profiles mentioned in the trending topics until September 24, 2009 resulting 106 million tweets, 4262 trending topics, 1.47 billion social relations, and 41.7 million profiles. Multiple analyses were conducted on collected data including relationship among Twitter users, number of followers vs. tweets, reciprocity, degree of separation, content, etc. By showing that the Twitter network has very low reciprocity, they claim that Twitter is more like an information spreading medium than the social network assuming that the following

relationship as the subscribing to Twitter content. They also compare the content topics on Twitter with other media like Google Trend and CNN Headlines to confirm their claim. One interesting finding in the paper is that over 85% of the trending topics are headlines or persistent news in nature. These works motivate us to understand more about tweets' content regarding news. We are interested in investigating if users actually discuss about news when spread it through Twitter network.

There is a large body of other research on the relationship between Twitter and news. Tsagkias et al. [50] discovered implicit links, where the hyper links do not exist, between news articles and social media. In other words, given a news article, they will find the social media contents that reference the given news article. For each news article, they generate multiple queries using document structure, explicitly linked social media, and term selection strategies. The queries then are used to retrieve the content from social media, which return multiple ranked lists of relevant content. The ranked lists are then merged in a single result. The retrieval step uses language model to estimate the likelihood of a social media content to generate the given news article. The experiments was conducted using Blogs08 and New York Times headlines collection by TREC, and different social media including Digg, Delicious, Twitter, NYT Community and Wikipedia. The results show that the query models built from the entirely news article content perform best. Motivated by this result, we make use of news headlines' structure to build query models for our research. However there are some differences in our approach. Since the headlines we are aiming to use from Google News contain a very limited amount of information, e.g. title, source,

and description, it would be not enough to build a language model for each headline. In addition, we are targeting only Twitter for relevant posts directly mentioning the given news headline. As Twitter messages are limited by 140 characters, the long queries may miss some important tweets. Thus we use a lexicon build from previous collected news headlines from Google News and use TF/IDF to extract the n most important words from the headlines to use as queries. We conduct experiment with different values of n and choose one that gives us most relevant results.

Zhao et al. [54] investigate the difference between social media and traditional news media by empirically comparing the content of Twitter and New York Times using topic modeling. The experiments were conducted on Edinburgh Twitter Corpus and New York Times during the period from November 11, 2009 and February 1, 2010 to compare these two in three dimensions: topic, topic category, and topic type. Topics was discovered in NY Times using LDA, and in Twitter using proposed Twitter-LDA which is used to generate tweets using the distributions of: available topics in Twitter, background words, and user topics. Topic categories in NY Times are estimated based on the category labels of the articles; in Twitter the topic categories are estimated based on the similarity of the Twitter topics and NY Times topics using JS divergence. Topic types including event-oriented, entity-oriented, and long-standing are manually assigned. The experimental results show that Twitter and traditional news media cover a same range of topic categories with different distributions of topic categories and topic types. Twitter focuses more on personal life and pop culture as well as celebrity and brands that traditional media do not cover.

Twitter users are more about retweeting the world event topics to spread the news instead of tweeting about them. The results from this work motivate us to investigate behaviors of Twitter users further. We know that generally there are certain topic categories in news media that do not get Twitter users' interests. However, we would like to see the differences in the level of interests among the topic categories and also for a same topic category we would like to see the differences in the level of interests among groups of Twitter users at different geographical locations.

2.4 Sentiment Analysis Using Twitter Data Stream

Since Twitter data stream contains a variety of opinions from users, the first Twitter mining task potentially provide valuable information is sentiment analysis. In fact, many companies have been studying Twitter data to discover the public opinions towards their products and services through general sentiment [2]. In literature, sentiment analysis using Twitter data stream becomes an interesting topic for researchers as it poses some challenges, e.g., handling a large collection of short messages with a lot of abbreviations. Go et al. [14] are one of the first groups doing sentiment analysis in Twitter. They used tweets with emoticons as labeled data to create different feature sets (Unigram, Bigram, Unigram + Bigram, Unigram + POS) for training classifiers (Naive Bayes, Maximum Entropy, and SVM) that can classify a certain tweet as *positive* or *negative*. Experiment results showed that unigram feature set perform equally or better others in all classifier training algorithms. Barbosa et al. [3] collected

tweets from three different websites detecting tweets' sentiment: *Twendz*², *Twitter Sentiment*, and *TweetFeel*³ as training data. Then they built two-step sentiment classification system in which tweets will go through the first classifier for being classified as *subjective* or *objective*. Ones with *subjective* label will go through the second classifier to be classified as *positive* or *negative*. The features are created using meta information of words in tweets, e.g. POS, and Twitter's syntax. Experiment results showed that SVM classifiers performed best with Unigram and proposed feature sets however the number of features in the proposed feature set is much smaller than in unigram. Agarwal et al. [2] developed a system to classify tweets into three classes *positive*, *negative*, and *neutral*. They proposed the tree kernel based model to represent tweets to combine many categories of features. They used unigram model as the baseline (average accuracy of 71.35%) and conducted experiments with another four different feature sets. Tree kernel based model outperformed the baseline however classifiers built from unigram and sentiment features, e.g. POS and emoticons, gave the best results (average accuracy of 75.39%). Yelena et al. compared the sentiment on the same set of topic among blogs, reviews, and Twitter [29] to answer some questions, e.g. *Do Twitter users react differently or the same to a specific news as blogger?* They found that for the same set of topics, different source has different sentiment. They also found that classifiers build from Twitter training data is generalizable which means that they can use to classify items in other sources as well as the do classifiers

²<http://twendz.waggeneredstrom.com>

³<http://www.tweetfeel.com>

trained from original training data. Their another work on comparing sentiment on the same topics including politicians, issues, and events from Twitter messages and YouTube⁴ comments [30]. The experiment results were consistent with their previous finding that the amount of sentiment expressed is different across media. They concluded that the choice of social media to analyze determines the results.

In our work, we use sentiment classifiers to classify each headline and its related tweets into either *positive* or *negative*. For each headline we calculate the proportion of *positive* and *negative* tweets which could indicate the reactions of Twitter users to that headline. Then we compare the reactions to each type of headline (*positive* or *negative*) among countries in all categories of news headlines.

2.5 Related Works from TREC

In recent years, Twitter has become an interested research topic in Text Retrieval Conference (TREC)⁵. The first TREC Microblog track was introduced in 2011⁶ where participants are provided with a corpus of approximately 16 million tweets collected over two weeks from Jan 24, 2011 to Feb 8, 2011. The participants participate in the Realtime Adhoc Tasks where they build realtime search systems replying a query in form of a topic, e.g. *2022 FIFA soccer*, with a list of relevant tweets ordered from newest to oldest starting from the the time the query is sent. Among the submitted works to TREC 2011 Microblog track we found that one from Tao et al.

⁴<http://www.youtube.com>

⁵<http://trec.nist.gov/>

⁶<https://sites.google.com/site/microblogtrack>

[48] is related to our work. Original queries submitted to their system for searching tweets are treated as topics containing a set of concepts (keywords). The concepts then are annotated the names using named-entity recognition (NER) service DBPedia Spotlight⁷. The named entities are used to search other related corpora for related entities which are added to the original topic. One of the corpora is the news articles written in the same time frame with the Twitter corpus containing news articles' titles and abstracts. Titles and abstracts are extracted for another round of NER if they contain named entities identified in the original topics. The new identified entities will be added to the original topic creating the topic profile. Concepts in the topic profile are assigned with a certain weight depending on the source of entities to create the final profile. The original queries with related entities in their profiles are used to search Twitter and news articles corpora again for related tweets or news articles which are then used for later query expansion. New identities extracted from search results are added to the final profile with certain weight. The final profile of a topic is translated into the Indri query language syntax to search the indexed tweets in the original corpus. Although the results from automatic methods for query expansion are better than the baseline where the original queries are used to search the index of original tweets, they are outperformed by the result of manual run. Both this work and our work have to deal with the problem of language gap between the query and Twitter messages which may cause less tweets to be found. The methodologies used in this work utilizing NER to build profile for the original query and for query expansion

⁷<http://spotlight.dbpedia.org>

sion suggest us an alternative way to generate queries from Google News to search Twitter.

Other TREC works close to ours are those who participated in TREC 2009 Blog Track [6], especially ones working on *Top Stories Identification Task*. In this task, the participants use *Blog08* collection including data of over 1 million blogs collected from Jan 14, 2008 to Feb 10, 2009 and a large sample of New York Times news headlines covering all articles published by NYT during the same timespan of *Blog08* to i) identify the top news stories for a given day, and ii) provide blog posts related to a given news story. McCreadie et al. [28] use a weighting model named DPH which calculates the relevance score of a document for a query based on term frequency within the document, term frequency within the collection, query term frequency within the query and query term frequency within all queries to search for related blog posts for each headline. The search is then repeated for several time at different days creating day oriented ranking results. Then they make the final result by merging the day oriented results and selecting the top blog posts. The results outperform TREC median for both α -NDCG@10 and IA-P@10. To select supporting blog posts for a given headline, Lee et al. [25] search the collection for relevant posts using KL-divergence language model for relevance score between a headline and a blog post. From the results they select top 10 posts using either Feed-Based Selection which chooses posts from as many blog feeds as possible or Cluster-Based Selection which groups the posts in the results into 10 clusters using Kmedoid an J-Divergence then selects 1 from each cluster. These methods of finding relevant blog posts for a

given headline perform well in the blogosphere, however in our case these may not perform well as tweets are very short, limited to 140 characters, and Twitter users use many abbreviations which is difficult for applying language models.

CHAPTER 3 DISCOVERING TARGET STAKEHOLDERS OF FIRM'S TWEETS

3.1 Background and Motivation

Since it was developed in early 1980s, *stakeholder theory* has been widely accepted and used in business practice. Many big companies, e.g. J&J, eBay, Google, Lincoln Electric, and AES, have been successfully applying stakeholder theory to run their businesses [11]. In their work [12], Freeman et al. define stakeholder as “*any group or individual who can affect or is affected by the achievement of the organization’s objectives.*” Thus, it is very important for the management to create methodologies to manage relationships with stakeholders in the strategic planning of firms. Among different ways to maintain the good relationships with stakeholders, it is vital for firms to be able to communicate well and clearly with their stakeholders about firms’ business goals to gain their support [47]. As a result, firms have developed many different strategies to communicate with their stakeholders. Some companies prefer an *integrated approach* where they identify a set of values regarding what they are doing and consistently convey the messages about those values with different emphases for different stakeholder groups. This approach helps balance the interests of stakeholders which would definitely benefit the companies in the long run [42]. Other companies use *corporate social responsibility (CSR) communication* by getting involved in the CSR activities to gain positive attitudes and reactions from their stakeholders [32] which leads to more purchases, seeking employment, and

investment [9].

To communicate with their stakeholders, firms have been using many different traditional methods, e.g. corporate news websites or press releases, to deliver their messages [10]. These methods can help stakeholders get updates about firms, however, they are mostly one way communications and also it takes time to compose and publish such messages leading to delay between the time some events happen and the time stakeholders get notified. Also, they can only meet the expectations of several stakeholder groups [19]. To overcome those limitations, many firms are utilizing social media platforms, like Facebook and Twitter, to communicate with their stakeholders as they help firms establish two way communication channels with stakeholders in real time. In research investigating how Fortune 500 companies engage stakeholders using Twitter [40], Rybalko et al. found that 170 out of 500 companies have an active Twitter account and among those companies, 61% were classified as dialogic communication which is defined as any negotiated exchange of ideas and opinions [31].

In this research, we are attempting to uncover the Twitter-based stakeholder communication strategy of firms. Specifically, we would like to explore which groups of stakeholders a firm is focusing on via its tweets. Also we would like to see how the stakeholder communication strategy of a firm changes overtime as well as the difference among the strategies of firms based on their characteristics like industry. The result will potentially provide valuable business intelligence to market analysts who would like to discover firms' social media strategies and behaviors.

3.2 Research Questions

As firms are increasingly using Twitter messages to communicate with their stakeholders, our focus is on the question *Do they have a specific stakeholder communication strategy via Twitter?*. We quantify the communication strategy in terms of the relative distribution of tweets over various stakeholder groups. To answer this broad question, we pose some methodological questions such as *How do we discover the strategies via tweets?*, *What is the relationship between firm's characteristics and the target stakeholder of its tweets?*, and *Can we predict the stakeholder groups based on the tweet's content and/or firm's characteristics?* If we can develop effective methods for predicting target stakeholder groups, we would like to explore several other interesting problems such as time based variations of target stakeholder groups at firm-level and industry-level.

3.3 Methodology

Our framework for data collection and classifier training is illustrated in Figure 3.1. We first gather information about selected companies like name, revenue, ranking, Twitter accounts, etc. from various public sources. Then we retrieve all tweets for those companies from the time their Twitter accounts were created. Next, we conduct the preliminary content analysis on a sample of collected tweets to discover the target stakeholder groups of those tweets. The results from content analysis were used for tweet labeling. We employed judges from a crowd sourcing service, oDesk, for annotating each tweet with its target stakeholder groups. We then used annotated

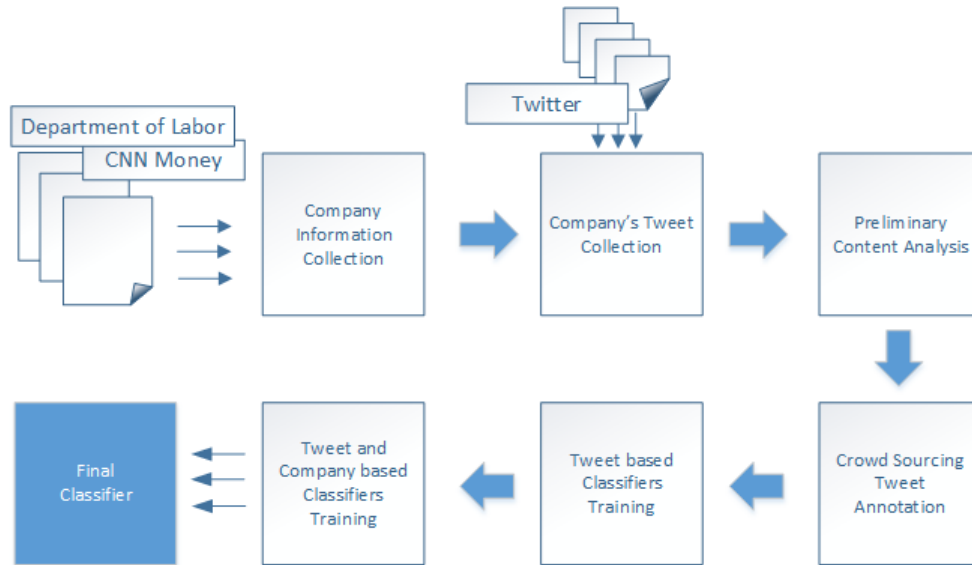


Figure 3.1: Framework for Data Collection and Classifiers Training

data to train classification models to automatically predict target stakeholder groups of a certain tweet. Finally we combine firms' characteristics like industry information with the features extracted from tweets content to see how additional firm-level information contributes to the performance of classification models.

3.4 Data Collection and Annotation

3.4.1 Data Collection

In this work, we are focusing on Fortune 100 companies from 2011, however our framework for collecting data is able to work with any company as long as it has a Twitter account. We extract the Twitter account information of Fortune 100 companies from CNN Fortune 500+ Web Application¹. Among 100 companies, 82 have

¹<https://www.fortune500-app.com/>

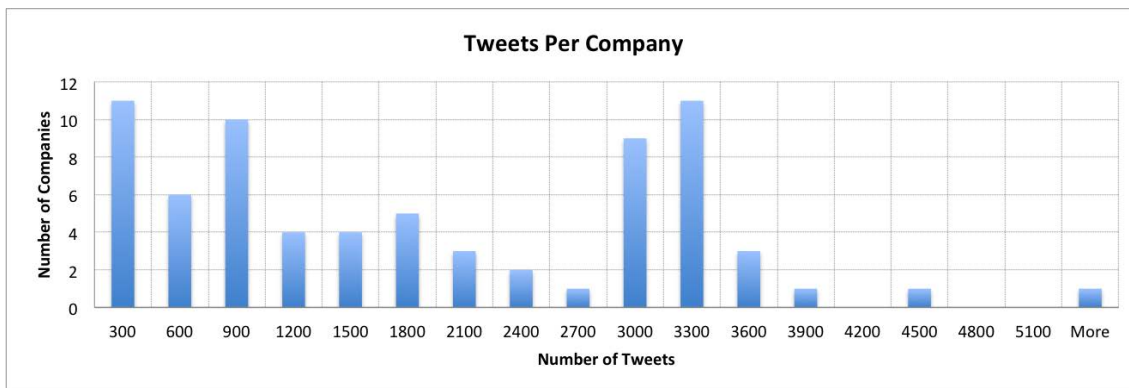


Figure 3.2: Distribution of Tweets per Company

Twitter accounts in which 72 accounts are active and have tweets. We use Twitter REST API² to get tweets from each company from the time of its first tweet until February 09, 2012. In total, we collect 128,374 tweets for 72 companies. The distribution of the number of tweets per account is illustrated in Figure 3.2. Most companies have from 1000 to 3500 tweets ($M=1782.972$, $SD=1304.06$). The most active company (Walgreen) posted 6077 tweets while the least active company (Comcast) had only 49 tweets.

3.4.2 Data Annotation

3.4.2.1 Preliminary Content Analysis

We conducted tweet content analysis to understand how semantics of tweet content relates to each of the target stakeholder groups. We identified firms' five main stakeholder groups: *Consumer*, *Investor*, *Employee*, *Government*, and *Comm-*

²<https://dev.twitter.com/docs/api>

nity. We classify tweets based on their content into classes according to their target stakeholder groups as follows:

Consumer-Focused: Tweets addressed to the Consumers usually provide information about new products, support for products sold, promotional campaigns, new locations (offices/stores) etc. Tweets in this class sometimes mention a company's customer-focused news/events/reports and include responses to the customers regarding their complains/comments.

Some examples:

- Self-service printing from Android smartphones arrives at FedEx Office. Learn more:
- deals: Amazing deal - only \$19 for a print, scan and copy printer by HP DJ-1051
- Recall is on Great Value steamable mixed vegetables & sweet peas contact Pictsweet Co. @ 1-800-367-7412 x417, or return for full refund.
- HP customers turning to Dell. 79% of respondents in IDG Research report indicate considering Dell for PCs
- Sorry to hear that. Make sure to reach out to should you have any more problems.

Investor-Focused: Tweets targeted to the investors of a company usually mention company's business plan like opening of new offices/stores, new product development projects, executive hiring, and merger/acquisition. Some investor-focused tweets describe company's performance like the revenue, stock price, market share, and advantages over the competitors as well as the state of the the industry.

Some examples:

- Walmart Reports Second Quarter EPS of \$0.97, Ahead of First Call Consensus;
Raises Full-Year EPS Guidance
- Total production for Angola Block 15 has exceeded 1 billion barrels.
- Lockheed Martin gets \$107 million contract
- We're excited to open our new 33,000 sq. ft. Mission Support Center today
in Clinton, MS!
- Dell plans to expand Silicon Valley staff for R&D

Employee-Focused: Tweets in this category allow for information sharing and communication within the company. Such tweets may cover company-related updates (from management), collaboration between employees to support customers, as well as employee appreciation messages.

Some examples:

- can you help? RT : - can you tell me what the UK Address is for printer
toner recycling please
- Congratulations to CIO Rob Carter, named to Fantasy Executive League! |
- Good luck on today's Jeopardy Tournament of Champions semi finals!
- Become the "CEO of You" to build your personal brand: In a career spent
managing corporate reputation, I've learned...
- PICS: Reunion. Now off to dinner!

Government-Focused: Tweets addressed to the Government (agencies) usually

cover issues such as jobs, taxes, and security. They may mention how government's policies have an effect on the company's business and how the company's business provides values to the nation.

Some examples:

- NYT advocates for short-term political gain on tax issue, misses long-term economic gain for U.S. [Read more on our blog](#)
- Our Kearl project also favors energy security: Canada supplies 20% of US oil imports & holds the world's largest reserves of oil sands.
- As MT Governor Schweitzer said about our project to move equipment thru his state to Canada, "It's jobs, jobs, jobs."
- Report: firms operating in the paid \$41 mil. in state taxes & \$35.6 mil. in local in 2009
- Our CEO talked about how fixing education can help fix other major challenges like our economy & global competitiveness

Community-Focused: Tweets in this category provide information about company activities or company-sponsored community activities to support the environment, education, health, children, and charity/philanthropy. Sometimes such tweets mention some news/report on these topics.

Some examples:

- 334 teacher fellowships over the past 27 years! [More on our 2011 Community Impact Award from |](#)
- RT : pledged \$1 million to the Japanese Red Cross to assist with the relief

and recovery

- ExxonMobil Community Summer Jobs Program partners w/ 60 nonprofits & welcomes newest class of Dallas-Fort Worth interns
- FedEx Response to Earthquake in Japan
- Thanks for spreading the word about our efforts to make food healthier and healthier food more affordable.ĴL

3.4.2.2 Annotating Data

Initially, we chose 5 companies in different industries for preliminary analysis and experimentation. Table 3.1 summarizes the information about these initial companies.

Company Name	Number of Tweets	Fortune500 Ranking
Wal-Mart Stores	3187	1
Exxon Mobil	712	2
Dell	848	41
Lockheed Martin	1581	52
FedEx	771	73

Table 3.1: Initial Companies

For each initial company, we randomly sample 100 tweets for annotating. Since a company may address its tweet to multiple stakeholders, each tweet may have more than one label. For example, the tweet *Media Advisory: Walmart Announces Opening*

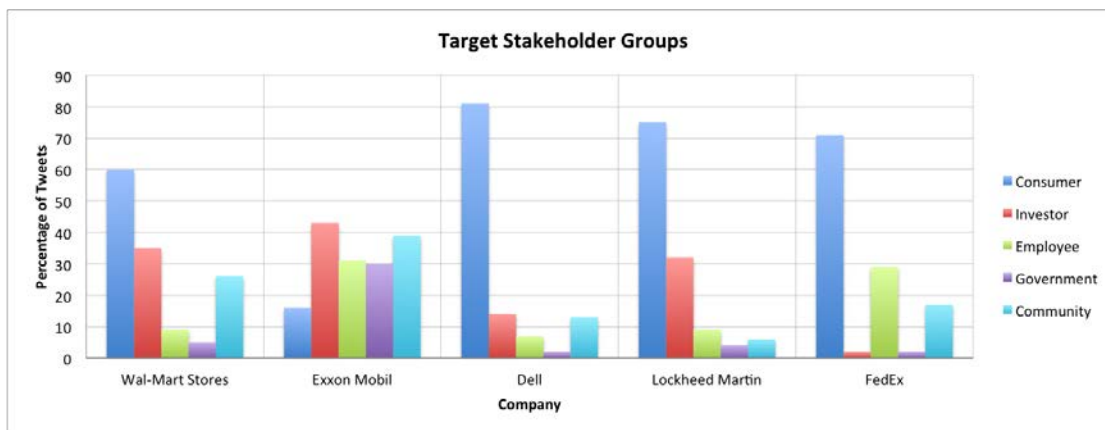


Figure 3.3: Target Stakeholders in Initial Companies

of Temporary Visitor Center in Bentonville <http://walmarturl.com/9Z2cB7> may be sent to both *consumers* to tell them the new location they can visit and *investors* to inform them the company's business plan. We manually annotate 500 tweets ourselves following the class definition in the previous section. Figure 3.3 shows the overall annotation results for each company. Most companies use a large portion of their tweets to communicate with consumers, especially Dell with over 80% of its tweets addressed to the consumers. For other stakeholder groups, the focus varies from one company to another. Based on Figure 3.3, it is clear that different companies have different stakeholder communication strategies.

In the next step we use crowd sourcing service from oDesk³ for annotating 500 sampled tweets corresponding to the initial companies. We use our labeling results as the gold standard to hire and evaluate the results of oDesk's contractors. We selected

³<http://www.odesk.com>

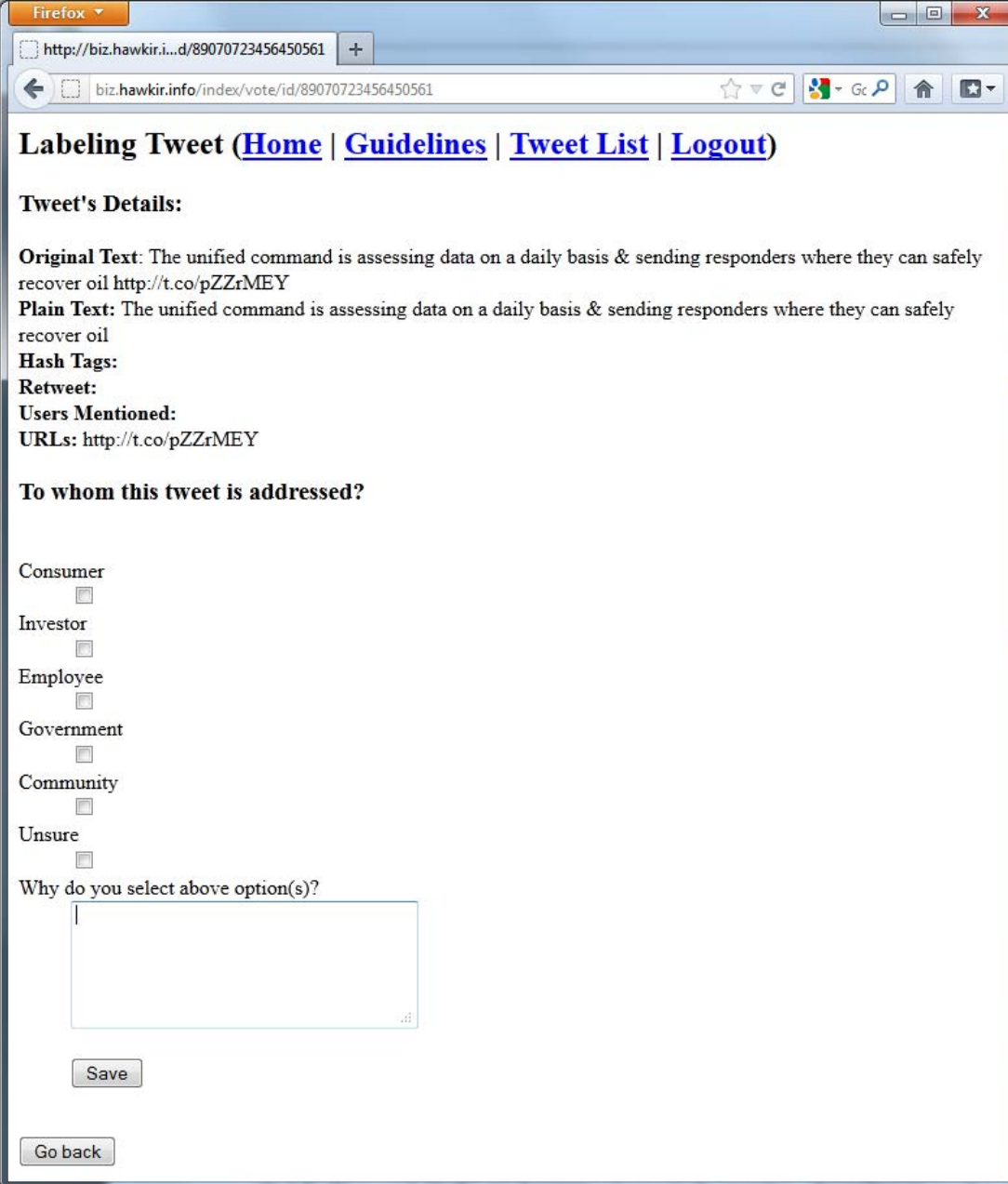
oDesk over other crowd sourcing services like Amazon Mechanical Turk⁴ because with oDesk we are able to test and hire the best applicants. In our case, after we posted the job on oDesk, there were 25 applicants. We asked them to read the annotating guidelines carefully and try to label 15 tweets. We rejected 17 applications because the applicants did not follow the guidelines and did not do well on the test. We hired 4 applicants among the remaining 8 ones to work on the job as they showed good performance, 2 applicants mislabeled 3 tweets and the other two mislabeled 4. The hired contractors then annotated 500 tweets using our online labeling system⁵. As illustrated in Figure 3.4, besides annotating the label(s) for each tweet, contractors have to type in the reason explaining their decision with the label(s).

After getting results from oDesk's contractors, we calculate the percentage of overlapping labels between gold standard and each contractor (judge). Also we calculate percentage of overlapping labels among each pair of contractors. Two sets of labels for a tweet from two different judges are considered overlapping if over 50% of labels in each set are identical. As shown in Table 3.2 judge #2 and judge #3 have the highest percentage of overlapping labels with the gold standard (80%). In addition, the percentage of overlapping labels among themselves is also high (71%). As a result, we select them for the next rounds of annotating tweets in which we will incrementally label sampled tweets for 72 companies.

In the next round of labeling, selected oDesk contractors label 50 sampled

⁴<http://www.mturk.com/mturk/>

⁵<http://biz.hawkir.info>



The screenshot shows a Firefox browser window with the address bar containing the URL `http://biz.hawkir.info/index/vote/id/89070723456450561`. The page title is "Labeling Tweet (Home | Guidelines | Tweet List | Logout)".

Tweet's Details:

Original Text: The unified command is assessing data on a daily basis & sending responders where they can safely recover oil <http://t.co/pZZrMEY>

Plain Text: The unified command is assessing data on a daily basis & sending responders where they can safely recover oil

Hash Tags:

Retweet:

Users Mentioned:

URLs: <http://t.co/pZZrMEY>

To whom this tweet is addressed?

- Consumer
- Investor
- Employee
- Government
- Community
- Unsure

Why do you select above option(s)?

Figure 3.4: Online Labeling System

	Gold Standard	Judge #1	Judge #2	Judge #3	Judge #4
Gold Standard	100%	62%	80%	80%	78%
Judge #1	62%	100%	66%	57%	59%
Judge #2	80%	66%	100%	71%	76%
Judge #3	80%	57%	71%	100%	69%
Judge #4	78%	59%	76%	69%	100%

Table 3.2: Percentage of Overlapping Labels among Judges

tweets for each of the remaining 67 companies, except for Comcast which has only 49 tweets. The oDesk contractors did not agree on labeling of 573 out of 3349 tweets (17%). We got the much better agreement in this round (83%) in comparison with the first round (71%). For tweets with labeling disagreement, we used the third judge and decide the final labels based on the majority vote. In the next step of data annotating process, we removed 19 irrelevant tweets which are either too short, containing only a URL, or non English. This process finally gives us a data set of 3380 labeled instances. Figure 3.5 summarizes the ratio of label (Y or N) for each stakeholder group. The ratios are found to be very imbalanced. We can see that most tweets are targeting consumers (78.80%) while very few appear to be communicating with employees (4.80%) and government (2.51%).

3.5 Experiments

3.5.1 Feature Description

From our content analysis on tweets as well as firms' characteristics, we identified a feature set including 5 tweet-based features (*text*, *hashtag*, *usermention*, *retweet*, and *URL*) and 9 firm-based features (*revenue*, *profit*, *industry division*, *ranking group*,

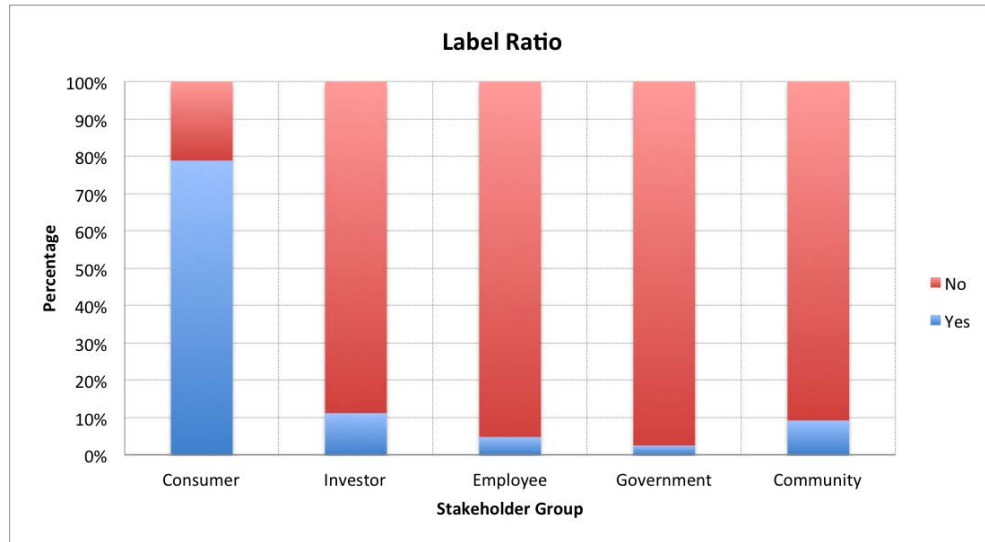


Figure 3.5: Label Ratio of each Stakeholder Group

and average percentage of tweets for each stakeholder group) which we will use for training the classifiers. Details about each feature are as follows:

- **Tweet-based features**

- **Text:** The plain text extracted from a tweet after the original text is cleaned, e.g. removing URLs, and # characters, etc. The plaintext is later on converted into a vector of words from which the top 1000 tokens are kept based on their ranking on information gain.
- **Hashtag:** A binary feature indicating if the tweet contains one or more hashtags.
- **Usermention:** A binary feature indicating if the tweet mentions some other Twitter users or not. We observed that in many cases, tweets targeting consumers or employees included other Twitter users in the content.

- **Retweet:** A binary feature indicating if the tweet is a retweet or not. We find that in some situations, retweet is used to forward a message among different departments of a firm.
- **URL:** A binary feature indicating if the tweet contains a URL or not. We notice that when a company announces something new, its tweets usually include one or more URL to a more complete version of the announcements.

- **Firm-based features**

- **Revenue:** Revenue of firm for the year 2010
- **Profit:** Profit of firm for the year 2010
- **Division:** The industry division that the company is categorized into. To obtain this information, we first use a firm’s ticker symbol to lookup its Standard Industrial Classification (SIC) code at the U.S. Securities and Exchange Commission.⁶ Then we gather information about company’s industry, industry group, and industry division from U.S. Department of Labor.⁷
- **Consumer:** The percentage of tweets targeting consumers. This information is obtained from the labeled data set.
- **Investor:** The percentage of tweets targeting investors.
- **Employee:** The percentage of tweets targeting employees.

⁶<http://www.sec.gov/edgar/searchedgar/companysearch.html>

⁷http://www.osha.gov/pls/imis/sic_manual.html

- **Government:** The percentage of tweets targeting government.
- **Community:** The percentage of tweets targeting community.
- **Group:** This feature identifies if a firm is in one of four groups based on its ranking on Fortune 100. The first group includes the first 25 companies, the second groups contains the next 25 companies, the third group has companies ranked from 51 to 75, and the last group includes the remaining companies.

We reduced the complexity of a multilabel classification problem by making 5 different binary classifiers corresponding to the 5 stakeholder groups, namely Community, Consumer, Employee, Investor, and Government. Each classifier classifies a given tweet into the corresponding stakeholder group or not. The final results would be the combination of output labels from those 5 classifiers. Tweets will be labeled with all the labels from the binary classifiers. As a result, we developed 5 different data sets for training appropriate binary classifiers. All of data sets have the same set of tweets. In a data set for a specific classifier, tweets labeled with that classifier will be marked Y , and N otherwise.

3.5.2 Evaluation

Figure 3.6 illustrates the relationship among outcomes of a classifier and the actual data. Classifier’s performance is usually measured by either *Precision* or *Recall* depending on the usage purposes.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

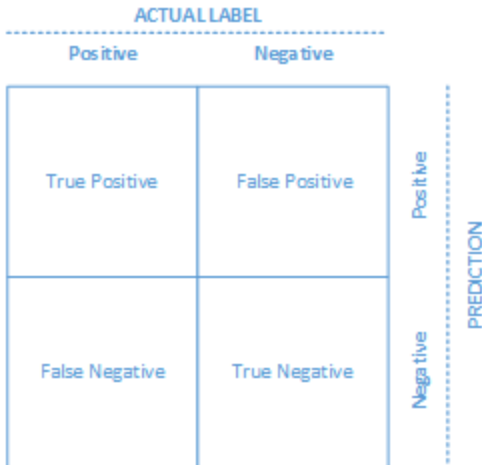


Figure 3.6: Prediction and Actual Data

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

As illustrated in Figure 3.5, our data set is very imbalanced. For example, approximately 80% of tweets are targeting consumers. Therefore we have to find a metric that can fairly reflect the performance of trained classifiers. We selected F-Measure as the main measurement of classifiers' performance as it considers both precision and recall.

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

We use *repeated random sub-sampling validation* to measure performance of our trained classifiers. For each algorithm, we train and test models 20 times. In each train-test iteration, we randomly split the original dataset into train set (90% of original dataset) and test set (10% of original dataset). The overall performance of a classifier is the average F-Measure from 20 iterations.

3.5.3 Classifiers Using Tweet-based Features

In order to investigate the predictive power of tweet-based features, we removed all the company based features from the data set and used only tweet based features for training classifiers. We started with three basic classifiers training algorithms, namely NaiveBayes, Decision Tree (J48), and Support Vector Machine (SMO). The training and validating of models follows the process described above. Table 3.3 summaries the evaluation results of trained classifiers.

	Naive Bayes	Decision Tree	Support Vector Machine
Community	0.52	0.21	0.19
Consumer	0.83	0.76	0.60
Employee	0.48	0.03	0.13
Government	0.72	0.35	0.54
Investor	0.24	0.20	0.21

Table 3.3: F-Measure of Base Classifiers

Naive Bayes classifiers outperform others for all stakeholder groups. However the performance on Investor label is still low. Support Vector Machine classifiers are slightly better than Decision Tree ones but our statistical tests show that there is no statistical difference.

In addition to the base classifiers, we applied several heuristics in order to improve performance for ones with the imbalance dataset. We selected Cost Sensitive Classifier and Threshold Selector to use with the base classifiers. Table 3.4 shows the

best results of heuristic classifiers.

	Heuristic Classifier
Community	0.46
Consumer	0.72
Employee	0.48
Government	0.69
Investor	0.22

Table 3.4: F-Measure of Heuristic Classifiers

We found that the Heuristic classifiers have slightly lower performance than the base ones. However our statistical tests show that there is no significant difference between F-Measure of these two groups.

3.5.4 Classifiers Using Firm-based Features

In order to examine the firm-based features, we removed all tweet-based features and use only company-based features to build classifiers. Table 3.5 shows F-Measure from 4 classifier building algorithms: NaiveBayes, Decision Tree, Support Vector Machine, and Logistic Regression.

Except for the consumer group, Naive Bayes classifiers again outperform others. Support Vector Machine classifiers perform well in consumer group but perform very poorly in other groups. Overall, we found that performance of classifiers trained using firm-based features is significantly lower than one from classifiers built using tweet-based features. This is not surprising since firm-level features are not specific

	NaiveBayes	Decision Tree	SVM	Regression
Community	0.38	0.00	0.00	0.19
Consumer	0.89	0.90	0.90	0.90
Employee	0.20	0.00	0.00	0.17
Government	0.12	0.00	0.00	0.00
Investor	0.35	0.16	0.00	0.25

Table 3.5: F-Measure of Firm-based Classifiers

to the tweet being classified but are the same for all tweets by the same firm, hence providing a coarse information source. We tried to use add some heuristics to those based classifiers in an attempt to improve their performance however the results were not as we expected, they are even worse than the ones from base classifiers.

3.5.5 Classifiers Using both Tweet-based and Firm-based Features

As we see that both tweet based and company based features have some predictive power to predict the target stakeholder groups of a given tweet. In this section, we present several ways to build classifiers using both types of features to see if we can achieve better performance.

We first put all tweet-based and company-based features together in one dataset then built classifiers from that dataset. The experiment results show us that naive integration tweet-base features and firm-based features does not provide additional advantages. In fact, its performance is worse than the ones from classifiers trained using either tweet-based or company-based features. In order to leverage the complement nature of firm-based and tweet-based features, we designed a new system

using a two-level process to build a classifier. The first level finds the optimal tweet-based classifier and the second level combines the output of the tweet-based classifier with several firm-level variables. As illustrated in Figure 3.7, we split the input data set A into set B (90%) and set C (10%). We put C aside for later testing. We continued splitting set B into train set D (90% of B) and validation set E (10% of B). We then apply our mechanism to build two level classifier as shown in Figure 3.8. We use D and E to train multiple classifiers including base algorithms and combination of base algorithms and heuristics. The best training algorithm will be use to train on B and test on C. In the next step we use trained classifier to assign a probability for each instance in B and C. Then we remove all the tweet based features from B and C resulting the train set and test set with only company based features with the probabilities from the tweet based classifier. Finally we train a logistic regression model on B and test on C. The whole process is repeated $n = 20$ times. Table 3.6 shows the averages of final results when the above mentioned training process is applied to each of the labels of stakeholder groups. Also, Figure 3.9 shows performance of all classifiers in our experiment. We observe that classifiers trained with *Naive Bayes* show a strong performance. Performance achieved by heuristically tweaking learning parameters (i.e., cost and threshold) is not significantly different from the baseline ($p\text{-value} \geq 0.05$ from $t\text{-test}$) Naive Bayes that uses default values. Except for the tweets corresponding to the *Government* label, the final classifier that uses both tweet-based and firm-based variables outperforms the best classifiers achieved using just the tweet-based features.

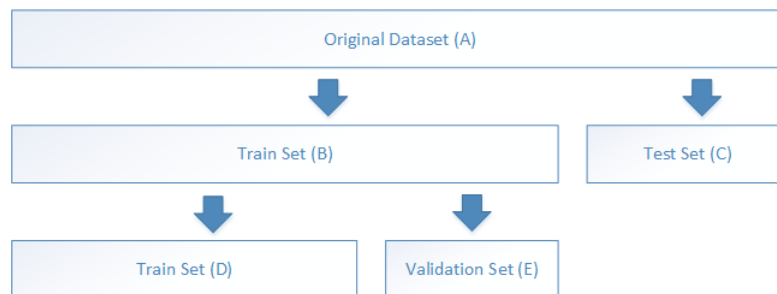


Figure 3.7: Splitting Data

	Combining Classifier	
Community		0.62
Consumer		0.83
Employee		0.73
Government		0.18
Investor		0.59

Table 3.6: Tweet-based and Company-based Classifiers

3.6 Analysis

3.6.1 Relative Contribution of Features

In general we found that both tweet-based features and firm-based features have some predictive power with all data sets of stakeholder groups. As illustrated in Figure 3.7, in all groups, the output from tweet-based classifier, *Probability*, play a significant role in the final results ($p\text{-value} \leq 0.05$). Of the tweet-based features, we found that features extracted from the tweet's text have more values than the others. Non-text features usually either appears in a large portion of tweets, e.g. 63% of tweets contain URLs, 48% of tweets are retweets, or too few tweets contain them,

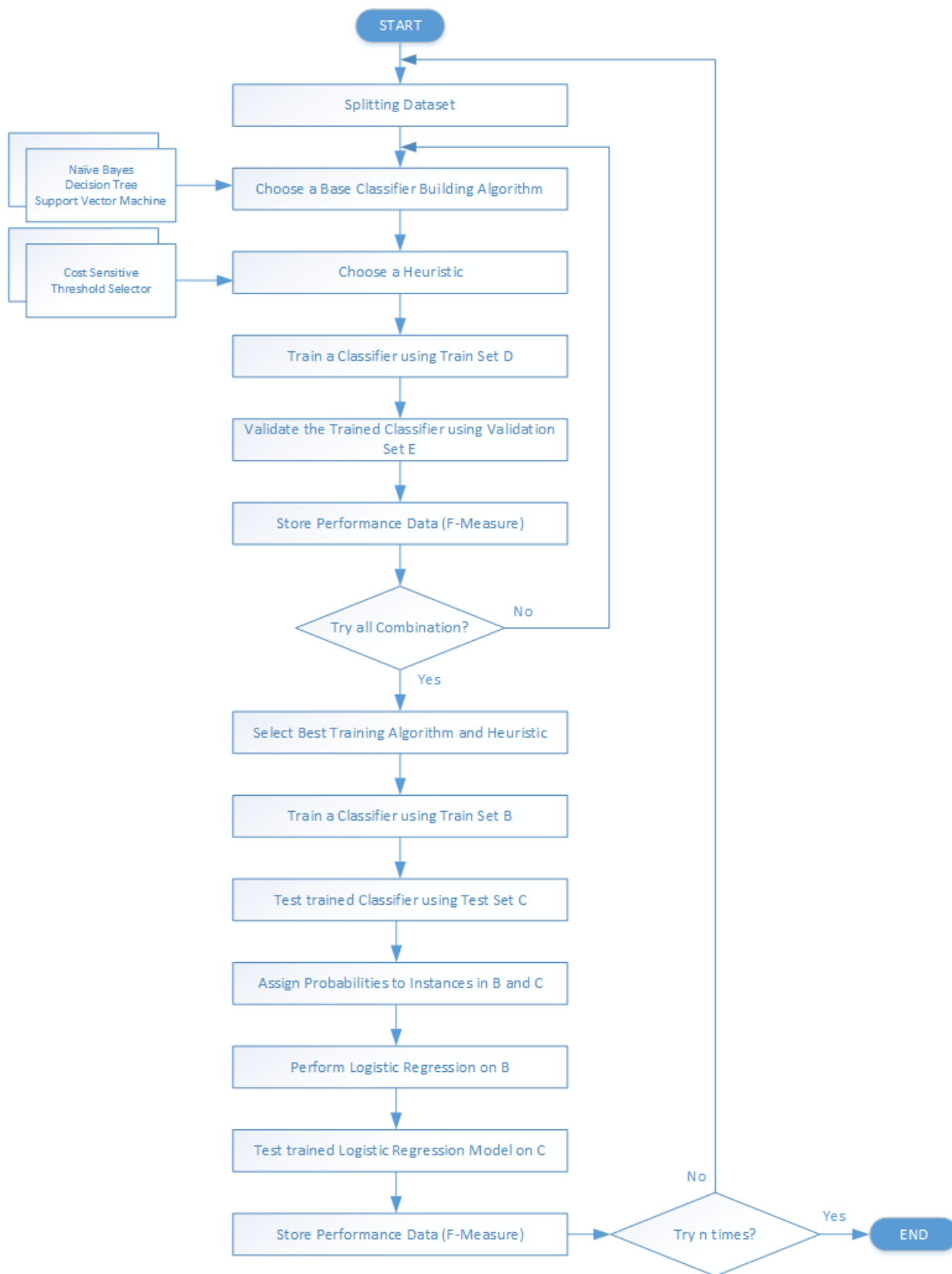


Figure 3.8: Combining Tweet-based and Firm-based Features

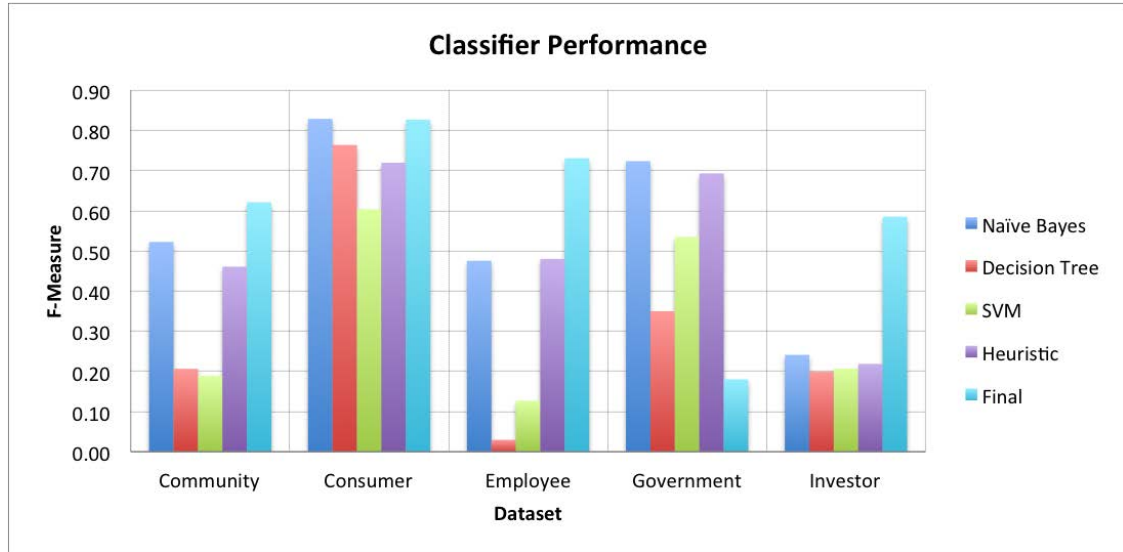


Figure 3.9: Classifier Performance for Different Stakeholder Groups

for example 10% of tweets mention other Twitter users. Thus they do not contribute much in classifying a tweet.

For the firm-based features, our experiment results from logistic regression models show that for some labels, the firm-based values are statistically significant ($p\text{-value} \leq 0.05$). Table 3.7 shows the firm-based features that are statistically significant for each label. We can see that 5 firm-based features are statistically significant in *Investor* label which help boost the performance of the corresponding classifier by approximately 145%.

3.6.2 Application on All Tweets

We have built and evaluated classification models to predict the target stakeholder group for tweets. In this section we are applying our best classifiers on over

Label	Firm-based Feature	p-value	Effect Direction
Community	Revenue	0.012	-
	% of tweets to consumers	0.002	+
	% of tweets to investors	0.003	+
	% of tweets to community	0.000	+
	Probability	0.000	-
Consumer	% of tweets to consumers	0.000	+
	Probability	0.000	+
Employee	Revenue	0.044	+
	Probability	0.000	-
Government	% of tweets to investors	0.040	+
	% of tweets to government	0.000	+
	Probability	0.000	-
Investor	Industry (Manufacturing)	0.001	+
	% of tweets to consumers	0.025	-
	Industry (Finance)	0.003	+
	Industry (Transportation)	0.016	+
	Industry (Wholesale)	0.016	+
	Probability	0.000	-

Table 3.7: Statistically Significant Firm-based Features

128,000 tweets we collected from 72 out of the Fortune 100 firms. Each tweet is classified by all 5 classifiers. The final label set of a tweet is the combination of results from 5 classifiers. We report the percentage of tweets targeting each stakeholder group during 2011 for 5 initial companies in Figure 3.10. It is clear to see that each company has a different distribution of the percentage of tweets for each stakeholder group. Walmart main focuses are consumer and community groups. Exxon Mobil has consumer group as its main focus and it has pretty balance number of tweets targeting the remaining group. This is very different from other companies. Most of Dell's tweets are targeting consumer and investor groups while a smaller number of tweets focuses on community. Besides consumer group as its main target, Lockheed Martin spends a considerable number of tweets to target investor and government groups. FedEx main targets are consumer and investor groups. It also has a significant number of tweets targeting employee group in comparison with other companies. We also observe that the distribution changes over time. Figure 3.11 illustrates the stakeholder communication strategy among different industries. We can see that while all industries spend a large percentage of their tweets on consumers, each industry has different focuses on other stakeholder groups. For example, *wholesale trade* focus more on investor and community groups, *services* focuses more on investor group, and *manufacturing* focuses more on investor and government groups. It is interesting to see a spike in community group in *retail trade* from week Sep 19, 2011 to week Oct 10, 2011. This is the time *Walgreens* ran a campaign to donate flu shots to the community, as a result the company posted a large number of tweets to promote the

campaign. We also plot the volume of tweets for each stakeholder group of initial companies and industries in Figure 3.12 and Figure 3.13 respectively.

3.7 Summary

In this chapter, we described the details of our work with company's tweets, *Discovering Target Stakeholders of Firm's Tweets*. We presented our methodology for collecting data of Fortune 100 companies from public resources and gather all tweets of those companies. Then we present our work on content analysis to identify the target stakeholder groups of tweets. We also described our methodology to have tweets annotated using crowdsourcing services. Finally we showed our feature set and experiment design to build classifiers to predict target stakeholder for tweets where we used base classifiers, heuristics, and our new method to combine results from tweet-based classifiers with company-based feature as the inputs for the final classifier. The experiment results show that the combined features help improve performance in term of F-Measure for most of the stakeholder datasets. As a future direction, we would like to find a better way to handle the imbalance class problem as some stakeholder data sets have a very small portion of of tweets in one class and the rest in the other. Also we would like to extend the data set to a larger number of companies, e.g. Fortune 500 or S&P 500.

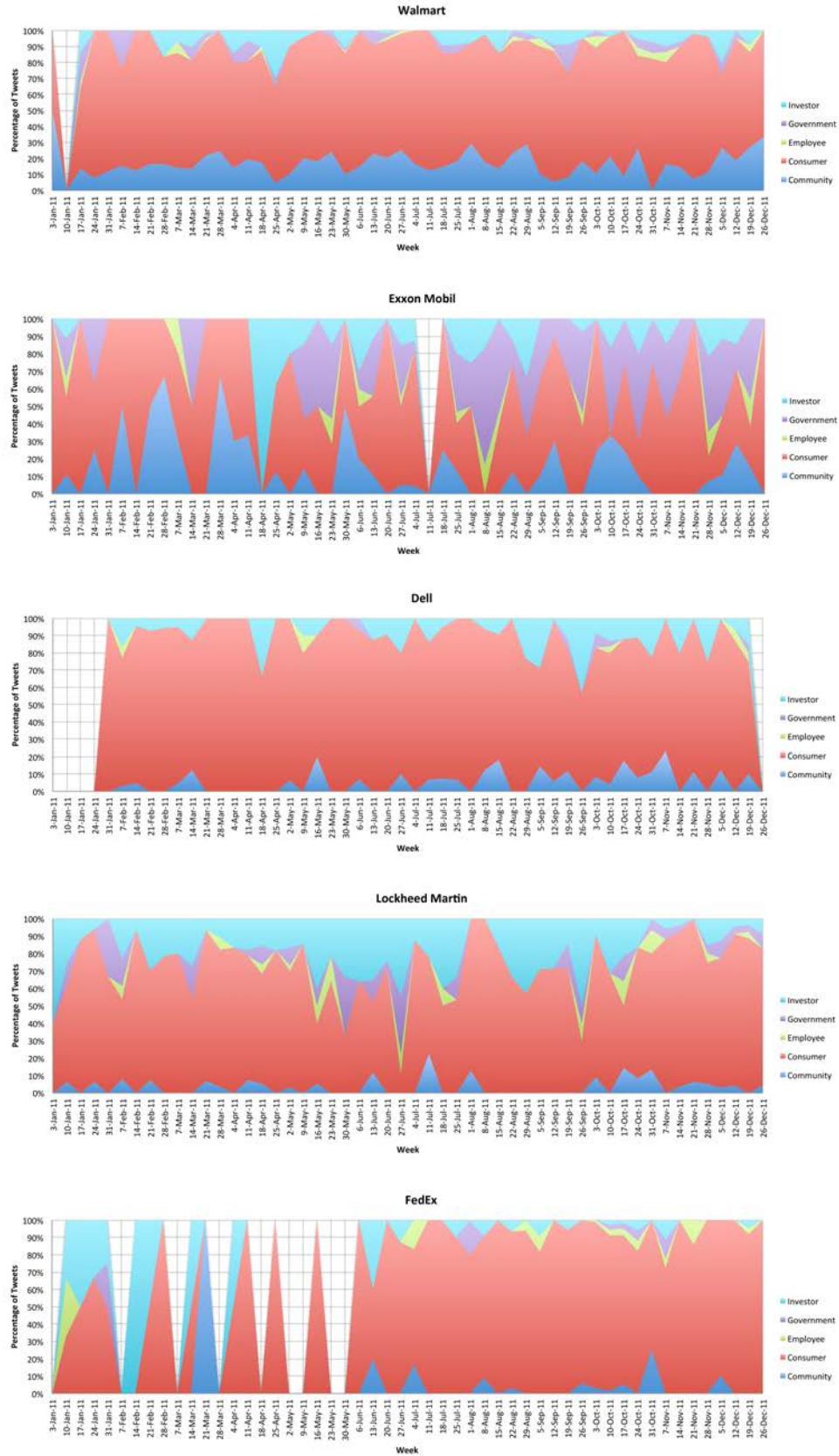


Figure 3.10: Stakeholder Communication Strategy of Initial Companies

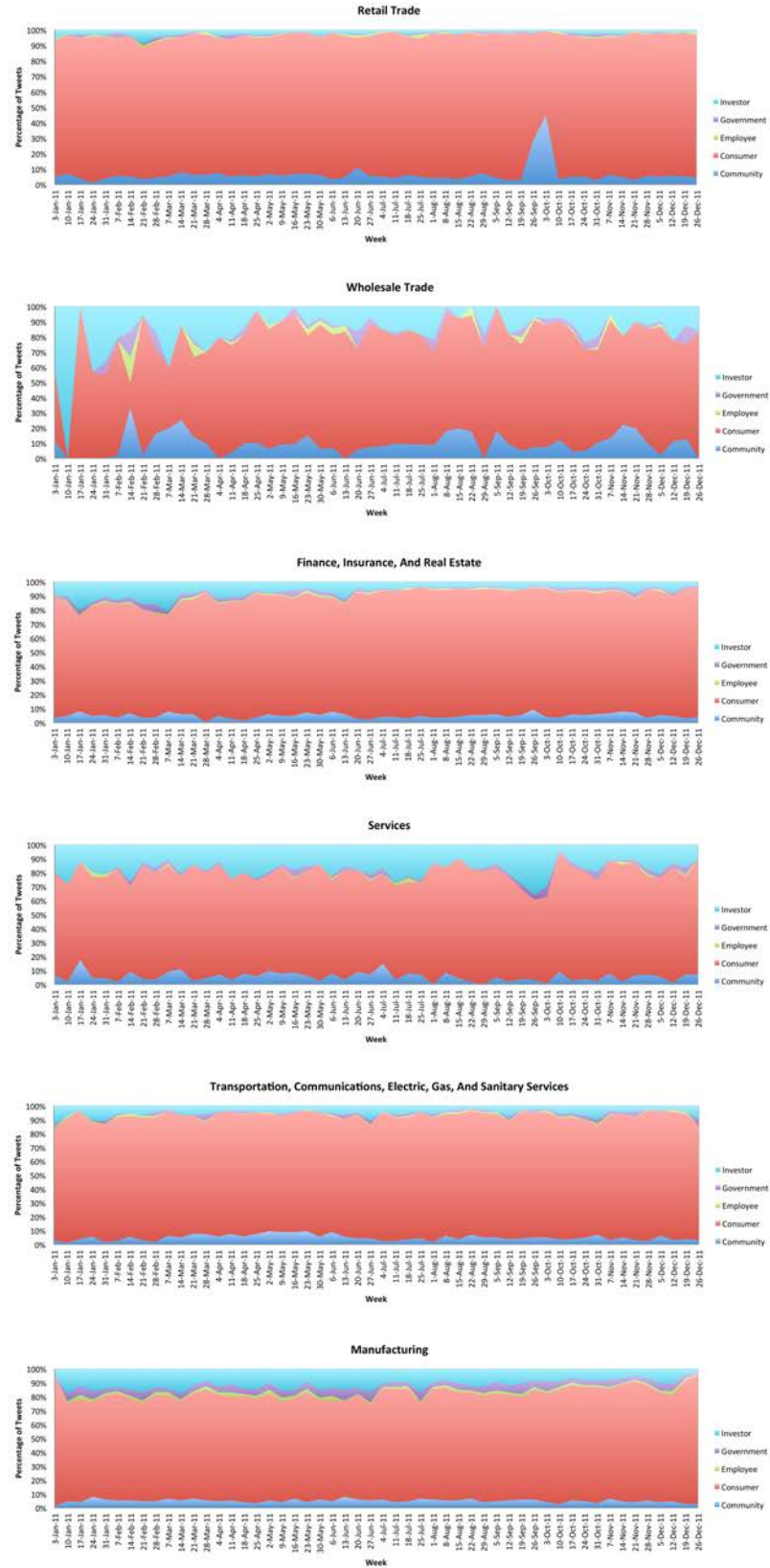


Figure 3.11: Stakeholder Communication Strategy of Different Industries

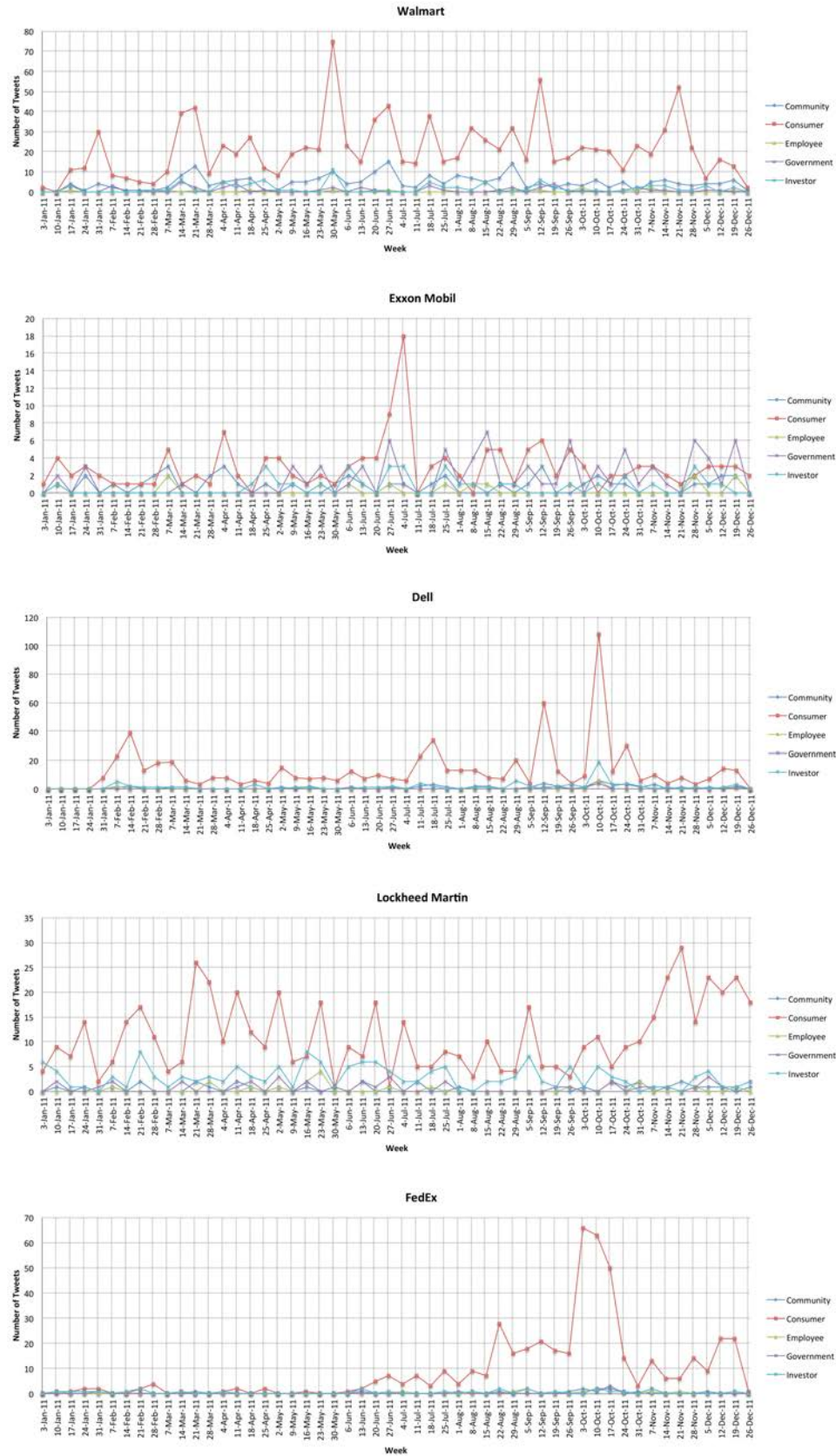


Figure 3.12: Stakeholder Communication Strategy of Initial Companies by Volume

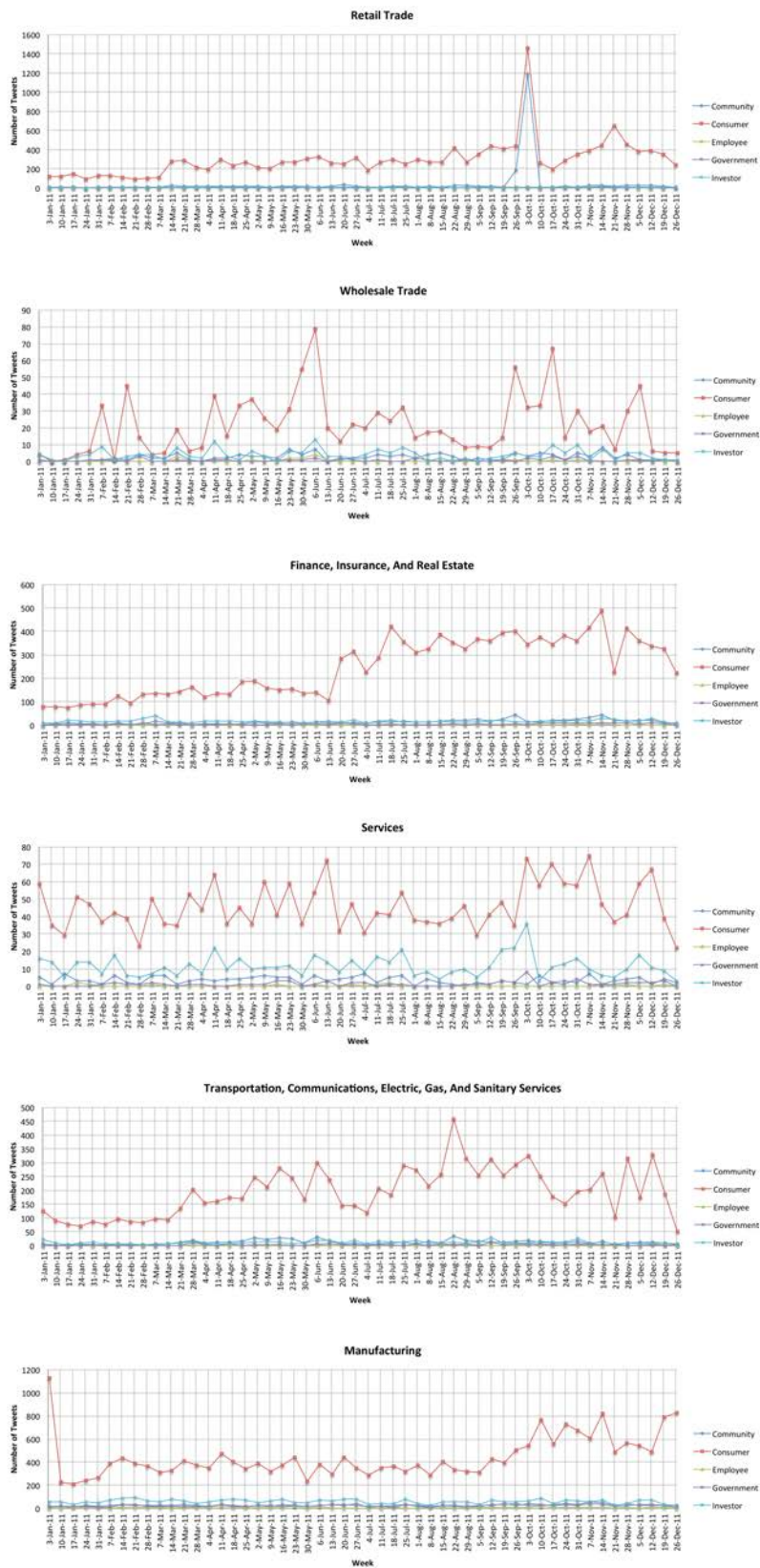


Figure 3.13: Stakeholder Communication Strategy of Different Industries by Volume

CHAPTER 4 DISCOVERING PUBLIC REACTIONS TO NEWS HEADLINES

4.1 Background and Motivation

It is very important for news organizations to understand their readers to serve them better [51], e.g. for only publishing news articles relevant to readers' interests. Also it is important for policy makers to understand the population through their reactions to certain type of news leading to appropriate changes in policies. Traditional survey methods have long been used for public opinions about news articles, TV programs, recent events, etc. In such methods, researchers, like ones from the Pew Research Center, usually have to prepare questionnaires, select the sample of population, conduct interviews either by phone, email, or in person, then analyze the survey data for results. While such methods can collect a portion of the public opinions, they have some drawbacks. First, it takes time for preparing the survey materials making the research work lag behind the time that news or events happen and the results may not correctly reflect the true reactions from the public. Second, these methods are not scalable as researchers have to identify the current sample of population then contact each individual for data. Third, the data only reflect opinions of a small portion of the public. Fourth, when people know that they are subjects of a survey, they may not completely express all of their views. Last but not least, these methods would be expensive as they require a lot of manual works.

Social media platforms, like Twitter, allow millions of people to freely and

anonymously express their personal opinions. Recent research from the Pew Research Center [44] show that about 19% Twitter users in the U.S. post personal updates 12% share links to news stories, and 16% post general live observations one or more times per day. Given the size of Twitter’s user base, we can see that Twitter’s data stream is possibly an excellent source for investigating public opinions. Although Twitter’s data stream may be noisy and contain abbreviations and many other stylistic variants, it has several advantages in comparison with traditional surveying methods. It provides opinions of much larger proportion of the population in real time and it allows researchers to automate the process of gathering and handling data with much less expenses. These motivate us to propose a new framework to passively gather public opinions regarding news or events by mining data from Twitter’s data stream.

We first investigate the level of public interest in different topic categories of news. We measure level of interest expressed by groups of readers from different parts of the world. We also explore public response to those news headlines by analysing tweet content. Here again we compare responses from different parts of the world. We also compare responses to news that is ‘local’ (i.e., from the same country) to news that is foreign. Our expectation is that by just examining tweets, we are able to understand the similarity and differences in public interests and reactions to topic categories news among different groups of news readers.

Our research has potential in many real life applications. For example, we can help the news organizations have insights into their audience interests to have better content publishing strategies to fit with different audience groups. We contribute

methods that passively and cheaply analyze public response to news (and thereby complement structured surveys). Our methods may be used for different purposes and will better scale than traditional surveys. We can also gather related tweets to the news articles as the readers' feedback to those news articles.

4.2 Research Questions

In this work, we are trying to find answers to several research questions. We know from the Pew Research Center that some Twitter users share links about news stories; however we would like to explore this behavior further by asking *Do the news readers actually discuss what they read in addition to sharing links?* Then we ask *How do the discussions vary across broad categories of news? How do the discussions vary among different categories of news within a geographical location? What is the difference among the discussions on the same news category across reader groups at different geographical locations?* In order to find answers to above research questions, we also ask several methodological questions: *How can we find the tweets mentioning/discussing a specific news article?* and *How do we analyze a user's reaction to news from her tweet messages?* Answers to the methodological questions will help us develop a framework to systematically discover and understanding the reaction of public to news headlines.

4.3 Methodology

We designed a framework, as illustrated in Figure 4.1, to address our research questions. We started with collecting news headlines and key metadata from sources.

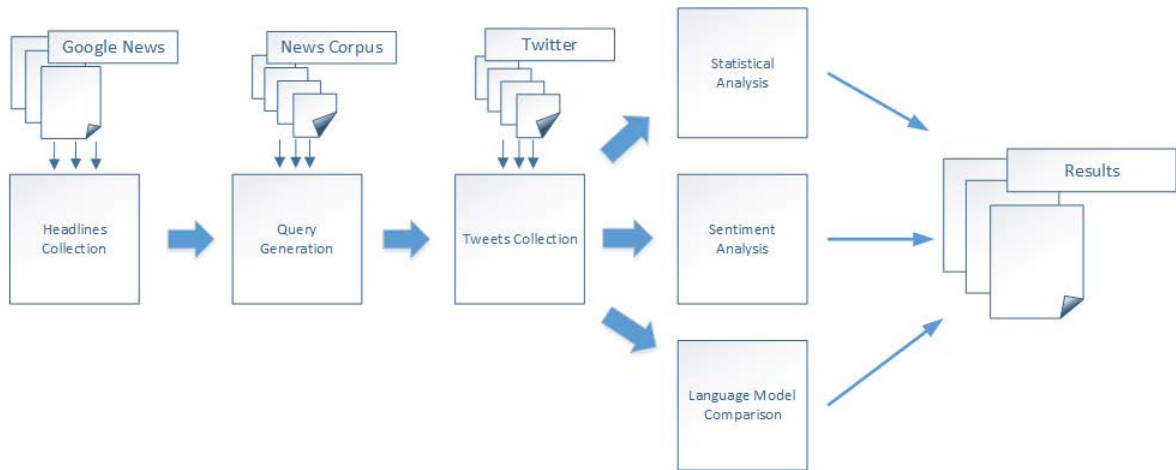


Figure 4.1: Framework for Analyzing Twitter Response to News Headlines

In our work, we selected Google News as a source for extracting news headlines for several reasons. First, Google News is a news aggregator that automatically pulls up to date headlines from many different news sources from all over the world then groups the similar headlines together into categories¹. Second, Google News is available in many different locations and languages where each local version of Google News is tailored to fit with audience in that location. Last but not least, Google News provides RSS feeds which are extremely helpful for us to extract and categorize headlines. Other options would have been Huffington Post² or World News³; however we choose Google News given its availability and ease of collecting data. In the next step, we collected related tweets for each headline. We use information from a headline

¹http://news.google.com/intl/en/about_google_news.html

²<http://www.huffingtonpost.com>

³<http://www.wn.com>

together with external information to generate a query to search Twitter for relevant tweets. We then perform some statistical analysis to gain insights and comparison of Twitter users' interests in news. We also applied sentiment analysis on tweets to measure the reaction of Twitter users to news. Finally, we use language model comparison to compare reactions to the same news headline from different groups of news readers. The rest of this chapter provides details about each step in our framework.

4.4 Data Collection

Motivated by the previous work on the international comparison of trending Twitter topics in English by Wilkinson et al. [52], we focus our work on five different countries: Australia, India, South Africa, United States, and United Kingdom as they represent different geographical locations and we can collect a reasonable number of English headlines and tweets. Of course we can collect headlines and tweets from other English speaking countries as well; however selecting above countries with the clear geographical separation would help us avoid the limitation of tweet collection process based on Twitter APIs that may confuse the country of a tweet when we issue queries for searching tweets in a specific country. In this work, we collect headlines and related tweets from these countries for further analysis. Also in order to compare the interests and reactions of news readers in different countries for the same news headlines, we perform cross-country tweet retrieval in which we use headlines from one country and search for related tweets in the remaining countries. Tweet retrieval

	Latitude	Longitude	Radius (Miles)
Australia	-27.00	133.00	1300
India	20.00	77.00	1000
South Africa	-29.00	24.00	600
United States	38.00	-97.00	1500
United Kingdom	54.00	-2.00	400

Table 4.1: Geocode Information for the Countries in our Dataset

process happens as follows. From a news headline, we generate a query of k keywords. Then we use Twitter Search API⁴ with these keywords to search for relevant tweets. To only retrieve tweets for a specific location, we attached a *geocode* including a tuple of $(latitude, longitude, radius)$ to a query. For each country of interest, we obtained the latitude and longitude of its central point from The World Factbook⁵ and the its radius from Google Maps⁶. Details about the geocodes for all 5 countries are shown in Table 4.1. Twitter Search API also allows searching for tweets that are less than 7 days old which enables us tracking relevant tweets for headlines that are less than 7 days old. For example, at October 8th, we are still able to search for tweets mentioning about headlines in October 1st.

⁴<https://dev.twitter.com/docs/api/1.1>

⁵<https://www.cia.gov/library/publications/the-world-factbook/>

⁶<http://maps.google.com>

Country	Headline	
	Title	Category
Australia	Wooden spoon looms for spineless Warriors - The West Australian	Sport
	URL vulnerability forces Australia Post service offline - ZDNet	Technology
	Challengers win in Georgia polls - The Australian	World
India	After medicines, Rajasthan govt toys with idea of free tests - Times of India	Health
	Akhilesh Yadav slams FDI in retail, says Mulayam will decide on Mamata's invite - NDTV	National
	Always wanted to direct Rani Mukhreej: Anurag Kashyap - Indian Express	Entertainment
South Africa	Cops kill man wielding toy gun - iAfrica.com	National
	Amplats says security worsens at its S.Africa mines - Reuters Africa	Business
	Records tumble in Solar Challenge - Independent Online	Technology
United States	Iran's President Ties Drop in Currency to Sanctions - New York Times	World
	DT in talks to merge T-Mobile USA with MetroPCS - Reuters	Technology
	Salmonella in Netherlands and US from Dutch smoked fish - BBC News	Health
United Kingdom	Buying Time for the Environment by Creating a Dust Cloud in Space - OilPrice.com	Science
	Runaway teacher agrees to UK return - The Press Association	National
	Dragon's Den star Duncan Bannatyne's health scare was not a heart attack - The Sun	Entertainment

Table 4.2: Example Headlines from Different Countries Collected on October 2, 2012

4.4.1 Headlines Collection

Google News organizes related headlines in each country into categories, namely, *Business*, *Entertainment*, *Health*, *National*, *Sport*, *Science*, *Technology*, and *World*. Headlines are automatically updated multiple times a day. Each Google News headline includes a title, a brief description, an original source, and the time it was updated.

In order to collect headlines, we use our own automated software to hourly query Google News RSS for each country (Australia⁷, India⁸, South Africa⁹, United States¹⁰, and United Kingdom¹¹) over a period of 15 days from September 30, 2012

⁷<https://news.google.com/news/feeds?cf=all&ned=au&hl=en&output=rss>

⁸<https://news.google.com/news/feeds?cf=all&ned=in&hl=en&output=rss>

⁹https://news.google.com/news/feeds?cf=all&ned=en_za&hl=en&output=rss

¹⁰<https://news.google.com/news/feeds?pz=1&cf=all&ned=us&hl=en&output=rss>

¹¹<https://news.google.com/news/feeds?cf=all&ned=uk&hl=en&output=rss>

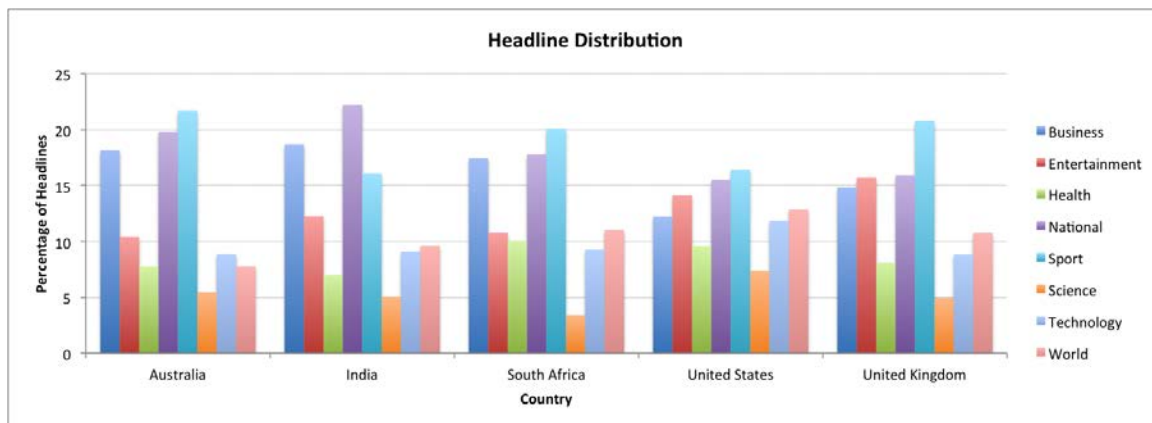


Figure 4.2: Headline Distribution

to October 14, 2012. In total, 30,974 headlines in 8 categories were collected for all 5 countries. Table 4.2 shows some example headlines with titles and categories from all five countries collected on October 02, 2012. Figure 4.2 shows the distribution of headlines in each category of each country. In general, headlines cover more on business, national, and sport. However, we can see that the categories of focus vary from one country to another. For example, more headlines in the United States mention about science and technology while headlines in India top in the national category. Table 4.3 shows the difference in percentage between the top ranked category and the bottom ranked for each country. It is clear to see that except for India, the most covered category accounting for from 16% to 22% of headlines in all other four countries is sport and the least covered category accounting for from 3% to 7% of tweets is science. Except for the U.S., the difference between the top ranked category and the least ranked one is approximately 17%. The gap for those categories in the U.S. is much smaller, about 9.5%.

Country	Top Ranked Category (%)	Bottom Ranked Category (%)	Difference (%)
Australia	22.57 (Sport)	5.44 (Science)	17.14
India	22.47 (National)	5.13 (Science)	17.34
South Africa	21.13 (Sport)	3.39 (Science)	17.74
United States	16.79 (Sport)	7.28 (Science)	9.51
United Kingdom	21.62 (Sport)	4.78 (Science)	16.84

Table 4.3: Differences in Top Ranked and Bottom Categories Ranked by Frequency

Figure 4.3 shows more details about headline distribution for each country during the headline collecting period. In general, the numbers of headlines in most categories follow the same pattern, going down at the weekend (October 6th, and October 13th) and going up at the beginning of the week (October 1st and October 8th). The number of headlines about national category is very high in India and it even goes up at the weekend. As we already saw before, science category has the least number of headlines in all countries. The number of headlines about business in Australia drops most over the weekend, from over 20% to under 10%. While there is an imbalance in the distribution of headlines in different categories in other countries, headlines in the U.S. seem to be least imbalanced among the categories.

4.4.2 Tweet Collection

In the next step, we collected the relevant tweets for each news headline. We used headline information to generate Twitter search queries then leveraged the Twitter Search API, which allows us to search for tweets within last 7 days, to search for and retrieve related tweets. This section details our strategies for query generating and tweets collecting.

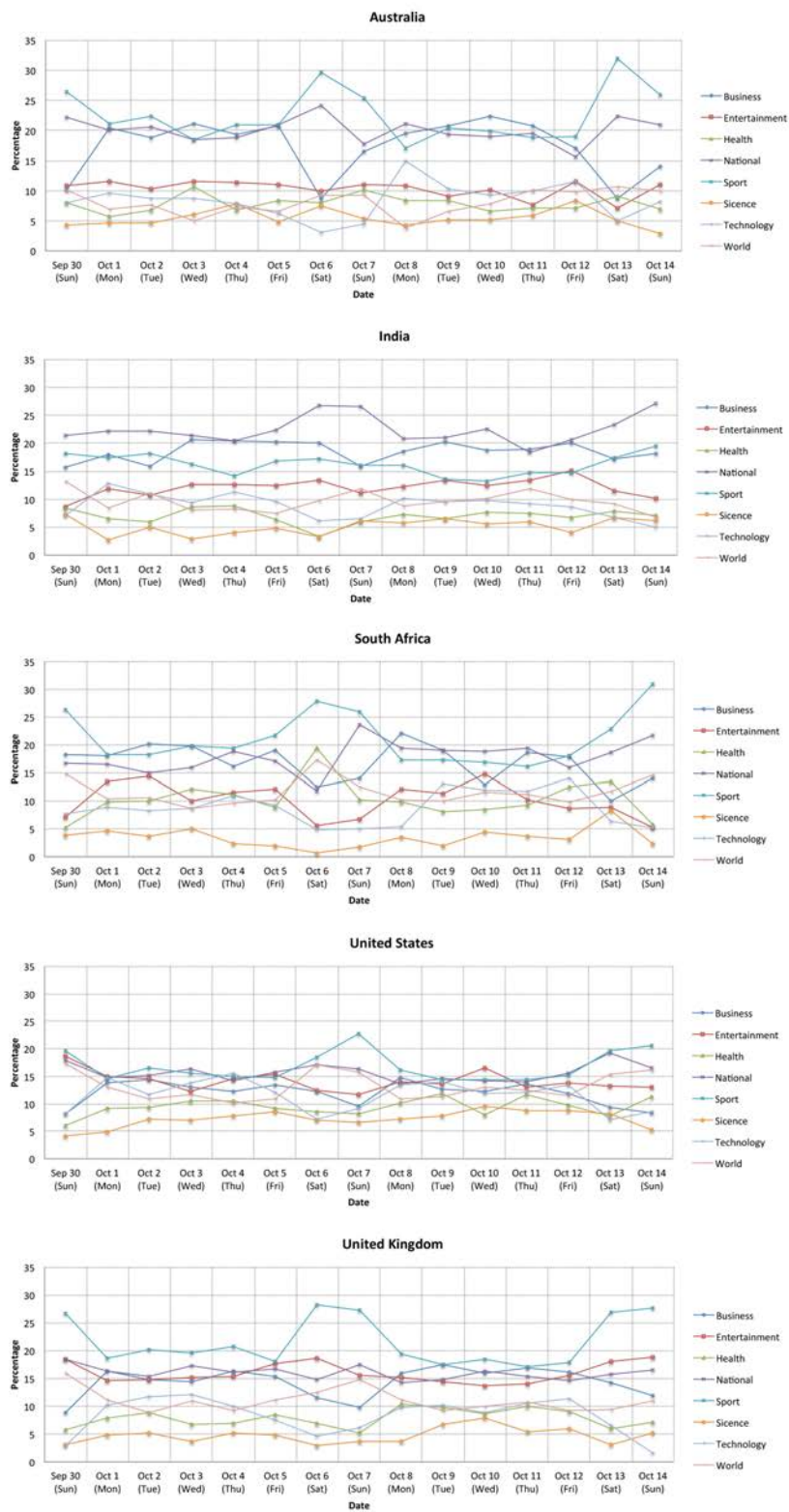


Figure 4.3: Distributions of Headlines Collected from 09/30/12 to 10/14/12

4.4.2.1 Query Generation

In order to search for tweets relating to a certain news headline, one can use the whole headline's title as query. However, there are some drawbacks for doing so. First, search using a whole headline's title may return all relevant tweets but many related tweets will be missed as they may not contain every word from the title, e.g. tweets generated from a social plugin function from news sites consist only part of the title and the links to the original sources. Therefore searching using a whole headline's title may give low recall.

We improved the possible low recall issue by selecting the top k keywords from different parts of headlines to generate Twitter search queries. To rank words in headlines, we employed the well known TF-IDF ranking [27]. Stop words are removed from text in headlines then the remaining words are stemmed using Porter Stemmer creating tokens. The TF score for each token is calculated by the frequency of token in the document. For the IDF score, we used 5 independently built collections of news corpora, one for each country by crawling news articles in each country using links from Google News. All tokens in the corpora are assigned with an IDF score. Table 4.4 summarizes the collected corpora.

Tokens from headlines that are not in the corpora are assign the max IDF score from the appropriate corpus.

In the next step, we have to select the best value for k for generating queries. In our training set experiments we found that queries with more than 5 keywords from headlines usually give a small number of relevant tweets, similar to queries using the

	U.S.	UK	SA	IN	AU
# of News Articles	8,085	5,509	3,541	5,203	3,457
# Tokens	15,388,689	11,753,475	6,100,888	7,297,167	6,749,786
Max IDF	3.91	3.74	3.55	3.72	3.54

Table 4.4: News Headline Corpus

AU: Australia, *IN*: India, *SA*: South Africa, *U.S.*: United States, and *UK*: United Kingdom

k	2	3	4	5
Avg. # of Returned Tweets	108	128.8	107.3	92.8
Avg. # of Relevant Tweets	90.6	107.5	93.2	92.8
Avg. Relevant Ratio	83.90%	92.85%	93.25%	100.00%

Table 4.5: Search Results with Different Values of k

whole title.

We randomly selected 10 headlines and generated multiple queries of 2, 3, 4, and 5 keywords for searching Twitter. Table 4.5 shows the summary of search results.

We can see that when $k = 5$, all retrieved tweets are relevant, however in comparison with different values of k , the number of returned tweets is much lower. When the value of k is 2 or 3, the relevant ratio and number of returned tweets are acceptable. In our work, we search for tweets dated a week after the date of headlines as allowed by Twitter. We also found that when headlines aged, queries with $k = 3$ reduces the precision. We also try to use the fixed percentage of top words in headlines

Headline	Query
World T20 preview: Australia take on a demoralised South Africa - Firstpost	Australia Africa World T20
Zuma almost tastes sweet victory in leadership race - Times LIVE	Zuma tastes race leadership
Syria violence: Aleppo souk burns as battles rage - BBC News	Syria burns rage battles
Mitt Romney, struggling, makes a new effort to connect - Los Angeles Times	Mitt Romney connect new
Foreign multinationals happily plugged into our energy grid - Sydney Morning Herald	Foreign grid energy plugged

Table 4.6: Query Examples

as queries, e.g. 80% of top tokens. But the performance was not better. Thus $k = 4$ is the optimal empirical choice. Table 4.6 shows some examples of queries generated using our proposed method with $k = 4$.

4.4.2.2 Tweet Search and Retrieval

Once we are able to generate queries from headlines, we can search for tweets using Twitter’s Search API. Since Twitter Search API allows searching for tweets that are less than 7 days old, we started collecting from October 08, 2012 to October 21, 2012 where the tweets collection duration covers all the headlines collected from September 30, 2012 to October 14, 2012. Our automated software ran daily at 23:00 local time to gather tweets for each country in the within-country retrieval process. The process searching and retrieving tweets for a headline kept running until one of the following condition occurs:

- Headline is more than 7 days old
- No new tweets are retrieved on 3 consecutive days.

Figure 4.4 illustrates the number of daily tracked headlines and new collected tweets. Since Twitter Search API allows searching for tweets less than 7 days old,

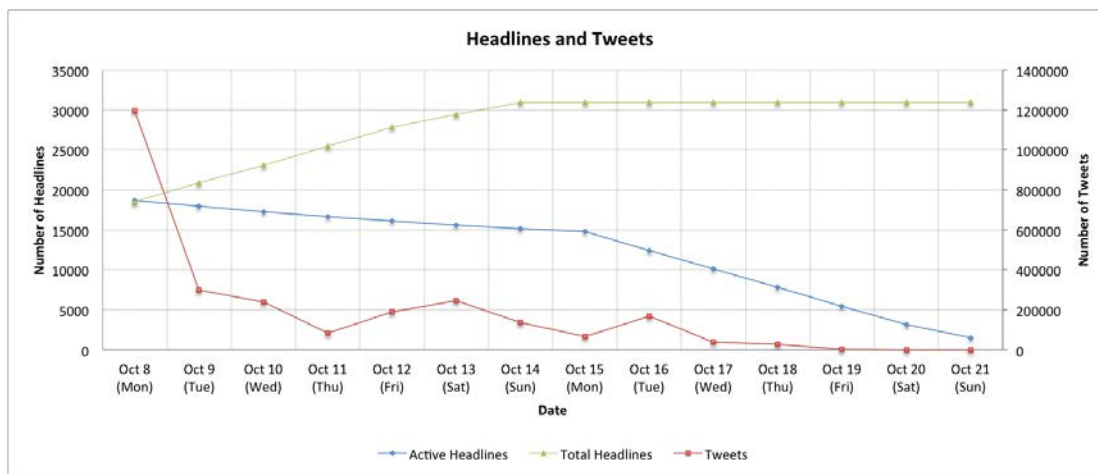


Figure 4.4: Headlines and Tweets from 10/01/12 to 10/21/12

The Number of Tweets Shown on 10/08/2012 is the Total Number of Relevant Tweets for Headlines from 10/01/2012 to 10/08/2012

when we started collecting relevant tweets for headlines from October 8th, we are able to search tweets for headlines from October 1st. The number of tweets shown in October 8th is the total number of relevant tweets for headlines from October 1st to October 8th. The number of active headlines reduced slightly from October 8 to October 15 because tweets collected for a number of headlines meet the stopping criteria. The number of active headlines declined significantly from October 15 to October 21 because the headlines collection process stopped, no new headlines were added. Also many headlines met the stopping criteria. The number of tweets varied from October 8 to October 15 then declined significantly from October 16 to October 21. Figure 4.5 shows the number of headlines collected from October 1st to October 14th and their relevant tweets.

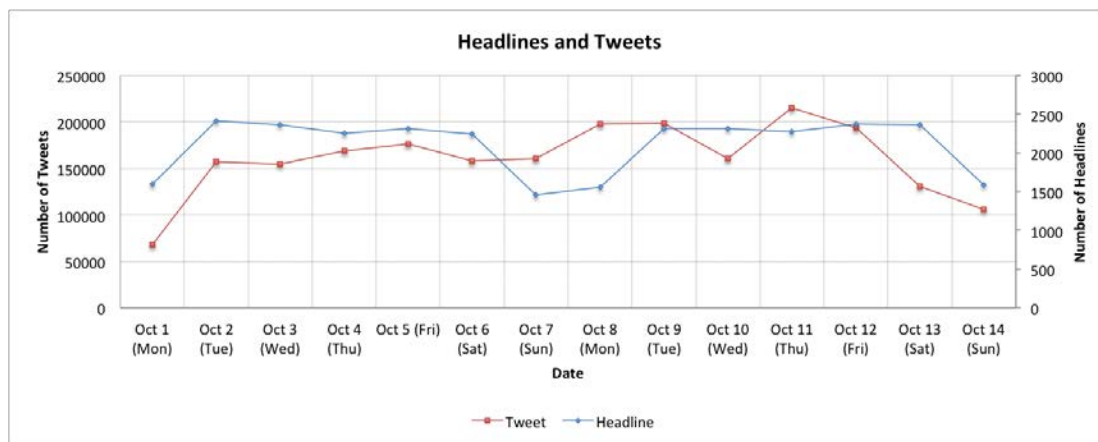


Figure 4.5: Headlines and Tweets from 10/01/12 to 10/14/12

For the cross-country tweets retrieval, we use the following headline selection strategy to select the most popular (tweet frequency) headlines. Each day starting October 8th we select the top 10 headlines receiving the most tweets in the previous 3 days from the within-country tweet retrieval for each country. Then we use these headlines to search for tweets in the remaining countries. We repeat this for each day upto Oct 21. Table 4.7 shows the number of headlines in each country that have tweets from the remaining countries.

In total, we collected 2,552,465 tweets where 2,307,161 tweets were from the within-country tweet retrieval process and 245,304 were from the cross-country tweets retrieval process. As illustrated in Figure 4.6, majority of tweets (approximately 88% of the dataset) are from the U.S. and UK. Tweets from South Africa contribute only about 1% of the dataset. On average, each headline (excluding ones without related tweets) in the within-country collection process has approximately 138 tweets

Country	Number of Headlines
Australia	79
India	78
South Africa	65
United States	98
United Kingdom	94

Table 4.7: Number of Headlines with Relevant Tweets in Cross-country Tweet Retrieval

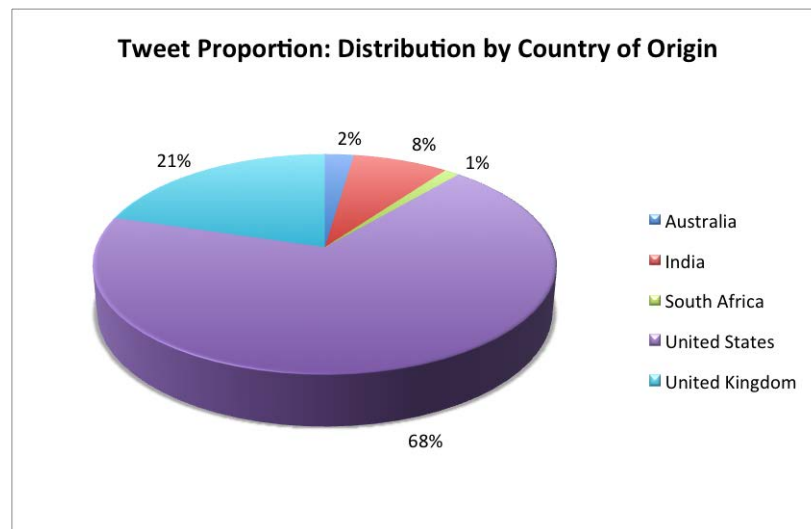


Figure 4.6: Tweet Proportion: Distribution by Country of Origin

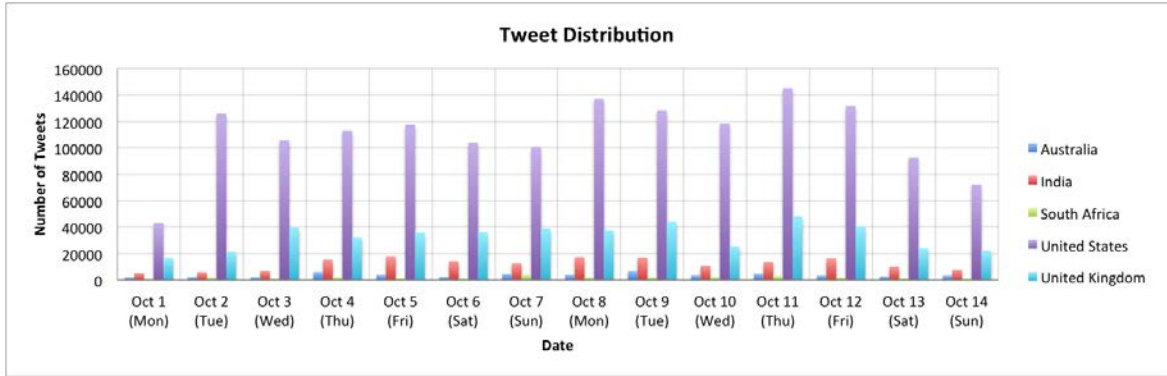


Figure 4.7: Within Country Tweet Retrieval: Distribution over Date and Country

Headline	Category	Country	Retrieval Type	# of Tweets
Barack Obama's team hopes veep debate halts Mitt Romney's momentum - Newsday	National	United States	Within Country	8652
Mitt Romney blasts Barack Obama despite drop in unemployment - Newsday	National	United States	Within Country	7303
Breast Cancer Awareness Month - Mid Valley News	Health	United States	Within Country	7015
Angry Birds Star Wars reportedly on the way - TechRadar UK	Technology	United Kingdom	Cross Country	5870
Apple to host October 23 event, iPad mini expected - IBNLive	Technology	India	Cross Country	5480
Mila Kunis named Sexiest Woman Alive - Channel 24	Entertainment	South Africa	Cross Country	5426
Eat Pink around town for Breast Cancer Awareness Month - Los Angeles Times	Health	United States	Within Country	5416
Obama-Romney's latest issue: Big Bird - USA TODAY	Entertainment	United States	Within Country	5284
Angry Birds does Star Wars - Sydney Morning Herald	Entertainment	Australia	Cross Country	4411
EU wins Nobel Peace Prize: Who, me? - Telegraph.co.uk	World	South Africa	Cross Country	4178

Table 4.8: Headlines with Largest Number of Tweets

($min = 6$, $max = 8652$), each headline in cross-country retrieval has 592 tweets ($min = 1$, $max = 5780$). Table 4.8 shows the headlines with largest numbers tweets in within-country retrieval and cross-country retrieval methods. Figure 4.7 illustrates the distribution of collected tweets over date and country.

4.5 Analysis

4.5.1 Within-Country Analysis

4.5.1.1 Distribution of Tweets Over Headlines

Of 30,974 headlines 16,636 (53.71%) headlines have at least 5 related tweets. Figure 4.8 illustrates the percentage of headlines having at least 5 related tweets for each category in each country. It is very clear to see that most headlines, approximately 80%, in all categories in the United States have related tweets and the percentage of headlines in all categories in South Africa having related tweets is comparatively low. This indicates that there is an active response to a majority (close to 80%) of the United States news independent of category. Interestingly, the most interest generated is in Technology category. South Africa is at the other end in term with the highest Twiter response for National news (35% headlines received at least 5 tweets). For other countries, the percentage of headlines having related tweets varies from one category to another. Headlines in India and United Kingdom have more related tweets than those in Australia.

4.5.1.2 Statistical Analysis

In order to answer our first research question *Do the news readers actually discuss what they read in addition to sharing links?* We perform some statistical analyses on the collected tweets. We assume that if a user's intent is only to share a news headline, she can just simply copy and paste the headline's title and its link to her tweet. Therefore if the words in tweets are a subset of words in the

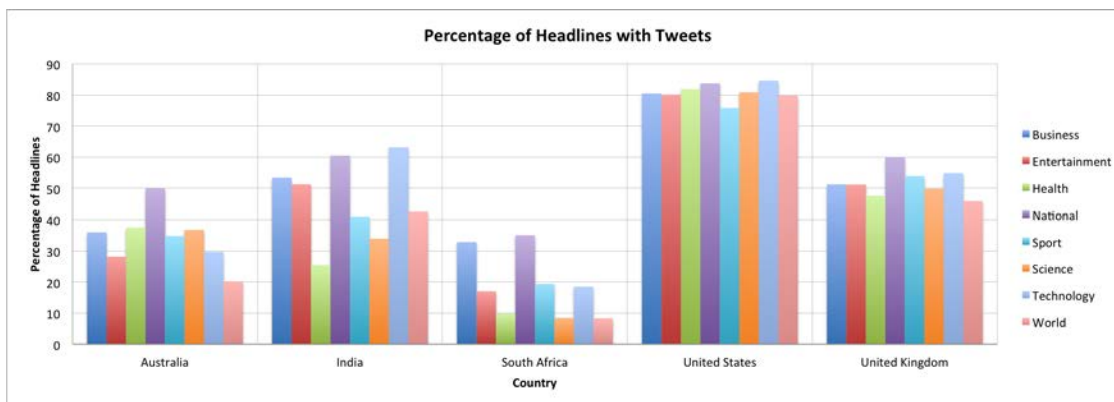


Figure 4.8: Percentage of Headlines with at Least 5 Related Tweets

headline's title, Twitter users only share the headlines, otherwise they are saying something about them. If the tweet contains new and non trivial words then we may infer that they actually say something non trivial about the headline. We removed trivial words, URLs, and Twitter specific syntax like #, @, etc. then count the number of non trivial words different between the remaining text and the headline's title. For example, with the headline *Florida man dies after winning roach-eating contest* and the tweet *Yet another reason to avoid Florida: Man dies after live roach-eating contest in Fla. <http://t.co/JylIJlh0>*, the different non trivial words are *another, reason, avoid, live, Fla.* Figure 4.9 shows the percentage of tweets containing URL, Retweet, Usermentions, and having additional content, containing more than one non trivial word different from the headline's title, for each country. We can see from the data that over 60% of tweets in all countries contain URLs and approximately 90% of tweets from all countries have additional content in comparison with the related headlines' titles. Table 4.10 shows some examples of additional content from tweets.

Headline	<i>More Suspicious Voter Forms Are Found - New York Times</i>
Tweet	<i>More Suspicious *GOP* Voter Forms Are Found http://t.co/ErrYYXu4 THEY CAN ONLY WIN IF THEY CHEAT!!! WE MUST STOP THESE SOCIOPATHS.</i>
Additional Content	<i>cheat!!!,win,*gop*,stop,sociopaths</i>
Headline	<i>Gigaba "ambushed" by SAA resignations - Fin24</i>
Tweet	<i>RT @user1: RT @user2: RT @user3 Gigaba "ambushed" by SAA resignations http://t.co/1cJM2nhu My suspicion is the board was ambushed and respond ...</i>
Additional Content	<i>respond,suspicion,board,ambushed</i>
Headline	<i>West coast rail debacle blamed on Whitehall brain drain - The Guardian</i>
Tweet	<i>RT @user: RT @user: Wrong kind of civil servant? West coast rail debacle blamed on Whitehall brain drain http://t.co/UKgSRQJl</i>
Additional Content	<i>civil,wrong,kind,serverant?</i>
Headline	<i>World Twenty20 2012: West Indies back on top of the world at last after ... - Telegraph.co.uk</i>
Tweet	<i>WEST INDIES WORLD TWENTY20 CHAMPIONS!!!</i>
Additional Content	<i>champions!!!</i>
Headline	<i>Fumble-prone Vick carried a football everywhere - Philadelphia Inquirer</i>
Tweet	<i>#Philadelphia - Fumble-prone Vick carried a football everywhere http://t.co/Vz8YIBJE #Eagles</i>
Additional Content	<i>eagles</i>

Table 4.9: Examples of Tweets with Additional Content

We can see that the additional content in the examples shows some forms of reactions of Twitter users to the news headline. In table 4.10 we present the top 50 non trivial words from tweets that have at least 5 different non trivial words from their related headlines' titles. It's obvious to see some words like *obama*, *romney*, *president*, and *debate* in the list as the events in which two presidential candidates *Barack Obama* and *Mitt Romney* had few debates for their campaigns are big events during the data collection period in October. Also we can see that words describing people's mood like *love*, *great*, and *good* in the list.

Figure 4.10 shows the percentage of tweets with additional content with different number of different non trivial words between tweets and their related headlines. Obviously the percentage of tweets with additional content reduces when the number of different words increases. However even with the number of different words of 5, the percentage of tweets are still greater than 70%. Given the limitation of 140 characters of tweets, 5 words can account for a reasonable portion of a tweet. The statistic from our data shows that Twitter users actually mention something about the headlines they tweet. Moreover, this trend remains consistent across countries.

Word	Frequency	Word	Frequency
<i>news</i>	79704	<i>time</i>	21469
<i>new</i>	71258	<i>good</i>	21291
<i>out</i>	62616	<i>final</i>	20794
<i>over</i>	49833	<i>see</i>	20735
<i>first</i>	44821	<i>years</i>	19634
<i>more</i>	42475	<i>people</i>	19189
<i>now</i>	38586	<i>man</i>	18272
<i>one</i>	37158	<i>mini</i>	18071
<i>win</i>	37010	<i>still</i>	17931
<i>today</i>	33193	<i>october</i>	17537
<i>2012</i>	32957	<i>next</i>	17364
<i>against</i>	32576	<i>team</i>	17179
<i>obama</i>	32495	<i>police</i>	16892
<i>romney</i>	29056	<i>two</i>	16709
<i>world</i>	27367	<i>president</i>	16687
<i>live</i>	26506	<i>here</i>	16220
<i>watch</i>	25600	<i>check</i>	15938
<i>back</i>	24362	<i>full</i>	15730
<i>beat</i>	24354	<i>debate</i>	15680
<i>game</i>	24141	<i>won</i>	15227
<i>cup</i>	23989	<i>down</i>	14888
<i>video</i>	23679	<i>former</i>	14594
<i>day</i>	23182	<i>love</i>	14495
<i>last</i>	21751	<i>star</i>	14493
<i>big</i>	21682	<i>great</i>	14484

Table 4.10: Most Frequent Words in Additional Content

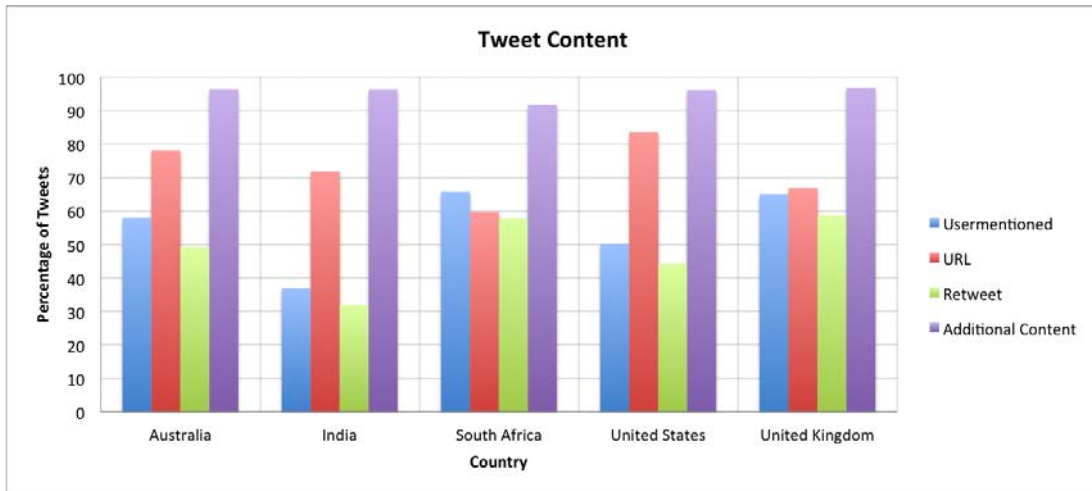


Figure 4.9: Tweet Content

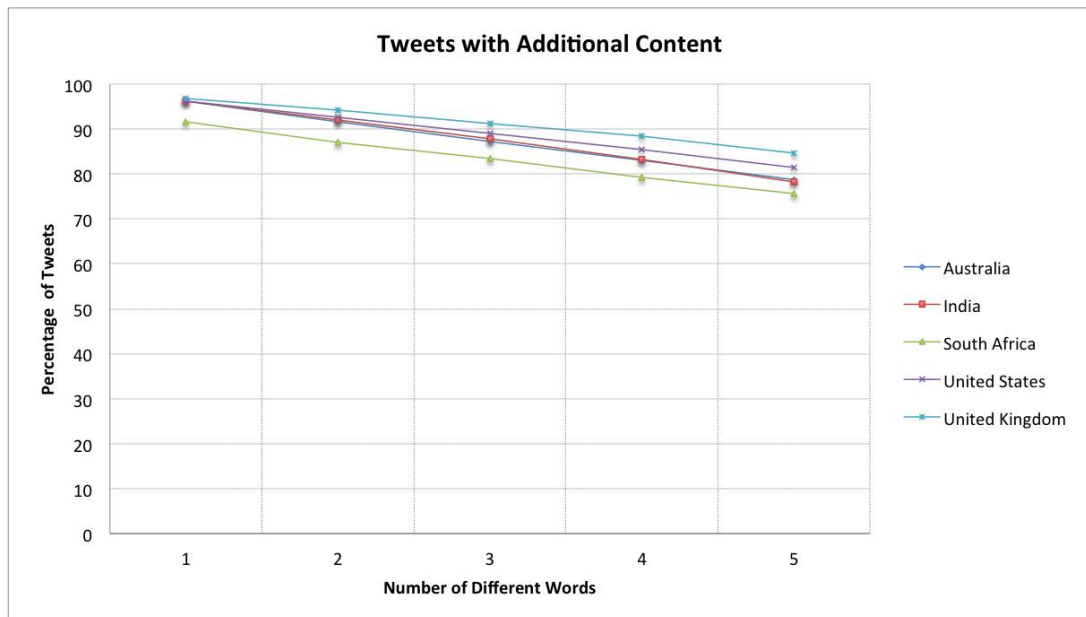


Figure 4.10: Tweets with Additional Content

In order to answer the research question *How do the discussions vary across broad categories of news?*, we examine the volume of tweets for a week from October 8, 2012 to October 14, 2012 for each categories of news from all 5 countries. For each week day, we calculate the percentage of tweets of that day in each category of news, the results are shown in Table 4.11. We visualize these by plotting the trends in Figure 4.11 where vertical axis shows the percentage of tweets, the horizontal axis shows the date, and the each line represents one category of news. For example, on October 9, 8% of all tweets are in business category, 19% of all tweets are in entertainment category, 4% of tweets are in health category, so on and so forth. The horizontal line at 15% divides category lines into 2 groups, the upper group with higher percentage of tweets, and the lower group with the least percentage of tweets. National, entertainment, and sport are the top three categories with the most tweets. The percentage of tweets for each category varies from one day to another. We quantify the comparison of these categories by performing the t -tests. We report the pairs of categories with no significant difference ($p > 0.05$) in the percentage of tweets in Table 4.12. Three categories in the upper group in Figure 4.11 have no significant difference. Also there is no significant difference among the categories in the lower group.

For the question *How do the discussions vary among different categories of news within a geographical location?*, we first plotted the percentage of tweets for each categories of news for each country in Figure 4.12. For each country, the vertical axis represents the percentage of tweets, the horizontal axis represents the date, and

	Oct 8 (Mon)	Oct 9 (Tue)	Oct 10 (Wed)	Oct 11 (Thu)	Oct 12 (Fri)	Oct 13 (Sat)	Oct 14 (Sun)
Business	10	8	14	6	7	5	9
Entertainment	29	19	18	18	15	19	20
Health	8	4	10	9	8	3	5
National	15	25	18	19	16	24	21
Sport	21	16	17	20	17	26	20
Science	3	7	7	6	10	5	5
Technology	9	11	9	13	16	8	7
World	5	11	8	9	11	9	12

Table 4.11: Percentage of Tweets in each News Category

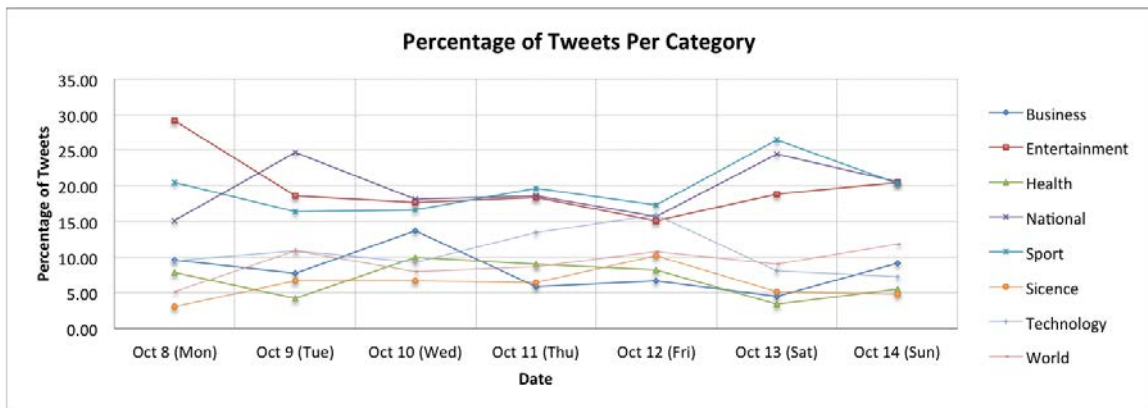


Figure 4.11: Percentage of Tweets in each News Category

	Business	Entertainment	Health	National	Sport	Science	Technology	World
Business								
Entertainment	0.000							
Health	0.397	0.000						
National	0.000	0.961	0.000					
Sport	0.000	0.957	0.000	0.998				
Science	0.181	0.000	0.581	0.000	0.000			
Technology	0.161	0.001	0.029	0.000	0.000	0.010		
World	0.487	0.000	0.093	0.000	0.000	0.025	0.346	1.000

Table 4.12: p-values for Comparison among News Categories (Non Significant Results are in Bold)

the lines represent categories. For example, in the United States on October 11, about 21% of all tweets are in national category, 19% of all tweets are in sport category, 14% of all tweets are in technology category, etc. Then for each country, we perform t -test to compare the difference among categories in term of tweet volume percentage as shown in Table 4.13. Pairs with $p\text{-value} < 0.05$ are significantly different in term of the percentage of tweets. Overall, the percentage of tweets varies from one day to another in all countries. In Australia, there are more tweets mentioning sport and national categories. In fact their $p\text{-values}$ indicate that they are significantly different from other categories. World, science, and technology categories have the least tweets. The t -test results show they are not significantly different from each other but significantly different from the categories with most tweets. There are no significant differences among the remaining categories. In India, the science category has the least percentage of tweets and also it is significantly different from other categories except health one. Except for the health category, the world category is significantly different from the remaining ones. National, entertainment, and sport categories have more percentage of tweets and there is no significant differences among them. It is interesting to see that the technology category is significantly different from the health and science categories. In South Africa, national, sport, and business have most tweets and they are not significantly different from each other but they are different from the rest. Health and science categories have least tweets and they are not different from each other. In the United States, national and entertainment categories are not significantly different and they have most tweets. Their $p\text{-values}$

indicate that they are significantly different from other categories. The sport category lies in the middle of the trend lines, except for the entertainment category, it is significantly different from the remaining categories. The technology category is not different from national and science ones but different from the rest. It is interesting to see that in the United Kingdom, sport is the most popular categories which most of the tweet percentage. Except for the entertainment category it is significantly different from the remaining ones. Science, technology, and health categories are in the group at the bottom with least percentage of tweets.

We quantitatively answered the research question, *What is the difference among the discussions on the same news category or news of reader groups at different geographical locations?*, using the same methodology. First, we plot the percentage of tweets in the same categories from different countries together as shown in Figure 4.13 where for each category of news, the vertical axis shows the percentage of tweets, the horizontal axis shows the date, and lines represents each country. For example, in entertainment category on October 11, about 40% of all tweets in United Kingdom are in this category, 10% of all tweets in India are in this category, 12% of tweets in the United States, 7% of tweets in Australia, and 3% tweets from South Africa are in this category. Then we calculate the *t*-tests to measure the significant differences of tweet percentage among the countries in the same news category. The significant test results are shown in Table 4.14. We can see that for each category of news, the percentage of tweets mentioning that category is very different from one country to another. It is interesting to see that in the business category, South Africa has the

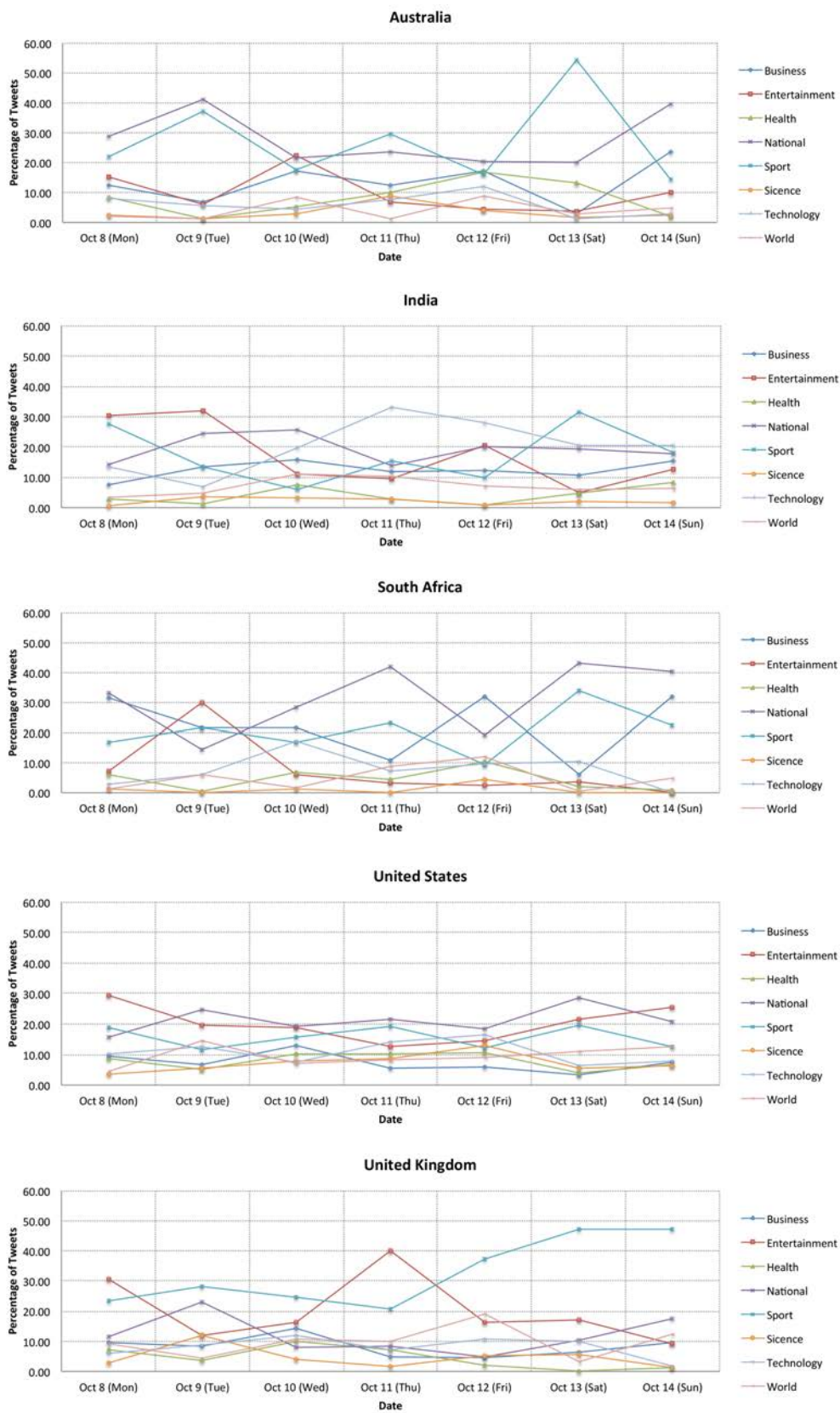


Figure 4.12: Tweets in Different Countries

Australia								
	Business	Entertainment	Health	National	Sport	Science	Technology	World
Business (3)	1.00							
Entertainment (4)	0.37	1.00						
Health (5)	0.16	0.61	1.00					
National (1)	0.01	0.00	0.00	1.00				
Sport (2)	0.05	0.02	0.01	0.93	1.00			
Science (8)	0.01	0.05	0.08	0.00	0.00	1.00		
Technology (6)	0.03	0.21	0.41	0.00	0.01	0.15	1.00	
World (7)	0.01	0.08	0.15	0.00	0.00	0.62	0.36	1.00

India								
	Business	Entertainment	Health	National	Sport	Science	Technology	World
Business (5)	1.00							
Entertainment (4)	0.28	1.00						
Health (7)	0.00	0.02	1.00					
National (2)	0.01	0.65	0.00	1.00				
Sport (3)	0.21	0.99	0.01	0.63	1.00			
Science (8)	0.00	0.01	0.14	0.00	0.00	1.00		
Technology (1)	0.05	0.57	0.00	0.80	0.55	0.00	1.00	
World (6)	0.00	0.04	0.07	0.00	0.02	0.00	0.01	1.00

South Africa								
	Business	Entertainment	Health	National	Sport	Science	Technology	World
Business (2)	1.00							
Entertainment (5)	0.02	1.00						
Health (7)	0.00	0.46	1.00					
National (1)	0.14	0.00	0.00	1.00				
Sport (3)	0.74	0.02	0.00	0.06	1.00			
Science (8)	0.00	0.14	0.05	0.00	0.00	1.00		
Technology (4)	0.01	0.98	0.22	0.00	0.00	0.02	1.00	
World (6)	0.00	0.56	0.78	0.00	0.00	0.05	0.34	1.00

United States								
	Business	Entertainment	Health	National	Sport	Science	Technology	World
Business (7)	1.00							
Entertainment (2)	0.00	1.00						
Health (6)	0.70	0.00	1.00					
National (1)	0.00	0.72	0.00	1.00				
Sport (3)	0.00	0.11	0.00	0.02	1.00			
Science (8)	0.95	0.00	0.65	0.00	0.00	1.00		
Technology (4)	0.09	0.00	0.14	0.00	0.02	0.08	1.00	
World (5)	0.21	0.00	0.33	0.00	0.01	0.19	0.60	1.00

United Kingdom								
	Business	Entertainment	Health	National	Sport	Science	Technology	World
Business (5)	1.00							
Entertainment (2)	0.03	1.00						
Health (8)	0.07	0.01	1.00					
National (3)	0.19	0.12	0.02	1.00				
Sport (1)	0.00	0.06	0.00	0.00	1.00			
Science (7)	0.08	0.01	0.93	0.02	0.00	1.00		
Technology (6)	0.93	0.03	0.09	0.17	0.00	0.10	1.00	
World (4)	0.49	0.05	0.05	0.50	0.00	0.05	0.46	1.00

Table 4.13: p-values for Comparison within each Countries (Significant Results are in Bold)

highest percentage of its tweets while United States has the lowest one. In fact the *t*-test results show that in this category, the percentage of tweets in South Africa is significantly different from ones in United States and United Kingdom. Similarly, percentage of tweets in this category from India is significantly different from ones in the United States and United Kingdom. There is no significant difference from each pair of other countries. In the entertainment category, the United States has the highest percentage of its tweets and South Africa has the lowest percentage. The significant tests show that there are significant differences between United States and South Africa and between United States and Australia. United States and Australia have the same percentage of their tweets mentioning headlines in health categories. India has lowest percentage of its tweets in this category. The *t*-test results confirm that there are significant differences between United States and India and also United States and South Africa. For other pairs, there are no significant differences. South Africa and Australia have highest percentage of their tweets in national categories while United Kingdom has lowest percentage. We also found from *t*-test results that the percentage of tweets of United Kingdom is significantly different from other countries in this category. In the sport category, United Kingdom and Australia have the highest percentage and United State has the lowest percentage. Except for Australia the percentage of tweets in United Kingdom in this category is significantly different from the remaining countries. The United States have highest percentage of its tweets mentioning about headlines in science category while South Africa has very low percentage. The results from significant tests shows that except for United King-

dom, the percentage of tweets from the United States in this category is significantly different from one in the remaining countries, also there is a significant difference among ones from Australia and South Africa. In technology category, India has the highest percentage of tweets while Australia has the lowest one. We also found that the percentage of tweets of India in this category is significantly different from other countries. Also there is a significant different from the percentage of tweets in Australia and one in the United States. In the world category, except for one in United Kingdom, the percentage of tweets from United States in this category is significantly different from ones in other countries. Also there is a significant difference between the percentage of tweets in United Kingdom and the one in Australia.

4.5.2 Cross-Country Analysis

4.5.2.1 Cross-Country Tweet

Table 4.15 presents the results for cross-country tweets retrieval. The row indicates the source country for the headline. The column indicates the country from which the tweet discussing the news originates. Cell values indicate number of tweets. Thus, for example, there were 812 SA tweets that discussed AU news headlines. Diagonal entries are homogenous in that the headline and the tweets are from a single country. Column *Heterogenous* sum all values but homogenous for each country in the same row. As the headlines selected for cross-country tweet retrieval are ones ranked top in the country of origin, we can see that the number of tweets from the country of origin is much larger than those from other countries, for



Figure 4.13: Tweets from Different Countries in the Same Category

Business					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.76	1.00			
SA	0.09	0.05	1.00		
U.S.	0.07	0.01	0.01	1.00	
UK	0.11	0.03	0.01	0.63	1.00

Sport					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.16	1.00			
SA	0.30	0.49	1.00		
U.S.	0.08	0.66	0.16	1.00	
UK	0.45	0.02	0.04	0.01	1.00

Entertainment					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.15	1.00			
SA	0.62	0.10	1.00		
U.S.	0.01	0.54	0.02	1.00	
UK	0.06	0.64	0.05	0.99	1.00

Science					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.26	1.00			
SA	0.06	0.17	1.00		
U.S.	0.03	0.00	0.00	1.00	
UK	0.47	0.12	0.04	0.18	1.00

Health					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.13	1.00			
SA	0.18	0.83	1.00		
U.S.	0.94	0.02	0.07	1.00	
UK	0.19	0.80	0.97	0.07	1.00

Technology					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.00	1.00			
SA	0.51	0.01	1.00		
U.S.	0.03	0.03	0.25	1.00	
UK	0.29	0.01	0.89	0.19	1.00

National					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.05	1.00			
SA	0.52	0.03	1.00		
U.S.	0.11	0.44	0.06	1.00	
UK	0.00	0.03	0.00	0.01	1.00

World					
	AU	IN	SA	U.S.	UK
AU	1.00				
IN	0.11	1.00			
SA	0.70	0.32	1.00		
U.S.	0.01	0.15	0.05	1.00	
UK	0.04	0.24	0.09	0.93	1.00

Table 4.14: p-values for Comparison of Different Countries

in the Same Category (Significant Results are in Bold)

AU: Australia, *IN*: India, *SA*: South Africa, *U.S.*: United States, and *UK*: United Kingdom

	AU	IN	SA	U.S.	UK	Heterogenous
AU	14210	2601	812	17003	7154	27570 (65.9%)
IN	3178	45478	3235	30479	17435	54327 (54.5%)
SA	2759	4712	9517	21465	15623	44559 (82.4%)
U.S.	4381	10637	2903	201885	23990	41911 (17.2%)
UK	7326	13707	5873	50031	105606	76937 (42.1%)

Table 4.15: Cross-Country Tweet Retrieval

AU: Australia, *IN*: India, *SA*: South Africa, *U.S.*: United States, and *UK*: United Kingdom

example for the headlines in the United States, the number of tweets gathered from the United States is almost five times larger than the total number of tweets collected from the remaining countries. Table 4.16 shows an example of cross-country tweet retrieval where we use one headline in the United States and search for tweets in other countries mentioning about the same content with the headline.

4.5.2.2 Sentiment Analysis

In order to measure and compare the reactions to the same news headline. We performed the sentiment analysis on headlines and their relevant tweets collected from both within-country and cross-country processes. For each country, we selected two headlines, one in national categories and the other in sport category, that have relevant tweets from that country and also from the remaining country. Also we selected headlines that potentially cause both agreement and disagreement from their relevant tweets. Table 4.17 lists the selected headlines. For each headline, we randomly sample

Headline (United States)	
Florida man dies after winning roach-eating contest - CNN	
Country	Tweet
Australia	The Miami Herald: "Man collapses, dies after winning roach-eating contest in Broward" http://t.co/riU4lij5 Just to win a python? :(RT @user Winner of roach-eating contest collapses and dies after eating dozens of the live bugs: http://t.co/5AumEiJ9 A real person did this - so that makes it sad. Especially for his family. But it also makes me very ill! http://t.co/E33DK9ne RT @user: Man wins roach-eating contest, dies. http://t.co/Xr8g9hi7 Hate to say I told you so...@user: Winner of roach-eating contest in FL dies after downing dozens of live bugs @amp; worms: http://t.co/SZPkw7Ra
India	#Florida man dies after winning #roach-eating contest http://t.co/qNwecPp5 that's what you get for eating the poor things! @user What a bugging story. RT @user Man eats insects in contest. In order to win a free...snake. Then he dies. http://t.co/RvjL8ug Not funny "@user: Man dies after winning roach-eating contest http://t.co/9aArwEXP " USA man dies after winning cockroach-eating contest http://t.co/EKVNb6qx ganday yukh Man dies after live roach eating contest.....good fuckin day white fools
South Africa	What people will do for fame. Tragic @user: #Florida man dies after winning #roach-eating contest http://t.co/9scVUWIX This dude's a moron—@gt; @user: #Florida man dies after winning #roach-eating contest http://t.co/e5RbqcPR Revenge of the roach: Man #dies after eating cockroaches in a competition held in #Florida #USA in http://t.co/yMCR5zjf "@user: Florida man dies after eating roaches and worms in contest http://t.co/9iXWA2Gq "* wat did he expect?? super powers?*" http://t.co/6YC4JaAR Florida man dies after eating roaches and worms in contest http://t.co/f7XLZBor they call muslims uncivilised! http://t.co/6YC4JaAR
United Kingdom	@user: #Florida man dies after winning #roach-eating contest http://t.co/iKeVk2cd " omg!!!! This is nuts Man dies after winning roach-eating contest http://t.co/wLseNsA9 "We only just met him, but we were all very fond of him". @user: #Florida man dies after winning #roach-eating contest http://t.co/6S0GcBE Irrefutable fact: Americans can always "out gross" This guy lived life to the MAX! @user: #Florida man dies after winning #roach-eating contest http://t.co/am3oGQEV Only in America... RT @user: #Florida man dies after winning #roach-eating contest http://t.co/AW5BFcLo

Table 4.16: Cross-country Tweet Examples

Headline	Country	Category	Within Country Tweets	Cross Country Tweets
1. Gillard's "misogynist" Abbott blast echoes around world - NEWS.com.au	Australia	National	860	25
2. Australia grind it out in second half to win 18-10 over New Zealand in ... - NEWS.com.au	Australia	Sport	106	170
3. "Jub Jub" judgment postponed - Eyewitness News	South Africa	National	635	53
4. Countdown to 2013 Orange African Cup of Nations Begins - AllAfrica.com	South Africa	Sport	201	29
5. Arvind Kejriwal has more "evidence" against Robert Vadra? - Zee News	India	National	1045	215
6. Humour: Tamil Nadu erupts in celebrations after West Indies T20 World Cup victory! - Cricket Country	India	Sport	1571	955
7. Occupy protesters chain themselves to pulpit of St Paul's Cathedral - Scotsman	United Kingdom	National	431	187
8. England 5 San Marino 0: match report - Telegraph.co.uk	United Kingdom	Sport	1488	1302
9. Obama Big Bird ad: a mistake, or shrewd? - Christian Science Monitor	United State	National	2425	627
10. Florida: Winner of Roach-Eating Contest Dies - New York Times	United States	National	1734	361

Table 4.17: Selected Headlines for Sentiment Analysis

25 tweets from the same country with the headline and 25 tweets from other countries.

We manually label headlines and tweets as *positive*, *negative*, or *neutral* by three judges. Table 4.19 summarizes the sentiment annotation and Table 4.18 shows some examples of labeled tweets.

To compare the reactions between within-country tweets and cross-country tweets. We calculate the *agreement ratio* as:

$$\text{agreement ratio} = \frac{\# \text{ of agreed tweets}}{\# \text{ of total tweets}}$$

Tweet	Sentiment
Watching the football Australia VS New Zealand. 18-10 Australia's winning! Woop woop #ausvnz	Positive
RT @solphenduka: The Jub Jub judgment has been postponed to Friday due to rain. Rain ???? Does the court's roof leak ? C'mon. THIS is a joke	Negative
Big Bird - Obama for America TV Ad: http://t.co/xY3FXmKL	Neutral
Well said! RT @chrismurphys: Hartcher SMH; A silly self important fart in global thunderstorm for Gillard PM #auspol http://t.co/dmOxUiOA	Negative
RT @Amul_Coop: West Indies won a World Cup after 33 years. Congratulations to the new T20 Champions! http://t.co/puVYpHUY	Positive

Table 4.18: Sample Tweets with Sentiment Labels

Headline	Sentiment	Country	Within Country Tweets			Cross Country Tweets		
			Positive	Negative	Neutral	Positive	Negative	Neutral
1	Negative	Australia	6	16	3	9	16	0
2	Positive	Australia	22	0	2	21	2	2
3	Negative	South Africa	0	25	0	3	22	0
4	Positive	South Africa	25	0	0	25	0	0
5	Negative	India	0	25	0	0	25	0
6	Positive	India	24	1	0	23	1	1
7	Negative	United Kingdom	1	24	0	0	25	0
8	Neutral	United Kingdom	8	3	14	6	3	16
9	Neutral	United State	1	14	10	2	14	9
10	Negative	United States	1	22	2	0	25	0

Table 4.19: Sentiment Annotation Results

Headline	Sentiment	Country	Agreement Ratio	
			Within Country	Cross Country
1	Negative	Australia	0.64	0.64
2	Positive	Australia	0.88	0.84
3	Negative	South Africa	1.00	0.88
4	Positive	South Africa	1.00	1.00
5	Negative	India	1.00	1.00
6	Positive	India	0.96	0.92
7	Negative	United Kingdom	0.96	1.00
8	Neutral	United Kingdom	0.56	0.64
9	Neutral	United State	0.40	0.36
10	Negative	United States	0.88	1.00

Table 4.20: Agreement Ratios

Table 4.20 shows the *agreement ratios* of within-country tweets and cross-country tweets for each headline. We can see that for *positive* and *negative* headlines, the *agreement ratios* between two type of tweets are very close and pretty high except for the one from Australia *Gillard's 'misogynist' Abbott blast echoes around world - NEWS.com.au* whose within-country tweets have all three types of sentiment and cross-country tweets have high number of both *positive* and *negative* ones. The *neutral* headlines have lower agreement ratios for both within-country and cross-country tweets. For the within-country tweets, we can also see the slightly differences in the *agreement ratios* among the countries. For example, the one for *positive* headlines in Australia is lower than ones from remaining countries. For the *negative* headlines, ones from Australia and the United States are lower than the rest. South Africa has ratios of 1.0 for both *positive* and *negative* headlines. *Agreement Ratios* for both *positive* and *negative* headlines from Australia are lower than the ones in other countries. For cross-country tweets, the *agreement ratio* for the *negative* headline in Australia is the lowest one, for other *negative* headlines, the ratios is high, especially one in India, United Kingdom, and United States with the value of 1.00. Another interesting observation is that the *positive* headline in South Africa has ratios of 1.00 of both within-country and cross-country tweets.

4.5.2.3 Language Model Comparison

Language modeling has been used for long in information retrieval [37] where each document is modeled by a probability distribution. For a certain query, a lan-

guage model assigns a probability of the query being generated by that language model. Relevant rankings of documents for a query is based on the probabilities documents assign to the query. To compare how close two documents are, we can compare their language models. One of many different methods to compare the language models is using Kullback-Leibler Divergence (KLD) [21] which asymmetrically measures the difference between two probability distributions. Two language models are close if the KLD is close to 0. In this research we apply KLD to compare the reactions on the same news headline from different locations. We used cross-country tweets to build language models. For a certain headline from a country, we use all the related tweets from that country to build the source language model. We used tweets collected from remaining countries for the same headline to build target language models, one for each country. Then we compute KLD scores for each pair of countries as shown in Table 4.21.

	AU	IN	SA	U.S.	UK
AU	0	6.73	6.69	6.42	6.36
IN	5.68	0	8.16	4.38	5.22
SA	7.20	8.25	0	7.04	7.31
U.S.	5.55	5.85	5.67	0	4.40
UK	5.60	5.85	7.53	4.98	0

Table 4.21: KL Divergence Score Averaged by Headline

It is obvious that there is a distance between each pair of countries as the KLD

scores are all greater than 0. Here we can compare the distances from a language model of one country to ones of the remaining countries to see how different each pairs is. We can see that the language models generating tweets in the United States and United Kingdom are closest together, $KLD(US||UK) = 4.40$ and $KLD(UK||US) = 4.98$. Language models generating tweets in South Africa and India are much more different, $KLD(SA||IN) = 8.25$ and $KLD(IN||SA) = 8.16$. Also language model for tweets in South Africa is very far from those that generating tweets for other countries. Language model for tweets in Australia seems to have the same distance to language models for tweets in other countries. Language models for tweets from the United States and India have the closer distance to those generating tweets in other countries. The results from language model comparison suggest that there should be a different in the reactions to news headlines between countries with high values of KLD.

4.6 Summary

This chapter presents entirely our work on *Discovering Public Reactions to News Headlines*. We described our data collection process from collecting Google News headlines, queries generation methodology and tweet retrieval from Twitter. Our statistical analyses addressed most of the research questions we raised, namely *Do the news readers actually discuss what they read in addition to sharing links? How do the discussions vary across broad categories of news?, How do the discussions vary among different categories of news within a geographical location?, and What is the*

difference among the discussions on the same news category or news of reader groups at different geographical locations? We found that over 70% of tweets actually discuss something non trivial about the related headlines. Also there are differences in the level of interest of each category of news, for example, entertainment, national, and sport categories have more interest than the other. Additionally, in each category of news, there are significant differences in term of the level of interest among some countries. We also performed sentiment analysis and language model comparison in order to have additional measurements of the differences of interests in and reactions to headlines from different countries. For the future works, we would like to improve the tweets collection process as with the current framework we are able to collect a sufficient number of tweets for five countries but the collection time was so long that will make it difficult to scale up the process to more countries. Also we would like to extend the work on other languages which may let to much more interesting results when we are able to compare the reactions to and interests in news headlines among countries of different languages.

CHAPTER 5 CONCLUSIONS

This thesis discovers entities' behavior through mining Twitter data stream. Specifically, we conduct two exploratory research works, namely *Discovering Target Stakeholders of Firm's Tweets* and *Discovering Public Reactions to News Headlines*, to understand behavior of firm and news readers via their Twitter messages.

In the first work, we seek to uncover the twitter-based stakeholder communication strategy of firms. We designed a framework to do experiments with Fortune 100 companies of 2011. We collect information about companies, e.g. rankings, revenues, profits, industry, Twitter accounts, etc. from public resources. Then we performed content analysis to understand the underlying target stakeholder groups of each tweet. In the next step we developed a metric to recruit and evaluate works from a crowdsourcing service, oDesk, for our data annotation. In the experiment step, we proposed a feature set including tweet-based and company-based features for training classifiers that can automatically predict the target stakeholder groups of a certain tweet. We built classifiers using tweet-based and company-based features separately with base classifiers, namely NaiveBayes, Decision Tree, and Support Vector Machine, and Heuristics. We also developed our methodology to combine results from tweet-based classifiers with company-based features to generate the final data sets that takes into account both company-based and tweet-based features. Classifiers trained using the combined features outperformed ones trained using base and heuristic algorithms with separated feature sets. Experiment results provide answers

to our research questions that firms have communication strategies via their tweets and the strategies vary from one firm to another. In addition firms' characteristics have some predictive power in predicting tweet's stakeholder groups.

In the second work, we investigate how readers from different parts of the world react to news headlines through their Twitter messages. We started by collecting news headlines from Google News for 5 countries: Australia, India, South Africa, United States, and United Kingdom. We develop a method to generate Twitter search queries for each headlines using TF-IDF ranking. Then we use the queries to search Twitter for headlines' related tweets in each country. We also did cross-country tweet collecting where we use headlines from one country to search for tweets in the remaining countries. In the next step, we perform statistical analyses on the collected data to answer our research questions, such as news readers are actually mention something about the news they read in stead of just share the link, and the discussions in Twitter about news do not follow the coverage of news and they vary from one news category to another category and from one country to another country. We also performed sentiment analysis on tweets and headlines as language model comparison between tweets from two different locations on the same headlines in order to have additional measurement of the difference of interests in and reactions to news headlines from different countries.

For the future works, we would like to find a better way to handle the imbalance class problem for the first research as some stakeholder data sets have a very small portion of of tweets in one class and the rest in the other. Also we would like to extend

the data set to a larger number of companies. For the second research, we would like to improve the tweets collection process. Although with the current framework we are able to collect a sufficient number of tweets for five countries the collection time needed make it difficult to scale up the process to more countries. In addition, we would like to extend the work to other languages which may lead to more interesting results.

Besides possible answers to the questions we raise, this thesis makes the following general contributions:

- We presented our mining processes from Twitter data stream collection to crowd sourced data annotation to classification models training strategies to result analysis for understanding firms and news readers.
- We proposed a new mechanism to combine different types of features to train classifiers with significant performance improvement.
- We showed some potential real world applications using our research results.

APPENDIX A SUMMARY OF OTHER COMPLETED WORKS

A.1 Spam Detection in Online Classified Advertisements

Online classified advertisement sites such as Craigslist, Ebay Classifieds, Adsglobe, Adpost, Adoos, ClassifiedsForFree, and Oodle are becoming increasingly popular. According to market researcher Classified Intelligence, the U.S. market for online classified advertisement was \$14.1 billion in 2003, and it has increased quickly since then. Online advertisement sites have attracted a huge number of posts and visits. Craigslist, for instance, receives about 50 million new posts every month¹, and is ranked the 7th most visited site in the U.S. and the 35th most visited site in the world, according to Alexa². Due to its popularity and commercial potential, online classified advertisement domain is a target for spammers. Spammers typically post fake ads on these sites to cheat buyers. For example, many posts on Craigslist offer items with too-good-to-be-true price. Spammers also use techniques such as keyword stuffing to mislead search engines. Spam posts have become one of the biggest issues in the online classified advertisement domain.

Previous approaches for Web spam detection typically use link-based features and content-based features such as n-gram ones to differentiate spam and non-spam pages. However, since online advertisement posts rarely link to each other, link-based features do not help in this particular domain. In terms of content, a key characteristic

¹<http://www.craigslist.org/about/factsheet/>

²<http://www.alexa.com/>

that discriminates spam from non-spam advertisement posts is that the spam posts often contain deceiving information. For instance, a spam advertisement post could attract buyers by asking an unrealistically low price. This characteristic cannot be captured by content-based features. Therefore, traditional approaches for Web spam detection do not work effectively in this domain.

Having identified the problem, in this paper we propose a new approach taking into account the particular characteristics of the online classified advertisement domain. Specifically, we propose a novel set of features particularly designed for this domain. For instance, in order to determine if the asking price for a car is reasonable, we extract various features of the car (e.g., brand name, model, and year) from the advertisement post. We then exploit external resources such as Kelley Blue Book (KBB)³ to get an estimated price for that car and compare it with the asking price.

We demonstrate the effectiveness of our approach via experiments on a dataset containing Craigslist advertisement posts. Compared to the baseline using traditional n-gram features alone, our approach achieves improvements of 59% and 52% in terms of precision and recall, respectively. In terms of F-1 measure, our approach is 55% better than the baseline. Our work was accepted to be published in Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality '11) in conjunction with the 20th International World Wide Web Conference in Hyderabad, India [49].

³<http://www.kbb.com/>

A.2 Belief Surveillance with Twitter

There is a long-standing recognition that social and bio-behavioral scientists and policy makers need accurate and up-to-date information about the broad spectrum of beliefs and opinions voiced in the population (Cummings et al, 2004). As an example, having an accurate estimate of the frequency of people who believe the HPV increases the risk of cervical cancer (Mosavel & El-Shaarawi, 2007) or that using deodorant increases the risk of cancer (Gansler et al., 2006) allows public health scientists to decide whether there is a need to mount a special health campaign to correct those beliefs. Large-scale survey approaches using mail, telephone, and special websites can provide useful data, but such approaches, by definition, having already formulated the content of their questions, do not tap the naturally occurring opinions or beliefs expressed by people. Moreover, typically there are always time-delays between preparation of the survey questions and administration. The development of a methodology that assesses the prevalence of the naturally occurring expression of beliefs and opinions would be of substantial benefit to behavioral scientists. In this paper, we will present a novel method that captures the content of Twitter messages in situ to assess agreement and disagreement (or support and opposition) and even doubt concerning a series of beliefs about (sometimes controversial) causes of illnesses and their treatments. Our methods though demonstrated and tested here in the health care domain are also broadly applicable to beliefs in other domains.

Twitter as a social medium well suits our goals. Tweets, when meaningful, tend to be pithy and to the point. However, Twitter also offers its own challenges

such as the presence of highly abbreviated language including spelling variations, an abundance of posts that are fairly low in information content and the presence of spam. Despite these vagaries, patterns and trends aggregated from Twitter can be meaningful. This is indeed the rationale behind several studies and implementations involving this social media [17, 23, 16]. There are several novel aspects in our research. First we examine beliefs using concise statements as probes. A statement represents a particular hypothesis or idea in the form of a binary (directional) relationship between two concepts. We explore two types of belief statements; those related to *causes* of ill health and those related to their *treatment*. These two basic categories of information are typically sought by individuals touched in some way by disease or ill health. Quite naturally these also form the basis of conversations in social media such as to raise awareness of key medical developments. Another innovative angle is that we are interested in exploring beliefs regarding factual, fictional and debated notions. Using measures of belief, disbelief and doubt that we propose, we compare public attitude towards our factual, fictional and debatable probe statements. We also create a novel dataset where tweets are human annotated both for relevance to our probe statements and also the position taken (support, oppose, doubt etc.). Finally a significant portion of our research is to see if we can use off-the-shelf tools to develop automatic classifiers that successfully replicate the annotations made by the human assessors. We keep our classification strategy intentionally general so that tweets about new probe statements (ideas) may be analyzed automatically. We obtain several interesting results. For example, public belief in our debatable statements (0.63) though

lower than in our true statements (0.83) is still quite high. Simultaneously, disbelief in fictional statements is quite low (0.27) compared to belief in such statements (0.45). Finally, in line with our motivations we propose methods for discovering beliefs in Twitter data beyond our probes. For example, there is significant discussion on skin products causing aging and milk causing osteoporosis, both statements are fictional. Our work was accepted to be published in Proceedings of the ACM Web Science 2012 [5].

A.3 Discovering Health Beliefs in Twitter

Social networking websites and social media are an integral part of our daily life now-a-days. Our views and opinions on a specific topic or the world in general are largely molded by not only traditional information sources (e.g. news, literature, etc.) but also by social media (e.g. Twitter, Facebook, blogs, etc.). Recent survey shows that almost 13% of online adults use Twitter⁴, which generates over 1 billion tweets⁵ per week from over 500 million users around the globe⁶. Online presence of individuals may be active or passive, where one can contribute to and/or seek information from various web sources. In the United States, 74% of adults use the internet with 61% of them looking online for health information and 6% of users posting health-related information on the internet⁷. Hence the use of social media for tracking and using

⁴<http://bit.ly/mwmzOp> (links to PewInternet.org)

⁵<http://blog.twitter.com/2011/03/numbers.html>

⁶<http://twopcharts.com/twitter500million.php>

⁷<http://bit.ly/3b8Np4> (links to PewInternet.org)

health information is as important as traditional approaches for tapping into various biomedical issues.

There is a long-standing recognition that social and bio-behavioral scientists and policy makers need accurate and up-to-date information about the broad spectrum of beliefs and opinions voiced in the population [8]. As an example, having an accurate estimate of the frequency of people who believe the HPV increases the risk of cervical cancer [33] or that using deodorant increases the risk of cancer [13] allows public health scientists to decide whether there is a need to mount a special health campaign to correct those beliefs. Large-scale survey approaches using mail, telephone, and special websites can provide useful data, but such approaches, by definition, having already formulated the content of their questions, do not tap the naturally occurring opinions or beliefs expressed by people. Moreover, typically there are always time-delays between preparation of the survey questions and administration. The development of a methodology that assesses the prevalence of the naturally occurring expression of beliefs and opinions would be immensely useful. Motivated by this, we recently proposed, in a research note, the novel function of belief surveillance and demonstrated how this could be done using Twitter [5].

The surveillance methods we proposed involve specific propositions that we call probes. A probe is a statement presenting a directed, binary relationship between two key concepts. An example is *smoking causes cancer*. In our prior work we studied belief surveillance for 32 probes and showed, for example, that although factual probes (e.g. *smoking causes cancer*) generally garner high degree of belief, there is still

considerable doubt regarding some false probes (e.g. honey treats allergies). Quite alarmingly, we find several debatable (e.g. Actos causes bladder cancer) and false statements also generate high level of belief among Twitter users.

Our prior work was mostly limited to the belief analysis of manually selected probe statements. We did not fully explore an automatic approach for identifying health beliefs in Twitter. In this paper, we extend our prior work with analysis of new beliefs for probes mined automatically from Twitter using two data-driven approaches. Automation is necessary to be able to scale our methodology to handle surveillance of beliefs as they arise. In summary, we ask the following new questions in this paper.

- What kinds of health beliefs are revealed by the naturally occurring discussions on Twitter? In particular we mine beliefs related to a set of health-hashtags and also a set 30 diseases and 20 drugs. Thus we are able to ask: What is the public perceptions on a health belief X ? What are the public perceptions of known effects (or side-effects) of drugs? What beliefs are observed regarding cures of diseases using prescription and OTC drugs? For this, we extend our earlier methods for mining new beliefs from Twitter.
- Which health beliefs are most prevalent in Twitter conversations? Thus we will be able to determine if the discovered beliefs are more or less common in this population.

Our work was accepted to be published at AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text [4].

APPENDIX B TWEETS ANNOTATING GUIDELINES

B.1 Introduction

The purpose of this task is to identify one or more groups of audience a tweet is targeting. The definition and examples for each group of target audience is illustrated in the table below.

Given a tweet, your duty is to label it with **one or more groups** (*Consumer, Investor, Employee, Government, and Community*). If you are not sure which group to label the tweet, you can select *Unsure* option. You are required to type in the brief information why you decide to pick your choices.

B.2 Annotating Tweets Online

There are total 500 tweets need to be labeled via our online labeling system. You will be given an account to login to the system at <http://biz.hawkir.info>

Once you logged in, a screen that lists the companies, each with the total number of tweets and the unlabeled ones will appear like the figure below.

Click on the company you want to label its tweets, a screen with all the tweets for that company will appear.

Labeled tweets are marked with the green signal [DONE], unlabeled ones are marked with the red signal [NOT LABELED]. Click on any tweet to label it, the screen for labeling will appear. Select the appropriate checkboxes that reflex your choice, type in why you select them and click on Save button to finish labeling that

Audience Class	Description	Example
Consumer	Tweets addressed to the Consumers usually provide information about new products, support for products sold, promotional campaigns, new locations (offices/stores) etc. Tweets in this class sometimes mention company's customer-focused news/events/reports and include responses to the customers regarding their complains/comments.	<ul style="list-style-type: none"> • Self-service printing from Android smartphones arrives at FedEx Office. Learn more: • deals: Amazing deal - only \$19 for a print, scan and copy printer by HP DJ-1051 • Recall is on Great Value steamable mixed vegetables & sweet peas contact Pictsweet Co. @ 1-800-367-7412 x417, or return for full refund. • HP customers turning to Dell. 79% of respondents in IDG Research report indicate considering Dell for PCs • Sorry to hear that. Make sure to reach out to should you have any more problems.
Investor	Tweets targeted to the investors of a company usually mention company's business plan like opening of new offices/stores, new product development projects, executive hiring, and merger/acquisition. Some investor-focused tweets describe company's performance like the revenue, stock price, market share, and advantages over the competitors as well as the state of the industry.	<ul style="list-style-type: none"> • Walmart Reports Second Quarter EPS of \$0.97, Ahead of First Call Consensus; Raises Full-Year EPS Guidance • Total production for Angola Block 15 has exceeded 1 billion barrels. • Lockheed Martin gets \$107 million contract • We're excited to open our new 33,000 sq. ft. Mission Support Center today in Clinton, MS! • Dell plans to expand Silicon Valley staff for R&D
Employee	Tweets in this category allow for information sharing and communication within the company. Such tweets may cover company-related updates (from management), collaboration between employees to support customers, as well as employee appreciation messages.	<ul style="list-style-type: none"> • can you help? RT : - can you tell me what the UK Address is for printer toner recycling please • Congratulations to CIO Rob Carter, named to Fantasy Executive League! • Good luck on today's Jeopardy Tournament of Champions semi finals! • Become the "CEO of You" to build your personal brand: In a career spent managing corporate reputation, I've learned... • PICS: Reunion. Now off to dinner!
Government	Tweets addressed to the Government (Agencies) usually cover issues such as jobs, taxes, and security. They may mention how government's policies have an effect on the company's business and how the company's business provides values to the nation.	<ul style="list-style-type: none"> • NYT advocates for short-term political gain on tax issue, misses long-term economic gain for U.S. Read more on our blog • Our Kearl project also favors energy security: Canada supplies ~20% of US oil imports & holds the world's largest reserves of oil sands. • As MT Governor Schweitzer said about our project to move equipment thru his state to Canada, "it's jobs, jobs, jobs." • Report: firms operating in the paid \$41 mil. in state taxes & \$35.6 mil. in local in 2009 • Our CEO talked about how fixing education can help fix other major challenges like our economy & global competitiveness
Community	Tweets in this category provide information about company's activities or company-sponsored community activities to support the environment, education, health, children, and charity/philanthropy. Sometimes such tweets mention some news/report on these topics.	<ul style="list-style-type: none"> • 334 teacher fellowships over the past 27 years! More on our 2011 Community Impact Award from • RT : pledged \$1 million to the Japanese Red Cross to assist with the relief and recovery • ExxonMobil Community Summer Jobs Program partners w/ 60 nonprofits & welcomes newest class of Dallas-Fort Worth interns • FedEx Response to Earthquake in Japan • Thanks for spreading the word about our efforts to make food healthier and healthier food more affordable.^LL

Figure B.1: Target Audience

tweet.

Remember to logout when you are not working with the system.



Figure B.2: Login Screen



Figure B.3: Home Screen



Figure B.4: Company Screen

Labeling Tweet ([Home](#) | [Guidelines](#) | [Tweet List](#) | [Logout](#))

Tweet's Details:

Original Text: Chevron CEO John Watson Speaks Before U.S. Congress Today on Gulf of Mexico Incident, America's Energy Future:
<http://bit.ly/9bv6lj>

Plain Text: Chevron CEO John Watson Speaks Before U.S. Congress Today on Gulf of Mexico Incident, America's Energy Future:

Hash Tags:

Retweet:

Users Mentioned:

URLs: <http://bit.ly/9bv6lj>

To whom this tweet is addressed?

Consumer

Investor

Employee

Government

Community

Unsure

Why do you select above option(s)?

Figure B.5: Labeling Screen

REFERENCES

- [1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization, UMAP'11*, pages 1–12, Berlin, Heidelberg, 2011. Springer-Verlag.
- [2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media, LSM '11*, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [3] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10*, pages 36–44, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [4] Sanmitra Bhattacharya, Hung Tran, and Padmini Srinivasan. Discovering health beliefs in twitter, 2012.
- [5] Sanmitra Bhattacharya, Hung Tran, Padmini Srinivasan, and Jerry Suls. Belief surveillance with twitter. In *Proceedings of the ACM Web Science 2012, WebSci 2012*, pages 55–58, New York, NY, USA, 2012. ACM.
- [6] Macdonald Craig, Ounis Iadh, and Soboroff Ian. Overview of trec-2009 blog track. In *In Proceedings of TREC 2009*. NIST, 2010.
- [7] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, pages 115–122, New York, NY, USA, 2010. ACM.
- [8] K Michael Cummings et al. Are smokers adequately informed about the health risks of smoking and medicinal nicotine? *Nicotine Tob Res*, 2004.
- [9] Shuili Du, C.B. Bhattacharya, and Sankar Sen. Maximizing business returns to corporate social responsibility (csr): The role of csr communication. *International Journal of Management Reviews*, 12(1):8–19, 2010.
- [10] Stuart L Esrock and Greg B Leichty. Organization of corporate web pages: Publics and functions. *Public Relations Review*, 26(3):327 – 344, 2000.

- [11] R. Edward Freeman, Andrew C. Wicks, and Bidhan Parmar. Stakeholder theory and the corporate objective revisited. *Organization Science*, 15(3):364–369, May/June 2004.
- [12] R.E. Freeman. *Strategic management: a stakeholder approach*. Pitman series in business and public policy. Pitman, 1984.
- [13] Ted Gansler et al. Sociodemographic determinants of cancer treatment health literacy. *Cancer*, 2005.
- [14] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.
- [15] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.
- [16] Bernardo A Huberman et al. Social networks that matter: Twitter under the microscope. *First Monday*, 2008.
- [17] Akshay Java et al. Why we twitter: understanding microblogging usage and communities. In *Proc. of WebKDD/SNA-KDD*. ACM, 2007.
- [18] Giora Keinan, Avi Sadeh, and Sefi Rosen. Attitudes and reactions to media coverage of terrorist acts. *Journal of Community Psychology*, 31(2):149–165, 2003.
- [19] Sora Kim, Jae-Hee Park, and Emma K. Wertz. Expectation gaps between stakeholders and web-based corporate public relations efforts: Focusing on fortune 500 corporate web sites. *Public Relations Review*, 36(3):215 – 221, 2010.
- [20] Roger D. Klein. Audience reactions to local tv news. *American Behavioral Scientist*, 46(12):1661–1672, 2003.
- [21] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):pp. 79–86, 1951.
- [22] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web, WWW '10*, pages 591–600, New York, NY, USA, 2010. ACM.
- [23] David Laniado et al. Making sense of twitter. In *Proc. of the ISWC '10*, 2010.

- [24] Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN '10*, pages 1–10, New York, NY, USA, 2010. ACM.
- [25] Yeha Lee, Hun-Young Jung, Woosang Song, and Jong-Hyeok Lee. Postech at trec 2009 blog track: Top stories identification. In Ellen M. Voorhees and Lori P. Buckland, editors, *TREC*, volume Special Publication 500-278. National Institute of Standards and Technology (NIST), 2009.
- [26] Kristen Lovejoy, Richard D. Waters, and Gregory D. Saxton. Engaging stakeholders through twitter: How nonprofit organizations are getting more out of 140 characters or less. *Public Relations Review*, 38(2):313 – 318, 2012. <ce:title>Strategically Managing International Communication in the 21st Century</ce:title>.
- [27] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.
- [28] Richard M. C. McCreddie, Craig Macdonald, Iadh Ounis, Jie Peng, and Rodrygo L. T. Santos. University of glasgow at trec 2009: Experiments with terrier. In *TREC*, 2009.
- [29] Yelena Mejova and Padmini Srinivasan. Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter, 2012.
- [30] Yelena Mejova and Padmini Srinivasan. Political speech in social media streams: Youtube comments and twitter posts. In *Proceedings of the 3rd Annual ACM Web Science Conference, WebSci '12*, pages 205–208, New York, NY, USA, 2012. ACM.
- [31] Kent M.L. and Taylor M. Building dialogic relationships through the world wide web. *Public Relations Review*, 24(3):321–334, 1998.
- [32] Mette Morsing and Majken Schultz. Corporate social responsibility communication: stakeholder information, response and involvement strategies. *Business Ethics: A European Review*, 15(4):323–338, 2006.
- [33] Maghboeba Mosavel et al. "I have never heard that one": young girls' knowledge and perception of cervical cancer. *J Health Commun*, 2007.

- [34] Akiko Murakami and Tetsuya Nasukawa. Tweeting about the tsunami?: mining twitter for information on the tohoku earthquake and tsunami. In *Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion*, pages 709–710, New York, NY, USA, 2012. ACM.
- [35] Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. On using the real-time web for news recommendation & discovery. In *Proceedings of the 20th international conference companion on World wide web, WWW '11*, pages 103–104, New York, NY, USA, 2011. ACM.
- [36] Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. Terms of a feather: content-based news recommendation and discovery using twitter. In *Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11*, pages 448–459, Berlin, Heidelberg, 2011. Springer-Verlag.
- [37] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98*, pages 275–281, New York, NY, USA, 1998. ACM.
- [38] Ana-Maria Popescu and Alpa Jain. Understanding the functions of business accounts on twitter. In *Proceedings of the 20th international conference companion on World wide web, WWW '11*, pages 107–108, New York, NY, USA, 2011. ACM.
- [39] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web, WWW '11*, pages 249–252, New York, NY, USA, 2011. ACM.
- [40] Svetlana Rybalko and Trent Seltzer. Dialogic communication in 140 characters or less: How fortune 500 companies engage stakeholders using twitter. *Public Relations Review*, 36(4):336 – 341, 2010.
- [41] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web, WWW '10*, pages 851–860, New York, NY, USA, 2010. ACM.
- [42] Eileen Scholes and David Clutterbuck. Communication with stakeholders: An integrated approach. *Long Range Planning*, 31(2):227 – 238, 1998.
- [43] Aaron Smith and Joanna Brenner. Twitter use 2012. 2012.

- [44] Aaron Smith and Lee Rainie. 8% of online americans use twitter. 2010.
- [45] Brian G. Southwell, Vanessa Boudewyns, Yoori Hwang, and Marco Yzer. Entertainment tonight? the value of informative tv news among u.s. viewers. *Electronic News*, 2(3):123–137, 2008.
- [46] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 841–842, New York, NY, USA, 2010. ACM.
- [47] A. Svendsen. *The Stakeholder Strategy: Profiting from Collaborative Business Relationships*. Berrett-Koehler Publishers, 1998.
- [48] Ke Tao, Fabian Abel, Claudia Hauff, and Geert-Jan Houben. Wistud at trec 2011: Microblog track: Exploiting background knowledge from dbpedia and news articles for search on twitter. In *Working Notes, The Twentieth Text REtrieval Conference (TREC 2011) Proceedings*. NIST, 2012.
- [49] Hung Tran, Thomas Hornbeck, Viet Ha-Thuc, James Cremer, and Padmini Srinivasan. Spam detection in online classified advertisements. In *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*, WebQuality '11, pages 35–41, New York, NY, USA, 2011. ACM.
- [50] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Linking online news and social media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 565–574, New York, NY, USA, 2011. ACM.
- [51] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya. Towards better tv viewing rates: exploiting crowd's media life logs over twitter for tv rating. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, ICUIMC '11, pages 39:1–39:10, New York, NY, USA, 2011. ACM.
- [52] David Wilkinson and Mike Thelwall. Trending twitter topics in english: An international comparison. *J. Am. Soc. Inf. Sci. Technol.*, 63(8):1631–1646, August 2012.
- [53] PAUL WILLIAMS and JULIE DICKINSON. Fear of crime: Read all about it? *British Journal of Criminology*, 33(1):33–56, 1993.

- [54] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11*, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

- [55] Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, and Vwani Roychowdhury. Information resonance on twitter: watching iran. In *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, pages 123–131, New York, NY, USA, 2010. ACM.