ALGORITHMS ON SOCIAL MEDIA DATA AND EMOTION

by

KAMİL TOPAL

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Dissertation Advisor: Gültekin Özsoyoğlu

Department of Electrical Engineering and Computer Science

CASE WESTERN RESERVE UNIVERSITY

MAY 2017

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis/dissertation of

Kamil Topal

candidate for the **Doctor of Philosophy degree** *.

Committee Chair

Dr. Gültekin Özsoyoğlu

Committee Member

Dr. Mehmet Koyutürk

Committee Member

Dr. Andy Podgurski

Committee Member

Dr. Gürkan Bebek

Date of Defense

04/03/2017

*We also certify that written approval has been obtained for any proprietary material contained therein.

To my wife Selda Topal,

and my son Ediz M. Topal.

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to thank my research advisor Dr. Gültekin Özsoyoğlu, for his constant guidance, mentorship, and thoughtful candor in discussion of all things academic. He has introduced the field of social media data analysis to me and helped me a lot through all my obstacles, with patience, wisdom, and kindness.

I must also thank Dr. Mehmet Koyutürk, who technically was not my advisor, but has helped me as much as an advisor would. Aside from research, Koyutürk family (Yalın, Armağan, Günnur) has been our relatives with their support and friendship.

I would like to thank Dr. Gürkan Bebek, who offered me to work on Chapter 6 (Network model disease diffusion). I have learnt a lot from him and his research group.

I would like to thank to Dr. Andy Podgurski for being my defense committee member for his helpful insight and suggestions.

My graduate work has been funded by Ministry of National Education (MEB) of Turkey.

Special thanks to my previous and present lab mates. We all worked a lot and shared great memories.

Finally, my deepest gratitude is for my family. Mainly, my wife, Selda has always been there for me, thank you for your love, patience, persistent support, and encouragement. This dissertation is simply impossible without you. Also, we are both lucky to have Ediz, our baby boy, such an amazing feeling! He became greatest source of our energy of life. I thank to my whole family; my mother, father, and brothers for their

love and support. Despite the distance, they always show their care, understanding and they always keep me on track.

Algorithms on Social Media Data and Emotion

Abstract

by

KAMİL TOPAL

Internet-based social media websites allow people to discuss, comment, or tweet their opinions, experiences, feelings, and ideas on a "topic" of their choice in an interactive way. These comments can be started by an article, a question, an event, or an incident. The comments can be anything from very useful and important to irrelevant, uncivilized, or undesired comments. People can use some of these comments for making decisions when, say, buying products, watching a movie, renting an apartment, choosing a restaurant, or voting in an election. Moreover, social media is a good source of viral contents to obtain "hidden information" in a society. Thus, mining and visualizing data on content sharing sites is an important research topic that is becoming ever more significant as the use of internet spreads across the world.

The purpose of this thesis is to analyze, or mine, social media comments in terms of (i) the behavior of topics and topic shifts in large numbers of (e.g., in thousands) discussion trees, (ii) the emotion levels expressed in comments, and interaction of emotions in comments and topic shifts, and (iii) the use of social media data, more specifically, twitter data, in predicting contagious disease diffusion.

1 Introduction

Internet-based social media platforms allow individuals to discuss, comment, or tweet their opinions, experiences, feelings, and ideas on a "topic" of their choice in an interactive way. These comments are a treasure trove, containing useful information for a huge variety of purposes: people can and do use these comments for making decisions when, say, buying products, watching a movie, renting an apartment, choosing a restaurant, or voting in an election. Moreover, social media is a good source of viral contents to obtain "hidden information" in a society. The purpose of this thesis is to analyze, or mine, social media comments in terms of

- The behavior of topics and topic shifts in large (e.g., in thousands) discussion trees,
- The emotions expressed in comments, and interaction of emotions in comments and topic shifts, and
- The use of social media data, more specifically, twitter data, in predicting contagious disease diffusion.

Thus, mining and visualizing data on content sharing sites is an important research topic that is becoming ever more significant as the use of internet spreads across the world. Chapter 2 summarizes papers relevant to the research presented in Chapters 3 to 6. Next we briefly summarize contributions of this thesis that are fully presented in Chapters 3 to 6. Chapter 7 is conclusions and future research.

1.1 Effects of Emotion and Area on Topic Shifts in Social Media Discussions

While comments in a social media discussion start, usually, around the topic of an article, it is not uncommon for the discussion topic to shift during comments and replies of the "discussion", sometimes drastically and abruptly, other times slightly, away from the original topic. This has been a problem, and, in fact, due to large numbers of unrelated, inflammatory, or uncivilized comments, as well as the large numbers of comments on some popular articles, news websites and blogs have started to eliminate their comment sections [1]. This is an unfortunate action as readers of a discussion, not just the commenters, gather highly useful information by the simple act of reading these, sometimes informed, comments, and form more informed opinions themselves—a significant loss for both readers and those websites and blogs that eliminate comments from their software systems.

We hypothesize in Chapter 3 that there are three causes for topic shifts in comments:

- Emotion levels,
- The specific area of the article (e.g., sports, politics, cancer, etc.), and
- The structure of the "discussion" (comment) "tree".

These three factors collectively play a role on topic shifts in comments; and, understanding their roles in more depth can lead to building better automated comment viewing software systems that help readers sift through large numbers of comments and gather information more effectively.

Motivated by our main hypothesis, Chapter 3 studies the phenomena of topic shifts in article-originated social media comments. We attempt to identify quantitatively the

effects on topic shifts of comments' (i) emotion levels (of various emotion dimensions), (ii) topic areas, and (iii) the structure of the discussion tree. We show that, with a better understanding of the topic shift phenomena in comments, automated systems can easily be built to personalize and cater to the comment-browsing and comment-viewing needs of different users: users can be provided with options in real-time to

a. Selectively view and reply to comments or discussion threads that are "on the topic", or within a range of either the original article or a specific, possibly shifted, comment of interest within the discussion tree,
b. Link and view discussions of interest in temporal order even when they belong to different discussion threads within the discussion tree,
c. Prune the discussion tree in real-time by specifically eliminating those discussion threads that are of no interest to them, or
d. View comments from all over the discussion tree that may have shifted from the original topic in a certain way, such as shifted to a certain "drifted topic".

1.2 Emotion Analysis of IMDb Movie Reviews

It is common knowledge that, usually, moviegoers, called users or reviewers in the rest of this chapter, utilize movie ratings and reviews in selecting their next movie to see/watch. And, sometimes movie reviews and ratings do not help users make the right choices, as evidenced by their emotional feelings after watching the movie.

Chapter 4 analyzes the relationships between (i) movie ratings and emotions, and (ii) movie genres and emotional responses based on the movie reviews. We show that all highly rated movies, which are those with ratings of 7 or above out of 10, have emotion

maps with low or medium emotion levels, and rarely extreme emotion levels. Clustering allows effective genre-specific grouping of movies, meaning genre-specific movies have specific emotion map patterns as identified by the aggregated emotion maps of the groups that they belong to.

In summary, users' reviews and ratings for a movie are strongly tied to their emotions evoked by the movie. Based on this observation, Chapter 4 discusses another possibility: reviewers' movie scores and reviews can be analyzed and enhanced with respect to their emotion content, aggregated and projected onto a movie, resulting in an emotion map or emotion signature for each movie. Moviegoers can then make a decision on which movie to watch next by selecting those movies having emotion maps with desirable emotion map patterns. One can then build automated software tools that (i) analyze movie reviews and ratings, and (ii) provide an emotional signature for a movie as evidenced from the reviews and ratings of that movie. Clearly, once emotion maps for all movies are at hand, one can build a personalized movie recommender system for users: a user can "submit" his/her desired emotion state and, possibly, the desired genre of the movie to watch, and the recommender system selects movies in that genre with "similar" emotion signatures.

1.3 Detecting Subtopics in Social Media Discussions and Emotional Change

Chapter 5 introduces a different point of view on grouping social media comments based on their perspectives on the original topic. Discussions in a large set of comments, say, in thousands, may contain interesting and/or useful discussion threads on subtopics, regardless of whether or not the discussion is started via an article. In chapter 5, our goal

is to identify and investigate features of "discussion threads on subtopics" in social media discussions, in terms of both the characteristics, and the emotional state of subtopic's comments. Our approach is to

- Use an existing topic modeling technique, namely, the Latent Semantic Analysis (LSA) [1], to identify automatically subtopics and related comments in a social media discussion;
- Measure and analyze the emotional "state" of each subtopic discussion, and
- Extract correlations between emotions and subtopic characterizations as identified by the LSA, and, finally;
- Develop and evaluate a model for identifying "emotionally consistent comments" within a given subtopic discussion with respect to the subtopic's emotional signature.

Clearly, such an approach may result in a practical tool where a list of subtopics can be automatically extracted in a discussion with large numbers of comments, users can identify subtopics that are of interest to them, and then, before reading the related comments, have emotionally inconsistent comments eliminated from the selected set of comments.

1.4 Predicting Contagious Disease Diffusion via Network Based Model of Social Media Big Data

Finally in Chapter 6, we improve Elkin's [1] thesis by adding airline activity to predict contagious disease diffusion. We use social media data, namely, Twitter data, to forecast Influenza Like Illnesses (ILI) diffusion. Influenza (flu) outbreaks affect

approximately 5-20% of U.S residents annually [2]. Depending on the severity, hospitals can experience an unexpected rush of patients. Predicting future outbreaks and understanding how they are spreading from location to location can improve patient care provided. Recently, mining social media big data provided the ability to track patterns and trends across the world. While various studies reported disease correlations with social network data, in this study, we are interested in analyzing social media micro-blogs and geographical locations to understand how disease outbreaks spread over geographies, while enhancing forecasting of future disease outbreaks. We use Twitter data as our social media data source, influenza-like illnesses (ILI) as our disease epidemic, and states in the United States as our geographical locations. We present a novel network based model to make predictions about the spread of diseases a week in advance utilizing social media big data. We show that flu related tweets align well with ILI data from the Centers for Disease Control and Prevention (CDC) ($p < 0.049$). We compare this model to earlier approaches that utilized airline traffic, and show that ILI activity estimates of our model were more accurate. Our disease diffusion model yields accurate predictions for upcoming ILI activity ($p < 0.04$), and we predict the diffusion of flu across states based on geographical surroundings at 76% accuracy. Our equations and procedures can be translated to apply to any social media data, other contagious diseases, and geographies to mine large datasets.

2 Background

In this section, we review the literature related to the research presented in Chapters 3 to 6. We organize the discussion of each study in separate sections.

2.1 Topic Shifts

The notion of topic shift has been studied in the field of web community discovery [3] via focused/topical crawlers [4], to identify those web pages (i.e., documents) that "stay on the topic at hand" [5], using an information retrieval model, usually a vector space model [4], that characterizes the topic of each web page and the distance between two pages that specifies the amount of topic shift. This approach is used in many other environments, e.g., O'Hare et al [6] apply sentiment analysis to financial blog corpus and identify topic shifts among documents in that corpus. Liu et al study topic drift on micro blog posts by using Latent Dirichlet Allocation model [7]. Knights et al detect topic drift with Compound Topic Models [8], to see how a topic evolves and changes to a different topic over a specific time. Our study borrows from these studies in that we also use the vector space model. However, we are mostly interested in the causes of topic shift, and add emotion to our model.

Vector space model has numerous advantages over its alternatives; it can extract the knowledge from text itself without using any lexicon, and performs very well measuring similarity between texts [9].

2.2 Emotion Models

In a recent study, Hasan et al [10] build and use a system, called EMOTEX, to extract emotions from Twitter data to label data for training, EMOTEX uses Twitter hashtags, without an effort to annotate data for any form of learning.

Sentiment Analysis, or Opinion Mining, aims to find the polarity [11] of sentiments, and detects their subjectivity [12] via Natural Language Processing techniques. These techniques usually produce two or three labels for documents, e.g., positive, negative, or neutral sentiments, together with a score ranging between two polarities of, say, -1 and 1. There are free, academic, or commercial tools available for sentiment analysis [13] [14]. There are two main models for representing emotions. The Circumplex model [15] characterizes emotions in two dimensions: activation and pleasure with General Inquirer [16] database containing lexicon of emotions, and more than 100 categories and 11,000 words.

The Hourglass model [17], the most recent emotion categorization proposed by Cambria et al [17], uses a more advanced model, and has four independent, but concomitant, dimensions, namely, *pleasantness*, *attention*, *sensitivity*, and *aptitude* (Figure 3.1). Each dimension captures a different type of emotion: (a) *Pleasantness* captures the user's "amusement level" with interaction modalities, (b) *Attention* captures interaction contents, (c) *Sensitivity* captures the comfort level of the user with interaction dynamics, and (d) *Aptitude* captures the user's confidence in interaction benefits. As also seen in Figure 3.1, each of the four dimensions of the Hourglass model have six levels of activation, which collectively characterize the emotional state of an individual. As an

example, *pleasantness* has six different activation levels, namely, *ecstasy* (the most pleasant), *joy*, *security*, *pensiveness*, *sadness*, and *grief* (The least pleasant). To form a dataset, Cambria et al. first create the AffectNet dataset [17] by blending ConceptNet [18] and WordNet-Affect [19] datasets. Then, they apply truncated singular value decomposition on AffectNet, and use dimension reduction on AffectNet by finding the best approximation.

Finally they use the k-means approach to cluster Sentic space to the HourGlass model. SenticNet 3.0 [20] database, which is publicly available, has more than 30,000 words and phrases that are already scored (in the range of [-1, 1]) for all dimensions. A snapshot of the SenticNet database is in Table 3.1. SenticNet database also has polarity scores for each word. In this thesis, we use HourGlass of Emotion Model to identify emotions from the text.

2.3 Finding Subtopics with using Topic Modeling

2.3.1 Non-Probabilistic Topic Models

Each subtopic in this approach is a "latent topic" obtained via Latent Semantic Analysis (LSA). LSA [21] basically uses dimension reduction, namely, Singular Value Decomposition to find latent topics. It takes term-document matrix in the vector space representation. Non-negative Matrix Factorization [22] is an unsupervised learning model to see tem-document matrix into two matrices (dimensionally lower than original matrix), basis of matrix and coefficient of matrix with the property of all matrices have non-negative elements. This property gives part-based decomposition.

2.3.2 Probabilistic Topic Models

Probabilistic Latent Semantic Indexing (pLSI) [23] introduced by Hoffman is built upon Latent Semantic Indexing. pLSI models each document as a small number of mixture model. After, Latent Dirichlet Allocation (LDA) is another unsupervised algorithm to detect topics in probabilistic way. LDA also models each document as a probability distribution over the topics by assigning another probabilities over the words, because a word may occur in different topic with different weights. Hierarchical Topic Modeling [24] models the topics in hierarchically and the number of topics parameter is not needed to figure out the topic model. Pachinko allocation [25] is also a topic model that improves LDA by modeling correlations between topics in addition to the word correlations which constitute topics. We have not implemented this tool for our dataset, but it is a good idea to capture how different the subtopics are.

2.4 Social Media Data and Influenza Like Illnesses Surveillances

Center for Disease Control and Prevention (CDC) maintains one of the most-trustworthy data sources for influenza statistics: the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) [26]. ILINet provides data from a combination of over 2,900 outpatient healthcare providers across the United States. Each week the CDC compiles the number of patients these providers have seen due to ILI and produces a measure of ILI activity for each of the 50 states, the District of Columbia, and New York City (NYC). These levels are reported via FluView [27] (Figure 6.1). Then using the standard deviation of the data, each state is assigned an activity level between 1 and 10, with 1 meaning there is hardly any ILI activity occurring and 10 meaning ILI activity is extremely high.

Social media has proven itself worthy of making accurate predictions about big data. In particular, Twitter (San Francisco, CA) has consistently proven itself to be a popular source for big data analysis among researchers. Over 645 million active Twitter users collectively post an average of 58 million tweets (micro-blogs no more than 140 characters long) per day [28]. Easy access to peoples' thoughts, moods, and opinions allowed researchers to develop prediction algorithms for social patterns and concepts. Earlier studies have utilized data collected from social media to predict many patterns including happiness [29], stock market activity [30] [31], NFL game outcomes [32], public opinion polls [33] and the spread of contagious diseases [34] [35] [36] [37] [38] [39] [40] [41] [42] [43].

Earlier work that analyzes social media data to predict contagious disease activity has taken the form of various approaches. Sadilek *et al.* focused on a single individual's health by examining tweets from the individual's friends, where the individual is mentioned [34]. Hirose and Wang looked at all tweets from the United States as one collective unit, instead of separating the tweets based on specific locations [35]. Lampos and Cristianini (2010) proposed a textual analysis method to predict ILI levels without utilizing any geographic data [40]. Lampos *et al.* (2010) created *Flu Detector* to map twitter data for 49 urban centers and predict ILI levels on a daily basis. However, they have not addressed how disease is spreading or made future predictions [41]. Achrekar *et al.* presented a new architecture to crawl through Twitter and make national-level predictions based on tweets and on an auto-regression model (ARX) [36]. More recently, Broniatowski *et al.* introduced an influenza detection algorithm based on time-series analysis focusing on the United States nationally, as well as NYC only [37]. A geolocation system named *Carmen*

was introduced to garner tweets associated with NYC [38]. They combined global positioning system (GPS) information associated with tweets (which only a small percentage of tweets have), and they used users' public biographic profiles. Vaccination sentiments were analyzed by creating the social network of 40 thousand users to understand spatio-temporal sentiment towards a new vaccine [39]. However, this approach was confined to a limited number of users with GPS locations. Paul and Dredze (2011) mined Twitter for public health information data that correlates with public health metrics and knowledge. In their analysis they provided a fairly coarse geographic analysis, and provided national flu trends [43]. Lamb *et al.* (2013) applied NLP methods to collected Twitter dataset to find more accurately associated tweets to flu. Their results focus on national levels of ILI activity, and did not take geographies into consideration [42].

Previously, the global spread of epidemics without social media predictions have also been studied [44]. Brockmann and Helbing (2013) represented the complex patterns of disease spreading processes with homogeneous wave propagation patterns. Using the 2009 H1N1 influenza pandemic and 2003 SARS epidemic data, they were able to model spread of these diseases all over the world by studying airline traffic patterns rather than geographic distances. They showed that effective distance, a probabilistically motivated distance measure, reliably predicted spatial origin and disease arrival times for these two instances. However, this measure was unsuccessful returning similar predictions within a country.

3 Effects of Emotion and Topic Area on Topic Shifts in Social Media Discussions

3.1 Introduction and Background

Topic shifts in a social media discussion quite commonly take away from the discussion at hand and create problems. In fact, due to large numbers of unrelated, inflammatory, or uncivilized comments, as well as the large numbers of unrelated and/or shifted comments on some popular articles, news websites and blogs have started to eliminate their comment sections [45]. This is an unfortunate action as readers of a discussion, not just the commenters, gather useful information by the simple act of reading these, sometimes informed, comments, and form more informed opinions themselves - a significant loss for both readers and those websites and blogs that eliminate comments from their software systems.

We hypothesize in this study that there are three causes for topic shifts in comments: emotion levels, the specific area of the article (e.g., sports, politics, cancer, etc.), and the structure of the "discussion" (comment) "tree". These three factors collectively play a role on topic shifts in comments; and, understanding their roles in more depth can lead to building better automated comment viewing software systems that help readers sift through large numbers of comments and gather information more effectively.

Motivated by our main hypothesis, in this chapter, we study the phenomena of topic shifts in article-originated social media comments. We attempt to identify quantitatively the effects on topic shifts of comments' (i) emotion levels (of various emotion dimensions),

(ii) topic areas, (iii) the structure of the discussion tree. We show that, with a better understanding of the topic shift phenomena in comments, automated systems can easily be built to personalize and cater to the comment-browsing and comment-viewing needs of different users: users can be provided with options in real-time to (a) selectively view and reply to comments or discussion threads that are "on the topic", or within a range of either the original article or a specific, possibly shifted, comment of interest within the discussion tree, (b) link and view discussions of interest in temporal order even when they belong to different discussion threads within the discussion tree, (c) prune the discussion tree in real-time by specifically eliminating those discussion threads that are of no interest to them, (d) view comments from all over the discussion tree that may have shifted from the original topic in a certain way, such as shifted to a certain "drifted topic", or (e) visualize on-the-topic/shifted comment sets. This chapter only discusses items (b), (d), and (e), as items (a) and (c) are left as future research topics.

For our experimental studies in this chapter, we have collected about 580,000 news article comments on ten topics in different areas (though, we only discuss results of six topics in this chapter), and analyzed the effects of three factors on topic shift: (i) the comment's location within the discussion tree--in terms of both the level and path of the comment within the tree, (ii) comments' emotion dimensions (i.e., sensitivity, aptitude, attention and pleasantness) and the associated emotion levels (e.g., for the sensitivity dimension, the six levels are rage, anger, annoyance, apprehension, fear, and terror), and (iii) the topic area (e.g., sports, politics, or health). We have found that:

- In terms of a comment's location in the discussion tree, the first comment of the discussion tree sets the tone for all of its descendants: if it is on the topic, usually,

the descendant comments in its discussion subtree also stay on the topic. In rare occasions where a descendant comment, say *c*, is off-topic (i.e., has a topic shift), regardless of the location of *c* in the discussion tree, most (~85%) of the descendant comments of *c* also end up having topic shifts of varying degrees.

- The role of emotion on topic shifts, as one would expect, is very significant: different emotion levels in different emotion dimensions cause differing degrees of topic shifts: highly emotional comments (such as those with high sensitivity dimension scores, e.g., *rage* and *terror* emotion levels) shift away from their original topics with very high frequency (around 90% of the time). And, comments with high emotion levels in emotion dimensions *sensitivity* and *aptitude* are associated with higher topic shift frequencies, as compared to comments with high emotion levels of *attention* and *pleasantness* dimensions.
- The role of the topic area on topic shifts is also quite significant: Topic areas such as sports or politics are more prone to higher levels topic shifts in comments (perhaps because they evoke higher levels of emotions on commentators) than other topic areas such as health (also perhaps because they evoke lower levels of emotions on commentators). This leads us to believe that all topic areas can easily be pre-classified as *high*, *medium*, or *low* emotion level provoking topic areas. Automated tools can then be built to help users identify (and take actions such as perhaps not view) comments with certain types of topic shifts, taking into account this classification and other factors.
- Topic shifts can easily be predicted via unsupervised or supervised learning techniques with around 80% accuracy based on the comments' emotion levels.

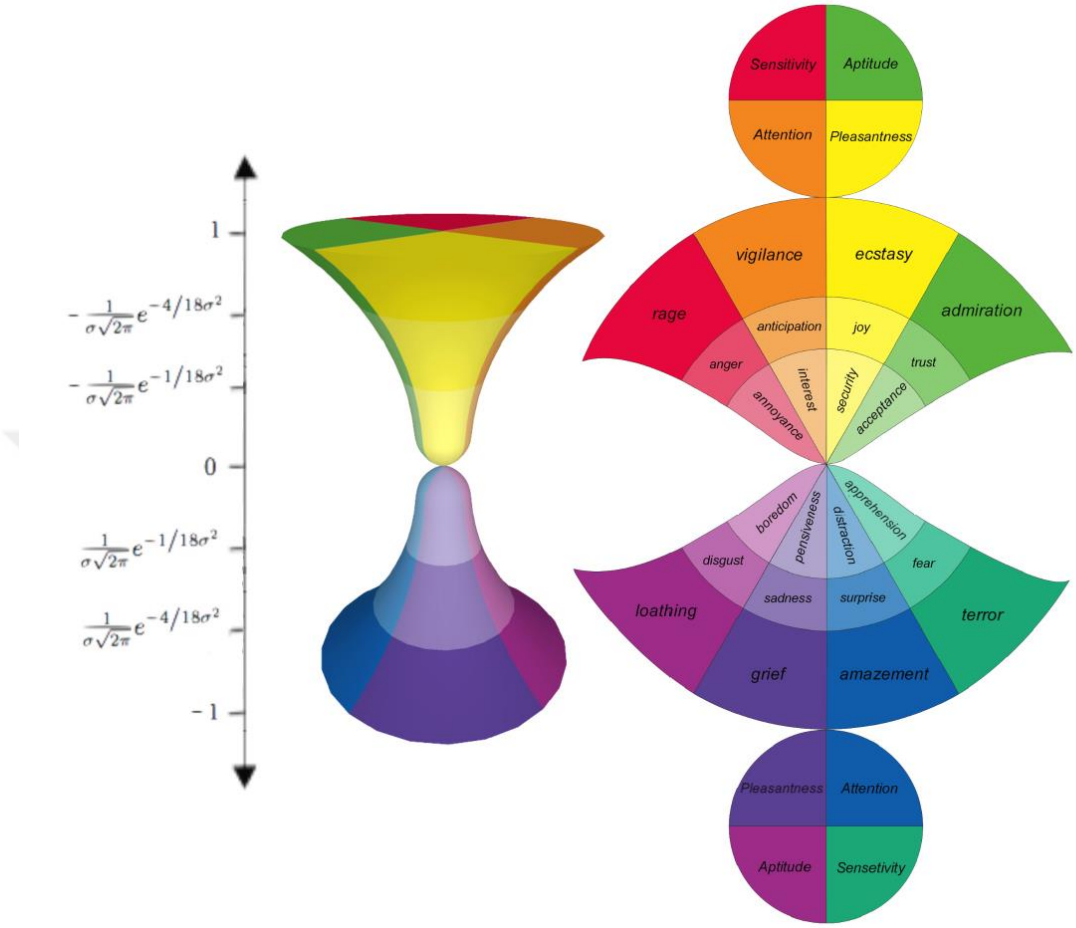

Figure 3.1- The HourGlass of Emotion Model including four concomitant, dimensions, namely, *pleasantness*, *attention*, *sensitivity*, and *aptitude*

In identifying topic shifts between an article/comment and another comment, we take an approach similar to topic shift detection in web community discovery, with a number of provisions, namely, discussion tree structure, revised comment similarity score functions in discussion trees, and emotion dimensions.

In social media discussions, commenters' emotions influence their comments, which in turn cause abrupt or slowly-changing topic shifts from one comment to another. To this

end, there is a need to identify/classify emotions of commenters, and investigate the causal effects of different emotions.

The Hourglass model [17], uses an advanced model, and has four independent, but concomitant, dimensions, namely, *pleasantness*, *attention*, *sensitivity*, and *aptitude* (Figure 3.1). Each dimension captures a different type of emotion: (a) *Pleasantness* captures the user's "amusement level" with interaction modalities, (b) *Attention* captures interaction contents, (c) *Sensitivity* captures the comfort level of the user with interaction dynamics, and (d) *Aptitude* captures the user's confidence in interaction benefits.

As also seen in Figure 3.1, each of the four dimensions of the Hourglass model have six levels of activation, which collectively characterize the emotional state of an individual.

Table 3.1- A snapshot of SenticNet 3.0 database

	Pleasantness	**Attention**	**Sensitivity**	**Aptitude**	**Polarity**
accept	0.972	0	0	0.894	0.622
acceptability	-0.997	0	0.878	-0.957	-0.944
acceptable	0.258	0.357	-0.088	0.351	0.293
acceptableness	-0.997	0	0.878	-0.957	-0.944
acceptance	0	0	0	0.3	0.1
acceptation	0.577	-0.77	0	0.517	0.621
accepted	0.683	-0.553	0	0.552	0.596

In this chapter, we use the Hourglass Model to classify the emotion dimensions and levels of commenters.

3.2 Methods

3.2.1 Discussion Trees

We present an abstract representation of the structure of a social media discussion via its discussion tree. We identify the main characteristics of (article-based and size-free) social media discussions as follows:

a. Each comment is either about the original article, or a reply to another comment/reply.
b. Each comment has a timestamp, indicating the posting time of the comment.
c. Each sequence of comments is either about the original article (represented by the topic t) or another "parent" comment (represented by its own topic).
d. Each comment sequence has a nesting level, which is equal to the number of comments preceding that comment in the reply-chain that contains the comment.

3.2.2 Vector Space Model

We use the vector space representation of the article and comments. The article and each comment are tokenized and stemmed using the Porter stemming algorithm [46]. Word tokenization involves removing characters from words (such as punctuations), and attaching a unique id for each word. Stemming and stop-word removal are applied. Related words are mapped to the same stem by removing their inflections. Stop-words are common in sentences, and add grammatical, but no context, value, and thus, they are not useful to determine keywords for a topic. Some researches do use stop-words, however. There is no

universal stop-word list in the literature. In our study, for stop-word removal, we employ a comprehensive list from the web [47]. Then, we use a modified version of the vector space model [48] as follows. Term Frequency for each document (article story and comments) is calculated using the Cornell SMART system's smoothened version [48]. Let t be a term in document d, where d is a comment or the article. Then the term frequency $TF(t,d)$ is computed as follows.

$$n(t,d) = \text{frequency of term } t \text{ in document } d$$

$$TF(t,d) = \begin{cases} 0, & if\ n(t,d) = 0 \\ 1 + \log(1 + n(t,d)), & otherwise \end{cases} \quad (1)$$

We compute the inverse domain frequency of each term t across all documents, $IDF(t)$, to scale up the effects of terms that occur in many comments or the article.

$$IDF(t) = \log(\frac{1 + |D|}{|D_t|}) \quad (2)$$

Here, D denotes the document collection (in our case, the set of comments and the article), D_t denotes the set of documents containing t, and log() is a dampening function. Note that this analysis is performed independently for each article and its associated comments; i.e., the set D is unique for each article. We then compute the relative frequency $x_d(t)$ of term t in document d as

$$x_d(t) = \frac{TF(t,d)}{IDF(t)} \quad (3)$$

Clearly, Equation 3 is the opposite of the standard approach of $TF() * IDF()$ where rare terms are considered important, and "rewarded" by the $IDF()$ factor. In our case, however, the “universe” of the documents for each article is composed of the article

and its associated comments. Therefore, a term that is frequent in this collection of documents indicates relevance to the article, whereas rare terms signal decrease in importance. In other words, our premise is that important words are not usually rare. For example, if the article is about Ebola virus dissemination, than the relevant comments are more likely to have the terms "Ebola", "hospital", "health", etc.. If we multiply $TF(\,)$ with $IDF(\,)$ scores, then these frequent terms will get smaller weights, which is an undesirable effect. On the other hand, dividing $TF(\,)$ by $IDF(\,)$ assigns more weight to these frequent terms. Furthermore, if a comment includes a term that is rarely used in other comments as well as the article, dividing $TF(\,)$ by $IDF(\,)$ lowers that term's weight. Note that uninformative words are already removed in the stop-word removal stage of analysis.

3.2.3 Similarity Scores

Since we have a vector representation of each comment and the article, we use the cosine similarity [4] (Equation 4) to calculate the topical similarity between the article and each comment. Namely, for a given article a and comment c, we compute the similarity between a and c as

$$C(x_a, x_c) = \frac{x_a x_c}{|x_a||x_c|} = \frac{\sum_{t=1}^{n} x_{a(t)} x_{c(t)}}{\sqrt{\sum_{t=1}^{n} {x_{a(t)}}^2}\sqrt{\sum_{t=1}^{n} {x_{c(t)}}^2}} \tag{4}$$

where x_a and x_c respectively denote the vector space representation of a and c, and n denotes the total number of terms. In choosing cosine similarity, we experimented with jaccard index [49], dice similarity [50], and cosine similarity in a small set of comments, manually judged their performance, and chose cosine similarity since it performed slightly better than the others in our environment.

Once we quantify the similarity between each article and comment, we use a threshold to distinguish between off-the-topic and on-the-topic comments. In order to set the threshold, we use the k-means algorithm [51] to create two different clusters for off-the-topic and on-the topic comment sets, and consider small centroid clusters as having shifted comments.

3.2.4 Emotion Modeling

To represent the emotional landscape of each comment, we use the SenticNet 3.0 database [20] containing a large collection of phrases. For each phrase in the database, there are five different scores (one for each of the four emotion dimensions, and a polarity score) in the range $[-1, 1]$. We map each comment to the SenticNet database by identifying all phrases that match the comment in SenticNet. Since each comment may map to multiple phrases in the database, we then aggregate the scores of each phrase to compute an emotion representation for the comment. For this purpose, for each emotion dimension, we compute the average of the absolute values of the respective dimension score across all matching phrases. This gives us a five-dimensional representation of the emotional landscape of the comment. We average the absolute values of the scores, since we are mainly interested in quantifying the "level" of emotionality of the comment, as opposed to quantifying the polarity of the emotion.

In the Hourglass of Emotion Model [52], there are four dimensions with scores in $[-1,1]$, namely, *pleasantness*, *attention*, *sensitivity*, and *aptitude*, each with six levels of activation that represent six different emotion levels. As an example, there are *grief*, *sadness*, *pensiveness*, *security*, *joy*, and *ecstasy* in the *Pleasantness* dimension. We eliminate polarity by taking absolute values, and, thus, the computed dimension scores are

in the range of 0 to 1, resulting in 3 distinct levels of activations in each dimension (instead of 6), with the emotion levels in each dimension symmetrically combined. E.g., if the pleasantness dimension score is less than -0.66 than it is *grief*; if it is between -0.66 and -0.33 then it is *sadness* and so on. After taking absolute values of all scores, we end up with three combined emotion levels for *Pleasantness*, namely, {*grief, ecstasy*} (which forms the "high" emotion level), {*sadness, joy*} (medium emotion level), and {*pensiveness, security*} (low emotion level).

3.3 Experimental Evaluation and Results

3.3.1 Dataset

Table 3.2- Topics and Number of Comments of our Dataset

News article topic	**# of Articles**	**# of Comments**
2016 U.S. Presidential Election Debates	11	83,604
Gun Laws Discussion	5	18,480
Immigration in U.S.	11	34,481
Supreme Court decision on LGBT Marriage	16	83,487
Media Brawl with politicians and journalist	9	89,243
Hillary Clinton Email controversy	9	93,580
Iran nuclear Deal with U.S.	13	58,388
Stocks	20	16,788
Planned Parenthood and Abortion	19	87,946
Economy	17	15,955
Total	**130**	**581,952**

Our data sets come from news article comments on various topics (Table 3.2). We collected

581,952 comments from 130 news articles in the period of June 2015 to September 2015 from blog comment hosting service Disqus [53] API. We chose Disqus for the following reasons. Disqus

- Provides a service for many major websites such as Politico, CNBC, ABC News, and The Washington Times.
- Allows replies to all comments and replies, which allows for deeper discussion trees shift behavior as a function of the structure tree.
- Allows the creation of discussion trees as JSON-formatted data.

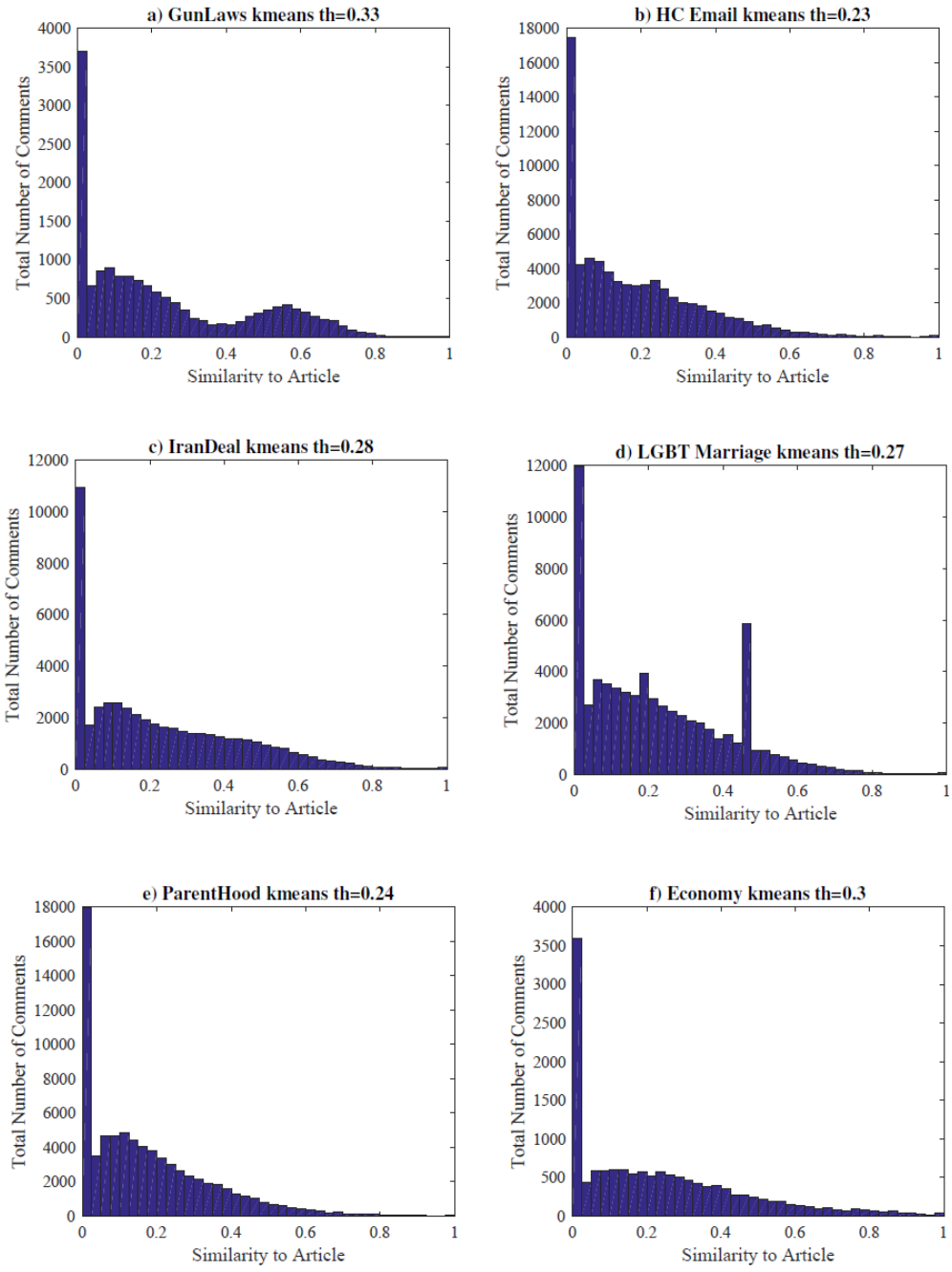

Figure 3.2 - Article-comment similarity distribution for six different topics over 40 equally divided buckets, x-axes represent similarity scores and y axes are the number of comments in a bucket

Our data preprocessing removes all spam comments in each data set. Each discussion tree and the level of each comment is extracted based on the reply relationship. Figure 3.2 lists

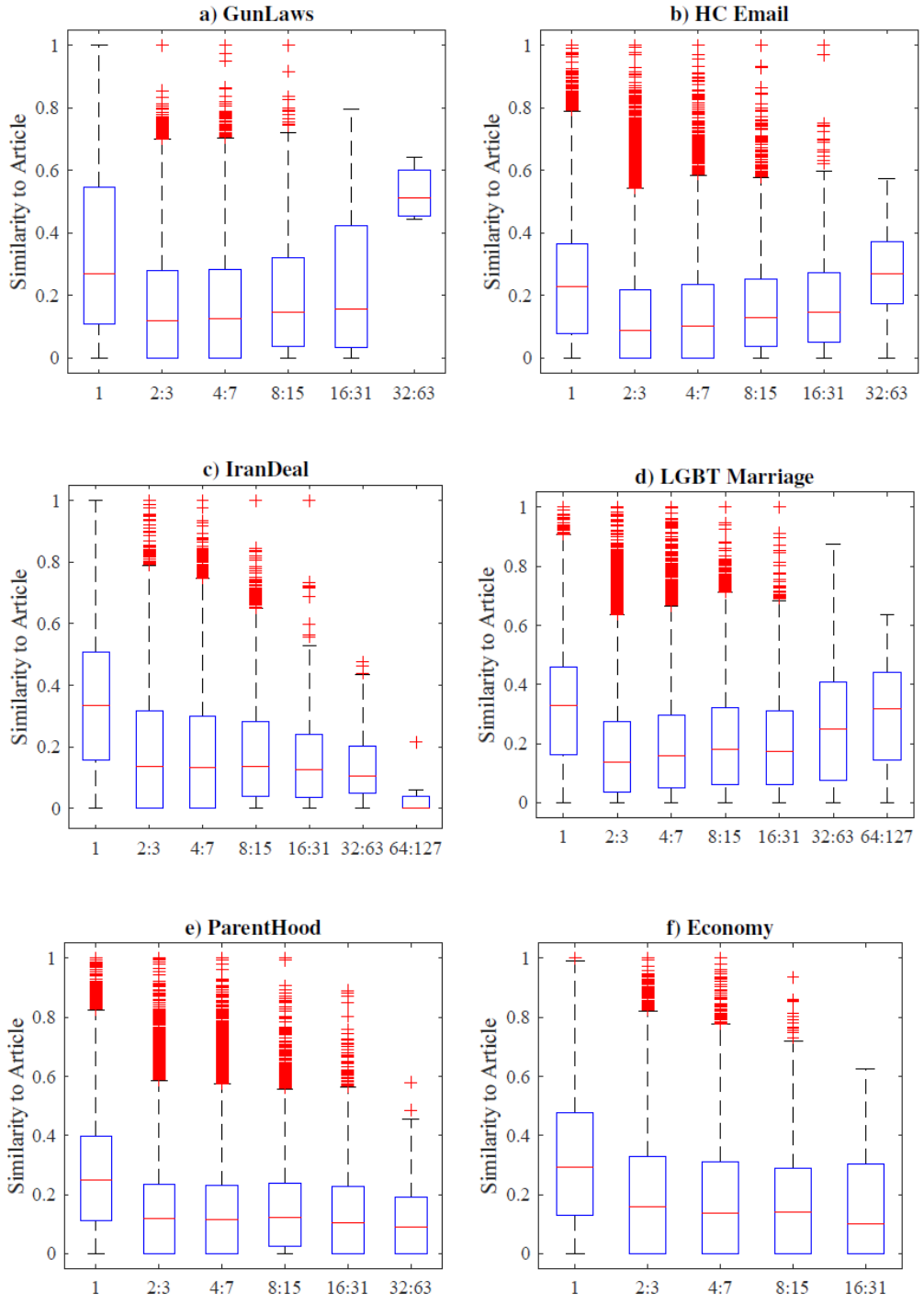

Figure 3.3 - The relationship between the level of an article in a conversation tree and its topical similarity to the article for six different topics. The x-axis shows the level of the comment in its respective conversation tree, where comments are binned logarithmically.

six different topic similarity distributions as histograms, each with 40 buckets, a title, and a threshold value (as determined by the k-means algorithm). Because of space limitations, we here focus on the results related to these six topics and their discussion trees.

3.3.2 Level-Based Analysis of Shifted Comments

Each comment has its level information: level 1 (root) nodes are comments directly on the article, and level 2 nodes are replies to a level 1 comment, and so on. We gather all comments about an article, and group the comments logarithmically according to their levels into buckets. Figure 3.3 (a)-(f) contain *box plots* [54], one for each bucket, summarizing the similarity distribution of comments in each bucket to the article. In each box plot of Figure 3.3 (a)-(f), the X axis represents the levels of comments represented by a bucket, and the Y axis displays five different statistics for similarity scores between comments within that bucket and the article, namely, minimum, first quartile, median, third quartile, and maximum. Note that the X dimension scale grows exponentially since the number of comments at higher level buckets decrease exponentially. Thus logarithmic bucketing provides a more balanced distribution of comments into buckets (but the number comments in a bucket still goes down with increasing number of levels). We have the following observations:

Observation 1. Among discussion tree levels for a topic, the level with the highest average similarity score to the article is level 1, i.e., the root.

That is, root-level comments (per discussion tree) are on the average the most similar comments to the article. Observation 1 is to be expected since topic shifts are likely to occur more frequently as the discussion thread continues.

There are exceptions to Observation 1 within our topics, namely, Gun Laws and Hillary Clinton Email Controversy. In these discussions, if the comment's level goes deep enough in the discussion tree, i.e., after level 32, surprisingly, the average of on-the topic comments exceed their root averages. This shows us that there are still on-the-topic (and, perhaps, useful) comments at deeper levels among the descendants of shifted comments.

Observation 2. Topical similarity to the article is highly variable for comments that are at different levels. Even in buckets that have low similarity to the article on average, there are still significant numbers of useful comments, as represented by their higher similarity scores. At least 25% of all comments in each bucket have similarity scores higher than 0.3, which are classified as on-the-topic comments by the k-means algorithm. This observation suggests that the level of a comment in the discussion tree may not be sufficient to predict the relevance or usefulness of a comment.

3.3.3 Dissimilar versus Similar Trees

We distinguish between "low similarity" and "high similarity" comment sets by employing a similarity threshold α defined as the maximum similarity score of the low similarity comment set.

Definition-Dissimilar/Similar Discussion Tree. Let $\beta_i$be the average similarity of all the comments in the discussion tree $T_i, 1 \leq i \leq n$. Then T_i, is a similar tree (to the article) if $\beta_i > \alpha$; otherwise it is a dissimilar tree.

In Figure 3.4, we dissect the analysis of Figure 3 into dissimilar and similar trees, by plotting the distribution of similarity to the article separately for comments in dissimilar trees and those in similar trees. For this purpose, we use the logarithmic bucketing of

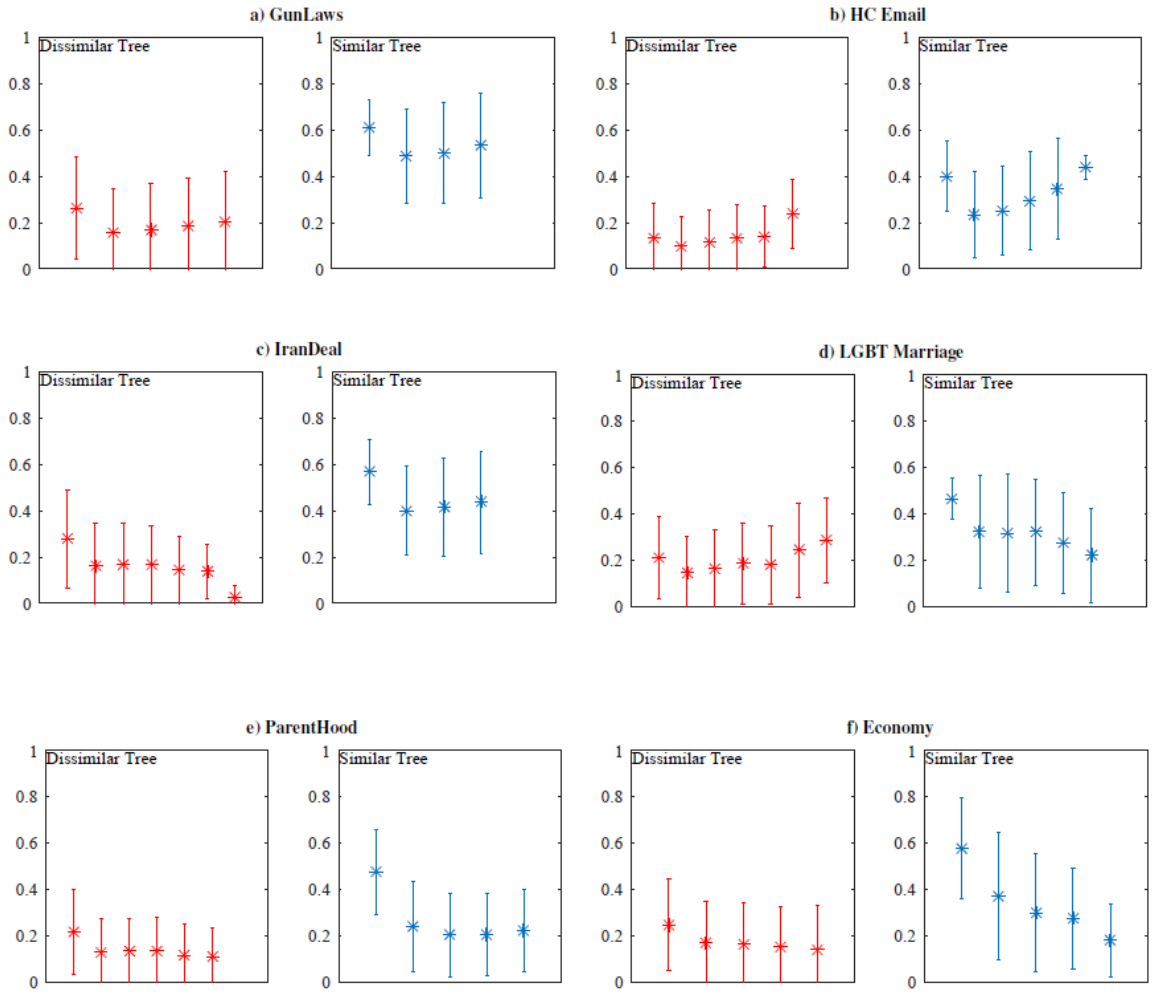

Figure 3.4 - The relationship between comment level and topical similarity to the article for Dissimilar versus Similar Trees. The x-axis shows the level of the comment in its respective decision tree, where comments are binned logarithmically.

Figure 3.4 (labels omitted for readability), and box plot similarity scores of similar/dissimilar trees per bucket, as error bars [55] for dissimilar and similar discussion trees. The purpose of this analysis is to understand whether trees are uniform in terms of the "topic shift" behavior of the comments they contain.

Observation 3. Similar trees have more on-the-topic comments even at deeper levels, compared to dissimilar trees. For all topics, similar trees have on-the-topic root comments

65% to 85% of the time. In comparison, these values decrease at least 20% for dissimilar trees.

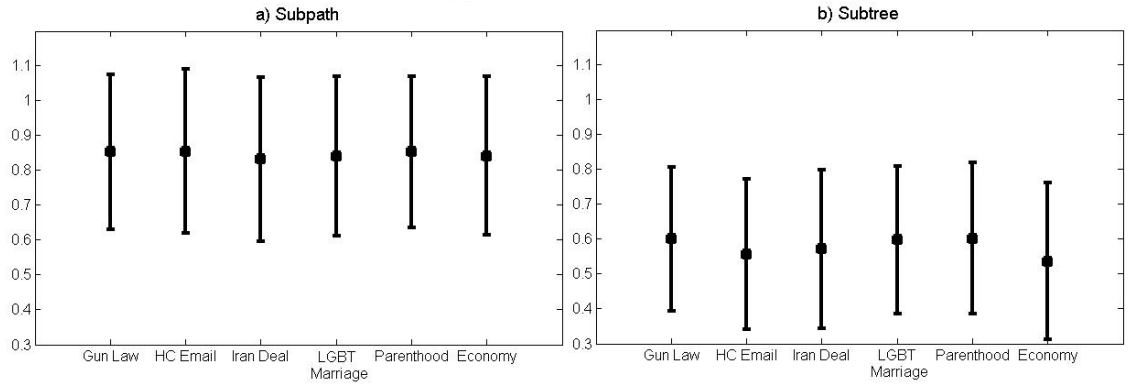

Figure 3.5 - Fraction of shifted comments that come after the first shifted comment in a) path b) subtree

Observation 4. In each discussion tree, the root comment for each article sets the tone, and mostly decides as to how the following discussions evolve. If a root stays on the topic, then the following comments usually also stay on the topic (i.e., result in a similar tree), at least 20% more frequently than those for a dissimilar tree. In comparison, for a dissimilar tree, there is up to 50% more shifted root comments, and this leads to, on average, a 40% higher topic shift.

3.3.4 Shift Behavior Within Paths and Subtrees of Discussion Trees

Motivated by the observation that the root comment appears to set the tone, we further investigate whether there could be deeper comments that initiate a topic shift. For this purpose, we extract each path, say, path *p*, in a discussion tree, and locate the first shifted comment, say, comment *c*, in *p*. We then divide the number of shifted comments after *c* in *p* by the total number of comments in the subpath from *c* to the leaf node of *p*, and obtain the fraction of shifted comments in a sub-path. Figure 3.5-a and b show, for each topic, the

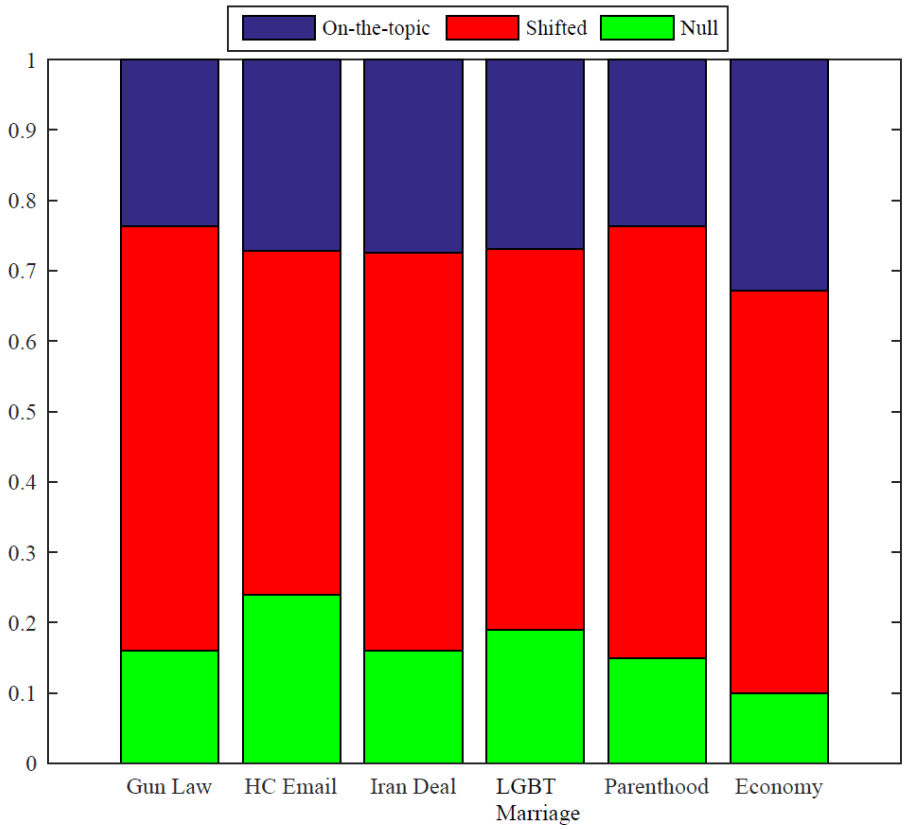

Figure 3.6 - On-the topic, Shifted, Null Comment percentages among different topics

mean and the standard deviation of the fraction of shifted comment over respectively all subpaths and subtrees that are rooted at a shifted comment.

Observation 5. In each path of the discussion tree, after the first shifted comment c, on average, 85% of the comments from c to the leaf comment shift away from the article's topic.

Observation 6. The fraction of shifted comments within subtrees with the first shifted comment as the root are 20% to 25% lower than paths on an average. The shifted comment averages in these subtrees are more than 55%.

Observation 6 is expected because if a comment is a descendant of a shifted comment, it cannot place another subtree. On the other hand, a comment can place multiple paths.

3.3.5 Relationship between topic shifts and the topic area

Some of the comments we extracted from Disqus do not match any phrases in SenticNet dataset. For example “Nope” or “So does Crump and Truz!” are comments for which there is no emotional score in the SenticNet 3.0 database. (We have refrained from adding our emotion scores for such comments, even though one can, after a diligent due-analysis, add a content-based emotion score for such phrases). Also, some of these comments are only stop-words and sarcastic words like “Crump” (as opposed to (Donald) “Trump”) (that would indeed evoke emotions on the replies and would shift them); we have similarly refrained from adding scores to such comments, and used a similarity score of 0 (i.e., “null emotion” comment) for such comments. Moreover, we removed these null emotion comments from our analysis. Figure 3.6 shows the percentages of on-the-topic, shifted, and null emotion comments for each topic.

Observation 7. After null score elimination, 64% to 72% of all comments are shifted from their original topics for all dataset.

The fraction of shifted comments varies topic by topic. For example, in *Hillary Clinton Email Controversy* 73% of comments are shifted away from the article topic. On the other hand, in economy news article this number drops to 67%.

Hillary Clinton Email Controversy, and *Supreme Court decision on LGBT marriage* news comments have more null comments than others.

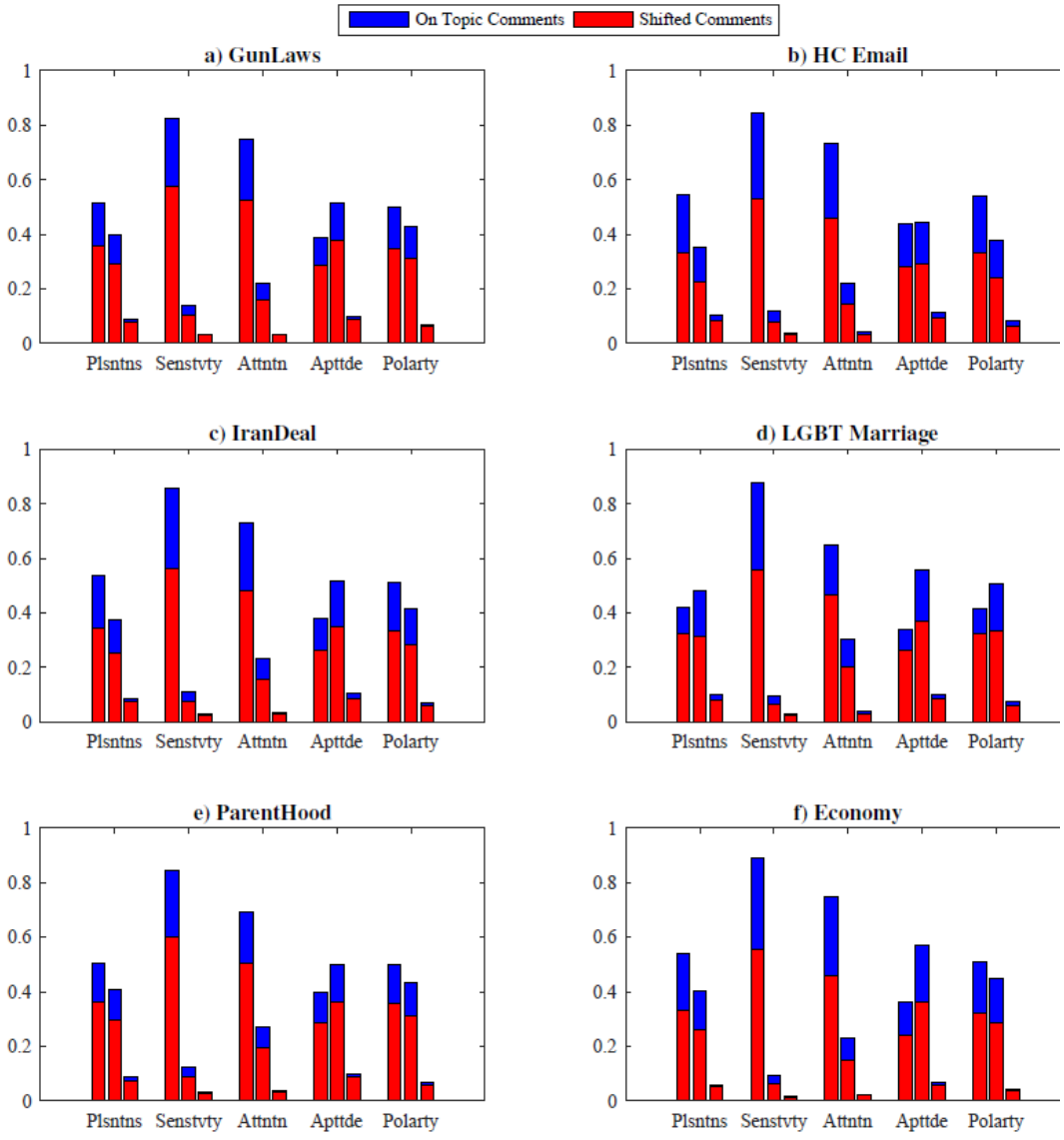

Figure 3.7 - Emotion bucket distribution and their shifted comment percentage, level 1, 2, and 3 distribution of emotional levels can be seen for each dimension, where we named dimensions Plsntns for *Pleasantness*, Senstvty for *Sensitivity*, Attntn for *Attention*, Apttde for *Aptitude*, and Polarty for *Polarity*.

Economy news comments have highest fraction of on-topic comments and the lowest fraction of null comments.

3.3.6 Effect of Emotion on Topic Shifts

This section analyzes the effects of emotion on comments. For each comment, we compute the four emotional dimension scores, namely, *pleasantness*, *sensitivity*, *attention*, and *aptitude*, as well as the similarity of the comment to the article, and compute the relevant statistics.

3.3.6.1 Emotion Statistics

Figure 3.7 lists emotion scores of comments (per topic and per emotion dimension) as histograms, where, for each dimension, there are three bars. (Note that, by taking absolute values, we reduce the number of emotion activation levels from 6 to 3). Within each bar of a given emotion dimension, we capture two emotion scores for that dimension, i.e., blue and red, representing the percentages of on-the-topic and off-the-topic comments, respectively, in that emotion dimension.

Observation 8. For most topics, more than 50% of comments have low levels of emotion, i.e., located in the leftmost bar for each histogram.

E.g., *Hillary Clinton Email Controversy* comments in Figure 3.7-b, for the sensitivity dimension, 85% of all comments is in the lowest level (i.e., *apprehension* or *annoyance* emotions). However, even though most comments have low emotion levels, they nevertheless evoke replies with high emotion levels, causing topic shifts in their replies.

In the histograms of each emotion dimension in Figure 3.7, the second bars (i.e., the medium level emotion category with scores in the range [0.33, 0.66]) have the second largest numbers of comments, with percentages ranging from 14% and higher. There are

some exceptions where most of the comments are located in this bar; e.g., for *LGBT marriage* comments (Figure 3.7-d), 56% of all comments are in the Aptitude dimension (*i.e., trust* and *disgust* emotions). The percentage of shifted comments in different dimensions of the medium level emotion category ranges from 63% to 76%.

And, finally, in Figure 3.7, the third bars (i.e., the high emotion level category) usually have the smallest number of comments, ranging from 2% to 12% of all comments. In comparison, 79% to 95% of all comments in this bar have shifted. For example, among *Gun Laws* (Figure 3.7-a) comments, the third bar of the sensitivity dimension (*rage* and *terror* emotions) have only 3% of all comments; but, 93% of these comments have shifted.

Observation 9. Almost in every dimension of all topics, the largest percentages of shifted comments are at the third (i.e., the highest emotion) bars.

As an exception, for *LGBT marriage* comments, in the Pleasantness dimension, 78% of the comments in the first (i.e., the lowest emotion) level (*serenity* and *pensiveness* emotions) have shifted; but in the third (i.e., the highest) level (*ecstasy* and *grief* emotions), this number is 79%; so, for *LGBT marriage* comments, high emotions do not affect topic shift itself.

Observation 10. Pleasantness and aptitude dimensions have, for all topics, more highly emotional comments (9% to 11%) than sensitivity and attention dimensions (3% to 4%).

Observation 11. The numbers of shifted comments increase for all topics when commenters choose words with higher emotion levels to express their opinion on a specific topic.

3.3.6.2 Discussions Driven by the First-Shifted Comment

We take, in each root-to-leaf path, the first shifted comment and the following comments (subtrees) in discussion trees, and measure the effect of these emotions on the

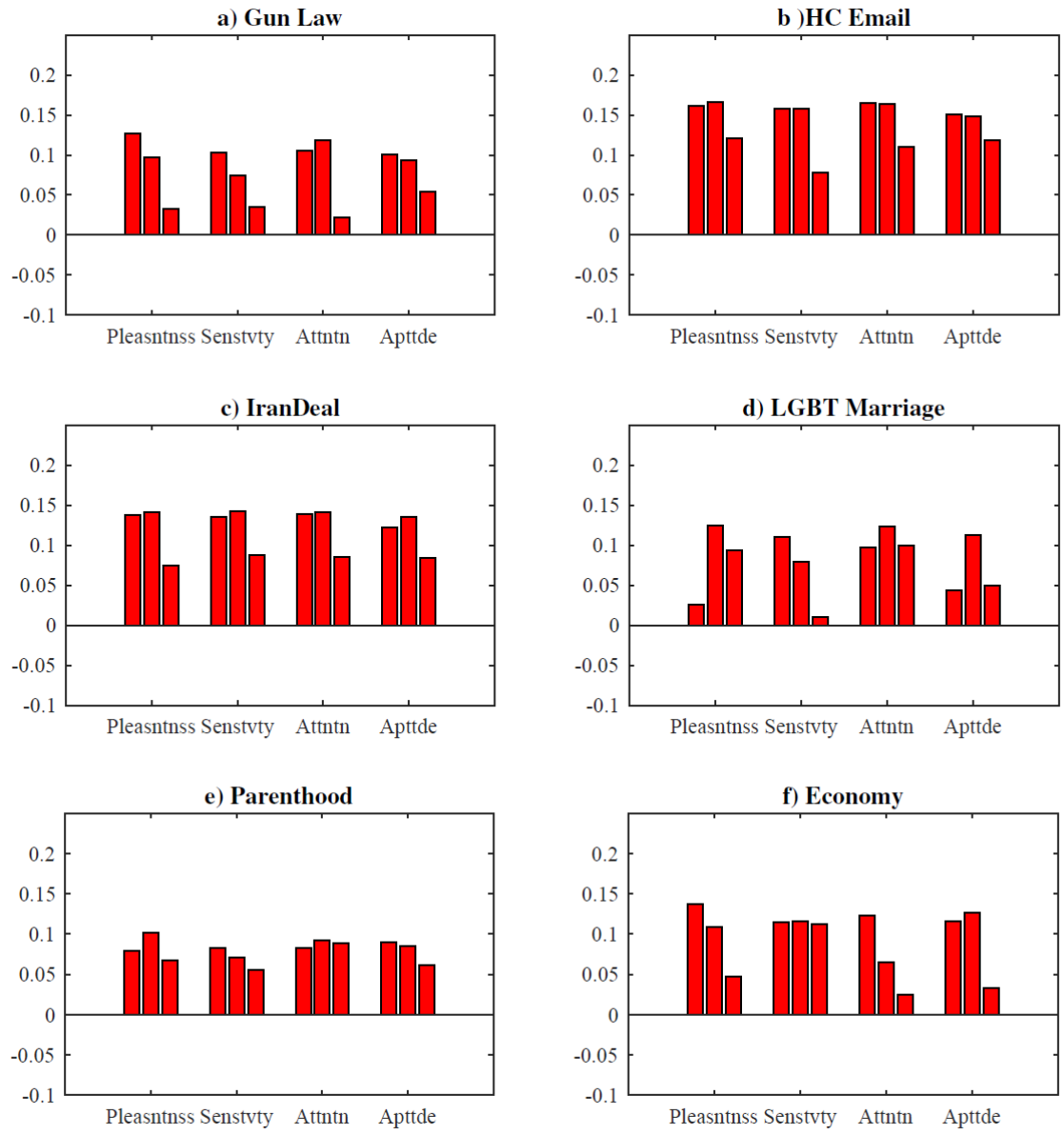

Figure 3.8 - Shifted comments change in response to comments that have larger Pleasantness scores, where we named dimensions Plsntns for *Pleasantness*, Senstvty for *Sensitivity*, Attntn for *Attention*, and Apttde for *Aptitude*.

topic shift and the following comments' emotions. First, we take all "first-shift" comments in a discussion tree just like we do in section 3.3.4, and have the following comments.

Observation 12. 90 % of the time, emotion levels of the first-shifted comments fall into the first and second (i.e., low and medium) emotion levels.

That is, the first-shifted comments do not contain high-levels of emotions per dimension.

Observation 13. Within the subtrees of all first-shifted comments, on-the-topic comments decrease almost more than 50% of the time for all topics (as compared to distribution shown in Figure 3.7).

Figure 3.8 summarizes, per topic, the topic shift percentage changes (increase or decrease) (Y dimension) as histograms where the X dimension represents, similar to Figure 3.7, the low-medium-high emotion bars of comments as a three-bar histogram (per emotion dimension). We see that, after a first-shift comment with pleasantness scores higher than .5 (i.e., *ecstasy* and *grief* emotions), all subsequent comments end up with higher percentages of shifted comments as compared to Figure 3.7.

For example, in *Hillary Clinton Email Controversy* (Figure 3.8-b), after a first-shift comment with high high pleasantness score, the topic shift in the comments of the following subtree increases more than 10%. This behavior of increased topic shift percentages is also observed, with very few exceptions (that occur in the *Economy* topic for our data sets), in the remaining emotion dimensions, leading us to conclude:

Observation 14. After a first-shift comment with emotion dimension scores higher than .5 for all dimensions, all subsequent subtrees end up with higher percentages of shifted comments as compared to Figure 3.7.

That is, first-shifted comments have a very large effect on decreasing the number of on-the-topic comments that follow, regardless of the emotion dimension.

3.3.7 Predicting a comment's shift by observing its emotion

Can we predict the topic shift in a comment by just looking its emotional dimension scores? To answer this question, we randomly take 80% of the data and train a decision tree model [56], and then predict the rest of the 20% of the data. We create a decision tree using pleasantness, sensitivity, attention, aptitude, and polarity scores to decide whether that comment is shifted or not. First, we create an $n \times 5$ matrix M such that each row represents a comment, and each column represents one of pleasantness, sensitivity, attention, aptitude, and polarity scores. Also, we know if a comment has shifted or not by looking its similarity scores to the article. We (a) set shifted comments to "0" and on-the-topic comments to "1" as "labels", (b) randomly select 80% of the data and create the decision trees by giving emotion scores and "labels", and (c) test the rest of the 20% of the data to see whether the comments are shifted or not. We use five-fold cross validation 100 times.

To improve the results, we add more fields to the decision tree. In addition to emotion scores, we add level information, and create an $n \times 6$ matrix M^{+} and repeat all the steps. Then we add another row "number of words in a comment" and an $n \times 7$ matrix M^{++}, and repeat all the steps. We take the average of precisions, and display the histograms in Figure 3.9.

Observation 15. Decision trees can predict shifted comments with precision higher than 75% for most of the topics by just looking at five emotion dimensions.

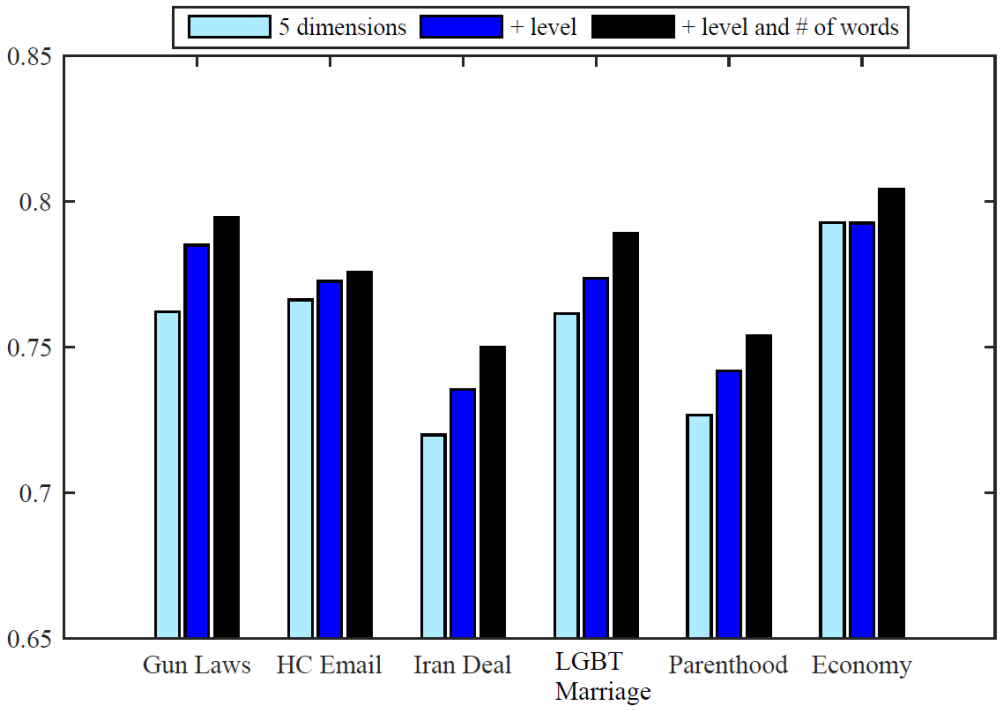

Figure 3.9 - Five-fold cross validation precision for decision tree predictions.

In other words, emotions of comments can tell whether or not a comment has shifted with at least 75% accuracy.

Observation 16. Adding level information to emotional dimensions increases the precision 1% to 2% for all topics.

As discussed in section 3.3.23.3.2, shifted comments are located at some specific levels; so this feature increases the predictions.

Observation 17. Adding the number of words in a comment as a new feature increases the precision 1% to 2% more.

This is because short comments may have wrong emotion scores at times; so, knowing the level and the number of words of comments adds 3% to 5% to the accuracy.

We have also used other learning techniques with similar results. In summary, learning techniques identify on-the-topic and shifted comments, even within the same (low, medium or high) emotion levels.

3.4 Conclusions

We have studied the effects on topic shifts of comments. To do so, we collect data from article originated comments, and we have modified tf-idf model to identify the closeness of comments and the article. Once we have the similarity scores of all comments, we distinguish off-the-topic and on-the topic comments and search the reasons behind why a comment go off-the topic. Understanding these phenomena on a real world experiment would be useful to eliminate undesired comments.

First, we look at topic areas and see size of off-the-topic comments varies topic by topic according to our observations.

Second, we inspect the structure of the discussion tree. We see that place of a comment in a discussion tree (tree level) makes effect on on-the-topic comments. First level comments are most likely to be on-the-topic comments, and the number fraction of on-the-topic comments are decreasing after the first level. However, at some point, the discussion recover. This point shows difference topic by topic. So, we conclude that there are still on-the-topic comments in deeper levels of the discussion trees.

Lastly, we look emotion levels (of various emotion dimensions) by using HourGlass of Model's dataset. Our observations show that there is a strong correlation between emotional levels of comments and their similarity scores. Majority of the on-topic-comments stay in neutral level of emotions. Also, we can predict whether a comment is on-the-topic or not by looking its emotional scores with more than 70% accuracy in every topic. Furthermore, we observe that a shifted comment with high emotional scores triggers strong emotions and following comments are also off-the-topic.

For future work, we plan to design, implement and evaluate automated tools that can identify and take actions on shifted comments, as guided by readers.

4 Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews

4.1 Introduction and Background

It is common knowledge that, usually, moviegoers, called users or reviewers in the rest of this chapter, utilize movie ratings and reviews in selecting their next movie to see/watch. And, unfortunately, sometimes movie reviews and ratings do not help users make the right choices, as evidenced by their emotional feelings after watching the movie. This is perhaps because users desire a certain emotional state after watching a movie, which does not match the emotions evoked by the selected movie.

Clearly, users' reviews and ratings for a movie are strongly tied to their emotions evoked by the movie. This chapter argues that

1. It may be useful for users in their decision-making process to choose the next movie to watch if a movie also comes with an (expected) "emotion signature" or an "emotion map", and,
2. Towards goal 1, we can build automated software tools that (i) analyze movie reviews and ratings, and (ii) provide an emotional signature for a movie as evidenced from the reviews and ratings of that movie.
3. Clearly, once emotion maps for all movies are at hand, if a user perhaps submits his desired emotion state and, possibly, the desired genre of the movie, it is easy to build a "personalized movie recommender system" for each user.

This chapter is a first step for goals 1 and 2 (but not for goal 3), and makes an attempt to analyze the relationships between

- Users' ratings for a movie and their emotions evoked by watching the movie as evidenced in their movie reviews and ratings, and
- Movies' genres and users' emotional responses from their movie reviews.

For our experimental evaluation, we used movie reviews from IMDb, the world's most popular content source for movie, TV and celebrity content [57] with reviews for more than 3.5 million movies. IMDb members provide reviews and usefulness scores for other reviews. However, the overall rating of a movie is calculated by IMDb's own rating algorithm, which, in turn, is based on reviewers' scores [58].

Our movie dataset contains 157,344 reviews for 134 movies, where (i) most movies were very highly rated, (ii) a small number of movies were very lowly rated, and (i) selected moves spanned 19 genres/genre combinations, with each genre having at least 10 movies. Then, using the top 100 "most useful"-ranked reviews of each movie in our data set, we extracted for each movie an "emotion map" that captures reviewers' emotions with respect to the four emotion dimensions of the Hourglass model of emotions of Cambria et al [17].

We have observed that (i) even for movies within the same genre, reviewers emotions may differ completely; that is, emotion maps of two very highly rated movies with the same genre labels may be drastically different. And, that (ii) emotion maps of very lowly rated (i.e., less than 3 out of 10) movies display minimal variation across the emotion scores of different reviewers; i.e., reviewers have very similar responses to a very low-rated movie.

Using the k-means clustering algorithm to cluster movies according to reviewers' emotions per dimension, we have observed that

- All highly rated movies, which are those with ratings of 7 or above, have emotion maps with low or medium emotion levels, and rarely extreme emotion levels.
- Clustering allows effective genre-specific grouping of movies, meaning genre-specific movies have specific emotion map patterns as identified by the aggregated emotion maps of the groups that they belong.
- Genre coverage of some clusters are quite complete, meaning some clusters have significant numbers of movies from all genres.
- Removing emotion dimensions from clustering changes the clusters significantly.

Table 4.1 - HourGlass of Emotion Model

	Pleasantness	**Attention**	**Sensitivity**	**Aptitude**
+3	Ecstasy	Vigilance	Rage	Admiration
+2	Joy	Anticipation	Anger	Trust
+1	Serenity	Interest	Annoyance	Acceptance
0				
-1	Pensiveness	Distraction	Apprehension	Boredom
-2	Sadness	Surprise	Fear	Disgust
-3	Grief	Amazement	Terror	loathing

Table 4.2 - Compound Emotions of Second Level

	Attention > 0	**Attention < 0**	**Aptitude > 0**	**Aptitude < 0**
Pleasantness > 0	Optimism	Frivolity	Love	Gloat
Pleasantness < 0	Frustration	Disapproval	Envy	Remorse
Sensitivity > 0	Aggressiveness	Rejection	Rivalry	Contempt
Sensitivity < 0	Anxiety	Awe	Submission	Coercion

Finally, we have redone the movie clustering with respect to all the reviews for a movie, not just the top 100 reviews of a movie. Our conclusion is that, within a cluster, the closest movies stay mostly the same regardless of the sizes of the emotion vectors used for cosine similarities. And, most similar movies still stay in the same cluster.

As seen in Figure 3.1 and Table 4.1, each of the four dimensions of the Hourglass of model have six levels of activation, which collectively characterize the emotional state of an individual. Moreover, more complicated emotions can be seen in Table 4.2 by combining different dimensions. In total, from Table 4.1 and Table 4.2, each review has 8 different emotions by looking four independent dimensions and their pairwise combinations.

4.2 Modeling Emotion in Moviegoer Reviews

To represent the emotional landscape of each review of a movie, we use the SenticNet 3.0 [59] database. For each phrase in the database, there are four different scores (one for each of the four emotion dimensions) in the range [-1,1]. We map each review to the SenticNet database by identifying phrases in the review to a phrase in SenticNet. Since each comment may map to multiple phrases in the database, we then use dimension reduction for each review to get a single review score per emotion dimension. The scores

of all reviews for a movie m per emotion dimension di forms the "emotion vector of movie m for emotion dimension d_i".

Let a movie m have n reviews, and the maximum number of SenticNet word instances in any review of movie m, duplicates included, be k. We create a k by n matrix to store all emotion scores of movie m for emotion dimension, say, d_i. If a review has less than k scores, we use bootstrapping [60] to k dimensions for that review. Next, obtain an emotion dimension-specific review score for a review of the movie m, we reduce the matrix dimension from k by n to 1 by n by using Singular Value Decomposition (SVD), as shown in Equation 5 and 6.

$$M_{d_i} = U_{d_i} \Sigma_{d_i} V_{d_i}^* \tag{5}$$

where, M_{d_i} is a $k \times n$ original matrix for each dimension, U_{d_i} is a $k \times k$ unitary matrix, Σ_{d_i} is a $k \times n$ diagonal matrix with non-negative real number on the diagonal where the diagonal entries are singular values such as $\sigma_{11} > \sigma_{22}$, and $V_{d_i}^*$ is $n \times n$ unitary matrix.

$$E_{d_i} = (U_{d_i}^{(1)})^T M_{d_i} \tag{6}$$

Since σ_{11} is the largest singular value, we take $U_{d_i}^{(1)}$ as first column of U_{d_i} and transpose it from $k \times 1$ to $1 \times k$ and multiply by M_{d_i}, then we get E_{d_i}a $1 \times n$ vector, which contains a single review score for each review, which we call the "emotion vector of movie m for dimension d_i", where d_i is one of dimensions *pleasantness, attention, sensitivity, or aptitude*.

An emotion map of a movie m is simply the four emotion E_{d_i} vectors of the movie, vertically stacked and represented as a heatmap that uses the same special red-green color coding scheme for all four emotion vectors.

4.3 Experimental Evaluation and Results

4.3.1 Dataset

Our dataset, collected using an IMDb API library [61], contains 157,344 reviews from (a) 116 highest-rated movies with large numbers of reviews at IMDb, and (b) 18 movies with low to very low reviewer scores. We also made sure that (i) our movie set had many genres, more specifically, 19 genres and genre combinations, and (ii) each genre category had at least 10 movies.

Note that (a) a movie may have multiple genre labels, (b) each review can be voted by other IMDb users for its usefulness, and (c) we collect individual user ratings of movies from reviews.

4.3.2 Emotion Visualization

To visualize the aggregated per-dimension emotion score of a movie's reviews, we use the movie's emotion map which is really a special heatmap [62], a two dimensional graphical representation of data with colors instead of numbers. The reasons for using a heatmap are: (i) multiple dimensions can be seen in one figure, and (ii) different patterns can easily be distinguished from each other via clustering.

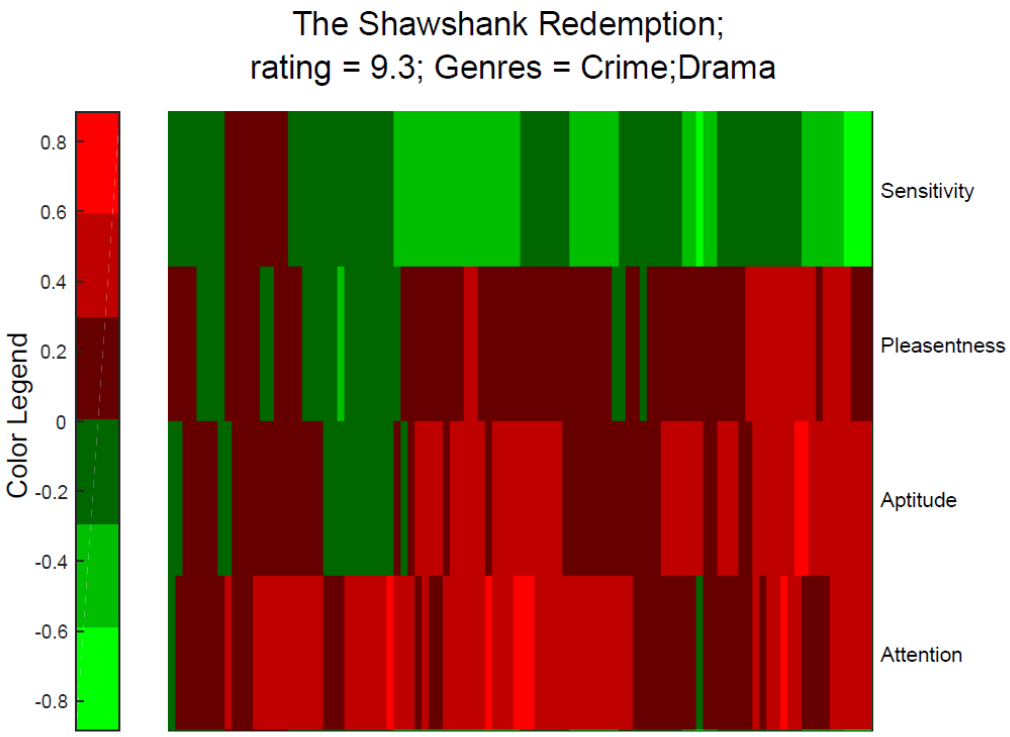

Figure 4.1 - Emotion Map of "The Shawshank Redemption".

There are various color schemes for heatmaps [63]. Usually, with respect to color wheel [64], complementary colors are chosen. For our movie emotion maps, we choose the red-green scheme with six levels that represent an emotion on an emotion dimension. An example view of a movie emotion map is shown in Figure 4.1. The four emotion dimensions form the rows of the emotion map (i.e., placed horizontally), and each review's scores per dimension form a column of the emotion map (i.e., placed vertically).

To save space, to construct a movie emotion map, we use the top 100 "useful" reviews of that movie within the reviews of the IMDb site. Moreover, to observe patterns within a movie emotion map, we hierarchically cluster reviews. For this reason, the original

“usefulness” orders are altered in a heat map; e.g., the most useful comment’s place in a heatmap is not necessarily the first column.

4.3.3 Results

As stated before, each review has 4 dimension scores. And, each review’s four emotion dimension scores are obtained as explained in Section III, by using the SenticNet database and Singular Value Decomposition (SVD).

4.3.3.1 Emotion Heat Maps

For each movie and a review, we compute the four emotion dimension scores, and create the corresponding movie emotion map. Figure 4.1, Figure 4.3, Figure 4.2, and Figure 4.4 contain emotion maps of different movies.

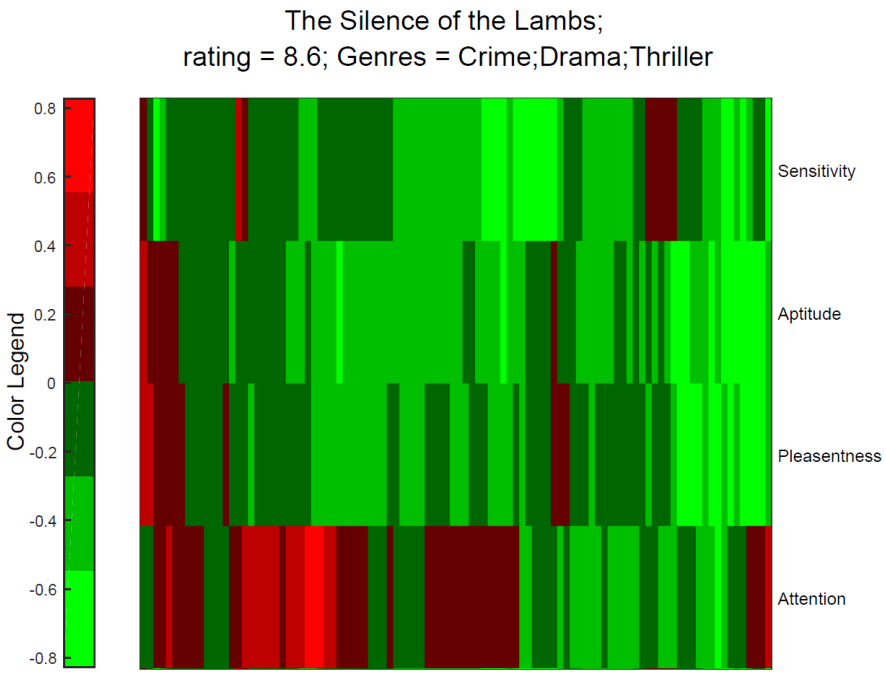

Figure 4.2 - Emotion Map of "The Silence of the Lambs".

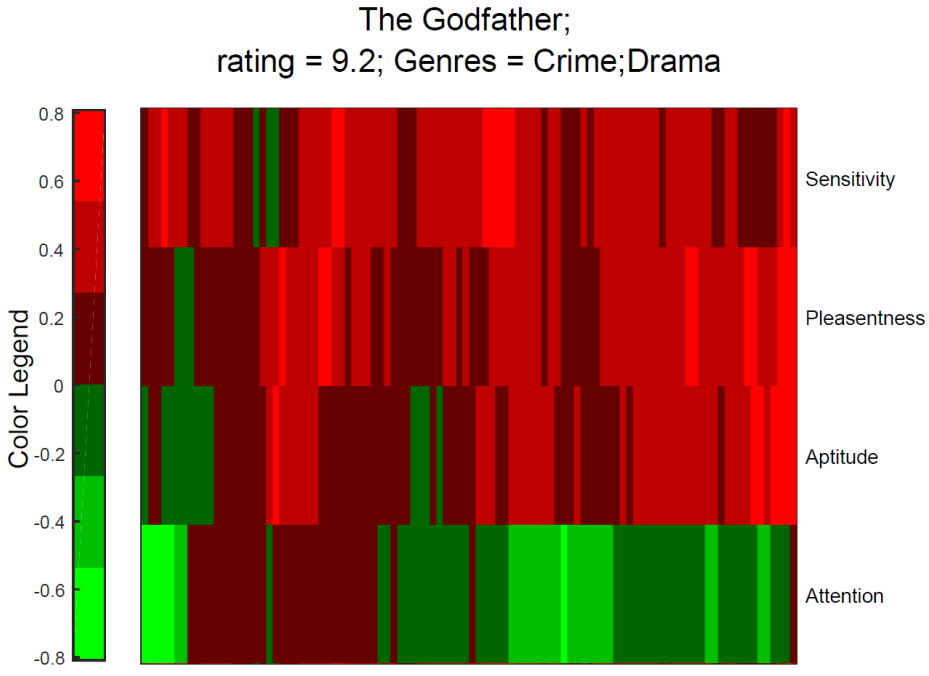

Figure 4.3 - Emotion Map of "The Godfather".

Figure 4.4 - Emotion Map of "The Hellcats".

Observation 1: The movie The Shawshank Redemption, from the most useful IMDb user reviews, evokes

- Mostly negative sensitivity dimension emotions, namely, apprehension and fear, and, only few users exhibit the terror emotion, and
- Mostly positive attention dimension emotions, i.e., interest and anticipation.
- The movie The Shawshank Redemption, rarely evokes the strongest emotion levels for all four dimensions (i.e., those close to +1 and -1).

Observation 2: Contrary to The Shawshank Redemption, the movie Godfather evokes

- Positive sensitivity dimension emotions, i.e., anger and annoyance;
- Negative attention dimension emotions, namely, distraction and surprise.

- Pleasantness and Aptitude dimensions look similar. Users exhibit joy and serenity; trust and acceptance emotions in different dimensions.
- Majority of reviews do not have polar emotions.

Observation 3: The movie The Silence of the Lambs evokes, for the majority of users, negative emotion levels in all dimensions.

Observation 4: The movie The Hellcats, which has a very low IMDb score, evokes very flat emotions and the same emotion score for all dimensions across all users.

Observation 5: Even for movies in the same genre, completely different reviewer emotions may be evoked. The movies The Shawshank Redemption and Godfather (Figure 4.1 and Figure 4.3) belong to the same genre category, namely, *Crime* and *Drama*, and their IMDb movie ratings are very close, namely, 9.3 and 9.2 respectively. Yet, their emotion maps are completely different.

Note that emotions displayed in a review may reflect the failed expectations of the reviewers about the movie or the true emotions of the user, independent of his/her expectations. For example, a crime movie may have an anger score in a review because of its content and the movie has a good score; on the other hand, the same movie may have an anger score in another review simply because the expectations of the reviewer has not been met.

Observation 6: Most extremely low-scored movies evoke extreme emotions. The movie *The Hellcats* has a lower score and its emotion map has just one review pattern (Figure 4.4). For this movie, almost every reviewer shows extreme emotions on each

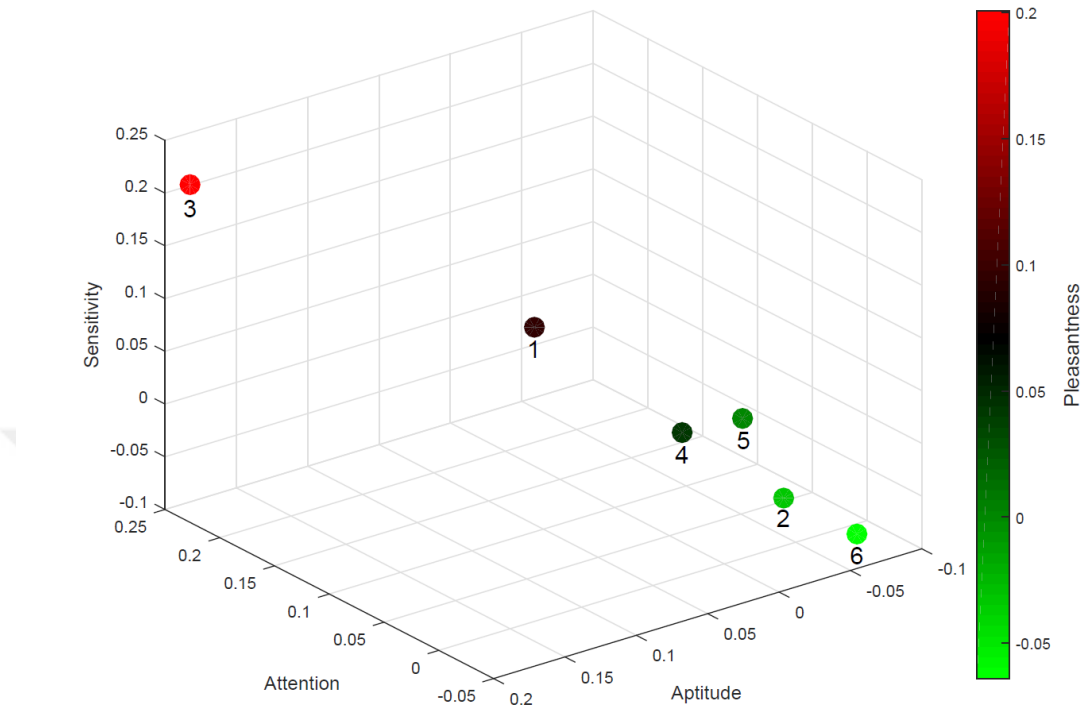

Figure 4.5 - k-means centroids. Since the data is four dimensional and the figure is three dimensional, the color within each circle indicates the score for the 4th dimension.

dimension. From *Pleasantness* and *Aptitude* dimensions, there are remorse emotions among the reviews.

4.3.3.2 Clustering Movies Based on Evoked Emotions

There are six emotion levels within a dimension in the hourglass of emotion model. To interpret the polarity of emotions, we name -1 and +1 emotions as level 1 (the low level), -2 and +2 emotions as level 2 (the moderate level), and -3 and +3 emotions as level 3 (the extreme level) emotions based on Table 4.1.

We use the k-means clustering algorithm to cluster movies according to reviewers' emotions per dimension. First, we remove low-rated IMDb movies, which are those with

ratings less than 7, from our dataset. This resulted in a movie data set of size 118. Again, for comparison purposes, the 100 most useful reviews per movie are used for clustering. There are 4 dimensions, and each movie has 400 reviewer scores across all four dimensions. After comparing the k-means algorithm with different k values, we have chosen $k = 6$. Figure 4.5 shows the centroids of each of the six clusters, numbered 1 to 6 and the position of each centroid computed as the median of the cluster members' scores visualized across three emotion dimensions, namely, *sensitivity*, *attention*, and *aptitude*. The median score value for the fourth dimension, namely, *pleasantness*, is represented by the color of each cluster centroid. Note that the extreme positions in each dimension are lost by taking the averages of dimension scores.

Next we investigate

- How close a movie is to its own cluster's centroid,
- The distance between clusters, and
- The distance between movies from each other for each run.

Observation 7: Overall, centroids of some clusters are very close; so movies in these clusters may be similar even if they are in different clusters. Moreover, since all low-scored movies (i.e., those with scores less than 7) are removed, most observed emotions are in levels 1 or 2.

Figure 4.5 shows centroids of each group, and their distances to each other. Groups 3 and 6 are the farthest groups among all groups. On the other hand, groups 2, 4, and 5 are very close to each other. Distance from each movie can be seen in Figure 4.6

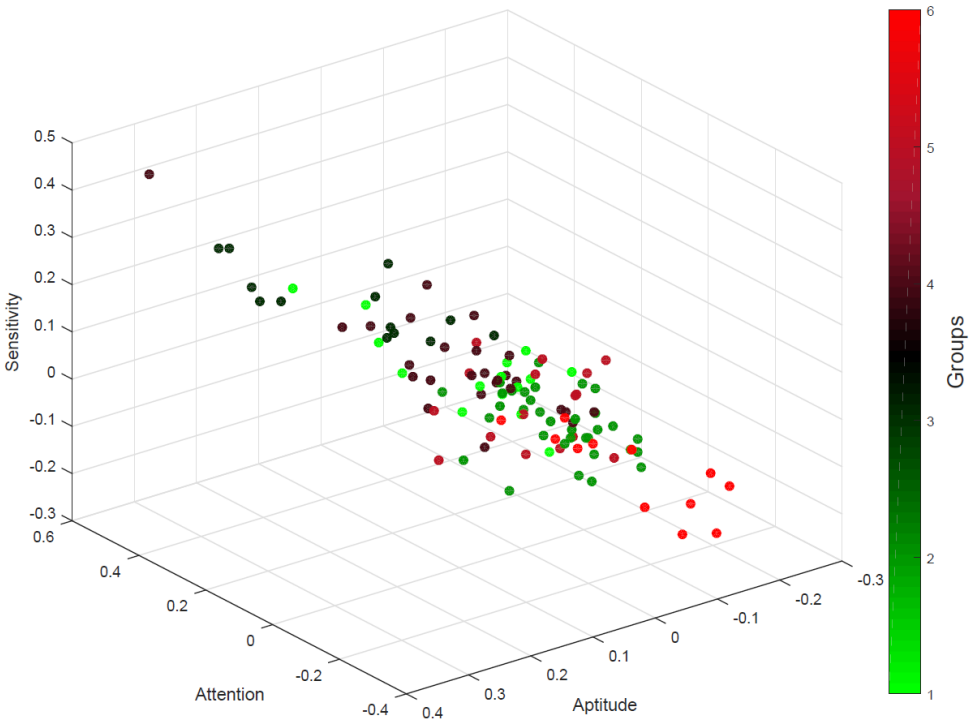

Figure 4.6 - Movies within each cluster. For visualization, the pleasantness dimension is removed from this figure. The color bar shows the group (i.e., cluster) for each movie.

. Table 4.3 shows the closest 9 movies namely to their group centroids in the dataset.

Next, we look into how genre distributions change among movies within a cluster. The total number of movies in the same genre in each group is divided by total number of movies of that genre. Figure 4.7 shows the results. As an example, 67% of all horror movies are in group 2, and there is no horror movie in groups 3, 5, and 6.

Observation 8: Movies belonging to certain genres are clearly distributed to specific groups, meaning genre-specific movies have specific emotion map patterns as identified by the aggregated emotion maps of the groups that they belong.

Table 4.3 - k-means groups, they are in ascending order in terms of Euclidean distance to belonging group centroids

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
V for Vendetta	Saw	The Shawshank Redemption	The Lord of the Rings: The Fellowship of the Ring	The Green Mile	Million Dollar Baby
The Godfather	The Avengers	300	The Matrix	Braveheart	Titanic
World War Z	Pirates of the Caribbean: The Curse of the Black Pearl	The Dark Knight	Fight Club	Frozen	Rocky
Spider-Man	Psycho	Pulp Fiction	The Dark Knight Rises	The Last Samurai	The Wizard of Oz
Inception	Avatar	Finding Nemo	Iron Man 3	12 Years a Slave	How to Train Your Dragon
Iron Man	Transformers	Interstellar	The Prestige	The Hunger Games	Kill Bill: Vol 1
Inglorious Bastards	Se7en	Cars	WALLE	Schindler's List	The Silence of the Lambs
American Beauty	Bowling for Columbine	Forrest Gump	Gladiator	Gangs of New York	March of the Penguins
The Usual Suspects	A Clockwork Orange	The Social Network	The Incredibles	District 9	Zeitgeist

As examples, group 6 has 33% of documentaries and half of the sport movies in the dataset whereas group 3 does not contain any documentary movies and only 13% of sport movies. There is a specific Crime genre distribution as well.

Observation 9: Some groups have significant number of movies from all genres, e.g., group 2, which has movies belonging to almost every genre.

This may be because each movie may have multiple genre labels and these movies may evoke the same or similar emotions, regardless of the genre.

Next we investigate the relationships between clusters and specific emotions. So far, we have looked all dimensions together. However, users may be interested in, say,

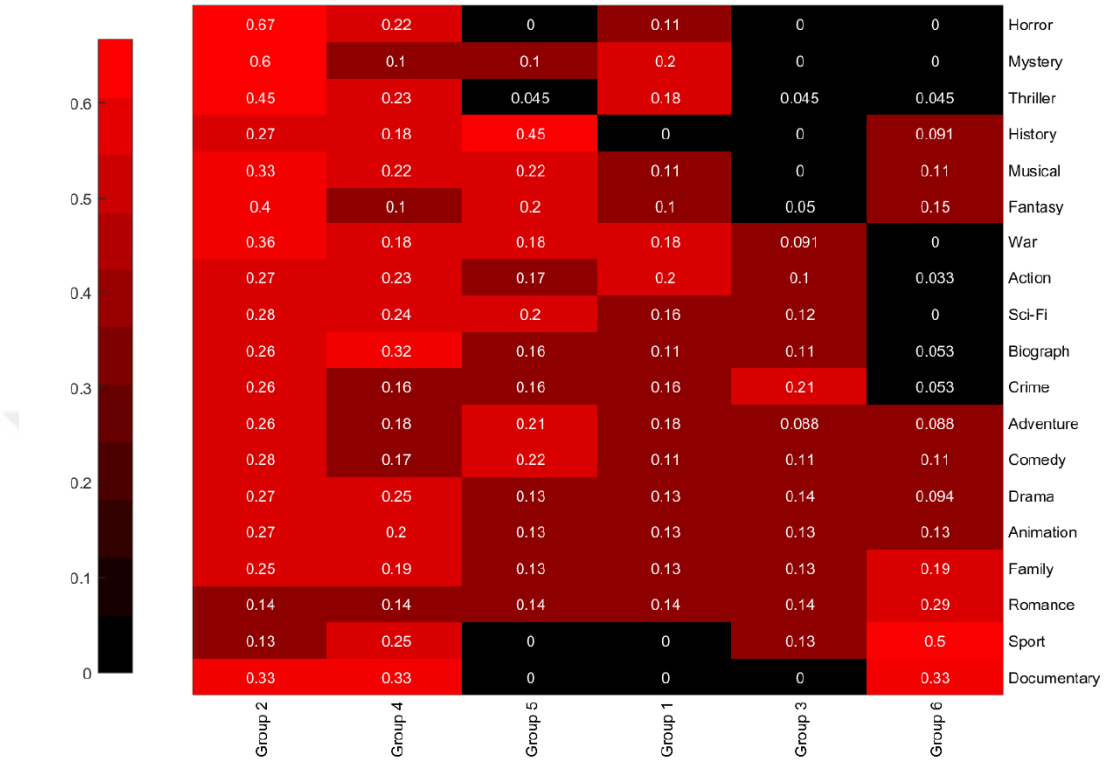

Figure 4.7 - Genre distribution of movies within the k-means (6-means) groups.

positive pleasantness movies. As also seen in Figure 4.1 and Figure 4.3, movies may have common as well as distinct emotions, and, by concentrating on only specific dimensions or specific emotions, movies may belong to different or the same categories. To investigate this possibility, we ran the k-means algorithm with the same parameters except with only two dimensions. To save space, below we discuss the results of only one clustering run.

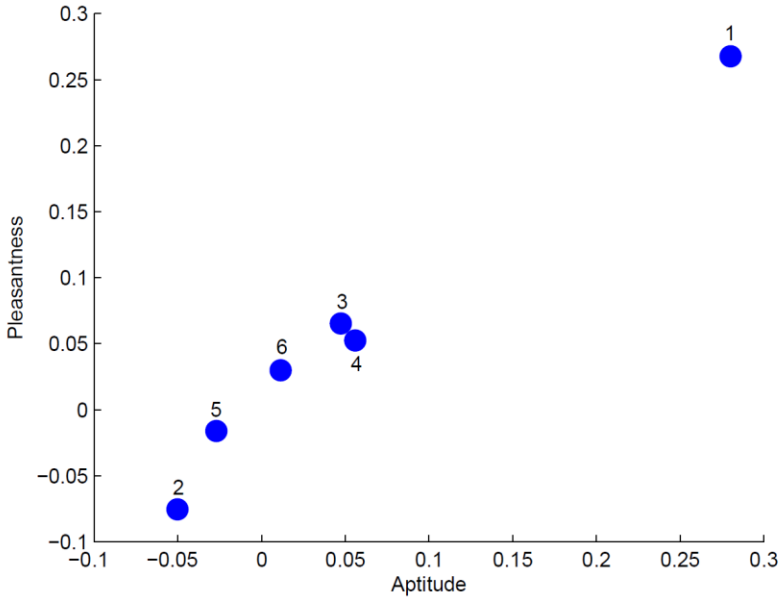

Figure 4.8 - Pleasantness vs Aptitude cluster centroids.

Observation 10: Running the k-means algorithm with only Aptitude and Pleasantness scores, we observe from Figure 4.8 that centroids of clusters are distinctively identified by some of the emotions love, gloat, envy, and remorse.

Observation 11: Re-clustering movies after removing emotion dimensions selectively from clustering changes the clustering patterns significantly, indicating different automated movie recommendation possibilities based on changes in clustering dimensions.

As examples, group 3 of Figure 4.5 and group 1 of Figure 4.8 are similar with respect to their *Aptitude* and *Pleasantness* scores. Group 3 in Figure 4.5 has 13 movies; on the other hand, group 1 of Figure 4.8 has 7 movies that exactly show up in group 3 of Figure 4.5 If specific emotions are desired, groups may change significantly.

4.3.3.3 Emotion Vector Similarities of Movies

Next, we change the dataset from the "top 100 useful reviews" to "all reviews", and compare pairwise similarities between movies as represented by their emotion vectors (please see section 4.2 for the definition of an emotion vector of a movie). Note that the number of reviews for each movie may be different, and the size of the input for a movie as defined by the number of reviews of the movie needs to be the same for a pair of movies in order to calculate their pairwise similarity. So, we use the following approach to force them to be the same. When calculating the similarity between two movies with, say, a and b numbers of reviews, we first need to compute each movie's emotion vector for the emotion dimension d. For the emotion vector computation, we always use $t = min(a, b)$ number of reviews from each of the two movies. As an example, the movie "The Dark Knight Rises" has 4,575 reviews, and the movie "Inception" has 2,777 reviews; thus, we take 2,777 most useful comments on "The Dark Knight Rises" and all of the reviews on "Inception" and calculate the cosine similarity of emotion vectors of the two movies for each of the four emotion dimensions. For emotion vectors, we take all of the dimensions, but each dimension can be taken independently or pairwise, as well.

One of the results is shown in Table 4.4 for the movie "V for Vendetta" which is the movie closest to group 1 centroid in Figure 4.5 and Table 4.3.

Table 4.4 - Similar and Dissimilar movies to "V for Vendetta"

Similar Movies	Dissimilar Movies
Inception	The Shawshank Redemption
The Godfather	Pulp Fiction
The Schindler's List	Interstellar
The Usual Suspects	The Silence of the Lambs
Iron Man	How to Train Your Dragon

Observation 12: The closest movies stay mostly the same regardless of the sizes of the emotion vectors used for cosine similarities. And, most similar movies stay in the same group.

Observation 12 is useful for an initial starting point to a customized movie portfolio for a user. If a user writes a review and rate a movie with a good score; we calculate his emotions; then our emotion maps can recommend to him other movies based on his past ranking/emotion trend using other similar user reviews.

4.4 Conclusions

Deciding a movie to watch is sometimes leading a wrong decisions by just looking at their genres and user ratings. IMDb is the leading website in terms of movie content and user reviews and these reviews are useful for moviegoers. We have introduced the notion of movie emotion maps for capturing the emotion content of movies via the emotions expressed in movie reviews. So, seeing the emotional levels of the movies will contribute more accurate decisions for moviegoers. To do so, first, we map the word matching of HourGlass of Model dataset in the reviews, and we get single scores of each HourGlass of

emotion dimension by using Singular Value Decomposition. Then, another challenge is showing these scores more conveniently to the moviegoers. We use special heat maps to show these results. We select the most useful reviews, which are voted by other IMDb users, and place them onto heat maps with clustered by their emotional levels. We see that similar movies in terms of genres and ratings, have different heat map layout. Adding this information could be useful into the website. We also introduce an algorithm to find emotionally similar movies to a given movie. In this way, one can find a movie with more parameters rather than having only genre and rating information. They do not even have to read all reviews to make a decision.

5 Detecting Subtopics in Social Media Discussions and Emotional Change

5.1 Introduction

Comments in a social media discussion start around a topic, and they branch into different aspects of the topic. Sometimes, these branches are related to the topic, and, other times, they completely shift into unrelated topics. In chapter 3, we discussed article-originated topics, defined the topic of a discussion as the topic of the article (as defined by a tf-idf-based vector representation of the article), and investigated topic shifts in comments with respect to the topic of the article.

Clearly, not all social media discussions are initiated by an article. Or, some discussions in a large set of comments, say, in thousands, may contain interesting and/or useful discussion threads on subtopics, regardless of whether or not the discussion is started via an article. In this research, our goal is to identify and investigate features of "discussion threads on subtopics" in social media discussions, in terms of both the characteristics and the emotional state of subtopic's comments. Our approach is to

1. Use an existing topic modeling technique, namely, the Latent Semantic Analysis (LSA) [21], to identify automatically subtopics and related comments in a social media discussion without initiated by an article,
2. Measure and analyze the emotional "state" of each subtopic discussion, and observe correlations between emotions and subtopic characterizations as identified by the LSA, and, finally,

3. Develop a model to identify "emotionally consistent comments" within a given subtopic discussion with respect to the subtopic's emotional signature.

We believe that such an approach may turn into a practical tool, in that users can go over the list of subtopics in a discussion with large numbers of comments, identify subtopics that are of interest to them, and then, if they wish to do so, perhaps eliminate emotionally inconsistent comments from the original set of comments in the chosen subtopic discussion.

Example. Consider the Reddit discussion [65] that starts with a news article, entitled "*CVS puts out a generic competitor to EpiPen at a 6th of the price*" about a cheaper medical device for injecting adrenaline to threat anaphylaxis, a serious allergic reaction. According to the article, the original device EpiPen costs more than $600, and CVS, a pharmacy company, introduces a generic version of EpiPen for $109.99. With more than 3,000 comments in this social media discussion, Reddit users discuss many subtopics with different points of views. For example, some users are talking about free market, capitalism; another group of users are talking about if pharmacist can switch the medication out to its alternative as without any comment in the prescription; another group is commenting about why the price is so expensive; yet another discussion threads are about deals and coupon information about the product. Moreover, there are some interesting and useful comments evoking different subtopics. For example, one user starts a subtopic thread by sharing his experience that he did not have EpiPen during his allergic reaction, and his grandfather (who used to sell animal organs to pharmaceutical companies, and knew pig kidneys were used to produce the treatment) cut off adrenal glands of animals,

and gave them to him and others. This specific user's comment starts a discussion thread on why the cure is very cheap, but the price of the device is so expensive. Another user mentions that synthetic drugs produced by pharmaceutical companies are expensive. One user blames research, development, and patent costs for the price difference. And, many more subtopic discussions can be found within the full set of discussions.

Clearly, in the above-summarized Reddit discussion, many comments are related to the original topic, and yet belong to semantically different subtopics. Moreover, these comments display different emotions, e.g., the user complaining about the prices display different emotions, and users sharing the coupon information display other types of emotions. In this study, we identify subtopics in a social media discussion and the emotional state of commenters. In more detail, we investigate the following issues within the context of social media discussions with different subtopics.

- Inspecting the "relatedness" of comments to a subtopic. Our approach is to associate a score for each comment specifying how close it is to each subtopic semantically,
- Inspecting the differences of subtopics within each other,
- Finding emotional status of each comment and subtopic,
- Finding *semantically mutual information* and *emotion* among subtopics within discussions. For example, can we distinguish why a subtopic is different from others by looking at the most important words and phrases in discussions of those subtopics?

In this research, we measure the emotional state of each comment via the HourGlass of Emotion Model [17]. We observe that

1. There are emotion-related and emotion-unrelated topic changes in discussions. We investigate the roles and degrees of different emotion dimensions in creating discussions on different subtopics.
2. Emotionally consistent comments form a more semantically meaningful subtopic discussion and distinguish from the subtopics having similar Latent Semantic Analysis scores. For this purpose, we sort the comments in a subtopic by their emotional distance to the subtopic emotion centroid.

5.2 Methods

Topic modeling is a text-mining tool for the discovery of hidden semantic structures in a document corpus [66]. Latent Semantic Analysis [21] is one of the most famous techniques to uncover hidden semantics among the corpus. The reason we choose LSA is because, in practice, LSA is faster and has almost similar accuracy rate with its alternatives [67]. Also, with a slight addition to LSA, more descriptive topic phrases can be extracted to identify topics rather than extracting single words [68].

LSA basically uses the Singular Value Decomposition technique to reduce the dimension of the data, and to find latent topics. First, the document-term matrix A is created, for which we use a tf-idf matrix where each comment is a row, and each word is a column. Note that stop-words are removed, and word stemming is not used. Let A be a $d \times t$ matrix where d is the number of documents (i.e., in our case, the number of comments),

and t be the number of terms (words) in the corpus. Our approach is to use each "latent topic" obtained via Latent Semantic Analysis (LSA) as a subtopic within a discussion.

We have also experimented with a document-term matrix by using only "nouns" instead of all words, with the thinking that (i) verbs, adjectives, adverbs, and conjunctions are commonly used for all topics, and, (ii) nouns are more descriptive of different subtopics; however, experimental results were negative, and, this approach is terminated.

In more detail, each latent topic, i.e., a discussion subtopic is obtained as follows. A is decomposed into

$$A = U\Sigma V^T \tag{7}$$

where U and V have the *left and the right singular vectors, respectively*, and Σ has the *singular values on the diagonal*. Next, we choose best rank – k approximation of A in the least-square sense to reduce the dimensions. In this framework, each dimension of k dimensional space, which is an eigenvalue from Σ with its corresponding vectors from U and V, is called a *latent subtopic*, i.e., a subtopic within our discussion set. Observe that, because Σ has sorted eigenvalues in descending order on the diagonal, the importance of subtopics are decreasing as well, which means, the most "important" latent subtopic is the first vector of U and V. We use U' and V' as first k vectors of U and V to describe latent topics for the rest of this chapter.

5.2.1 Finding the most important phrases for each subtopic

Each subtopic is defined via its "most important words", identified by LSA. LSA gives these words with their importance scores in the columns of the right singular vectors of V in equation 7. The first k leftmost columns of V reflect the terms' importance weights

over the subtopics. Moreover, since left singular vectors of U in equation 7 have weights for each comment over the subtopics, we use them to find the most important bigrams $\Gamma^{(\alpha)}$ in a subtopic:

$$\Gamma^{(\alpha)} = \sum_{i=1}^{n} U_i^{(\alpha)} \beta^{(i)} \tag{8}$$

where $\Gamma^{(\alpha)}$ is the bigram scores of subtopic α, n is the number of non-negative weighted comments in subtopic α; $U_i^{(\alpha)}$ is the left singular vector for comment i, subtopic α; β is the bigram corpus, and $\beta^{(i)}$ is the bigram frequency numbers in comment i. The intuition behind this equation is that larger weighted comments have more important bigrams. One can use the same approach for trigrams, or n-grams, as well.

5.2.2 Modeling Emotion in Reddit Comments

To represent the emotional landscape of each subtopic, we use the SenticNet 3.0 [69] database. For each phrase in the database, there are four different scores (one for each of the four emotion dimensions) in the range [-1,1] and these dimensions have six levels of activation, which collectively characterize the emotional state of a word.

We calculate E_c^j, emotional aggregated discretization vector for each comment c, for each HourGlass of Model dimension j where $j \in \{Pleasantness, Sensitivity, Attention, Aptitude\}$. First, we map each Reddit comment to the SenticNet database by identifying phrases in the comment to a phrase in SenticNet. Second, we categorize these scores into equally divided six groups by setting the corresponding group to 1 and 0 for others. (If the score is in between $[-1, -0.66]$, then 1 for first group; if the score is in between $[-0.66, -0.33]$, then 1 for second group, and so on). For example, if a word w in

comment c has score of -0.88 for the *Pleasantness* dimension, we create the discretization vector $[1, 0, 0, 0, 0, 0]$ for w since -0.88 falls into first group. Finally, we then aggregate the discretized scores of each word in the comment c. Now, we have the emotion frequencies for each comment. For example, if a comment c has four SenticNet matchings for the *Pleasantness* dimension, then its aggregated discretization vector E_c^j may look like $[1, 2, 0, 0, 1, 0]$ where j is *Pleasantness*. Note that, since SenticNet has four emotion different dimensions, $j \in \{Pleasantness, Sensitivity, Attention, Aptitude\}$.

To obtain the emotional characterization of each subtopic, we use the left singular vectors which include subtopic importance scores for each comment in following equation:

$$T_\alpha^j = \sum_{i=1}^{n} U_i^{(\alpha)} \mathrm{E}_i^j \tag{9}$$

where, T_α^j has the subtopic emotions for subtopic α, where $1 \leq \alpha \leq k$ for the SenticNet emotion dimension j; $U_i^{(\alpha)}$ is importance weight of i^{th} comment on α^{th} subtopic, n is the total number of comments in the subtopic; E_i^j is the aggregated discretized emotion vector for comment i and dimension j. The intuition behind this is that more relevant comments for a subtopic contribute more to the emotional information for the subtopic.

Finally, we normalize each T_α^j into $[0, 1]$ to compare the magnitude levels of emotions with each other.

5.2.3 Measuring Emotional Range within a Subtopic

Each comment has SenticNet word matching frequencies E_i^j. Recall that, each E_i^j has six discretized emotion frequencies. Since there are four different SenticNet

dimensions, there are 24 different emotions in total. Given a subtopic α and the set of comments belonging to the subtopic α, after normalizing the E_i^js, we take the standard deviations of each 24 emotions of the subtopic σ^{j_m} where m is emotions scores on dimension j , and m is $1 \leq m \leq 6$. If all σ^{j_m} for the subtopic α are less than a certain threshold, than we say that, subtopic α is emotionally consistent with all of its comments; otherwise, the subtopic needs further clustering. We use the standard deviation for emotions because a small standard deviation shows that the data points are closer to the mean of the data; whereas a high standard deviation shows the data is in a wider range. In our case, the scores are ranging in different emotions groups. So, small standard deviations indicate homogenous emotions over the comments.

5.2.4 Finding Emotionally Different Comments in a Subtopic

So far, we have obtained subtopics and comments. After obtaining the emotional ranges for all subtopic, chosen subtopics are "purified" as follows. First, we calculate the *emotional distance* of each comment in the subtopic to the *subtopic emotional centroid*. The furthest comments to the centroid are the most irrelevant comments of the subtopic in terms of emotions.

5.3 Experimental Evaluation and Results

5.3.1 Dataset

Reddit [65] is a web-based social media environment where people discuss news, internet content, or some specific ideas. Reddit has its own rules to organize discussions with the goals of (i) users staying on topic, (ii) no spams or advertising, (iii) no hate speech, and (iv) no personal attacks. Replies to all comments and replies are allowed which in turn

results in deep discussion trees. Reddit is the 11th most visited web-site in the US, and 25th in the entire world [70]. Reddit had 725 million comments in 2015 [71]. We have collected Reddit data via its API by using the Python Reddit API Wrapper (PRAW) tool [72]. Each article-based discussion has its own unique emotional status, and, thus, necessarily, our experimental results section discussions below are article-based. In the Experimental results section below, we have evaluated three article discussions, namely, Reddit discussions on the articles:

Table 5.1 - Most important Single Words and their importance scores for the first 6 Sub-Topics of Discussion A

Subtopic 1		Subtopic 2		Subtopic 3		Subtopic 4		Subtopic 5		Subtopic 6	
doctor	0.18	generic	0.35	generic	0.37	pharmacist	0.38	doctor	0.17	drug	0.22
drug	0.18	brand	0.3	market	0.32	free	0.35	know	0.15	pharmacist	0.21
generic	0.18	doctor	0.29	free	0.29	doctor	0.34	cost	0.14	cost	0.21
dont	0.16	name	0.26	brand	0.26	market	0.33	thats	0.13	free	0.18
get	0.15	pharmacist	0.19	name	0.23	work	0.22	epipen	0.12	market	0.17
people	0.15	prescription	0.09	epipen	0.16	capitalism	0.16	dont	0.12	doctor	0.16
pharmacy	0.15	ask	0.09	drug	0.13	know	0.11	people	0.08	price	0.15
pharmacist	0.15	patient	0.08	price	0.1	patient	0.08	drug	0.08	pharmacy	0.09
like	0.14	unless	0.07	competition	0.1	competition	0.07	im	0.07	patient	0.09
wa	0.14	adrenaclick	0.07	capitalism	0.09	ask	0.05	would	0.07	cv	0.08
one	0.13	know	0.07	adrenaclick	0.09	call	0.05	need	0.07	company	0.08
know	0.13	drug	0.06	patent	0.08	state	0.05	right	0.06	insurance	0.08
make	0.13	write	0.06	work	0.07	script	0.04	company	0.06	prescription	0.07
cost	0.12	written	0.06	cost	0.06	alternative	0.04	pharmacist	0.06	get	0.07
price	0.12	dispense	0.06	company	0.06	prescription	0.04	think	0.06	alternative	0.05
epipen	0.12	prescribed	0.05	fda	0.06	look	0.04	alternative	0.06	going	0.05
company	0.12	medication	0.05	government	0.06	advice	0.04	fda	0.06	without	0.05
cv	0.12	prescribe	0.05	epinephrine	0.06	prescribe	0.04	saying	0.05	epipens	0.04
work	0.11	script	0.04	product	0.05	required	0.04	system	0.05	go	0.04

1. Discussion A: "*CVS Puts Out Generic Competitor to EpiPen At A 6th of the Price*" [73].
2. Discussion B: "*LeBron: I'm coming back to Cleveland*" [74].
3. Discussion C: "*I made a chart showing the popular vote turnout in 2008, 2012 and 2016. Hillary didn't lose because the Republicans grew their base; she lost because the Democrats didn't come out to vote.*" [75].

5.3.2 Latent Subtopics

To fit the results on a table, we set k to six, which means first six subtopics are shown. Table 5.1 shows the most important words for each of the "first six most important subtopics" for Reddit discussion A. Recall that, the importance scores of these words are located in V'.

Observation 1: In the most important subtopic i.e., *subtopic 1*, as compared to the rest of the subtopics, importance weights of words start with lower weights and decrease slower.

Note that Latent Semantic Analysis assigns weights to comments, and a comment can belong to different subtopics. Since, the first subtopic has the highest eigenvalues (which means it is the most important subtopic), it is the union of the remaining most important subtopics. That is why words of *subtopic 1* also occur in other subtopics with higher weights. For example, the word "generic" has a score of 0.18 in *subtopic 1*, where it is 0.35 and 0.37 in *subtopics 2* and *3*, respectively, in discussion A. Similar results can be seen in discussion B and C (Appendix Table 1 and Appendix Table 3).

Observation 2: Bigram matchings are more informative in that they contain more semantically meaningful information as compared to single words. Examples are *medical*

advice, *intellectual property*, etc. In the bigram table (Table 5.2), one can observe the topic differences better. For example, for *subtopic 1*, 50% of the single words in a small sample, e.g., first 20 words, does not sufficiently identify the topic such as words “don’t”, “get”, “one”, “know”, “make”, etc.. In comparison, in the bigram table, 25% of the first 20 bigrams of *subtopic 1* is not informative (e.g., “*don’t know”, “don’t think”, “im saying, “pretty much”*, etc.). The first 6 subtopics of the bigrams table have more useful information than single words.

Table 5.2 - Bigrams Table for Discussion A

Subtopic 1		Subtopic 2		Subtopic 3		Subtopic 4		Subtopic 5		Subtopic 6	
brand name	2.37	brand name	3.81	free market	3.33	free market	3.42	dont know	0.84	free market	1.76
free market	1.64	name brand	1.68	brand name	3.27	dont know	0.63	dont think	0.51	insurance company	0.73
insurance company	1.41	active ingredient	0.61	name brand	1.41	brand name	0.62	im saying	0.44	pharmaceutical company	0.67
dont know	1.34	brand generic	0.61	generic drug	0.71	call doctor	0.6	pharmaceutical company	0.41	drug price	0.64
name brand	1.03	generic drug	0.6	active ingredient	0.63	market work	0.6	drug company	0.41	drug company	0.61
drug company	0.86	generic name	0.53	drug company	0.61	ask doctor	0.38	epi pen	0.4	call doctor	0.55
pharmaceutical company	0.82	name drug	0.5	brand generic	0.52	capitalism work	0.34	insurance company	0.33	new drug	0.45
dont think	0.81	write prescription	0.49	insurance company	0.51	name brand	0.33	dont need	0.32	get prescription	0.37
epi pen	0.78	call doctor	0.47	market work	0.49	doctor get	0.31	new drug	0.27	health care	0.35
call doctor	0.75	brand drug	0.45	name drug	0.48	pharmacist doctor	0.3	ask doctor	0.25	lower price	0.32
im sure	0.7	doctor prescribe	0.44	pharmaceutical company	0.47	medical advice	0.3	minimum wage	0.25	prescription drug	0.31
health care	0.68	generic alternative	0.42	generic alternative	0.44	doctor prescribe	0.27	call doctor	0.25	generic drug	0.29
drug price	0.66	dispense written	0.39	generic version	0.44	doctor office	0.26	drug price	0.23	independent pharmacy	0.28
active ingredient	0.66	prescription brand	0.39	generic name	0.42	perfectly free	0.26	doctor get	0.23	90 day	0.27
goodrx coupon	0.64	generic brand	0.38	brand drug	0.41	market action	0.25	dont even	0.22	company make	0.26
generic drug	0.64	doctor writes	0.37	drug price	0.41	pharmacy school	0.25	im sure	0.22	cost alternative	0.26
make sure	0.62	name epipen	0.36	fda ha	0.41	insurance company	0.24	make sure	0.21	ask doctor	0.24
minimum wage	0.61	prescribed generic	0.35	name epipen	0.38	pharmacist ha	0.23	doctor know	0.21	lower cost	0.24
im saying	0.6	slightly different	0.35	new drug	0.38	advice describe	0.22	lot people	0.21	health insurance	0.23
make money	0.59	ask doctor	0.34	generic epinephrine	0.37	doctor know	0.22	medical advice	0.21	rd cost	0.22
get prescription	0.57	dont know	0.33	slightly different	0.35	pharmacist dispensing	0.22	make sense	0.21	minimum wage	0.22

We then locate the most informative comments on each subtopic based on the scores of left singular vectors U in equation 7. For example, the highest weighted comment $\max(U^{(1)})$ in the first subtopic comment scores, named high 1, is the most informative

comment for the first subtopic. The following comments are the highest weighted (which we believe, are the most informative) comments for the first five subtopics:

- High 1:
    - *Why would a doctor write a prescription for just Epinephrine then? Does this happen regularly? I get that with some generics the active ingredient is the same but the other stuff that controls delivery/absorbtion rate can be slightly different. But surely the cheapest option with the same amount of active ingredient should be the default and prescribed in 99% of cases? Only prescribing the more expensive option if it's demonstrably better for that particular patient or if they specifically ask for it? I always thought that a non-generic was only superior for only a very small set of drugs and for people with particularly niche circumstances and those prescriptions should only be happening very very rarely. Not often enough for posts like this to exist. In the case they do need the brand name, the patient should know they are getting the more expensive drug and given a reasonable explanation why by their doctor. These warnings should not be necessary at all. Something must be fucking rotten. A doctor has a duty of care to their patients and that should include not fucking them over financially, especially when costs are this high. The justification being that it's not good for your mental health or physical health if you keep having to shell hundreds of dollars every time you need an epipen. Some people might even go without because of the high prices and put themselves at serious risk. We're not talking about the generic 30p paracetamol vs the brand name one that costs a few dollars. This hundreds of dollars, for a lot of people this price difference (even with insurance with deductibles/excess) is enough to cause some *serious* hardship. Even if they default to the expensive and change it if you ask for it, not everyone is educated about this and it's still fucked up. I would even go as far to say that it was just straight up corruption as their must be some kind of kickback involved for a doctor to do that.*
- High 2 (first two most informative comments on *subtopic 2*):
    - *Part of the issue is that, in most (all?) other cases, even if the doctor writes the prescription for a brand name, the pharmacist can, and probably will, give you the generic, unless the doctor checks "dispense as written."*
    - *Almost all doctors do. In my experiences, when there is a brand drug that has a generic alternative, 98-99% of the time, the generic is the drug that*

is prescribed/dispensed at the pharmacy. Even if a doctor prescribed 'Crestor' on their prescription slip, if they do not note to the pharmacist to DAW (dispense as written), the pharmacist has the right to give the patient the generic alternative. Sometimes doctors will require the pharmacist to dispense the exact brand they prescribe, and sometimes patients insist on the brand drug instead of the generic for whatever reason, but it's very rare.

- High 3 (the first two most informative comments on *subtopic 3*)
    - *The free market at work.*
    - *Competition + free market + ????= profit*
- High 4
    - *Oh look. The free market at work.*
- High 5 (first two most informative comments on *subtopic 5*)
    - *thats really good to know. I didnt realize they dont have a kids dose. I will have to bring this up to the doctor*
    - *I'm a doctor, I didn't know that. Dammit. It is possible that your doctors do not know this either.*

Observation 3: The most informative comments are mostly distinct and about different aspects of discussions.

For example, *Subtopic 2* discusses if doctors prescribe a brand name; *Subtopic 3* is about free market. Using these most informative comments (not only the first one, but the top *k*) may be useful in identifying subtopics in a discussion, instead of single words or bigrams, even though bigrams are most informative than single words.

Observation 4: Some comments belong to the same subtopic even though they are under different discussion trees and comment levels.

For example, first two comments belong *subtopic 2*, but they are under different discussion trees. So, bringing them together for the benefit of readers.

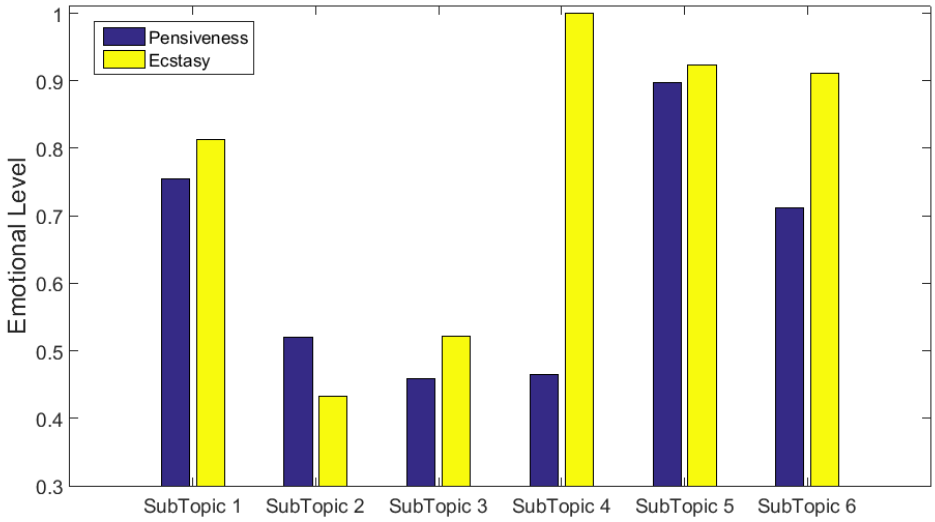

Figure 5.1 – Emotional Difference over the first six Subtopics

Observation 5: Some subtopics are not uniquely identified.

We see that *subtopic 3* and *4* have very similar properties (single words, bigrams, and comments) and they share the same comments, which means these subtopics are slightly different from each other. These close relationship can be identified and merged to find more unique subtopics.

Observation 6: In *subtopic 1*, importance weights of comments start with lower weights and decrease slower than rest of the subtopics.

This observation supports the *Observation 1*, which is that *Subtopic 1* covers rest of the subtopics. For example, the most informative comment of *subtopic 1* also has a similar score to *subtopic 2*. Recall that *subtopic 2* mentions about writing prescription and the highest weighted comment of *subtopic 1* also mentions about writing prescription with longer paragraphs.

5.3.3 Emotion Difference within Subtopics and Purifying the Subtopics

To extract the emotional differences between subtopics, we use T_{α}^{j}'s as we introduced in section 3. We look at each of the twenty-four different emotion levels in each subtopic, and we observe that most of the emotional levels are slightly different within subtopics. Our proposed algorithms can find emotions having significant differences on subtopics. In Figure 5.1, there are two emotions having different values in the same article discussions.

Observation 7: As subtopic changes, there are some emotional differences.

Since subtopics have common words, most of the emotions remain the same. For example, for discussion A 22 out of 24 emotions have very close levels. However, as seen in Figure 5.1, magnitude levels of emotions have different values within the subtopics, e.g., *Subtopic 4* has stronger "*Ecstasy"* emotion than *Subtopics 2* and *3*. This result is unexpected, because we find that subtopics 3 and 4 are not uniquely identified by LSA, but emotional levels show the difference between these topics.

Below, we have sorted comments by their emotional levels within the subtopics. We call this process "emotional purification".

Observation 8: Emotionally sorted comments show different importance results than LSA.

The following comments are those that are emotionally closest to their subtopic's emotion centroids.

- Emotionally two closest comments to the centroid of *subtopic 1*:
    - *"I honestly don't know many specifics. I just generally know there are tens of thousands of medications out there and it's the Pharmacists job to distribute them to you safely and, to some degree, cost effectively. I know*

*that Pharmacists go to 6+ year graduate programs, have to pass state licensing exams and be board certified, and make a median salary of 121,000 a year. There's a reason for those requirements and they do earn that salary. These people are a serious, experienced, knowledgeable part of your healthcare. I'm pretty sure you can request to discuss your medications with them if you do have any questions or concerns about your existing medications or possible alternatives. I really do think more people should take advantage of that available resource. I don't mean to say that your Psychiatrist or General Practitioner wouldn't know about Drug X deliberately somehow. Or that Drug Y they prescribe wouldn't treat whatever medical issue you have effectively. EpiPens, for example, do their job fine. They're just expensive. Doctors don't have the time to keep up with every new drug or generic available and what rules they meet, so they're subject to a bias of prescribing the medications they know of without knowing if there are alternatives available or the rules about them. Your Psychiatrist has to spend a lot of time keeping up with the constant changes to the DSM for example, so he can accurate diagnose and treat you for things. That's his part of the job. He needs a general familiarity with the prescription treatments available and their effectiveness, but it's just not his specialty. I'm just saying that the pharmaceutical industry is *hella* freaking complicated and we've got an entire specialized field of medicine dedicated entirely to navigating it. That's your Pharmacist's job. Mostly it bothers me when people treat a field like this too causally or don't recognize them for what they really do and know. People make that mistake with a lot of careers they don't know enough about."*

- *"Another Edit: First thank you to whoever gave me gold! Second, /u/Jahodac is right with pretty much everything he said here (https://www.reddit.com/r/news/comments/5njj2x/cvs_puts_out_generic_competitor_to_epipen_at_a/dcclw19/). Thank you for bringing all that up. There is a manufacturer discount from Mylan as I mention below that can save up to $300 (https://www.activatethecard.com). It doesn't work if you don't have insurance and personally for me, it only took $100 off but perhaps it was going towards my huge ass deductible. Some redditors have said that as the EpiPen generic is not an A/B generic, its injector is different and perhaps more difficult to use, so it may be worthwhile to get your PCP to prescribe both medicines for you in order to decide what you want. As far as the generic goes, if you download the coupon from the link below, it says out of pocket you can save $300, and since its cash price is only $109, the coupon should make it free for literally everyone. And there are programs like the $4/$10 deal that Walmart and others do. I know*

CVS/Target uses ScriptSave, and for some maintenance meds it brings it down to $4. Tl;dr: Read the links. Edit: I found this using the GoodRx app. This is one of the most versatile tools at the hands of patients, doctors, and pharmacy workers alike in keeping drug prices in check and it's FREE! I use it everyday I work whenever I see expensive drugs, and it helps like 80% of the time. It is NOT insurance, and the downside of using its discounts is that your payments won't go to your deductible, but if you're like me, and you'll never reach your deductible, you should give it a shot and compare it to how much insurance will pay. While it doesn't tell you how much your insurance will charge you, it does tell you generally the cheapest(Walmart, although I have heard from some coworkers that Walmart doesn't take GoodRx at all) and most expensive pharmacies (lol CVS most of the time) and provides its own discounts that sometimes beat insurance especially if you have a deductible. The Tips tab on GoodRx can help you find if there are manufacturer coupons you can sign up for. Manufacturer coupons only exist for brand name medications and typically only cover 30 or 90 day supplies, and the 90 day supply is almost always cheaper. GoodRx is accurate almost all the time in displaying coupons, but it isn't for the EpiPen, which just discontinued its full coverage program, instead replacing it with one that could "potentially" save you $300. Seriously, screw Mylan Pharmaceuticals. I'm deathly allergic to nuts and legumes so I have to have an epipen and I had to pay $535 out of pocket so this issue is personal to me. CVS Pharmacy Tech here! To clarify you have to get a prescription for ADRENACLICK OR EPINEPHRINE in order for us to fill the medication. There is also a patient assistance program that will allow you to receive Adrenaclick for free by the manufacturer Impax Laboratories. Here is the website (http://www.epinephrineautoinject.com) Present us the coupon at drop off with your prescription pretty please!"

- Emotionally two closest comments to the centroid of *subtopic 2*
    - *"Reporting from Argentina. Same here."*
    - *"Actually, the U.S. is similar! Unless you ask for a name brand, the doctor will prescribe you a generic unless they do not think it would be a good fit for you. It's usually at the doctor's discretion. The only difference (and it's a huge difference) is that some people do not have insurance so a generic is still insanely expensive for them."*
- Emotionally two closest comments to the centroid of *subtopic 3*

- *"'> Mylan is already a generic drug company.' But they own the patent on the brand name EpiPen delivery system. That is not a generic device/drug. Of course the Epinephrine is cheap as balls, but it was hard to replicate the auto-injector to get the consistent results of the EpiPen, IIRC.*
- *"I would have the doc specifically write for the generic just to be safe. I know one of my old neurologists used to specifically write for the brand name. Only when I switched to a new neuro did I start getting the generic and realize its the same thing for way less money."*

- Emotionally two closest comments to the centroid of *subtopic 4*
  - *"We know plenty about prescriptions. We are required to take some pharmacology training on top of the standard dispensing training. It is not allowed for me to answer your questions. If you ask me if taking Advil would work for a headache I have to have the pharmacist answer that question. Stupid but I assume it's to limit liability."*
  - *"'> Pharmacists should not be dispensing medical advice as you describe it'. 8 years of education to become a pharmacist, and most take additional years on top of that before they start doing compounding, and you want them to just count out pills?"*

Again, the difference between *subtopic 3* and *4* can be seen by looking their emotional comments. *Subtopic 3* talks about generic version of the drug, where *subtopic 4* talks about pharmacists. Adding emotions to the subtopics may lead more distinction over the subtopics.

5.4 Conclusions

In this chapter, we have studied to extract subtopics of social media comments starting around a topic; and analyze their differences. Our proposed method provides finding subtopics without a main article comparing our previous study. For this purpose, we collect several social media discussions including at least 3000 comments from Reddit.

We use and modify Latent Semantic Analysis to figure out the subtopics, their characterization, and their emotional behavior.

First, we find the most informative words, bigrams, and comments of each subtopic. These information maintain the subtopic characterization. One can eliminate or promote certain subtopic comments by looking subtopics' content.

Finally, we look their emotional levels and how much different they are from each other. We see that there are emotional differences between the subtopics and we use them to distinguish similar subtopics found by LSA. Also, we sort the comments in a subtopic according to their emotional status.

The observations show that our methods will be useful for the social media users to bring similar comments and discussions in order to extract and classify the information among thousands of comments.

6 Network Based Model of Social Media Big Data Predicts Contagious Disease Diffusion

6.1 Introduction and Background

Approximately 5-20% of U.S residents are affected by the flu (influenza) annually [2] and 200,000 of these are hospitalized for illnesses related to the flu [76]. This number is following an increasing trend [77]. If a hospital is ill-prepared for a rush of patients from influenza-like illnesses (ILI), this could poorly affect the care patients receive. It is crucial for hospitals to obtain estimates pertaining to the number of patients they will likely receive in order to maintain an adequate supply level to care for each patient.

In this chapter, we utilize data collected from social media to address a novel problem, where we investigate how diseases are spreading geographically.

Our contributions are two-fold. First, while extensive work has been presented utilizing time-series analysis on single geographies (e.g. large cities, regions or countries) [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] or post-analysis of highly contagious diseases [44], no previous work has provided a generalized solution to identify *how* contagious diseases diffuse across geographies, such as States in the USA. Secondly, due to nature of the social media data, various statistical models have been extensively used to address these problems. In this chapter, we present a new network-based approach to model disease activity across geographical locations.

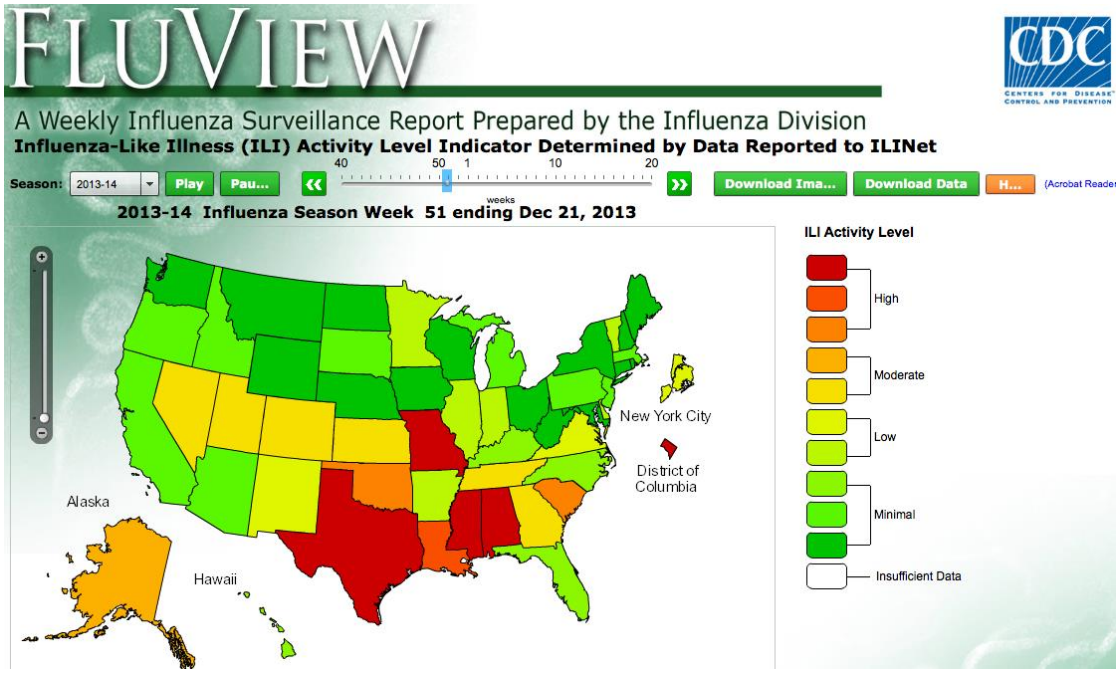

Figure 6.1 - A screen shot of CDC FluView [27] Influenza Surveillance System. CDC Levels are ILI activity levels indicating the prevalence of ILI in a particular state for a given week.

We use ILI as an example to make predictions about *future* ILI cases spreading across the United States. With enhanced forecasting about how outbreaks spread, we will improve upon disease control and prevention efforts.

Throughout this chapter, we talk about CDC levels. The levels we are referring to are the weekly ILI activity levels from ILINet (Figure 6.1) (1 meaning there is hardly any ILI activity occurring and 10 meaning ILI activity is extremely high).

Motivation for Modeling Disease Diffusion

In this chapter, we utilize data collected from social media to address a novel problem. We investigate how diseases are spreading geographically. We use ILI as an example to make predictions about future ILI outbreaks spreading across the United States.

While earlier studies focus solely on the number of ILI cases that arise, our work focuses on the spread of ILI and how the number of ILI cases in one geographical location influences, or is influenced by the ILI activity from surrounding geographical locations. We present a network model to analyze social media big data to generate predictions about future ILI activity based on geographical influences and how ILI diffuses across locations.

Figure 6.2 and Figure 6.3 illustrates why our particular focus is essential for understanding the complexity of ILI activity. Figure 6.2 shows cumulative ILI activity in the United States. The curve, which represents the national projection of weekly CDC levels, is calculated by multiplying each state's weekly CDC levels by its state population, and summing those values for all states each week. Regional ILI activity, represented by the box plots of state CDC levels, is using weekly CDC levels. Based on the same peaks and dips each week in the two graphs in Figure 6.2 and Figure 6.3, CDC levels align well with ILI activity, which is measured by using positive influenza lab tests

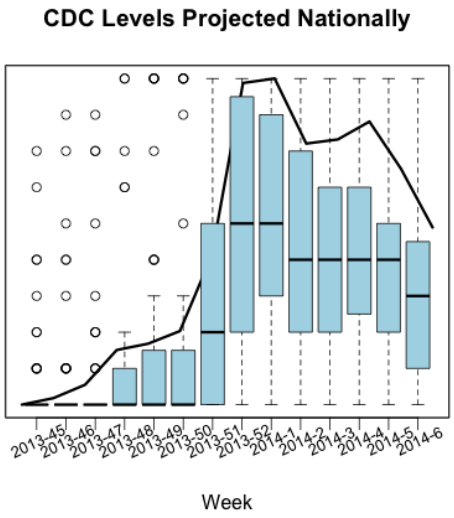

Figure 6.2 - The curve in graph A represents CDC levels projected to the national level. We multiplied the levels by their respective state populations, summed the values for all states for each week, and then divided those sums by the total population of the nation. The box plots represent the different CDC levels that each state had per week, which shows what the CDC levels looked like at the state level.

reported to CDC. But regionally (the box plots from Figure 6.2), there is much variation in ILI activity. By examining states in a geographical context and by looking at how ILI activity is spreading across the nation, we will have a better understanding as to where an increase in ILI activity will occur next.

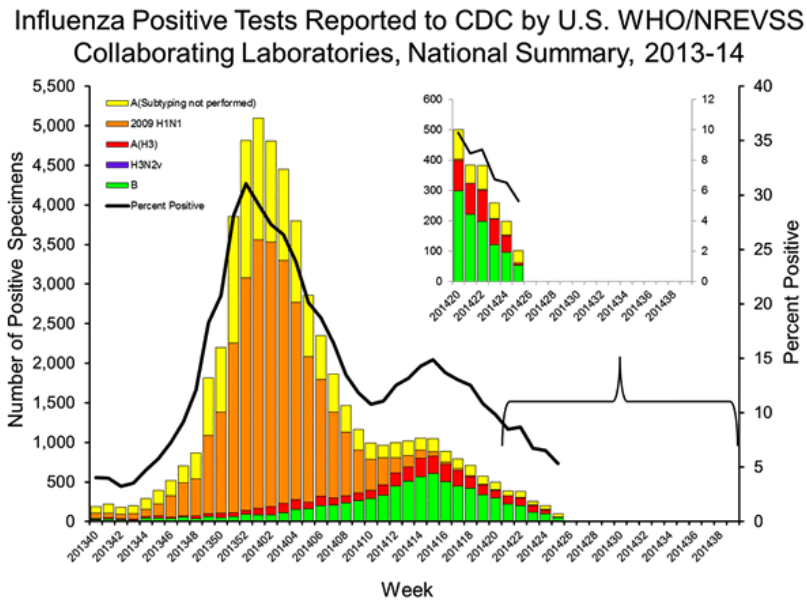

Figure 6.3 - A screen shot from the CDC FluView [27] surveillance system. The black curve shows influenza activity at the national level.

6.1.1 Social Media Datasets

In this study, we utilized Twitter (`www.twitter.com`) tweets. We selected Twitter as our main data source because of its public availability and ease of access. However, the approaches described in this chapter can be applied to other social media data, as well. Twitter provides two application programming interfaces (APIs): the representational state transfer (REST) API for historical twitter data and the Streaming API [78] for accessing tweets posted in real-time. In this study, we have utilized both interfaces to collect our data.

6.1.1.1 Content Filtering

To obtain our dataset, we first created an inclusion dictionary and an exclusion dictionary. The dictionaries were created by researching influenza-related terms from several sources [79] [80] [81] [82] [83], and by extensively researching through collections

of tweets. Twitter mainly contains colloquial, informal content, so we had to translate terms we obtained from more formal sources into their corresponding informal versions. For example, in the U.S., people are more likely to tweet about having a *headache* than they are to tweet about having *cephalgia*. In the inclusion dictionary, we included 11 key words or phrases that indicated the tweet content was flu-related, such as "fever"', "headache", "runny nose", "stuffy nose", "flu", etc. The exclusion dictionary included 63 key words or phrases that may be accepted by the inclusion dictionary, but should not be included in our database, such as "sick and tired of" and "Bieber fever".

Typically, the flu season in the United States begins in winter months, and peaks in January or February [84]. We collected tweets with their associated handles and dates from November 3, 2013 to February 8, 2014, which is right in the middle of the flu season in the United States. We filtered the tweets using both the inclusion and exclusion dictionaries.

In this study, our focus was developing a model to effectively predict upcoming disease activity based on how disease outbreaks spread across geographical locations. While natural language processing based approaches can produce a higher quality dataset of tweets [30] [32] [34], we chose to use a similar approach to Hirose and Wang [35] by filtering our collection of tweets with dictionaries to focus on the modeling aspects. We validated our data collection process by randomly selecting 5,000 tweets and manually reviewing their relevance to the flu. We observed that ~70% of the tweets were truly flu-related. This was sufficient for our purposes, and allowed us to focus our efforts on the model.

6.1.1.2 Location Filtering

Only a small percentage of tweets contain GPS location data, so we used user-specified biographic information to determine the locations associated with tweets. We extracted out the locations for each handle associated with our collected tweets. However, we had to take into account that many Twitter users input aliases for their locations. For example, “Cali” for California, “Philly” for Philadelphia, and “Bmore” for Baltimore etc. were used.

Finally, we queried the data to get the number of tweets each state had per week. An illustration of our entire data collection process is shown in Figure 6.4.

6.1.1.3 Comparison of Social Media Data with CDC ILI Activity

In order to ensure that the weekly tweet counts we have collected are representative of ILI activity, we compared our data to data from the CDC's influenza surveillance system, FluView [27] . We mapped our data to CDC levels by using a similar procedure to the one used for FluView. First, we had to normalize the data, similar to how CDC takes into account state variability. We made projections for the number of Twitter users per state by

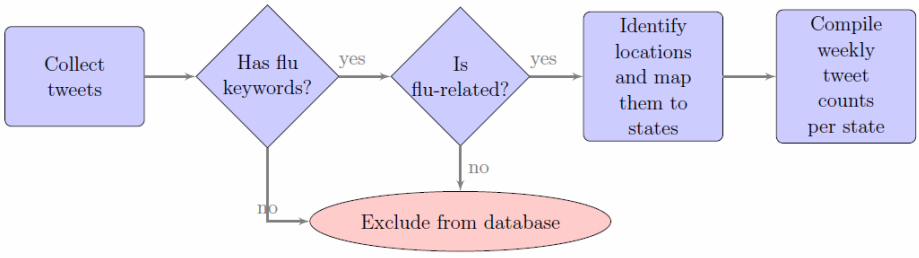

Figure 6.4 - A flow chart illustrating the process involved in obtaining our dataset of relevant tweets.

using the number of internet users per state as of June 2010 [85] and the percentage of Americans who were Twitter users in 2010 (7%) [86].

These statistics are the most current statistics that we had available to us. We were able to utilize our projections of the number of twitter users per state as described below because both statistics came from the same year, they provide strong approximations for the number of twitter users per state in 2010. According to census, households with a computer and internet has stayed mostly unchanged between 2010 and 2012 (76.7%-78.9%) [87]. Because there is not much change in the proportion of twitter users per state (Twitter had 49 million monthly active users on average during the second quarter of 2013, only up 2% from the previous quarter [88]), we can assume that the proportion in 2010 approximately stayed unchanged. We utilized weekly tweet counts by normalizing the data using the following equation:

$$T_{state,norm} = \frac{T_{state,orig}}{p_{usa} * I_{state}} * 100 \quad (10)$$

where $T_{state,norm}$ is a state's normalized weekly tweet count, $T_{state,orig}$ is a state's original weekly tweet count, p_{usa} is the percentage of Americans who were Twitter users in 2010 (which we stated above to be 7%) and I_{state} is a state's total number of internet users. Realistically though, the number of Twitter users is not the same percentage of the number of Internet users for each state within the United States, and the individual change in various states are not the same either. So, we adjusted the percentages to minimize the differences per state between CDC data and our data to get this final normalization formula:

$$T_{state,norm} = \frac{T_{state,orig}}{p_{state,adj} * I_{state}} * 100 \quad (11)$$

where $p_{state,adj}$ is the adjusted percentage of Internet users who are Twitter users per state. Next, we mapped the normalized data to disease levels similar to ILINet.

6.1.1.4 Influenza Like Illness (ILI) Levels

For a given time point, we determined the mapping from normalized data ($T_{state,norm}$) to CDC levels by using a similar method to that used by the CDC. We first calculated the mean and standard deviation of the normalized data. We then assigned a constant to the mapping calculations, which controls how wide or narrow the mapped levels are in regards to the range of weekly tweet counts that get assigned to each level. Level assignment was based on the following:

$$level = \begin{cases} 1, & T_{state,norm} \leq \overline{x} - (s * c) \\ 2, & T_{state,norm} \leq \overline{x} \\ 3, & T_{state,norm} \leq \overline{x} + (s * c) \\ 4, & T_{state,norm} \leq \overline{x} + (2 * s * c) \\ 5, & T_{state,norm} \leq \overline{x} + (3 * s * c) \\ 6, & T_{state,norm} \leq \overline{x} + (4 * s * c) \\ 7, & T_{state,norm} \leq \overline{x} + (5 * s * c) \\ 8, & T_{state,norm} \leq \overline{x} + (6 * s * c) \\ 9, & T_{state,norm} \leq \overline{x} + (7 * s * c) \\ 10, & otherwise \end{cases} \quad (12)$$

where $level$ is the calculated level that corresponds to CDC levels, $T_{state,norm}$ is a state's normalized weekly tweet count, $\overline{x}$ is the mean of all of the data, s is the standard deviation of all of the data, and c is constant that controls the width of the range of tweet counts to be assigned to each level (which is crucial since our normalization calculations result in a

much smaller-scaled dataset than the CDC dataset prior to level-mapping). We utilized $c = 0.15$, which maximizes the overlap between levels predicted and CDC data. The only value that varied among states was $T_{state,norm}$; the rest remained constant.

Next, we compared the tweet-based ILI activity levels (calculated with Equation 12) to CDC levels to measure how accurately social media data replicated that of the CDC. The mean of the differences was 0.05. After conducting a bootstrapping analysis, we found that a 95% confidence interval on the mean of the differences for all states was $(-0.162, 0.277)$. Moreover, the two one-sided test (TOST) for equivalence [89] also showed that tweet-based ILI levels are highly similar to the CDC data ($\varepsilon = 0.3, p < 0.049$).

6.2 Experimental Evaluation and Results

6.2.1 Modeling Disease Diffusion

We hypothesize that disease diffusion across geographical locations can be predicted by modeling social media activity. Specifically, we focus on how diseases such as ILI are spreading across the United States by observing tweets. In this study, we developed a variation of the Threshold Model of Diffusion in Networks [90] to predict flu activity across the United States.

6.2.1.1 Diffusion Model for ILI

The Threshold Model of Diffusion utilizes a weighted, directed graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. Each node has an assigned threshold, and when $\sum_{v \in Adj(u)} w_{vu} \geq t_u$ for w_{vu} being the weight of edge (v, u) and t_u being the threshold of node u, u becomes active. If $\sum_{v \in Adj(u)} w_{vu} < t_u$, u remains inactive. Initially,

some nodes are active. The weight of each edge $(v, u) \in E$ represents a factor indicating the influence node v has on node u. The idea of this model is that activations spread, or diffuse, throughout the network.

Our variation of this model tailors the network representation more towards the spread of ILI as it relates to social media data. We examine the idea that geographical locations, which are states in our example, are a factor in the spread of ILI. In a graph $G = (V, E)$ of the United States, where V is the set of nodes and E is the set of edges, there exists a node for each state, and there exists two directed edges between each neighboring state (Figure 6.5.A).

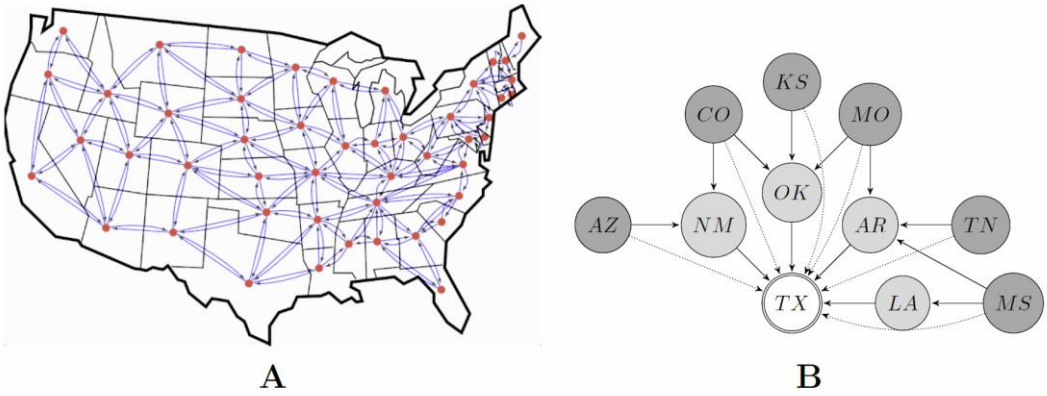

Figure 6.5 – **A**. An illustration of the planar state neighborhood network (where the network can be drawn without having any edges overlap) for the lower 48 states of the United States (which excludes Hawaii and Alaska) is shown. **B.** We expand upon network A to produce a non-planar network, where edges are added for each 2-hop neighbor of a state in order to obtain a more detailed idea about how ILI is spreading around each state. For example, in B, we added edges from 2-hop neighbors of Texas (TX) to Texas (illustrated by the dotted edges). Note that B serves as an example, and does not show every node or edge in our network.

We have expanded our model from Figure 6.5.A to also include 2-hop neighbors so that we can gain a better understanding about how diseases diffuse about on geographical locations. These additional edges, would increase the model accuracy by accounting for possible influential geographical neighbors 2-hops away, the epicenters of influenza, that might be spreading flu rapidly and causing other geographical locations to experience a rise in disease activity [91]. Some added edges are illustrated in Figure 6.5.B as the dotted edges, which connect 2-hop neighbors of Texas (TX) to Texas. The version of this expanded network that we use in our model includes directed edges between the lower 48 states of the United States and their 2-hop neighbors. In our model, we have two directed edges between every (state, 1-hop neighbor) and (state, 2-hop neighbor) pair: an incoming edge and an outgoing edge. The direction of the edge signifies how diseases are spreading and which states are influencing that spread.

Using our Twitter data, our weekly tweet counts demonstrated a linear correlation (using 0.7 as a threshold for the Pearson correlation coefficient) for 80% of all 105 state-neighbor relationships; the average Pearson correlation coefficient among all 105 pairs was 0.78. Because of this strong relationship, we used linear regression models to calculate edge weights for our network. For $(v, u) \in E$,

$$w_{vu} = \left(T_{v,t} * m\right) + b \tag{13}$$

where w_{vu} is the weight of directed edge $(v, u) \in E$ that goes from v to u, $T_{v,t}$ is the most-recent weekly tweet count for node v (which is the node that influences node u), and m and b are the respective slope and intercept of the best-fit linear regression equation that uses v to determine u. Note that, we calculate m and b by using the most-recent weekly tweet counts (using last 8 weeks) until week t. Our network is dynamic, since our variables

change on a weekly basis. The best-fit linear regression equation takes into account all of the weekly tweet counts between the two states corresponding to v and u, while $T_{v,t}$ is the latest weekly tweet count of v.

6.2.1.2 Predicting Upcoming Social Media Activity

We have designed the following steps to obtain predictions on ILI related social media activity:

1. Define graphs G_1 and G_2 from networks A and B respectively from Figure 6.5 as $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$. In other words, G_1 is the planar network representing the lower 48 states of the U.S. geographically, and G_2 is the network derived from G_1. Both graphs have the same set of vertices, V. However, $E_1 \subseteq E_2$ because E_2 contains edges corresponding to both 1-hop and 2-hop neighbors of V in G_1.
2. Using Equation 13 or edge weights, calculate the unscaled predicted tweet counts, $P_{u,t+1}$, using the most-recent weekly tweet counts until week t, using the following equation: $\forall u \in V$,

$$P_{u,t+1} = \frac{\sum_{(v,u)\in E_1} w_{vu}}{n} + \frac{\sum_{(v,u)\in E_2 - E_1} w_{vu}}{m} \quad (14)$$

where $P_{u,t+1}$ is a state's unscaled predicted tweet count for the upcoming week $(t+1)$, $(v,u) \in E_1$ refers to all incoming edges for state u that correspond to 1-hop neighbors, n refers to the number of incoming edges for state u that correspond to 1-hop neighbors, $(v,u) \in E_2 - E_1$ refers to all incoming edges for state u that correspond to 2-hop neighbors, m refers to

the number of incoming edges for state u that correspond to 2-hop neighbors, and w_{vu} is the weight of edge (v, u).

3. Now find c_s, the scaling factor. c_s is crucial for our calculations because adding $\frac{\sum_{(v,u)\in E_1} w_{vu}}{n}$ and $\frac{\sum_{(v,u)\in E_2 - E_1} w_{vu}}{m}$ causes the predictions to be greater than the actual levels. However, it is necessary to add both of those terms to obtain more detail about how outbreaks are spreading geographically. c_s is the ratio of the average of weekly unscaled predictions, $P_{u,t+1}$, for all states to the average of the actual weekly tweet counts for all states. The larger c_s is, the more influence geographical surroundings had on a location's disease activity in that particular time frame. c_s ranges between 1 and 2 per dataset. For example, the value of c_s for our collected data from the 2013-2014 season was 1.902.
4. Once c_s is found, find each of the lower 48 states' scaled predicted tweet counts using the following equation: $\forall u \in V$,

$$T_{u,t+1} = \frac{P_{u,t+1}}{c_s} \tag{15}$$

where $T_{u,t+1}$ is a state's scaled predicted tweet count (the value used for the actual predictions), $P_{u,t+1}$ is a state's unscaled predicted tweet count for the upcoming week (using Equation 14), and c_s is the scaling factor. $T_{u,t+1}$ is what we use for the actual predictions.

Since we already concluded that our tweet counts aligned well with CDC data in Section 6.1.1.3, the next step was to ensure that our predictions aligned with our tweet

counts. The mean of the differences between tweet counts and our predictions for each week was -0.00393, and the 95% confidence interval was $(-40.002, 39.99)$, which we found from a bootstrapping analysis (Figure 6.6). A TOST equivalence test with $\varepsilon = 40$ was also significant ($p < 0.025$).

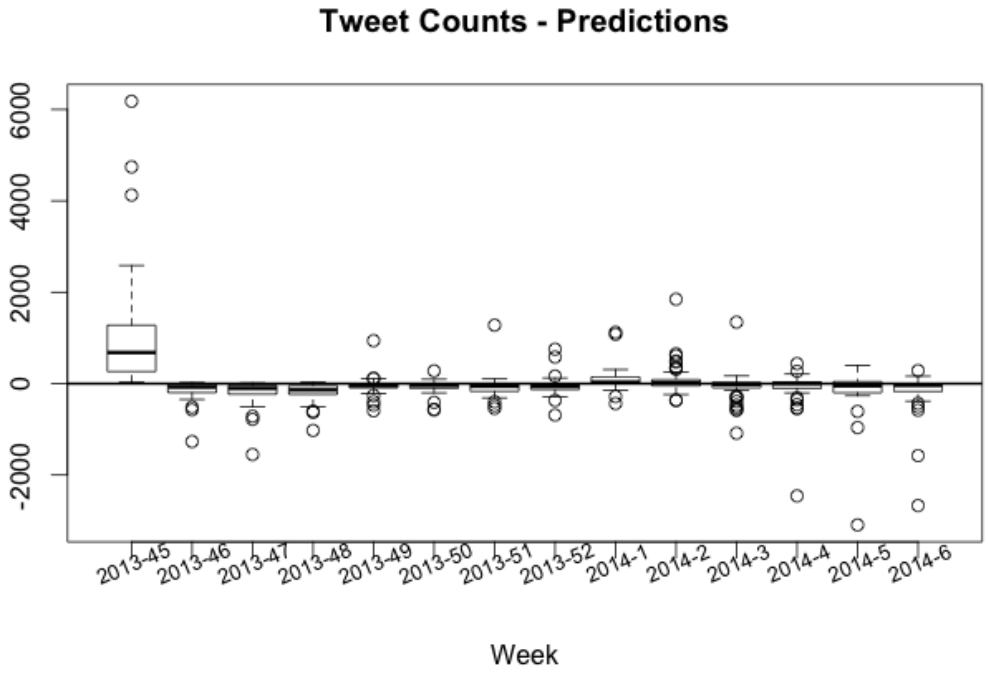

Figure 6.6- A box plot of the differences between tweet counts and our scaled predictions for each state per week. The mean was -0.00393 and the 95% confidence interval was $(-40.002, 39.99)$, which provides strong evidence that the mean of the differences is approximately 0.

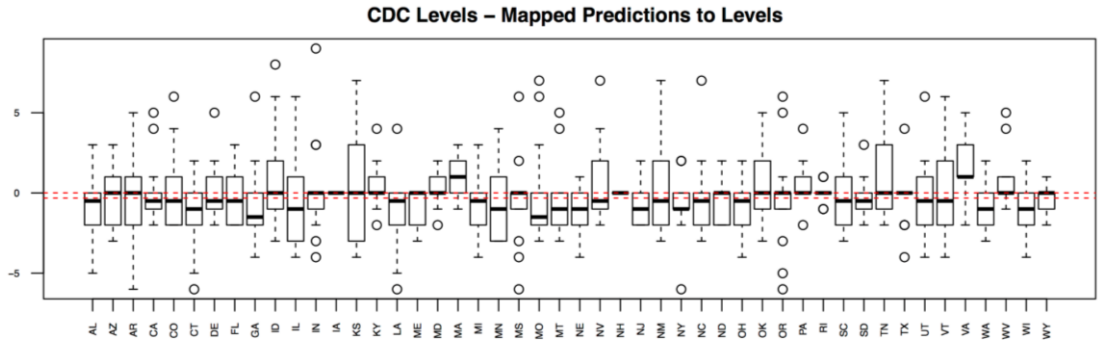

Figure 6.7 - A box plot of the differences between mapped levels associated with predictions and CDC levels. The two horizontal dotted lines represent the 95% confidence interval (-0.319, 0.006) for the mean of the differences, showing that the predicted levels align well with CDC data.

In Section 6.1.1.3, we showed that counts of ILI-related tweets align well with CDC data. Next, we wanted to compare our predictions directly with CDC data to make sure our predictions correctly portray ILI activity. In order to do this, we needed to first map our predictions to comparable CDC levels. We did this by using Equations 11 and 12, with our predictions in place of tweet counts. The mean of the differences between CDC levels and mapped levels associated with predictions for each state was -0.15. We found a 95% confidence interval on the mean of the differences to be $(-0.319, 0.006)$ from a bootstrapping analysis (See Figure 6.7). We also conducted a TOST equivalence test which showed significant overlap ($\varepsilon = 0.3, p < 0.04$). These results provide strong evidence for us to conclude that our ILI level predictions based on this model align with CDC levels.

6.2.2 Modeling Airline Traffic with Effective Distance

Brockmann and Helbing [44] have studied how contagious diseases spread all over the world by looking airport traffic instead of geographic distances. They utilized all flight

information from OAG (Official Airline Guide) [92] and created the effective distance of all cities all over the world by creating matrix P, the fraction of the passenger fluxes.

$$P_{mn} = \frac{F_{mn}}{F_n} \tag{16}$$

where $F_n = \sum_m F_{mn}$ and F_{mn} is the direct air traffic from node n to m which makes $0 \leq P_{mn} \leq 1$.

$$d_{mn} = (1 - \log P_{mn}) \geq 1 \tag{17}$$

where d_{mn} is the effective distance from a node m to node n. The idea behind this equation is a small fraction of air traffic makes a greater distance from two places. Finally they take the all pairs shortest path to calculate effective distance of all pairs.

$$D_{mn} = \min_{\Gamma} \lambda(\Gamma) \tag{18}$$

where $\lambda(\Gamma)$ is an ordered path $\Gamma = \{n_1, \dots, n_L\}$ as the sum of effective lengths along the legs of path. They used this table to predict the original place of 2009 H1N1 influenza pandemic and 2003 SARS epidemic. Also, they can predict the day of virus arrival to any airport in the world, once the outbreak location is given.

6.2.2.1 Predicting ILI with Effective Distance

We wanted to compare the effective distance model introduced in [44] with CDC levels to see how important air traffic is to predict ILI activity. In this model, we add an edge between states that have significant air traffic. We have not considered international flights because there is no synchronized flu levels throughout the world.

$$ED_{mn} = \begin{cases} 1, & if\ D_{mn} < \psi \\ 0, & Otherwise \end{cases} \tag{19}$$

Since, one of our goal is to see how much air traffic affects flu dissemination with the neighborhood model, we have created the effective distance of all states by using real passenger numbers from United States Department of Transportation [93] database. Next, we take the states that has effective distance less than a certain threshold ψ in ED_{mn}. Finally we are predicting the flu level by using Equation 20:

$$P_{u,t+1} = \frac{\sum_{(v,u)\in ED_{vu}} \frac{w_{vu}}{\log D_{vu}}}{r} \tag{20}$$

where r refers to all incoming edges for state u that corresponds to certain states have significant air traffic by looking Equation 19.

We have conducted TOST for equivalence [89] to see how different ψ would effect the CDC predictions, and varied the ψ across all ED values (Figure 6.8).

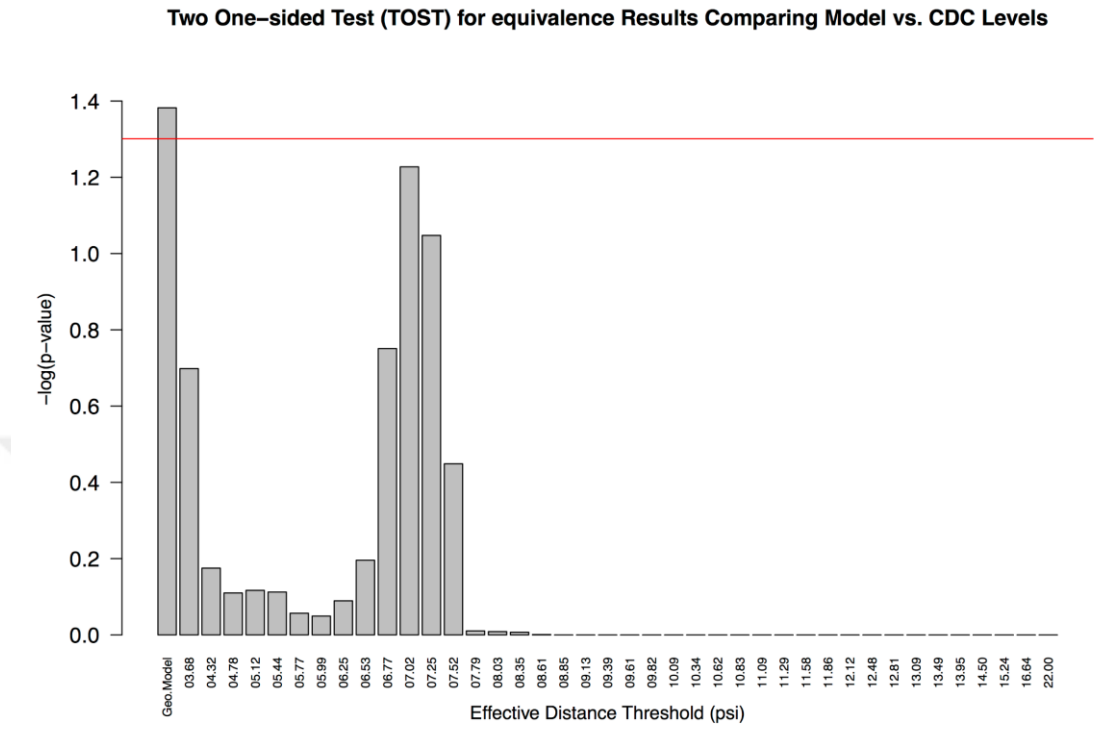

Figure 6.8 - A bar plot of the TOST for equivalence results across all ψ is shown. The horizontal red line is the significance threshold for 0.05. All effective distance models failed, whereas the Diffusion Model based on geographies produced significant results as shown in Figure 6.7.

6.2.2.2 Effective Distance only Model:

We first utilized the ED model to compare how important air traffic is to predict flu outbreaks. We have created the ED of all states by using real passenger numbers from United States Department of Transportation [93] database. This model was unable to generate significant predictions for ILI activity Figure 6.8. This was expected, as ED model was unable to find the spatial origin on a country resolution before [44]. Hence the Diffusion Model (Section 6.2.1.2) performed better than the ED model.

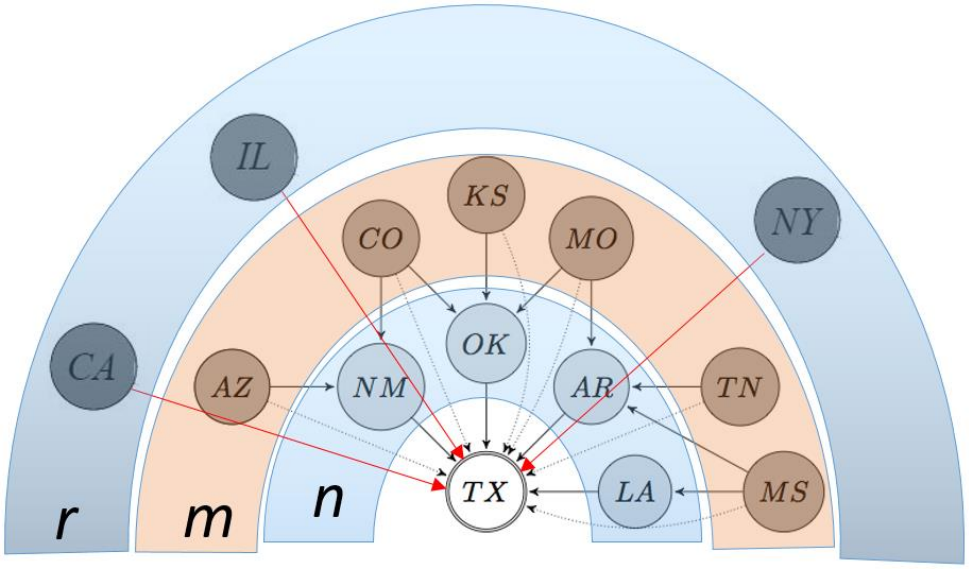

Figure 6.9 - Plugging Airline traffic into our Network Model, where n is the number of neighbors, m is the number of 2- hop neighbors, and r is the number of "important" airport neighbors (Equation 22).

6.2.2.3 Diffusion and Effective Distance Mixed Model:

We also introduced a mixed model with both ED and our original Diffusion Model based on geographies to compare how important air traffic is to predict CDC level of flu. In the mixed model, we add an edge between states that have significant air traffic but not close to each other. In this study, we only consider flights between US airports, and exclude international flights because there is no synchronized flu data or levels throughout the world.

After finding ED between all states, we take 1-hop neighborhood states, 2-hop neighborhood states, and the states that has effective distance less than a certain threshold ψ (See Equation 19).

We create $G_3 = (V, E_3)$ based on Effective Distances ED. V has same vertices with G_1 and G_2, E_3 is the set of edges representing effective distances.

$$E_3 = \begin{cases} ED_{uv}, & if\ V_u\ and\ V_v\ is\ not\ 1-hop\ or\ 2-hop\ neighbor \\ 0, & Otherwise \end{cases} \tag{21}$$

The weight of each edges in E_3 is calculated similar to E_1 and E_2.

$$P_{u,t+1} = \frac{\sum_{(v,u)\in E_1} w_{vu}}{n} + \frac{\sum_{(v,u)\in E_2 - E_1} w_{vu}}{m} + \frac{\sum_{(v,u)\in E_3} \frac{w_{vu}}{\log D_{vu}}}{r} \tag{22}$$

where r refers to all incoming edges for state u that corresponds to certain states have significant air traffic by looking Effective Distance table. Also, we are decreasing the effect of air traffic by dividing weight to logarithm of ED value, because geographical neighborhoods has more contribution to flu spread than air traffic.

While the mixed model has generated best results for $\psi < 4.32$ (TOST $p < 0.05$), which was identified by testing all ψ available, the predicted levels were clearly different

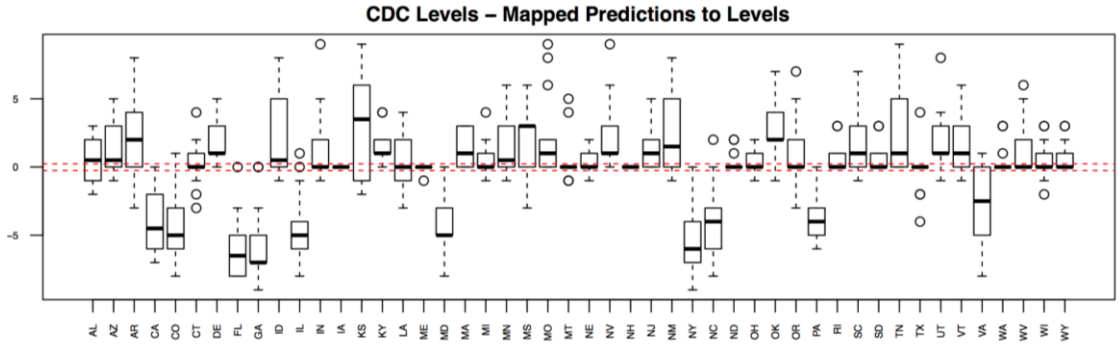

Figure 6.10 - A box plot of the differences between mapped levels associated with predictions from Geographical and Effective Distance Mixed Model and CDC levels. Compared to Figure 6.7, more than ten states report worse differences in contrast to CDC levels. The two horizontal dotted lines represent the 95% confidence interval (-0.25, 0.25) for the mean of the differences.

compared to CDC data, when airline traffic is introduced in certain states within this threshold (Figure 6.10). Based on these results we have excluded the effective distance measures and utilized the Diffusion Model for predicting the diffusion of ILI.

6.3 Conclusions

In this study we have shown that we can observe how a disease spreads using social media activity. We generated a network model to represent this activity and generated predictions that can estimate how diseases diffuse across geographical locations. The predictions provide forecasting for one week in advance. Interestingly, in our network representation of states, the two-hop neighbors accounted for additional influence that is beyond geographical neighbors. This might also reflect disease spreading over and across multiple states within a week of time.

We have compared this model to earlier approaches that used airline traffic to study contagious disease spread. However, airline traffic data, with our network model, was unsuccessful in predicting Influenza Like Illnesses (ILI) activity across geographies. We conclude that geographic distance is the main dynamic to predict ILI activity.

Our network model provides insight that can ultimately give hospitals a stronger warning about when outbreaks will occur in their particular areas. Using our model with information systems can help with disease control and prevention efforts. Knowing how a disease spreads would give health officials a better chance to maintain good health among citizens.

7 Conclusions and Future Work

7.1 Thesis Contributions

In this thesis, we have studied algorithms on (i) social media discussions to identify topic-related issues such as shifted comments or classifying comments based on subtopics, and (ii) social media big data to predict contagious disease diffusions. To find shifted comments, we collect data from article-originated comments, and modified the tf-idf model to identify closeness of comments and the article. Our approach is to obtain similarity scores of all comments to the article, distinguish off-the-topic and on-the topic comments, and search the reasons behind why a comment goes off-the topic. Understanding these phenomena on a real world experiment may be useful in building applications that classify and/or eliminate undesired comments. Our observations show that the fractions of shifted comments change topic by topic. The structure of the discussion tree plays an important role to locate shifted comments. We see that first-level comments are most likely to be on-the-topic, and, as expected, fraction of on-the-topic comments decrease after the first level. However, at some point, in deeper levels of discussion trees, surprisingly, the discussion's focus returns back to the original topic. Emotion levels are another feature of shifted comments. From the measurements of emotion levels of comments, we conclude there is a strong relationship between off-the-topic comment sets and emotion levels.

In chapter 4, we discuss how to measure and visualize emotion levels of movie reviews. The first challenge is to find emotions for each review. We map each word in a review to the SenticNet dataset, and save the scores. Since, multiple numbers of words in a review may match to SenticNet database words, we reduce emotion dimension scores of

a review to a single score. Then, for each movie, we visualize the emotion scores for the highest-ranked one hundred reviews via a heatmap. We compare similar movies having same genres and closer ratings. Plots show that there are differing emotion maps for these movies, and they differ from each other in terms of their emotion levels.

In chapter 5, we have extracted subtopics of social media comments around a topic; and analyze their differences. This study differs from our work in Chapter 3 in that, unlike chapter 3, our proposed methodology identifies subtopics within the discussion where the discussion does not originate from an article. For this purpose, we collect several social media discussions including at least 3,000 comments from Reddit. We use and modify the Latent Semantic Analysis technique to identify the subtopics, their characterization, and their emotion behaviors. We find the most "useful" words, bigrams and comments for each subtopic, allowing us to characterize a given subtopic. This approach may lead to applications where users can choose to read or skip certain subtopic comments by simply looking a subtopics' signature. Next, we look subtopic comments' emotion levels and how much they differ from each other. We observe that there are emotion differences between subtopics, and we use these differences to distinguish similar subtopics found by LSA.

In chapter 6, we show that one can observe and predict how a disease spreads using social media activity. We generate a network model to represent this activity, and produce predictions that can estimate how diseases diffuse across geographical locations. The predictions provide forecasting for one week in advance. Interestingly, in our network representation of states, the two-hop neighbors accounted for additional influence that is beyond geographical neighbors. This might also reflect diseases spreading over and across multiple states within a week of time. We compare this model to earlier approaches that

use airline traffic to study contagious disease spread. However, airline traffic data, without our network model, was unsuccessful in predicting Influenza Like Illness (ILI) activity across geographies.

7.2 Future Research

7.2.1 More on Effects of Emotion and Topic Area on Topic Shifts in Social Media Discussions

Improving chapter 3 and 5, linking and viewing discussions of interest in temporal order, even when they belong to different discussion threads within the discussion tree, is possible and can be analyzed using our proposed methods. Furthermore, one can view comments from all over the discussion tree that may have shifted from the original topic in a certain way, such as shifted to a certain "drifted topic".

In chapter 5, some of the latent subtopics look very similar. This close relationships can be identified, and their causes can be analyzed. One can also merge these discussions for more specific or unique subtopics. Also, on-the-topic comment set may be identified via Latent Semantic Analysis without a main article.

7.2.2 More on Social Media Big Data Prediction of Contagious Disease Diffusion—Additional Research Ideas

To improve chapter 6, one can consider the Susceptible (S), Infected (I), and Recovered (R) (SIR) model [94] in the prediction of disease diffusion. In epidemiology, if population exposes to a some certain type of illness, number of Susceptible (S), Infected (I), and Recovered (R) people in population can theoretically be calculated by the SIR model [94]. SIR model aims to model the disease movement throughout the population over the time. Under some assumptions, fixed number of population, if we know the average number of interaction between people, the time of recovery of infected people, than the model creates the probability distribution of susceptible, infected, and recovered people over the time (Figure 7.1). Since our proposed model in chapter 6.2.1, does not consider the illness recovery, the model cannot predict decreasing trend of disease diffusion. In left panel of Figure 7.1, an example of SIR model is shown. Green line represents the infected number of people over the time. Observe that, in the right panel of

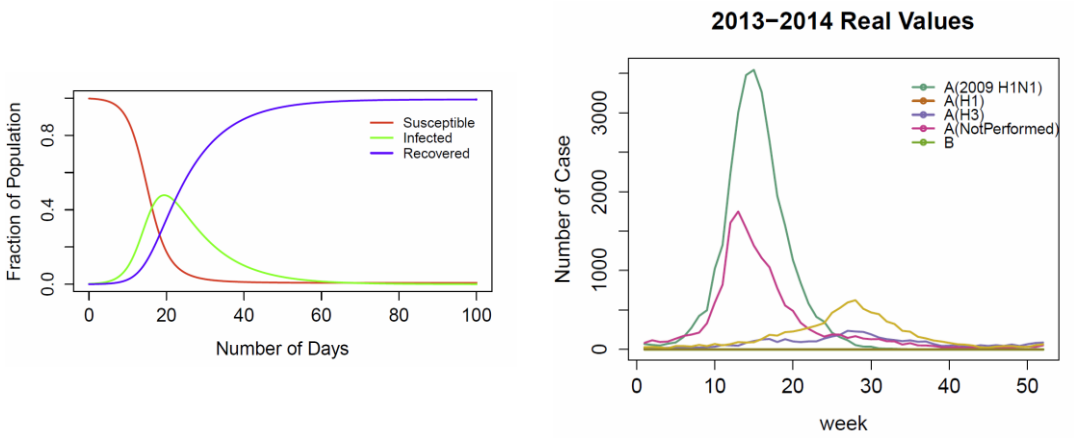

Figure 7.1 - SIR Model vs Flu related virus cases in 2013-2014.

Figure 7.1, the number of influenza test virus cases behave the same as the green line in the SIR model.

According to the Centers for Disease Control and Prevention (CDC), Influenza Like Illness (ILI) levels are calculated by combining multiple numbers of virus cases that are either concurrently or independently active during the year [27]. Our next research direction is to forecast the next 6 weeks' flu levels for each state without using the number of individual virus cases, because these numbers are not available for all states, and they are not under the same source. Our first line of approach will be decompose CDC Influenza levels into virus cases in order to forecast upcoming weeks better, because we can model the number of infected people over the time by using the SIR model. Then, combining these information will give the predicted CDC influenza levels.

To do this, we will use Non-negative Matrix Factorization (NMF) [22] to identify latent virus activity behaviors. Non-negative Matrix Factorization is an unsupervised learning technique to decompose original matrix into two matrices (dimensionally lower than original one), that are basis matrix and coefficient matrix. This property gives a part-based decomposition. Inspecting these two decomposed matrices is easier to see hidden properties:

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \tag{23}$$

where r is the number of features that we want to see, W is the basis matrix, and H is the coefficients of the original matrix V which is CDC levels of every state, weekly, in our case. Each row represents weeks and each column represents states in V. If we set r to number of active viruses, then in W, we get hidden important basis properties in the

columns. Our preliminary results show that the columns of W look like infected number of people over the time in SIR model (Figure 7.1). In coefficient matrix H, there are weights for each "virus" for each state; and, one can infer that similar weighted states share common viruses at the same time.

Once we find virus patterns, we can predict next week's virus levels on W by using coefficients to predict each state's flu levels. For this, we plan to use the Autoregressive Integrated Moving Averages (ARIMA) [95], a time series fitting model to find a model that characterizes previous and upcoming data points.

We plan to collect historical data on influenza levels for each state from CDC. We would like to evaluate the viability of predictions for the next six upcoming weeks by looking at n latest weeks' influenza levels.

Our research pan is as follows. We will

- Predict scores by just using the ARIMA model,
- Predict future levels by using a combination of the NMF and ARIMA models.
- Plot the heatmap of the coefficient matrix of results to see how similar state patterns are. Since CDC reveals the information two to three weeks later, we will use social media data to cover this period.

To get more accurate results from the social media, we will use natural language processing to eliminate false information. In summary, we plan to extend the contagious disease diffusion research by employing the SIR model to forecast more accurate disease propagation predictions in the long-term.

Appendix: Discussion A and B Results from Chapter 5

Appendix Table 1 – Single Word Weights of Discussion B: "*LeBron: I'm coming back to Cleveland*"

Subtopic 1		**Subtopic 2**		**Subtopic 3**		**Subtopic 4**		**Subtopic 5**		**Subtopic 6**		**Subtopic 7**	
fan	0.31	wa	0.16	fan	0.47	real	0.19	lebron	0.75	get	0.19	cavs	0.46
lebron	0.31	lebron	0.15	im	0.17	wa	0.14	cleveland	0.14	wiggins	0.15	get	0.32
heat	0.25	get	0.12	cavs	0.13	heat	0.12	good	0.1	cavs	0.13	gold	0.25
cleveland	0.23	love	0.1	happy	0.11	miami	0.12	wiggins	0.09	love	0.12	wiggins	0.22
im	0.18	wiggins	0.1	cleveland	0.1	year	0.12	james	0.09	year	0.08	love	0.22
cavs	0.18	like	0.08	miami	0.07	like	0.12	love	0.08	team	0.07	real	0.18
wa	0.17	really	0.08	coming	0.06	get	0.08	jersey	0.07	mention	0.06	fan	0.16
miami	0.15	good	0.08	home	0.06	one	0.08	fuck	0.07	flair	0.06	god	0.16
year	0.14	cleveland	0.08	bandwagon	0.05	team	0.08	back	0.05	next	0.06	fuck	0.14
get	0.14	going	0.07	jersey	0.05	think	0.08	respect	0.04	cleveland	0.06	oh	0.12
like	0.14	year	0.07	lebron	0.05	dont	0.08	confirmed	0.04	trade	0.05	life	0.11
fuck	0.13	home	0.07	base	0.02	life	0.07	man	0.04	pick	0.05	mention	0.09
team	0.12	im	0.06	bandwagoners	0.02	time	0.07	luck	0.04	win	0.04	trade	0.07
back	0.12	think	0.06	good	0.02	ha	0.07	article	0.03	going	0.04	jersey	0.07
going	0.12	article	0.06	deserve	0.02	love	0.06	welcome	0.03	gold	0.04	happy	0.06
good	0.11	right	0.06	real	0.02	wiggins	0.06	heat	0.03	heat	0.04	kevin	0.06
dont	0.11	back	0.06	rip	0.02	know	0.05	well	0.03	still	0.03	holy	0.05
go	0.1	time	0.06	excited	0.02	people	0.05	written	0.03	na	0.03	thank	0.05
think	0.1	one	0.06	team	0.02	post	0.05	thanks	0.03	gon	0.03	wait	0.05
one	0.1	would	0.05	burned	0.02	private	0.04	unburn	0.03	doe	0.03	shit	0.04

Appendix Table 2 - Bigram Weights of Discussion B: "*LeBron: I'm coming back to Cleveland*"

Subtopic 1		Subtopic 2		Subtopic 3		Subtopic 4		Subtopic 5		Subtopic 6	
heat fan	5.09	gon na	1.12	cavs fan	3.02	heat fan	2	lebron james	1.61	gon na	0.74
cavs fan	2.97	4 year	1.05	heat fan	2.67	4 year	1.02	heat fan	0.72	next year	0.7
4 year	2.33	coming home	1.02	cleveland fan	1.07	miami heat	0.83	lebron jersey	0.69	cavs fan	0.64
gon na	1.92	im coming	0.7	coming home	1.07	four year	0.79	lebron fan	0.53	mention wiggins	0.59
four year	1.46	four year	0.61	miami fan	0.99	next year	0.75	good luck	0.53	4 year	0.52
next year	1.45	next year	0.6	im coming	0.96	cavs fan	0.64	back cleveland	0.51	kevin love	0.46
cleveland fan	1.37	lebron james	0.55	im happy	0.78	dont think	0.64	love lebron	0.5	heat fan	0.45
im happy	1.37	going back	0.54	bandwagon fan	0.64	real life	0.48	4 year	0.5	four year	0.28
coming home	1.29	dont think	0.54	lebron fan	0.53	feel like	0.46	lebron wa	0.5	next season	0.27
miami heat	1.26	im happy	0.53	fan im	0.42	miami fan	0.4	miami heat	0.46	doesnt mention	0.27
lebron james	1.24	lebron wa	0.52	gon na	0.39	dont know	0.4	lebron ha	0.44	win championship	0.27
back cleveland	1.24	year ago	0.52	happy cleveland	0.39	year ago	0.38	lebron cleveland	0.4	didnt mention	0.27
dont think	1.1	back cleveland	0.51	back cleveland	0.38	gon na	0.37	cavs fan	0.39	dont think	0.26
miami fan	1.03	feel like	0.46	fan base	0.37	mention wiggins	0.35	lebron back	0.36	draft pick	0.26
going back	0.96	mention wiggins	0.45	lebron jersey	0.34	heat flair	0.32	respect lebron	0.35	1 pick	0.25
im coming	0.94	come back	0.43	4 year	0.33	look like	0.3	come back	0.35	get gold	0.25
come back	0.89	kevin love	0.36	miami heat	0.31	lebron wa	0.29	wiggins lebron	0.34	love trade	0.24
lebron wa	0.88	northeast ohio	0.36	coming back	0.28	last time	0.29	lebron cavs	0.33	wiggins love	0.24
feel like	0.88	good luck	0.35	real fan	0.25	kevin love	0.28	lebron going	0.32	miami heat	0.23
go back	0.86	well written	0.35	lebron james	0.24	northeast ohio	0.26	good lebron	0.32	cavs flair	0.21

Appendix Table 3 – Single Word Weights of Discussion C: "*I made a chart showing the popular vote turnout in 2008, 2012 and 2016…"*

Subtopic 1		Subtopic 2		Subtopic 3		Subtopic 4		Subtopic 5		Subtopic 6		Subtopic 7	
vote	0.4	vote	0.65	didnt	0.4	voter	0.31	didnt	0.34	wa	0.39	trump	0.58
people	0.23	popular	0.2	democrat	0.25	graph	0.27	people	0.28	didnt	0.27	voted	0.31
wa	0.21	didnt	0.19	trump	0.19	democrat	0.25	dont	0.26	graph	0.22	graph	0.21
didnt	0.2	state	0.15	voted	0.18	party	0.25	graph	0.22	misleading	0.11	misleading	0.13
trump	0.2	democrat	0.15	hillary	0.18	voted	0.24	want	0.15	bar	0.1	popular	0.12
hillary	0.16	electoral	0.14	bernie	0.18	turnout	0.2	misleading	0.14	vote	0.1	hillary	0.09
democrat	0.15	come	0.11	wa	0.16	million	0.17	democrat	0.13	turnout	0.1	read	0.09
voted	0.14	felon	0.08	candidate	0.13	number	0.14	read	0.13	come	0.09	clinton	0.09
would	0.14	party	0.08	come	0.12	third	0.14	know	0.11	win	0.09	win	0.09
candidate	0.14	college	0.08	lost	0.11	misleading	0.14	like	0.1	obama	0.08	number	0.08
dont	0.14	win	0.06	dnc	0.09	obama	0.13	come	0.1	read	0.08	bar	0.08
party	0.13	lost	0.06	clinton	0.08	data	0.12	think	0.1	popular	0.07	electoral	0.07
election	0.13	matter	0.04	wanted	0.06	republican	0.11	bar	0.09	like	0.07	chart	0.06
like	0.13	3rd	0.04	primary	0.06	election	0.11	make	0.09	chart	0.06	would	0.06
voter	0.12	count	0.04	supporter	0.06	bar	0.1	start	0.07	look	0.06	college	0.06
voting	0.1	million	0.04	sander	0.06	chart	0.09	chart	0.07	start	0.06	data	0.06
think	0.1	third	0.04	dems	0.05	2008	0.09	thats	0.07	black	0.06	start	0.05
state	0.1	counted	0.03	president	0.04	2012	0.08	data	0.06	point	0.06	zero	0.05
clinton	0.1	still	0.03	rigged	0.03	candidate	0.08	care	0.05	hillary	0.05	hate	0.05
get	0.09	california	0.03	want	0.03	3rd	0.07	zero	0.05	zero	0.05	im	0.05
bernie	0.09	convicted	0.02	blame	0.03	read	0.07	get	0.05	show	0.05	got	0.05

Appendix Table 4 – Bigram Weights of Discussion C: "*I made a chart showing the popular vote turnout in 2008, 2012 and 2016…"*

Subtopic 1		Subtopic 2		Subtopic 3		Subtopic 4		Subtopic 5		Subtopic 6	
popular vote	5.02	popular vote	5.5	didnt vote	2.73	third party	3.51	didnt vote	2.54	popular vote	2.11
didnt vote	3.94	electoral college	2.82	come vote	2.73	voter turnout	3.15	come vote	1.93	didnt vote	1.7
third party	2.95	come vote	2.63	democrat didnt	2.59	3rd party	1.52	democrat didnt	1.89	come vote	1.57
electoral college	2.71	didnt vote	2.55	didnt come	2.22	democrat didnt	1.47	didnt come	1.58	democrat didnt	1.3
come vote	2.31	democrat didnt	2.04	didnt want	1.39	come vote	1.39	people dont	1.54	didnt come	1.25
people voted	2.13	didnt come	1.94	voted trump	1.34	people voted	1.26	dont think	1.17	look like	0.97
democrat didnt	2.09	electoral vote	1.71	people didnt	1.06	didnt come	1.24	people didnt	1.09	bar graph	0.91
dont think	1.95	third party	1.2	vote hillary	1.02	party candidate	1.22	didnt want	1.02	bar chart	0.88
voter turnout	1.93	3rd party	1.1	people voted	0.95	didnt vote	1.2	dont know	0.93	wa going	0.78
vote trump	1.88	vote trump	1.03	didnt like	0.9	look like	1.11	bar chart	0.9	electoral college	0.73
people vote	1.85	win popular	0.98	vote trump	0.87	bar graph	1.08	people vote	0.89	hillary wa	0.72
voted trump	1.77	million vote	0.98	bernie supporter	0.85	party vote	1.08	lot people	0.86	turnout wa	0.72
didnt come	1.76	vote hillary	0.77	hillary wa	0.77	voted obama	0.96	dont vote	0.84	didnt want	0.71
many people	1.71	people vote	0.75	lost democrat	0.75	democrat voted	0.95	dont want	0.84	people didnt	0.65
vote hillary	1.71	swing state	0.72	hillary clinton	0.69	turnout wa	0.93	people voted	0.83	voter turnout	0.65
3rd party	1.7	vote state	0.69	third party	0.69	eligible voter	0.87	look like	0.77	obama wa	0.64
lot people	1.69	didnt want	0.67	3rd party	0.66	million vote	0.85	many people	0.77	vote hillary	0.59
people didnt	1.66	dont vote	0.67	want vote	0.63	bar chart	0.82	bar graph	0.74	vote trump	0.55
people dont	1.52	lost democrat	0.65	bernie sander	0.61	voted trump	0.79	didnt like	0.68	start zero	0.54
hillary wa	1.41	party candidate	0.63	vote didnt	0.59	registered voter	0.79	want vote	0.67	clinton wa	0.52

8 Bibliography

[1] L. S. Elkin, "Predicting diffusion of contagious diseases using social media big data," *Case Western Reserve University*, 2015.

[2] "Seasonal Flu," U.S. Department of Health & Human Services, [Online]. Available: http://www.flu.gov/about_the_flu/seasonal/index.html. [Accessed 03 08 2015].

[3] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A. S. Tomkins, "The Web as a graph : measurements , models , and methods," *Computer*, vol. 1627, pp. 1-17, 1999.

[4] C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[5] D. He, A. Goker and D. J. Harper, "Combining evidence for automatic Web session identification," *Information Processing and Management*, vol. 38, pp. 727-742, 2002.

[6] N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin and A. F. Smeaton, "Topic-dependent sentiment analysis of financial blogs," in *CIKM workshop on Topic-sentiment analysis for mass opinion*, 2009.

[7] Q. Liu, h. Huang and C. Feng, "Micro-blog Post Topic Drift Detection Based on LDA Model," in *Behavior and Social Computing*, Springer International Publishing, 2013, pp. 106-118.

[8] D. Knights and C. M. &. N. N. Mozer, "Detecting Topic Drift with Compound Topic Models," in *ICWSM*, 2009.

[9] P. D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research,* vol. 37, pp. 141-188, 2010.

[10] M. Hasan, E. A. Rundensteiner and E. Agu, "EMOTEX: Detecting Emotions in Twitter Messages," in *Academy of Science and Engineering (ASE)*, 2014.

[11] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *Human Language Technology and Empirical Methods in Natural Language Processing*, Stroudsburg, 2005.

[12] J. M. Wiebe and R. Ellen, "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," *Computational Linguistics and Intelligent Text Processing*, vol. 3406, pp. 486-497, 2005.

[13] S. NLP, "Sentiment Analysis," August 2013. [Online]. Available: http://nlp.stanford.edu/sentiment/.

[14] Alchemy, "Sentiment Analysis API," 2015. [Online]. Available: http://www.alchemyapi.com/api/sentiment-analysis.

[15] J. A. Russel, "A circumplex model of affect," *Journal of Personality and Social Psychology*, pp. 1161-1178, 1980.

[16] M. S. Smith, D. M. Ogilvia, P. J. Stone, D. C. Dunphy and J. J. Hartman, "The General Inquirer: A Computer Approach to Content Analysis.," in *American Sociological Review*, 1967, p. 859.

[17] E. Cambria, A. Livingstone and A. Hussain, "The hourglass of emotions," *Lecture Notes in Computer Science* , vol. 7403 LNCS, pp. 144-157, 2012.

[18] C. Havasi, R. Speer and J. Alonso, "ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge," in *RANLP*, Borovets, 2007.

[19] C. Strapparava and A. Valitutti, "WordNet-Affect: an Affective Extension ofWordNet," in *LREC*, Lisbon, 2004.

[20] SenticNet, "SenticNet Dataset," [Online]. Available: http://sentic.net/downloads/. [Accessed 20 9 2016].

[21] T. K. Landauer, P. W. Folt and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2, pp. 259-284, 1998.

[22] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems*, no. 1, pp. 556-562, 2001.

[23] T. Hofmann, "Probabilistic latent semantic indexing," *ACM SIGIR*, pp. 50-57, 1999.

[24] X. Wei, "Retrieval, Topic Models in Information," *Ph.D thesis*, 2007.

[25] L. Wei and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pp. 577-584, 2006.

[26] "Overview of Surveillance in the United States," Centers for Disease Control and Prevention, 16 07 2014. [Online]. Available: http://www.cdc.gov/flu/weekly/overview.htm. [Accessed 20 07 2014].

[27] "CDC: FluView," Centers for Disease Control and Prevention, [Online]. Available: http://gis.cdc.gov/grasp/fluview/main.html. [Accessed 17 07 2014].

[28] S. Brain, "Twitter Statistics," [Online]. Available: http://www.statisticbrain.com/twitter-statistics/. [Accessed 16 07 2014].

[29] J. Lee, "Twitter predicts when you are sad, hungover, late.," [Online]. Available: http://www.usatoday.com/story/news/nation-now/2014/03/11/twitter-sad-happy-late-work/6283655/. [Accessed 17 07 2014].

[30] J. Bollen, H. Mao and X.-J. Zeng, "Twitter mood predicts the stock market," *J. Comput. Science*, vol. 2, pp. 1-8, 2011.

[31] X. Zhang, H. Fuehres and P. A. Gloor, "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"," *Procedia - Social and Behavioral Sciences*, vol. 26, pp. 55-62, 2011.

[32] S. Sinha, C. Dyer, K. Gimpel and N. A. Smith, "Predicting the NFL using Twitter," *CoRR*, vol. abs/1310.6998, 2013.

[33] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *ICWSM*, 2010.

[34] A. Sadilek, H. Kautz and V. Silenzio, "Predicting Disease Transmission from Geo-Tagged Micro-Blog Data.," in *AAAI Conference on Artificial Intelligence*, 2012.

[35] H. Hirose and L. Wang, "Prediction of Infectious Disease Spread Using Twitter: A Case of Influenza," in *Proceedings of the 2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, Washington, DC, USA, 2012.

[36] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu and B. Liu, "Predicting Flu Trends using Twitter data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, 2011.

[37] D. A. Broniatowski, M. Paul and M. Dredze, "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic," *PLOS One*, 2013.

[38] M. Dredze, M. Paul, S. Bergsma and H. Tran, "Carmen: A Twitter Geolocation System with Applications to Public Health," in *AAAI Workshops*, 2013.

[39] M. Salathe and S. Khandelwal, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control," *PLoS Comput Biol*, vol. 7, p. e1002199, 2011.

[40] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social web," in *Cognitive Information Processing (CIP), 2010 2nd International Workshop on, IEEE*, 2010.

[41] V. Lampos, T. De Bie and N. Cristianini, "Flu detector-tracking epidemics on Twitter," in *Machine Learning and Knowledge Discovery in Databases*, 2010.

[42] A. Lamb, M. J. Paul and M. Dredze, "Separating Fact from Fear: Tracking Flu Infections on Twitter.," in *HLT-NAACL*, 2013.

[43] M. J. Paul and M. Dredze, "You are what you Tweet: Analyzing Twitter for public health.," in *ICWSM*, 2011.

[44] D. Brockmann and D. Helbing, "The hidden geometry of complex, network-driven contagion phenomena," *Science* , vol. 342, pp. 1337-42, 2013.

[45] D. Gross, "Online comments are being phased out," 21 November 2014. [Online]. Available: http://www.cnn.com/2014/11/21/tech/web/online-comment-sections/.

[46] M. Porter, "An algorithm for suffix stripping," in *Readings in Information Retrieval*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 1997, pp. 313-316.

[47] XPO6, "List of English Stop Words," 3 October 2009. [Online]. Available: http://xpo6.com/list-of-english-stop-words/.

[48] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management: an International Journal,* pp. 513-523, 1988.

[49] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau and A. Vanhoutte, "Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula," *Information Processing and Management,* pp. 315-318, 1989.

[50] M. Murguia and J. L. Villasenor, "Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classifications," *Annales Botanici Fennici,* vol. 40, pp. 415-421, 2003.

[51] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society C,* vol. 28, pp. 100-108, 1979.

[52] E. Cambria, A. Livingstone and A. Hussain, "The hourglass of emotions," in *international conference on Cognitive Behavioural Systems (COST`11)*, Berlin, 2011.

[53] "Disqus API," 2015. [Online]. Available: https://disqus.com/api/docs/. [Accessed 09 09 2015].

[54] N. Yau, "How to Read and Use a Box-and-Whisker Plot," FlowingData, 15 02 2008. [Online]. Available: https://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/. [Accessed 17 March 2016].

[55] H. Motulsky, "The link between error bars and statistical significance," GraphPad Software, [Online]. Available: https://egret.psychol.cam.ac.uk/statistics/local_copies_of_sources_Cardinal_and_Aitken_ANOVA/errorbars.htm. [Accessed 22 4 2016].

[56] L. Rokach and O. Maimon, "Decision Tree," *Data Mining and Knowledge Discovery Handbook,* pp. 165-192, 2005.

[57] IMDb, "Internet Movie Database," IMDb, [Online]. Available: http://www.imdb.com. [Accessed 15 05 2016].

[58] IMDb, "The vote average for film," [Online]. Available: http://www.imdb.com/help/show_leaf?votes. [Accessed 03 03 2016].

[59] SenticNet, "SenticNet," [Online]. Available: http://sentic.net/downloads/. [Accessed 03 03 2016].

[60] D. a. Freedman, "Bootstrapping Regression Models," *The Annals of Statistics,* vol. 9, pp. 1218-1228, 1981.

[61] imdb-pie. [Online]. Available: https://github.com/richardasaurus/imdb-pie. [Accessed 02 03 2016].

[62] C. Skuta, P. Bartunek and D. Svozil, "InCHlib - Interactive cluster heatmap for web applications," *Journal of Cheminformatics,* vol. 6, 2014.

[63] axismaps, "Color Brewer," [Online]. Available: http://colorbrewer2.org/. [Accessed 03 03 2016].

[64] C. Parkhurst and R. L. Feller, "Who invented the color wheel?," *Color research and application,* vol. 7, no. 3, pp. 217-230, 1982.

[65] "Reddit," [Online]. Available: https://www.reddit.com/.

[66] M. D. Blei, "Topic Modeling," [Online]. Available: https://www.cs.princeton.edu/~blei/topicmodeling.html. [Accessed 28 9 2016].

[67] L. H. Anaya, "Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers," UNT Digital Library, Denton, Texas, 2011.

[68] J. Chen, J. Yan, B. Zhang, Q. Yang and Z. Chen, "Diverse topic phrase extraction through latent semantic analysis," in *IEEE International Conference on Data Mining, ICDM,* 2006.

[69] SenticNet, "SenticNet," [Online]. Available: http://sentic.net/downloads/.

[70] "SimilarWeb," [Online]. Available: https://www.similarweb.com/website/reddit.com. [Accessed 13 11 2016].

[71] "Reddit in 2015," [Online]. Available: https://redditblog.com/2015/12/31/reddit-in-2015/. [Accessed 13 11 2016].

[72] "PRAW," [Online]. Available: https://praw.readthedocs.io/en/stable/. [Accessed 13 11 2016].

[73] Reddit, "Reddit: News," [Online]. Available: https://www.reddit.com/r/news/comments/5njj2x/cvs_puts_out_generic_competitor_to_epipen_at_a/. [Accessed 07 03 2017].

[74] Reddit, "Reddit: NBA," [Online]. Available: https://www.reddit.com/r/nba/comments/2afphe/lebron_im_coming_back_to_cleveland/. [Accessed 07 03 2017].

[75] Reddit, "Reddit:Data is Beautiful," [Online]. Available: https://www.reddit.com/r/dataisbeautiful/comments/5c5k4e/i_made_a_chart_showing_the_popular_vote_turnout/. [Accessed 07 03 2017].

[76] "Seasonal Influenza-Associated Hospitalizations in the United States," Centers for Disease Control and Prevention, 24 06 2011. [Online]. Available: http://www.cdc.gov/flu/about/qa/hospital.htm. [Accessed 18 07 2014].

[77] W. W. Thompson, D. K. Shay, E. Weintraub, L. Brammer, C. B. Bridges, N. J. Cox and K. Fukuda, "Influenza-associated hospitalizations in the United States," *JAMA : the journal of the American Medical Association*, vol. 292, pp. 1333-1340, 2004.

[78] T. Developers, "Twitter API Documentation," [Online]. Available: https://dev.twitter.comp/docs. [Accessed 16 07 2014].

[79] "Flu Symptoms & Severity.," Centers for Disease Control and Prevention, 2013 12 09. [Online]. Available: http://www.cdc.gov/flu/about/disease/symptoms.htm. [Accessed 16 7 2014].

[80] M. Reviewed by Kimball Johnson, "WebMD: Cold, Flu & Cough Health Center," [Online]. Available: http://www.webmd.com/cold-and-flu/flu-glossary. [Accessed 16 07 2014].

[81] "Symptoms of the Flu," U.S. Department of Health and Human Services, [Online]. Available: http://www.flu.gov/symptoms-treatment/symptoms/. [Accessed 16 07 2014].

[82] M.-W. Inc., Merriam-Webster's medical dictionary, Merriam-Webster, 2007.

[83] J. Rogers and O. Bodenreider, "Snomed CT," 2008. [Online]. Available: http://www.ihtsdo.org/snomed-ct. [Accessed 16 07 2014].

[84] "Flu season may be heading toward peak," CNN Health, 10 01 2014. [Online]. Available: http://www.cnn.com/2014/01/10/health/cdc-flu-activity/.

[85] "Internet World Stats, United States of America Internet and Facebook Users Stats," Miniwatts Marketing Group, 2010. [Online]. Available: http://www.internetworldstats.com/stats26.htm. [Accessed 16 07 2014].

[86] T. Webster, "Twitter usage in America," 29 04 2010. [Online]. Available: http://www.edisonresearch.com/twitter_usage_2010.php. [Accessed 16 07 2014].

[87] "Computer and Internet Access in the United States: 2012," U.S. Census Bureau, [Online]. Available: http://www.census.gov/hhes/computer/publications/2012.html. [Accessed 19 10 2014].

[88] J. Yarow, "Twitter Has A Surprisingly Small Number Of US Users," 04 10 2013. [Online]. Available: http://www.businessinsider.com/twitter-has-a-surprisingly-small-number-of-us-users-2013-10. [Accessed 16 07 2014].

[89] D. J. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *Journal of pharmacokinetics and biopharmaceutics,* vol. 15, pp. 657-80, 1987.

[90] M. Granovetter, "Threshold Models of Collective Behavior," *American Journal of Sociology,* vol. 83, pp. 1420-43, 1978.

[91] G. F. Pyle, The Diffusion of Influenza: Patterns and Paradigms, Rowan Littlefield, 1986.

[92] "OAG," Official Airline Guide, [Online]. Available: https://www.oag.com/. [Accessed 03 08 2015].

[93] "National Transportation Atlas Transportation networks, features and statistics," Office of the Assistant Secretary for Research and Technology, [Online]. Available: http://www.rita.dot.gov/bts/home. [Accessed 03 08 2015].

[94] M. Newman, Networks an Introduction, Oxford Scholarship Online, 2010.

[95] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 2014.