# REPRESENTING BLOGGER INFLUENCE ON ISSUE SENTIMENT AND OPINION AS AN EPIDEMIC MODEL

by

Michael J. Garrity
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Sciences and Informatics

Committee:

____________________________________ Dr. Edward Wegman, Dissertation Director

____________________________________ Dr. Kirk Borne, Committee Member

____________________________________ Dr. Igor Griva, Committee Member

____________________________________ Dr. Padmanabhan Seshaiyer, Committee Member

____________________________________ Dr. Kevin Curtin, Department Chair

____________________________________ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

____________________________________ Dr. Peggy Agouris, Dean, College of Science

Date: ______________________________ Spring Semester 2016
George Mason University
Fairfax, VA

# Representing Blogger Influence on Issue Sentiment and Opinion as an Epidemic Model

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Michael J. Garrity
Master of Science
George Mason University, 2011
Bachelor of Science
George Mason University, 2010

Director: Edward Wegman, Professor
Department of Computational Sciences and Informatics

Spring Semester 2016
George Mason University
Fairfax, VA

# DEDICATION

I dedicate this dissertation to the love of my life, Jenny.

## ACKNOWLEDGEMENTS

I would like to thank my wife, Jenny, as well as the rest of my family and friends for all their support to make this happen. I would also like to thank my director, Dr. Edward Wegman, along with the other committee members whom have greatly helped me in achieving my goal.

# TABLE OF CONTENTS

Page

## LIST OF TABLES

## LIST OF FIGURES

# ABSTRACT

REPRESENTING BLOGGER INFLUENCE ON ISSUE SENTIMENT AND OPINION AS AN EPIDEMIC MODEL

Michael J. Garrity, Ph.D.

George Mason University, 2016

Dissertation Director: Dr. Edward Wegman

Online blogging continues to be a popular way for people to share and discuss their opinions with others in the community. Blogospheres also provide a forum where authors can follow other bloggers as well as recommend and leave comments on individual posts. Over time, certain authors become more popular with high numbers of recommendations and comments, leading to an increased influence on connected bloggers' sentiment and opinion within the network. Given the large number of users and issues being discussed in any blogosphere, it is extremely difficult to easily gauge patterns and trends in public sentiment and opinion, as well as any level of influence certain bloggers may have.

In this application of text mining, I introduce new techniques to model the evolving sentiment and opinion independently as well to track the blogger's influence within the blogosphere. While it may seem logical at first to consider sentiment as a form

of opinion or derive an opinion based on the sentiment of the text, this analysis discovers numerous instances where authors either express subjectivity about a topic without any emotion, or where they incorporate sentiment words without any subjective viewpoints in their writings.

Additionally, this work also presents a social contagion model to represent a blogger's influence on neighboring authors within the blogosphere to determine where the overall level of sentiment and opinion converges to over time. This dissertation incorporates a modified epidemic model where sentiment and opinion represent infectious diseases and influence signifies an infection between individuals, differing from other recent work by incorporating the likely interactions of sentiment and opinion contagions within the same social network.

# INTRODUCTION

In this chapter, I begin with the background and motivation for my research. Next, I provide an overview of the general problem along with the objectives I aim to achieve. Finally, I outline the framework for the following sections of the dissertation.

## Background and Motivation

Current research shows the continued interest in investigating sentiment analysis, specifically regarding both automated discovery as well as any practical applications. Within this field, there also exists a subset of research examining subjectivity classification. However, within this field, opinions typically refer to “subjective expressions that describe people’s sentiment” towards various topics and issues [36]. Within this dissertation, I aim to demonstrate that an author can express an opinion without expressing any sentiment towards that entity, and vice versa. My goal is to explore beyond the idea that the terms *sentiment analysis* and *opinion mining* are used interchangeably [45] and discuss the value of measuring these two variables independently within this research to then determine if any dependencies exist between sentiment and opinion. The main focus of this research is to study the impacts of sentiment and opinion within popular authors’ online blogs and measure how these variables can then influence those same two factors with respect to other authors’ future posts within the same blogging website.

Today, blogging websites have become a popular source for authors to spread information and express their opinions about current events [13]. Issues of discussion include everything from politics, to local stories, sports, entertainment, and events around the globe. Over the last decade, people increasingly move to online sources to receive their news over print newspaper and television. In that same period, more authors have avoided writing opinion pieces on mainstream media news websites, in order to join various blogspaces to present their opinions in their own personal writing styles. Especially within political journalism, blogging has become the prominent form of expressing and sharing personal opinion [51]. While objective writing for political events is common for mainstream media since these sites follow common ethics and standards in journalism, there is more flexibility for objectivity within blogospheres. Thus, it is much more likely for the author's personal feelings and opinion to be mixed in with the objective reporting of any recent issues.

Additionally, blogging is now an ideal outlet for many concerned citizens to communicate and converse about relevant political issues [57]. Therefore, blogospheres continue to become nearly as popular as traditional media websites in regards to staying up-to-date on current topics, by allowing bloggers within the community to become both authors and readers, and provide an avenue to engage with one another to discuss and debate a variety of important issues. Throughout the rest of this dissertation, I continue to use the term *blogosphere*, which refers to the entire collection of documents, authors, and the network of connections between all bloggers within the community.

Blogospheres are unlike a regular news website where none of the authors interact with one another. Within a blogosphere, readers reply to other authors' posts and even have the opportunity to further delve into a topic or provide another viewpoint by becoming an author of blogs for others to read. As bloggers continue to share their writings with following bloggers, pieces of that information along with any associated opinions can quickly be shared with almost all active bloggers.

This concept of all bloggers being connected has been researched for many decades. The idea of 'six degrees of separation' dates back to 1909, where Guglielmo Marconi showed that entire global coverage within person-to-person connections could occur within less than six degrees [38]. Then in the 1960's, Stanley Milgram began researching the best method to measure the connectivity of American people to determine whether or not there existed a separation factor between any two strangers. In his findings, Milgram concluded that only a small degree of separation was needed to connect the entire population [41].

Since then, connectivity within a network has been examined using a much larger population such as the World Wide Web. In 2008, David Bradley published article summarizing the aforementioned research stated that even with as little as 31 contacts on a social media website, four degree yields over 200 million connections, as shown in Figure 1 [8], and five degrees could yield as many as nearly two billion contacts. While there may still exist both instances of single nodes as well as clusters of people that never may connect to others any degree, almost all other remaining nodes are connected to one another, showing how associated people are to one another online whether they realize it or not.

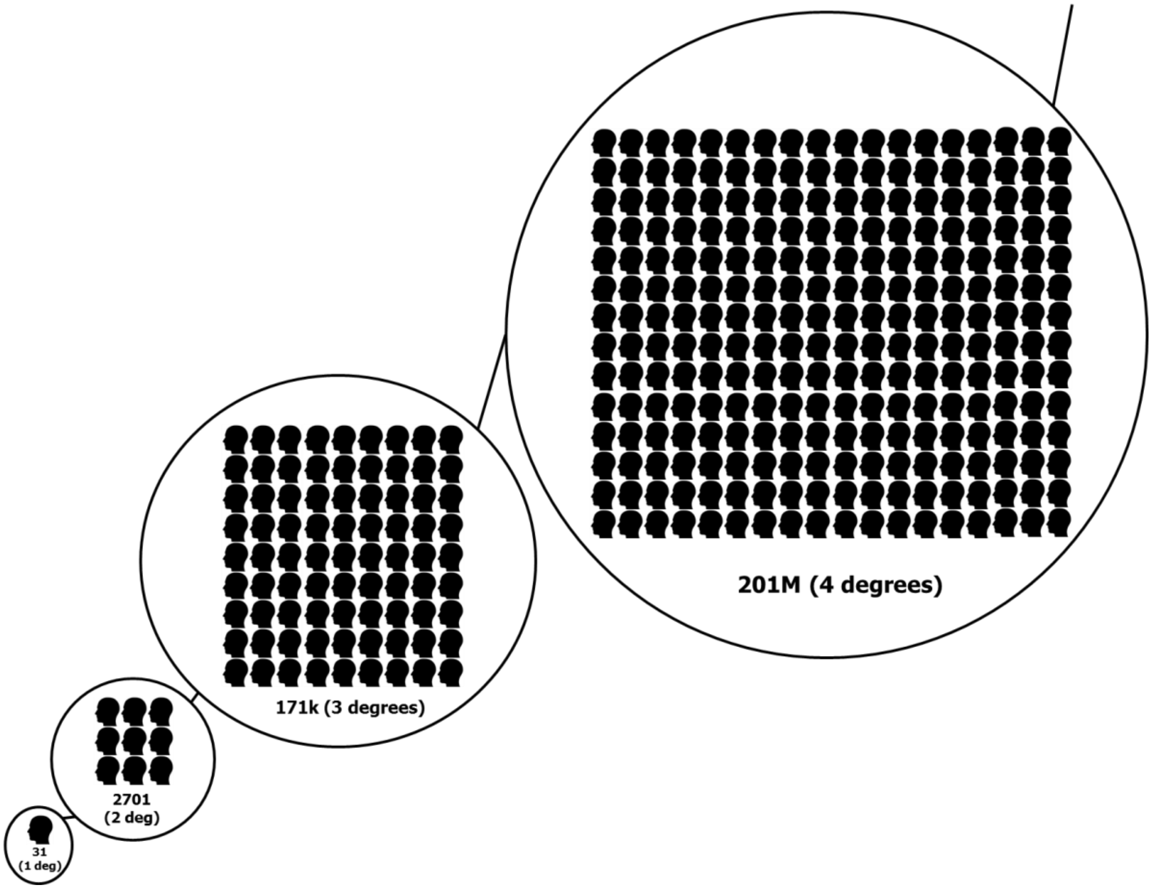


**Figure 1: Mass Connectivity Based on Degrees of Separation [8]**

In reality, there are only seven billion people in the world, most without accessibility to the internet. Therefore, it is typical that blogospheres are likely to have a few million unique visitors a month. However, even for blogospheres with a smaller number of readers and active members, there still exists a similar degree of separation for connections between two bloggers. Thus, one of the objectives of this dissertation is to research how to measure the rate of spread of influence based on the daily interactions of users within the community.

With tens of millions of blogposts available, it is highly improbable for each user to follow or read every author's documents. Instead, authors within the blogosphere are more likely to connect to other bloggers that may share similar viewpoints and post stories of interest. Therefore, I expect that more influential authors to have more connections and a larger presence with the blogosphere community. An author is observed as more popular than other bloggers based on a variety of factors on their blogs. The most important variables taken into account during this research include receiving a higher number of comments and followers, as well as individual blog recommendations in comparison to the rest of the community within the blogosphere.

These influential contacts are not only spreading neutral, objective information to their neighboring bloggers, but they are also passing on their own subjective views and sentiment on discussed topics. Even when the author's opinion and sentiment is subtle, it may possibly impact the sentiment and views of those reading the blog and writing future posts as well. This can lead to a spread of opinion and sentiment of a number of contacts across the blogosphere. While it may not be noticeable to the general reader, I hope to determine the hidden existence of this spread through the model.

Given the amount of subjectivity and author influence that exists while blogging compared to the mainstream media, blogospheres have become an especially important aspect in the world of politics [57]. In addition, the endless person-to-person interactions occurring within the blogosphere allows for a number of ways to represent the social network of the entire community. Hence, political blogospheres now present a new area of research opportunities that can go beyond the traditional text mining of mainstream news corpuses, thus the reason for pursuing the objectives within a blogosphere for this case study.

## Problem Description

This dissertation aims to provide a text mining method to measure sentiment and opinion within a document independently of one another. Additionally, this model tracks the evolving sentiment and opinion of the blogger along with author influence through the blogosphere. This is important due to the increasing number of people that become vital contributors of blogging sites across the internet, sharing their views which may eventually change over time.

Also, since the text within blogs is highly unstructured, it may only show subtle clues of subjectivity and sentiment. The overall goal within this dissertation is to model the sentiment and opinion over time of each author, as well as their immediate connections to determine an author's influence on those reading and recommending their blog.

The first step of the research is to model and visualize a dynamic stream of data for each blogger, and take into account not only the sentiment and opinion of each blog entry, but also the day the blog was posted, in order to track any changes in these variables. This provides a sense of how the blogger's sentiment and opinion on the given issue viewpoints evolve over time. In addition to tracking individual blogger views, I also analyze how respective bloggers interact with one another within the overall community. Finally, I model how the author's influence spread to other bloggers, which assists in developing a model for the influential spread of sentiment and opinion within the constructed social network.

## Research Objectives

Detecting which authors have the most influence on others' viewpoints within the blogosphere is challenging due to the following two major factors: 1) the immense volume of network connections due to the frequent number of interactions between bloggers, and 2) the subtle, and more likely hidden, relationships between an issue and the bloggers' sentiment and opinion. In order to better understand the influential impacts on bloggers' evolving sentiment and opinion, the resulting questions are answered:

1) <u>How can I model a blogger's sentiment and opinion separately for an issue over a specific time period?</u>

   For this problem, I first independently measure the issue sentiment and opinion of a single blog entry. To do so, I develop a model that measures the blogger's sentiment and opinion of two differing issue viewpoints within a blog as discussed in Chapter 3. The model first collects and investigates individual bloggers and issue content to extract key sentiment and representative sentences followed by evaluating an optimal scoring algorithm to independently score both sentiment and opinion values of each blog. Then additional techniques are used within Chapter 4 to best visualize and track the evolving changes in sentiment and opinion over time with the overall goal to determine any potential trends that may occur among multiple bloggers.

2) How do I model a blogger's influence of sentiment and opinion on the entire blogosphere community?

Utilizing the results from the developed model, the next step is to then correlate potential changes in the author's sentiment and opinion with neighboring bloggers in addition to the rest of the blogosphere. Within Chapter 5, the model provides a representative a social network based off the blogosphere to determine and measure how popular certain authors are and how closely connected they are to their neighboring bloggers. Once those two factors are determined, the model then calculates how quickly emotions and subjective views spread throughout the blogosphere, starting with the sentiment and opinion changes of the more prominent bloggers. This is done by modifying existing contagion models used to track the spread of viruses by substituting infected states with that of sentiment or opinion influence. In doing so, the model outputs an accurate representation of the existing sentiment and opinion states of the blogosphere.

By solely analyzing the text, this model does not only produce independent sentiment and opinion scores for each document, but it also measures blogger influence and the current and future states of sentiment and opinion within the entire blogosphere.

### Outline of this Thesis

The rest of the dissertation progresses as follows:  Chapter 2 provides a detailed literature review on past work related to this research. Chapters 3 through 5 explore the proposed solutions to the major research problems as well as provide an evaluation of the discussed techniques and methodology using collected data for the chosen case studies. Finally, the conclusion of overall results, summary of contributions, limitations, and future work are discussed in Chapter 6.

# LITERATURE REVIEW

In this chapter, I review previous literature related to the main research problems of determining the sentiment and opinion of an issue discussed within a blog as well as the influential spread of sentiment and opinion to the rest of the community within the blogosphere. I discuss how concepts from past studies relate to my aims and objectives as well as how this research either expands on or differs from previous research

## Related Work: Individual-Level Sentiment/Opinion Profiling

Most existing work in opinion mining from recent years is aimed at modeling and analyzing the evolution of either a topic and its associated sentiment or its associated opinion, but not both variables separately. This is an important factor I examine as the model determines if one of the two variables is more likely to spread versus the other. This model is different from existing research based on two main characteristics: 1) it results in a two-dimensional profile of a blogger's position on an issue based on the total level of expressed sentiment and opinion within a blog, and 2) it generates a visualization to discover key patterns in sentiment and opinion changes by analyzing the evolution of these two factors over the all blogs within the given period.

The initial process is to develop a model to score the opinion and sentiment for each blog. In doing so, I determine how people are feeling about a certain topic based on their writing. When reviewing past research, these two terms, sentiment and opinion (subjectivity), tend to be used almost interchangeably, however, this research shows that these two words can vary in meaning. Thus, one of the research objectives is to discover how to best separate any connection between these two terms into independent measurable variables.

In 1985, Randolph Quirk defined subjective language as a private state not open to objective observation or verification [46]. Later, Wiebe et al. provided further Natural Language Processing (NLP) by beginning to label sub-components of subjective writing to include the writer's attitude when discussing a specific topic [54][55]. Then in 2007, full list of subjective clues was introduced by Furuse et al. to extract opinion sentences from unstructured blogs [20]. This list identified opinion based on more than solely a positive or negative sentiment, but to also include additional subtle writing clues such as other themes of thought, value, and uncertainty. While this model still showed some interdependence between sentiment and subjectivity, the intent of my research is to show how opinion varies significantly when only using positive and negative terms, and that sometimes using sentiment alone does not significantly alter the overall objectivity of the writing.

Moreover in 2011, Chen et al. developed an opinion extraction model which scored subjective sentences by first using the opinion clues from the above research to define the possible existence of an opinion [11], and then by calculating additional feature functions to score the level of opinion for that specific sentence. For the model I am proposing in this research, it takes a slightly different approach by using supplementary and independent feature functions to score both the sentiment and opinion of each extracted sentence.

To expand on the first objective this model utilizes hidden associations between three independent factors: topic, sentiment, and opinion, and in doing so, leverages the previous research on topic modeling to uncover semantic structure by mining for meaningful hidden topic associations. The Topic-Sentiment Model created by Mei and Zhai in 2007 calculated sentiment of documents by modeling a mixture of predictions of both topics and sentiment [40], and Lin and He’s Joint Sentiment/Topic model in 2009 captured the sentiment at the document level opposed to the topic level [34]. Additionally in 2012, Lin et al. also proposed an aspect-based model to extract objective and subjective opinions, which included sentiment, for given topics by identifying the hidden relations between topic and opinion words [35]. Moreover in 2011, Jo and Oh developed an aspect and sentiment unification model for online review analysis by detecting both the sentiment and discussed topic simultaneously [32].

Previous research focus specifically on determining the hidden relations between topic words and either solely sentiment words or solely opinion words which sometimes even incorporate sentiment words into the subjectivity lexicon. To expand on this method for this research, I follow a similar methodology for finding the hidden associations in the model, but also investigate this for sentiment words from opinion words independent from one another. In doing so, I discover existing hidden relations between not just two factors, but all three variables (topic, sentiment, and opinion). Furthermore, I choose to use the word “issue” opposed to “topic” to focus on subjects that tend to have at least two different sides or viewpoints, which is a common occurrence in political blogospheres.

In addition, I examined past work analyzing the evolution of topic and sentiment in document collections. Most of this work continues to be one-dimensional, as it only examined either sentiment or opinion in relation to topic modeling, while the objective to model both sentiment and opinion requires modeling a two-dimensional score. Originally in 2005, Mei and Zhai explored temporal text mining and discussed a probabilistic approach to spatiotemporal theme pattern mining on web blogs, modeling the evolution of topics over time [40]. Further in 2011, Chen proposed a profile-based topic predicting model to estimate the occurrence of future topics in the political blogosphere [12], while around the same time, Blei and Lafferty offered a dynamic topic model to depict the evolution of topics over time by estimating the distribution of topics at various times [7].

To visualize the temporal sentiment analysis of the topic modeling described above, in 2007 Fukuhara, Nakagawa and Nishida proposed two graphs, each displaying the temporal change of either the sentiment associated with a topic or temporal change of topics with sentiment [19]. Finally in 2012, Zheng et al. suggested a topic sentiment trend model, which integrates the topic with sentiment and analyzes the temporal trend of the sentiment-topic to combine both the topic and sentiment evolution [58]. In conclusion, the past analyses only looked at either one-dimensional subjectivity or positive and negative sentiment over time associated with the topic. Within this model, I expand on the appropriate methodology to model and display any temporal trends within the data using a combination of all three factors (topic, sentiment, and opinion).

Additionally, I reviewed two recent visualizations that tracked sentiment corresponding to multiple topics over time. The first was a time-aware topic-sentiment (TTS) graphical model that analyzed the joint topic-sentiment evolution over time to quantitatively measure the polarity against the ground truth [31]. The second visualization, SentimentRiver, was used to analyze the dynamic sentiment of a twitter stream with the purpose of demonstrating the effectiveness of illustrating sentiments on specific topics discussed on social media [15]. While both of these models were successful in displaying topic sentiment, their focus was on visualizing a single variable. For my contribution, this research leverages techniques from both existing models to visualize two-dimensional scores of sentiment and opinion over time with the goal of evaluating the modified models to determine the best single visualization that incorporates both variables.

## Related Work: Blogger Influence and Sentiment Spread

The other significant component within the blogosphere modeled is a given author's influence to other connected bloggers. The literature review for this section first explores how past research has modeled the flow of information within a blogosphere, as well as any exposure of sentiment onto fellow bloggers.

In the past decade, there has been increased focus on visualizing sentiment change within the community on a given topic. Recall that the second research objective is to develop and visualize a social network analysis as well as any spread of influence within the blogosphere. In doing so, I expand on the recent research that has been published in this field. The differentiating piece in this work is that this model uses any influential measurements determined from the analysis of the social network, and uses those factors to output the overall influential spread of sentiment and opinion through the entire blogosphere.

The first group of past literature I reviewed was in regards to advancements in modeling social networks. This idea started initially in 1978 when Granovetter proposed a threshold model of collective behavior to determine deviance of individuals and thresholds where those individuals show the same selection patterns as one another [23]. Research then shifted into focusing on the online community, to include modeling blogospheres as social networks. In 2005, Herring et al. investigated patterns and characteristics within a blogosphere to include the interconnectedness of neighboring blogs and determined which predominant blogs were more central or important within the community [26]. Then by 2014, Gliwa, Zygmunt, and Koźlak analyzed various characteristics of groups within a blogosphere to identify social relations by considering different types of interactions, such as focusing on the classifications of comments directly addressing others within the blog and then calculating the sentiment of those blogs [21].

To continue the research of combining sentiment and social networks, West et al. presented a method to exploit the social network structure to incorporate sentiment analysis by implementing a signed-network analysis to model and jointly predict the social network in addition to the sentiment of person-to-person evaluations [53]. In this work, I expand even further by incorporating opinion into the social network separately from the fused sentiment influence, while at the same time, exploring different approaches to determine which nodes are more centralized as well as measure the edge distances connecting bloggers to one another.

The next collection of literature focuses on recent work involving modeling influence within a blogosphere. In 2013, Wang et al. proposed using relationship maps to characterize the different sentiments of each topic and the interactions between authors [52]. Also, Hui and Gregory created a blog model to quantify sentiment and influence within a blogosphere to apply the relation map visuals to a blogspace [30]. Their algorithm computes per-topic influence-weighted sentiment by measuring influenced based on web-ranking algorithm by capturing features such as follower numbers, topic relevance, comments, and followers. Finally in 2014, to expand on the use of thresholds in the information environment field, Sela, Oved, and Ben-Gal proposed a method to model the information spread in a connected world [47]. More specifically, they examined the spread of information by methods of word-of-mouth and through search engines. Their suggested methodology was to use thresholds of neighbors holding opinions for propagation to occur in the network.

Additionally in 2007, Song, Chi, Hino, and Teseng developed a way to identify "opinion leaders" in a blogspace by measuring blogger influence based on comparisons of original and reposted material (not on sentiment) [48]. Then in 2012, Sukute further examined the behavior patterns of the opinion leaders of various sources to determine any trends in their actions [49]. To expand on the above methodology to fit this blogosphere problem, the model and accompanying visualization examine similar factors that determine which bloggers are more influential than others as well as measure rates of susceptibility and influence.

Another idea described in previous research revolves around the concept of using viral models in regard to the spread of information. Originally in 2004, Dodds and Watts presented a model of contagion unifying and generalizing the spread of social influences and infectious diseases [16]. Following this research, Hill proposed evaluating the long-term spread of emotional states based on social network interactions using a classical epidemic model [27][28], thus incorporating a Susceptible-Infectious-Recovery (SIR) model [33] to show the spread of sentiment analysis in a social network. In determining a more advanced way to model sentiment propagation, Liu, Zhang, and Lan more recently proposed a newer model expanding on the SIR model developed from Hill's initial research [37]. In doing so, they defined a modified version of the infected state within the SIR model, by splitting that state into two unique states matching the opposing sentiments, positive/negative or optimistic/pessimistic. Thus, they portrayed the spread of sentiment contagion for both sentiments opposed to a single infected state.

Similar to the previous literature cited above, the model incorporates contagion-modeling techniques to track the spread of sentiment and opinion. To do so, I also incorporate the ideology of competition among viruses within the same environment as discussed in 2012 both by Beutel et al. [6] and Myers and Leskovec [44].

# CALCULATING THE BLOGGER'S ISSUE SENTIMENT AND OPINION (ISO) PROFILE SCORE

In this chapter, the objective is to separately score the sentiment and opinion as two independent factors within a blogger's single post on a given issue. To achieve this goal, representative sentences are extracted from the blogger's writing and classified based on both its level of subjectivity, from objective to weak or strong opinion, as well as the degree of sentiment as neutral, positive, or negative.

I start this section by presenting a formal definition of the problem. This is followed by the proposed methodology to discover and interpret clues within the text connecting issue keywords with related sentiment and opinion words. Next, I then build the Issue Sentiment and Opinion (ISO) model by merging and modifying existing scoring algorithms to determine the two-dimensional profile score of a blogger for the given issue viewpoints. Afterwards, I present the case study used throughout this dissertation, as well as review the experimental results using the discussed methodology, concluding with the next steps and ideas for future research.

## Problem Definition

This section contains a full set of definitions and variables that are referenced within this chapter. The problem of this chapter focuses on an individual blogger, which, for ISO scoring purposes, refers to a single author. The blogger's role outside of being an author is discussed in Chapter 5. The blogger is denoted by $B_i \in \Omega$, where $\Omega$ represents the blogosphere which is a set containing the entire community of *m* bloggers, $\{B_1, B_2, ..., B_m\}$.

In this chapter, there is an emphasis on a specific issue of interest, and within each issue, I declare a set of aspects corresponding to more than one perspective of the issue. For this study, I focus on splitting the selected issue into two well-defined, opposing viewpoints, $v_1$ and $v_2$, each composed of an independent collection of noun words. For the specific issue, each document (also referred to as a blog or blogpost) that the blogger, $B_i$ writes is denoted as $b_j^i$, where *i* matches the corresponding blogger, and *j* denotes the specific blog within $\{b_1^i, ..., b_n^i\}$, the collection of *n* blogs which are written by the blogger $B_i$ sequenced in chronological order. Each blog is composed into a group of individual sentences, $\{e_1, ..., e_n\}$, where each sentence contains a string of words $\{w_1,...,w_n\}$, which is the basis of extraction for indicating the level opinion and sentiment.

Using the proposed methodology, both sentiment and opinion scores are assigned for individual sentences through two scoring formulas, each composed of underlying feature functions, all of which is discussed in more detail below. The scoring functions are then propagated upwards to provide a top-level blog score, $b_1^i(o, s)$ for each of the two issue viewpoints, where *o* represents the opinion score and *s* represents the score for sentiment. The ISO scores for each blog are then used to visualize the blogger’s ISO at specific instances, and for the following chapters, the scores are used to help determine any patterns in sentiment and opinion changes over time.

## Issue Sentiment and Opinion (ISO) Scoring Model

For each sentence within the document, the blogger's score are based on either writing objective statements about a given issue or expressing a level of opinion using key subjective words. This also holds true for the amount of sentiment within each sentence, which falls into one of three categories: neutral, negative, or positive. Finally, the sentiment and opinion values are scored independently of one another for each sentence.

Prior to scoring each sentence, the initial step is to build two lists that represent the two defined issue viewpoints, $v_1$ and $v_2$. These serve as the model's issue features for the entire case study. To begin scoring each sentence, the first step is to identify common influential features that likely appear in the text when personal feelings and opinion are present. If any influential words are discovered, opinion and sentiment feature functions are then used to score sentiment and opinion associated with both $v_1$ and $v_2$. Therefore, I propose using a modified version of Chen's Opinion Scoring Model [11], by constructing the following high-level features to calculate the Issue Sentiment and Opinion (ISO) score for each sentence: issue features, influence features, opinion features, and sentiment features.

## Issue Features

Issue features found within the text detect the specific topics discussed by the blogger. These features are extracted from documents in the forms of noun words. To help identify which of the two viewpoints is referenced in a specific sentence, the issue feature function is defined as $f_{iss}(s)$, which detects the equivalent noun words and then returns a value corresponding to whether the sentence is more likely associated with viewpoint $v_1$ or $v_2$.

### Influence Features

Prior to further analyzing each sentence, the first step is to measure whether there is enough evidence to determine if any opinion and/or sentiment occur. Therefore if the influence feature function, $f_{inf}(s)$ produces specific value, the model declares that there exists enough influence to move forward with calculating an overall ISO score for that sentence. Otherwise, it skips searching for opinion and sentiment features and moves onto the next sentence.

### Opinion Features

As with extracting noun words to define the issue features, opinion features are determined through the use of other parts of speech (i.e. adjectives, verbs, and adverbs). For each part of speech, different words are used to express distinct sentiments and opinions in relation to the associated sides of the issue. Usage patterns of opinion features also are analyzed to determine the blogger's opinion by means of the following feature function: $f_{Oaav}(s)$, which measures the usage statistics of adjectives, adverbs, and verbs which are commonly used to express subjectivity. This function returns a higher value when an opinion is read. Similarly, another opinion feature function calculates the frequency of double adjectives used, and is referred to as $f_{2adj}(s)$.

Once the model evaluates the above functions, the next step is to calculate the overall opinion scoring function, $f_{opinion}(s)$, which incorporates each of the individual opinion feature functions previously discussed, producing a score for each sentence based on the total level of subjectivity.

### Sentiment Features

Similar to the opinion features, the same parts of speech are examined to find sentences where the blogger expresses sentiment. This feature function is represented as $f_{Saav}(s)$. In addition, the model also incorporates another sentiment feature function that calculates the dependency distance function, $f_{dep}(s)$. The dependency distance refers to the number of words between the sentiment and connected issue viewpoint word, where a value of zero represents an adjective word directly preceding the noun.

Finally, the sentiment scoring function, $f_{sentiment}(s)$ incorporates both of the feature functions to produce an overall sentiment score for each sentence within the document.

## Defining the Qualitative Features

In this section, I discuss in further detail each of the feature functions mentioned above starting with the issue features followed by the respective features for influence, sentiment, and opinion.

### Issue Function

Before determining any influence that may exist within the blog, the issue feature function determines the viewpoint within the sentence. The first step to define each viewpoint is to manually collect and label noun words from numerous blogposts on the specific topic to create a corpus for each of the two viewpoints. As viewpoints on specific issues change, it is important to maintain an evolving lexicon for evaluation of the issue feature function, thus the use of a frequency-based word weighting such as discussed in Hohman's analysis of streaming news documents [29]. The issue lexicon continues to increase at specific time nodes, which is used as the basis for classifying sentences from blogs posted as either $v_1$ or $v_2$.

Within the model, the first objective is to define if a word in the given sentence represents an issue. To optimize the classifying algorithm, WordNet v3.0 [18] is implemented into the model, providing the part of speech (POS) tag of the given word. Since noun words describe the issue viewpoints, the model collect the noun words within the sentence and use that against the viewpoint lexicons.

For the built-in classifier, the model incorporates a primal support vector machine (SVM) to determine the optimal hyperplane best separating the two viewpoints. Since the output of the SVM classifier produces a continuous value between 0 and 1, I elect to use 0.5 as a decision factor $f_{inf}(s)$, which returns either a value of either 0 or 1 for any value less than or greater than 0.5 respectively. For this function, a value of 0 represents $v_1$ and a value of 1 represents $v_2$. If the returned value is closer to 0.5, it shows that words from both sides of the issue are being used within the same sentence mainly because the model does not differ from subject versus object within a sentence. To address a similar issue with longer compound sentences, compound words (i.e. *and*, *or, because)* are considered sentence dividers, similar to a period or semicolon forming two smaller sentences, for which the model then applies the function to each of the new sentences.

### Influence Function

Prior to scoring a sentence, a simple algorithm is used to determine whether there is a possibility that a sentence expresses any potential influence, which is an effective way to determine if opinion or sentiment does exist within the sentence prior to using larger corpuses [11]. To do so, a corpus of words which best determines the existence of influence is created by using an evaluator corresponding to various judgment clues. These influence features consist of a collection of frequently used sentiment and opinion words with respect to journalism [20]. Table 1 provides a subset of the most common judgment clues found within the political blogosphere to also include example words that fall into each category.

A full list of word clues used for determining the presence of influence within each sentence is further described in [20], and synonyms of these words are found using WordNet v3.0 [18].

**Table 1: Sentence Influence Clues**

| Category: | Example Words: |
|---|---|
| Sentiment Judgment | Terrific, Awful, Horrifying |
| Emotion | Glad, Pleased, Concerned |
| Propositional Attitude | Should, Would, Ought |
| Thought | Think, Consider, Believe |
| Uncertainty | Doubting, Questioning, Wondering |
| Intensifier | Extremely, Tremendously, Awfully |
| Impression | Confusing, Bewildering, Surprisingly |
| Declarative | Possibly, Might, Probably |

If any single word from one of the judgment clues is located within the select sentence, then the influence feature function, $f_{inf}(s)$, returns a value of 1, indicating that the specific sentence is then be examined further to determine what level of opinion or sentiment exists. Otherwise, $f_{inf}(s)$ returns a value of zero, providing the model with two commands to move to the next sentence within the document. This iterative process continues until all sentences within the document have been scored. Note that if the sentence is skipped due to no judgment clues being found, that sentence is immediately scored a value of zero for both the sentiment and opinion representing both a neutral and objective sentence within the document. Additionally, this optimizes the rest of the algorithms within the model by reducing computation time and skipping over any opinion and sentiment functions since they would provide identical zero scores.

### Opinion Score

When describing an issue, a blogger either discusses topic viewpoints using only objective information or presents their views in the form of an opinion. If there is any existence of influence in the sentence, then $f_{inf}(s)$ equals 1, and the next step is to determine the level of opinion within each sentence. For more subtle or sparse opinions, the text is evaluated as weak subjective; while more noticeable attitudes return a higher value representing a stronger subjective score. Although opinion scores are categorized into three different groups, opinion is still scored as a continuous value opposed to a discrete score like the influence function.

The first step in evaluating the opinion score is to collect the most commonly used opinion words. For this research, opinion words are determined through the use of publically available corpuses, such as OpinionFinder [56] and MPQA Corpus [54][55]. Both corpuses include limited numbers of sentiment words, which are excluded as scoring for opinion is done independently. Therefore, another corpus is also used, SentiWordNet [2], based off of the existing WordNet [18], to remove words from the opinion word database that appeared in the sentiment corpus, while at the same time, remove opinion words from the sentiment lexicon.

Afterwards, both opinion feature functions are measured based on the opinion words (adjective, adverb, and verb) usage frequency found within the sentence. This function returns a continuous value of $f_{Oaav}(s)$ where a higher value represents a stronger subjectivity score. The second feature, $f_{advj}(s)$, measures the total number of adverb-adjective combinations, which are the instances where these words appear concurrently within that sentence.

Additionally, the use of multiple features provides a higher accuracy than individual functions as discussed in other related research [4]. Therefore in order to combine the above features into one score, the model trains itself based on a linear regression model using the discussed features. The combination of the two feature functions to score the opinion level of sentence on the opinion is as follows:

$$f_{opinion}(s) = \alpha_0 + \alpha_1 * f_{Oaav}(s) + \alpha_2 * f_{advj}(s). \tag{3.1}$$

This function returns a continuous value between 0 and 1, where a value of 0 would represent a truly objective piece and any value closer to 1 signifies how strong an opinion exists in the sentence. Weak subjective sentences are scored for sentences that score less than 0.5, while any score over 0.5 represents a stronger subjective sentence.

### Sentiment Score

Similar to the opinion score, this model utilizes two feature functions using the set of adjectives, adverbs, and verbs. As discussed above, the sentiment of words in the sentence is determined by utilizing the existing corpus of sentiment words, SentiWordNet [2]. This database scores these words as positive, negative, or neutral, which is then used to help calculate the overall sentiment of the sentence. In addition to the words in the latest sentiment corpus, I also augment the list to include trending and slang words that may not be in the lexicon as well as amend any words that have changing sentiment words that differ in the blogs versus what appears in the SentiWordNet lexicon database. For example, while certain words like *awful* and *egregious* have changed between sentiments over centuries and would not impact existing lexicons, currently use trendy words such as *snazzy* and *dicey* have more recent changes and need to be reflected [1]. The overall process is used to determine interchangeability of neighboring terms and/or provide new and different sentiment words within different time-spans.

Once the sentiment word database is established, two sentiment feature functions. The first feature function, $f_{Saav}(s)$, calculates the frequency of sentiment words and returns a continuous value between -1 and 1. The second sentiment feature function calculates the dependency distance function, $f_{dep}(s)$, returning a score based on the number of words between each sentiment word and the nearest issue viewpoint word [24]. The dependency path consists of a modified scoring algorithm based on a constructed dependency graph and shortest path algorithm, where scoring is based on an established list of grammatical dependency distances which determine any meaningful relationships, where a higher returned value indicates a shorter distance implying a stronger sentiment [24].

Then similar to the two opinion feature functions, the overall sentence-level sentiment function is calculated by using a linear regression model with the following feature functions:

(3.2)

$$f_{sentiment}(s) = \alpha_0 + \alpha_1 * f_{Saav}(s) + \alpha_2 * f_{dep}(s).$$

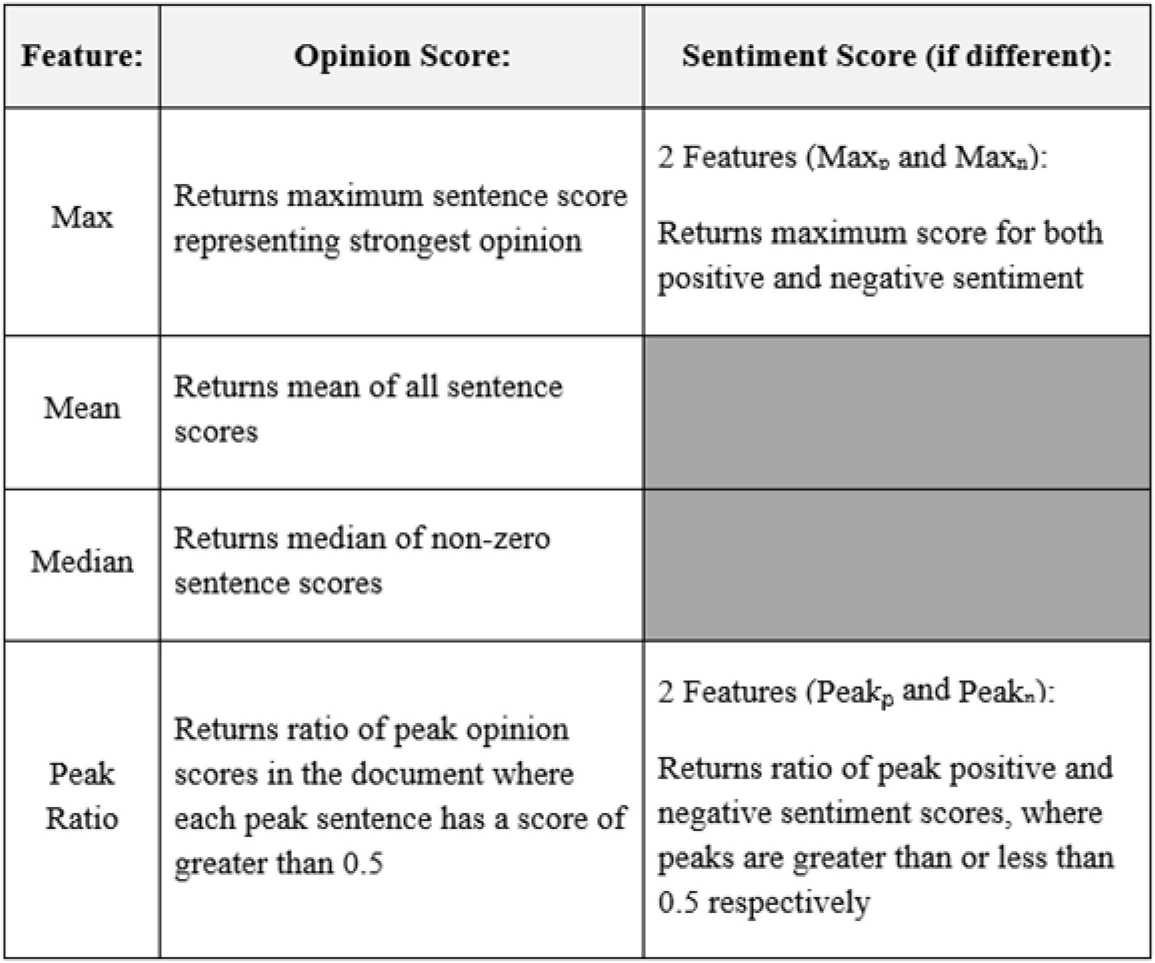
However, unlike the opinion function which returns values between 0 and 1, the sentiment scoring function returns a continuous value between -1 and 1, where a value of 0 represents a truly neutral sentence and any value closer to -1 and 1 indicates the existence of negative or positive sentiment, respectively.

### Blog-Level ISO Score

The next crucial step in the model is to calculate a single blog ISO score for each document based on the individual sentence scores.

To score each blog, I first analyze multiple statistics to determine which one performs best. While the median or mean of sentence scores for each blog seem useful, they tend to usually not perform well, especially each blog varies from five or six sentences to over one hundred sentences. Table 2, below, summarizes the statistics used to calculate the blog-level ISO score. A description of the best and worst performing statistics is discussed in the following analysis section.

**Table 2: Blog-Level Score Features**



| Feature: | Opinion Score: | Sentiment Score (if different): |
|---|---|---|
| Max | Returns maximum sentence score representing strongest opinion | 2 Features ($Max_p$ and $Max_n$):<br>Returns maximum score for both positive and negative sentiment |
| Mean | Returns mean of all sentence scores | |
| Median | Returns median of non-zero sentence scores | |
| Peak Ratio | Returns ratio of peak opinion scores in the document where each peak sentence has a score of greater than 0.5 | 2 Features ($Peak_p$ and $Peak_n$):<br>Returns ratio of peak positive and negative sentiment scores, where peaks are greater than or less than 0.5 respectively |

## Experimental Analysis

In this section, I first discuss the entire collection and analysis of blogosphere data using the scoring algorithms from the previous sections. I then review both sentiment and opinion scores for a set of documents at both the sentence and blog-level, providing key qualitative and quantitative results. Finally I conclude with a review of findings based on the model's performance for the specific case study.

For this research, MATLAB® is used for all or data collection and experimental analysis. Although there are other computational software tools more commonly preferred for text mining, my choice for using MATLAB® to develop the model is based on previous years of experience along with the following list of recommended reasons, in that this program [3]:

- Runs programs in a high level application-oriented language which is user-friendly and simple to understand
- Provides functions and algorithms come included from pre-programmed functions and toolboxes
- Allows for simple steps to generate various plots and visualizations from much larger datasets

In addition, MATLAB® is a primary focus on performing computational operations using high-dimensional matrices. Rafael Banch's complete work of *Text Mining with MATLAB®* provides additional methods to operate with collected unstructured data by using vector space models with complete list of models and applications to handle and explore text data available in his work [3].

## Data Collection

The main focus of this research revolves around the popular blogosphere *DailyKos* [14], which accounts for a variety of political and national issues. *DailyKos* receives hundreds of thousands of views on an average day and was ranked the second best blog by *Time* readers in 2009 [39].

The first step was to gather blogs for a specific issue of interest. For this case study, I started with a limited collection of 2,881 blogs, all related to the selected issue regarding the Gaza Strip, with two issue viewpoints focusing on Israel and Palestine. Since the metadata for document topic keywords were user-defined, I also included a variety of similar topic keywords into one category (i.e. Gaza, Israel, Palestine, Hamas, Jerusalem, etc.).

Afterwards, collected blogs were then sorted into six-month increments starting from January 2011 and ending at June 2015. Afterwards, a term frequency inverse document frequency (TFIDF) vector for each time step is used to determine the high-frequency words for that timeframe [29]. To do so, a random sample of 20% of the blogs within each time step is chosen, followed by manually labeling of appropriate noun words into one of the two viewpoints where possible.

To better explain this concept are visualized portions of the training set using Wordle, which creates word clouds that display the most common keywords [17]. Figure 2 shows what the most common words would be using a training set consisting of a random collection of 20% of the blogs over the entire timeframe (Jul 2012 – Jun 2015). Figure 3 displays a 20% training set for three selected periods within the overall timeframe.

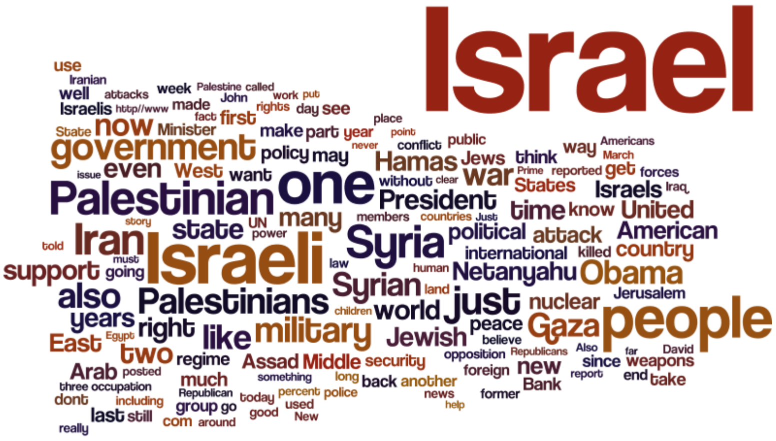


**Figure 2: Word Cloud (Entire Timeline) [17]**

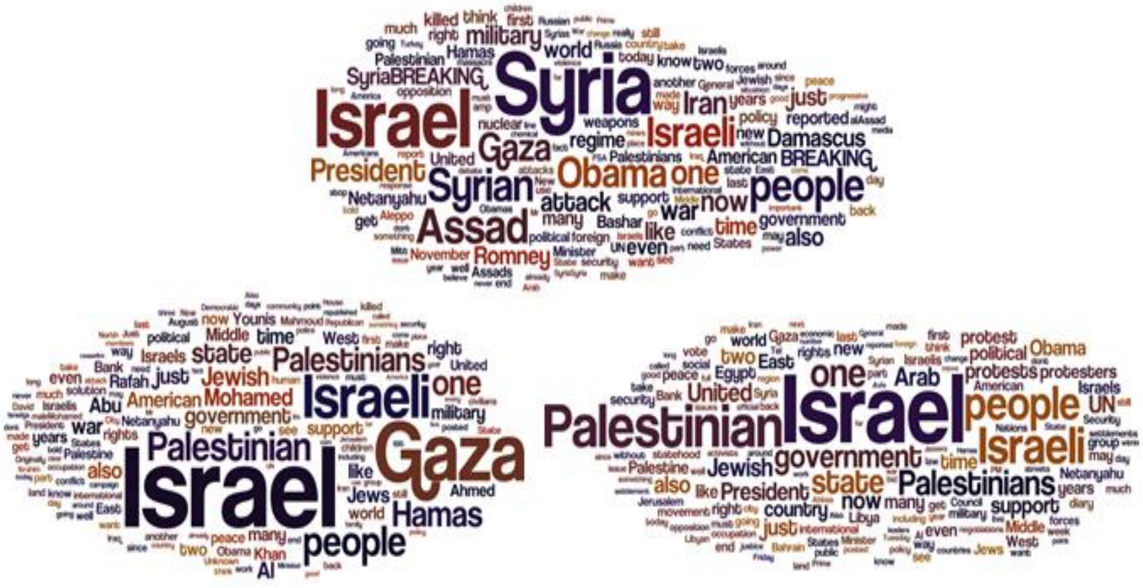


**Figure 3: Word Clouds for Jan-Jul in 2012 (Top), 2014 (Left), and 2015 (Right) [17]**

As initially expected, there exists a large collection of common words used to describe this case study no matter which timeframe is used. By manually separating these words, a set of two unique lexicons are constructed to differentiate the two issue viewpoints. The most frequently used keywords within each of the lexicons are shown below in Table 3.

**Table 3: Issue Viewpoint Frequent Keywords**

| **Israel ($v_1$)** | **Palestine ($v_2$)** |
|---|---|
| israel | palestin |
| netanyahu | moh/muh/mah |
| jewish | arab |
| forces | hamas |
| navy | islam |
| idf/fdi | muslim |

However, after taking a closer look at the TFIDF vectors for each time step, it was determined that a number of terms frequently are used during those specific periods, but which are rarely discussed over the much larger timeline. For example, during the 2012 presidential elections, Mitt Romney was discussed more often than in 2013 and beyond after he had lost the election in November 2012.

Additionally, there are more specific incidents that spurred more dialogue than normal, such as the attack at a school in the Gaza city of Rafah. According to Google Trends, this incident peaked interest for online searches and tweets in the summer of 2014 [22], likely prompting more blogposts on this topic within the blogosphere as well. While the city itself might not fall under either viewpoint, key people and terms related to that event are likely to be captured instead.

Figure 4 [22] provides a better understanding the trending discussions regarding the event in Rafah city of Gaza peaked for a couple months in 2014, but otherwise, there were limited discussions outside of that time period.

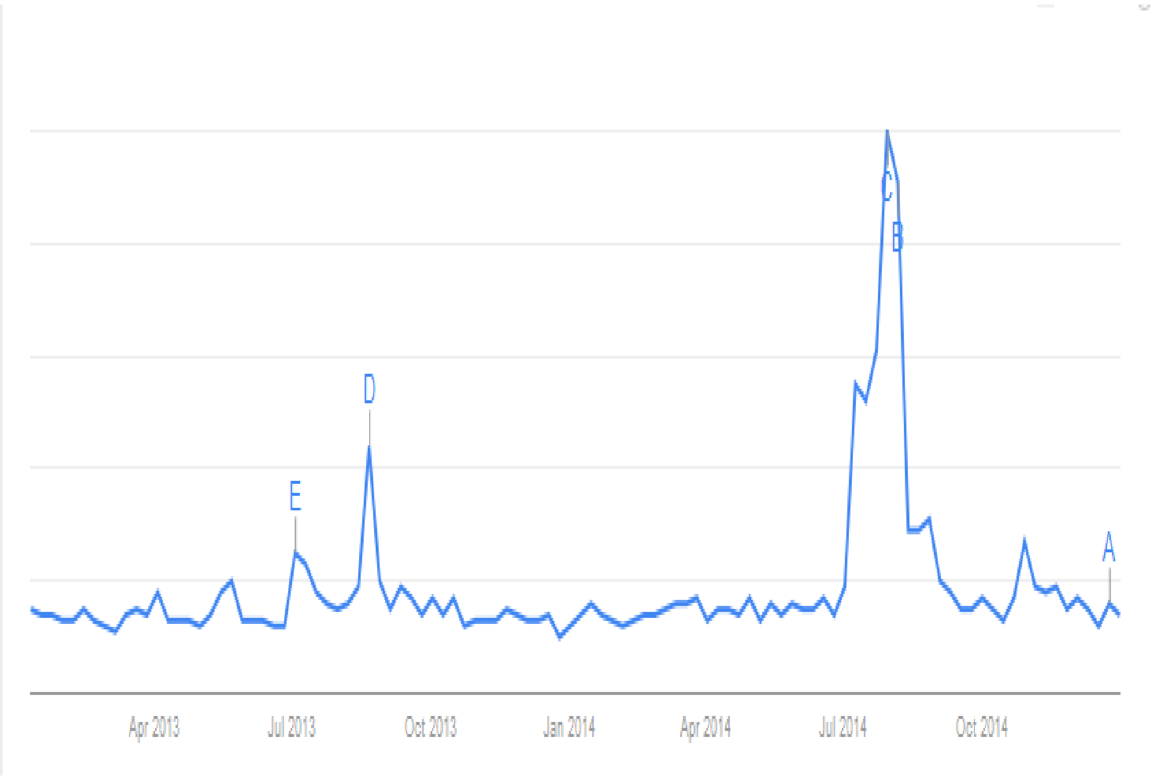


**Figure 4: Trending Discussions on the Gaza City of Rafah [22]**

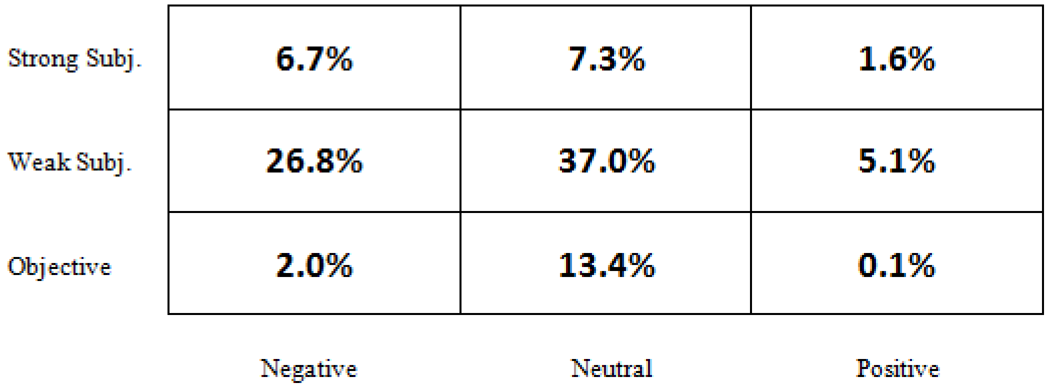
Thus if only 20% of randomly selected documents of the whole corpus is trained, many terms would likely be missed with manual classification of the two issue viewpoints. However, by splitting the corpus into segments and evaluating the TFIDF vectors for each period, a larger issue feature set is constructed to provide a more accurate picture of the bloggers' ISO scores.

## Sentence-Level ISO Scores

The selected corpus consisted of 58,615 sentences matching with one of the two issue viewpoints. Of those sentences, 68.8% corresponded to Israel ($v_1$) and 31.2%

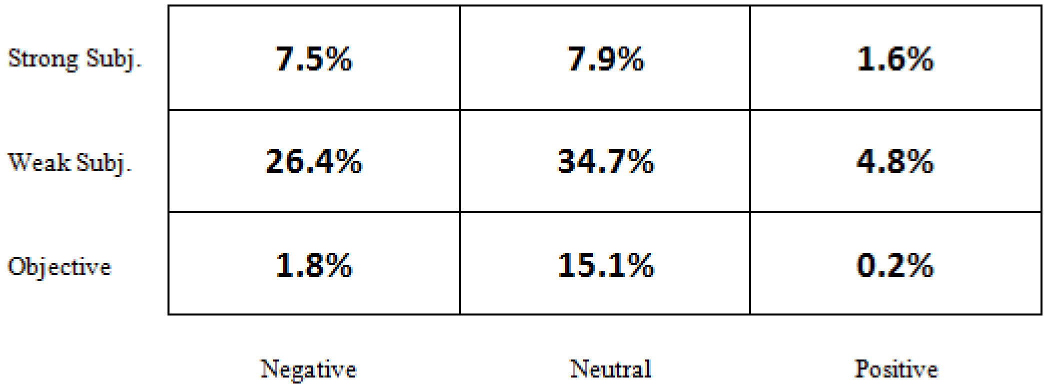
corresponded to Palestine (*$v_2$*). The following two figures, Figure 5 and Figure 6 show the breakdown of opinion and sentiment scores for each of the issue viewpoints.

Although bloggers were twice as more likely to discuss Israel topics in their documents, the breakdown of sentiment and opinion between the two issue viewpoints were very similar. The key result from the initial analysis is that only around 15% of all sentences were actually objective and neutral, showing that a large majority of bloggers do express some amount of personal feelings in writing. Secondly, on both ends of the spectrum, bloggers tend to be much more negative about the issue with roughly only 7% of sentences scoring a positive sentiment. In reality, this does make sense, since this case study focuses on a very sensitive topic surrounding the constant conflicts and unfortunate events within the region.



| | | | |
|---|---|---|---|
| Strong Subj. | 6.7% | 7.3% | 1.6% |
| Weak Subj. | 26.8% | 37.0% | 5.1% |
| Objective | 2.0% | 13.4% | 0.1% |
| | Negative | Neutral | Positive |

**Figure 5: Israel (*$v_1$*) Sentence-Level ISO Breakdown**



| | | | |
|---|---|---|---|
| Strong Subj. | 7.5% | 7.9% | 1.6% |
| Weak Subj. | 26.4% | 34.7% | 4.8% |
| Objective | 1.8% | 15.1% | 0.2% |
| | Negative | Neutral | Positive |

**Figure 6: Palestine (*$v_2$*) Sentence-Level ISO Breakdown**

Additionally, results also show the existence of sentences showing subjectivity but without any sentiment keywords, as well as fully objective sentences that still hold some level of negative or positive sentiment, although the latter of these two examples was much more uncommon in this case study. In further research I would like to improve on these findings by experimenting with other issues and blogospheres to help determine any additional steps needed in either the model or corpuses to analyze difference in subjectivity and sentiment while. At the same time, I also would want to further examine the relationship between sentiment and opinion to determine how to calculate the interdependencies between the two variables and whether one variable is more dependent the other but not vice versa.

To look beyond the numerical scores discussed above, Figure 7 provides a further understanding of the breakdown of sentiment and opinion scores by analyzing selected example sentences within the different categories for both viewpoints.

Issue viewpoints are marked light blue to represent issue viewpoint $v_1$ and dark blue is used to represent $v_2$. The keywords denoting the author's opinion are highlighted in yellow, while sentiment keywords are marked as red and green words corresponding to either negative or positive sentiment.

**Weak Negative ($v_1$):**

It is bad enough that Boehner is inviting a foreign leader to argue against our administrations delicate and historic negotiations with Iran in front of Congress

**Weak Negative ($v_2$):**

Instead of allowing the security to be turned over so that rebuilding from donor aid could continue Hamas uses what little funds available to not help rebuild Gaza, but to rebuild its military threat instead

**Weak Positive ($v_1$):**

In light of the visit by Prime Minister Benjamin Netanyahu to Washington and Obamas victory in November I believe that there is a great chance to resolve the Mideast Palestinian Israeli conflict

**Strong Negative ($v_1$):**

It would put even more pressure on Israel to explain why it is using such ferocious attacks on a civilian population it has illegally oppressed for so many years

**Objective Negative ($v_1$):**

IDF troops went house to house and killed between 140 and 275 people

**Objective Negative ($v_2$):**

The terror group Hamas has been accused of responsibility for the recent terror bombings of Fatah embers homes in Gaza

**Strong Neutral ($v_1$):**

I strongly believe in the right of Israel to exist

**Figure 7: Example Sentences**

Unfortunately, many of the observed sentences could not be classified as either issue viewpoint. There were two major reasons for this to occur. The first challenge is the use of pronouns in various sentences, while the second is the number of sentences using a single word from each viewpoint, especially in shorter sentences. To avoid these problems in future research, I propose enhancing the model by classifying the issue viewpoint of sentences based on the subject, object, or predicate of the sentence, as well as to incorporate further methods which relate pronouns to the correct viewpoints possibly based on the classification of previous sentences.

### Blog-Level ISO Scores

The next step is to take the sentence scores and propagate them upwards to create 2,881 blog-level scores for each of the two issue viewpoints. A score breakdown is shown in Table 4 based on using the mean of the peak-ratio sentence scores.

A full analysis showed that both the mean and peak-ratio scores provided the most insight on blog-level sentiment and opinion. Median scores were almost always at or near zero if I looked at the entire document, likely due to the fact that a majority of the middle of the documents were filled with objective, neutral information. Future research would expand the scoring based on a combination of all useful statistics as well as the previously mentioned notion that more personal feelings are expressed near the beginning and end of the document.

**Table 4: Average Blog-Level ISO Scores**

| | Israel ($v_1$) | Palestine ($v_2$) |
|---|---|---|
| Sentiment | -0.32 | -0.29 |
| Opinion | 0.31 | 0.25 |

Using the blog-level scores, I see that there is slightly more negative sentiment and strong opinion for issue viewpoint $v_1$. Also, since positive sentiment is not as common for all sentences, blog-level scores were nearly always neutral or negative. While single blog-level sentiment and opinion scores are simpler to calculate and easier to compare with one another, it unfortunately does not provide as detailed of a picture of the nine different categories, as will be shown in the sentence-level versus blog-level visualizations discussed in the next chapter.

## Conclusion

In this chapter, I propose scoring documents for opinion and sentiment independently with the text mining model. Experimental results show those cases where displaying opinion did not necessarily mean expressing a positive or negative tone, and occasionally, either sentiment is used to describe events without infringing on the text's objectivity. At the same time, I show that for this case study in particular, there is stronger frequency towards a greater negative sentiment for any level of opinion.

For future work, I propose increasing the number of issue viewpoints or breaking the two viewpoints down into further subtopics. I would also examine determining sentiment and opinion based on the subject and/or object of the sentence, to improve the accuracy of the model. Additionally, I would explore better building the blog-level ISO scores by experimenting with larger datasets, to include other blogospheres websites as well as across multiple topics within a single blogosphere.

To continue researching this specific problem, the next chapter visualizes existing ISO scores corresponding to specific bloggers over time and look for any trends or differences in sentiment and opinion compared with the overall blogosphere average. While blog-level calculations provide a single propagated score at each time step, this research more specifically examines a full analysis of sentence-level scores to provide a complete picture of each document to provide a dynamic picture of the blogger's sentiment and opinion scores.

# VISUALIZING THE EVOLUTION OF ISSUE SENTIMENT AND OPINION OVER TIME

During the previous chapter, the ISO model calculated the specific instances of sentiment and opinion levels for individual document and sentence scores and analyzed patterns in scores independent of associating scores with a specific blogger. To expand on that concept within this chapter, I examine the evolution of a specific blogger’s ISO profiles over time. With the proposed visualization techniques presented in this chapter, I analyze any trends in either sentiment or opinion scores for the issue viewpoints to determine whether either of these variables either fluctuate or stabilize over time. To achieve this, ISO scores are extracted for each specified blogger’s writings calculated from the scoring model discussed in the previous chapter. These scores are then used to build a dynamic visualization to track the blogger’s evolving ISO profile with the goal of utilizing the graphs to predict the blogger’s future sentiment and opinion regarding the specific issue viewpoint.

As with the last chapter, I start this section by introducing formal definitions for any newly introduced variables. I then present a two-dimensional colormap to track the changes of blog-level scores for the specific author. This is followed by a detailed analysis on another timeline visualization to track the blogger’s sentence-level ISO scores over time, with the goal to expand on existing models which track a single variable, usually sentiment, and discuss how each model can be visualized to combine both sentiment and opinion on the same visualization. Finally, I conclude with a comparison of the effectiveness of the proposed models transitioning to the following objective regarding the spread of blogger influence which is further discussed in the following chapter.

## Problem Definition

This section contains new definitions and variables that appear within this chapter. Recall that $b_i^1(o, s)$ represents the top-level blog score of document *i* for blogger $B_1$, where *o* represents the opinion score and *s* represents the score for the sentiment. The ISO scores are stored in chronological order for all *n* blogs. Each blog and related ISO scores are then associated with a time step based on the posted date $t_i$ within $\{t_0,..,t_z\}$, measured in days, where $t_0$ represents the day of the initial post and $t_z$ represents the final post's date. For this research, documents posted by a blogger within the same time step, are labeled as $t_{a.0}$, $t_{a.1}$, and so forth until reaching the end of the day ($t_{a+1}$). In addition to the above definitions, visualizations also have unique functions and variables which are further discussed during the respective subsection within this chapter.

## Blog-Level ISO Score Visualization

In this section, I introduce a method for modeling the three-dimensional problem (sentiment, opinion, and time) on a two-dimensional plot for the blog-level ISO scores. The first step is to construct a coordinate map using two of the variables, followed by a colormap to display the third dimension.

### Creating to a Two-Dimensional Cartesian Model

The ISO scores calculated in the previous chapter are first converted into a two-point (x, y) coordinate, where the sentiment represents values on the x-axis ranging from [-1, 1] and the opinion values represent values on the y-axis for the values [0, 9]. These values are partitioned on the grid into the three categories for both sentiment and opinion. Since the ISO scores are continuous in the respective intervals, the visualization is initially represented as a 3x3 grid. Within the grid, there exists nine 3x3 squares, each representing a combination of one sentiment, and one opinion corresponding to the x-axis and y-axis respectively, as seen in Figure 8, below. After constructing the grid, the next step is to implement an optimal colormap to track the ISO scores.

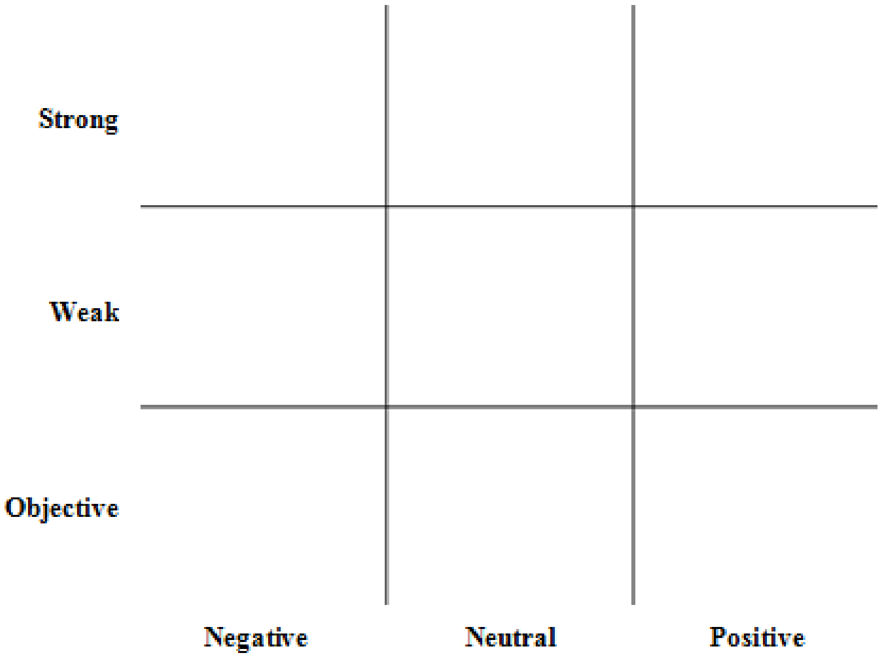


**Figure 8: 3x3 Cartesian Plots (No Data)**

### Implementing an Optimal Colormap

The next component is to represent the ISO profile of the blogger of interest $B_a$ using hot and cold zones similar to calculating a batter's hitting performance in Major League Baseball using standard strike zones. In a strike zone map, there exist nine regions colored from blue to red to represent cold and hot zones respectively [43]. For this implementation, the nine regions represented the sentiment and opinion combinations as discussed above, while the cold and hot zones indicating the frequency of blog entries within each specific category.

To begin the "strike zone" analysis, the 3x3 grid contains a value of zero for each region $Region(s_i, o_j)$ where $i$ = -1, 0, 1 and $j$ = 0, 1, 2 correspond to the sentiment and opinion categories of the nine regions. Next, the ISO score for each one of $B_a$'s blogs is determined by adding 1 to the corresponding region. Once the initial analysis of all of $B_a$'s blogs is complete, the normalized probability for each region is calculated, as shown in the equation below, where values closer to 0 represent a colder zone and values closer to 1 represent a hotter zone.

$$P[Region(s_i, o_j)] = \frac{Region(s_i, o_j)}{\sum_i \sum_j Region(s_i, o_j)} \tag{4.1}$$

After calculating the probability of each region, the colormap visualizes those regions as colors ranging from red to blue by implementing a set of RGB functions based on an existing sentiment visualization tool, SentimentRiver [31]. Within the SentimentRiver application, tweets are weighed to determine whether they hold a positive, neutral, or negative sentiment on a continuous scale. The model measures the variation of topic sentiment over time using a dynamic stream of tweet. The specific component of this model utilized in this visualization from SentimentRiver was the layer color gradient. More information regarding the full SentimentRiver model to include all components is found in [31].

In regards to this visualization, the layer color gradient provides a method to show a continuous colormap going from either red to blue based on the probability of the nine regions for sentiment and opinion. These colors are determined using the following RGB function, using *m* = *max(p)* to represent the maximum region probability, and *a*, which equals the probability of the nine regions ranging from 0 to *m*.

$$RGB_{opinion}(t) = \begin{cases} (255, 0, 255 * (1 - a/m), (a > 0.5 * m) \\ (255 * (1 - a/m), 0, 255), (a < 0.5 * m) \end{cases}. \tag{4.5}$$

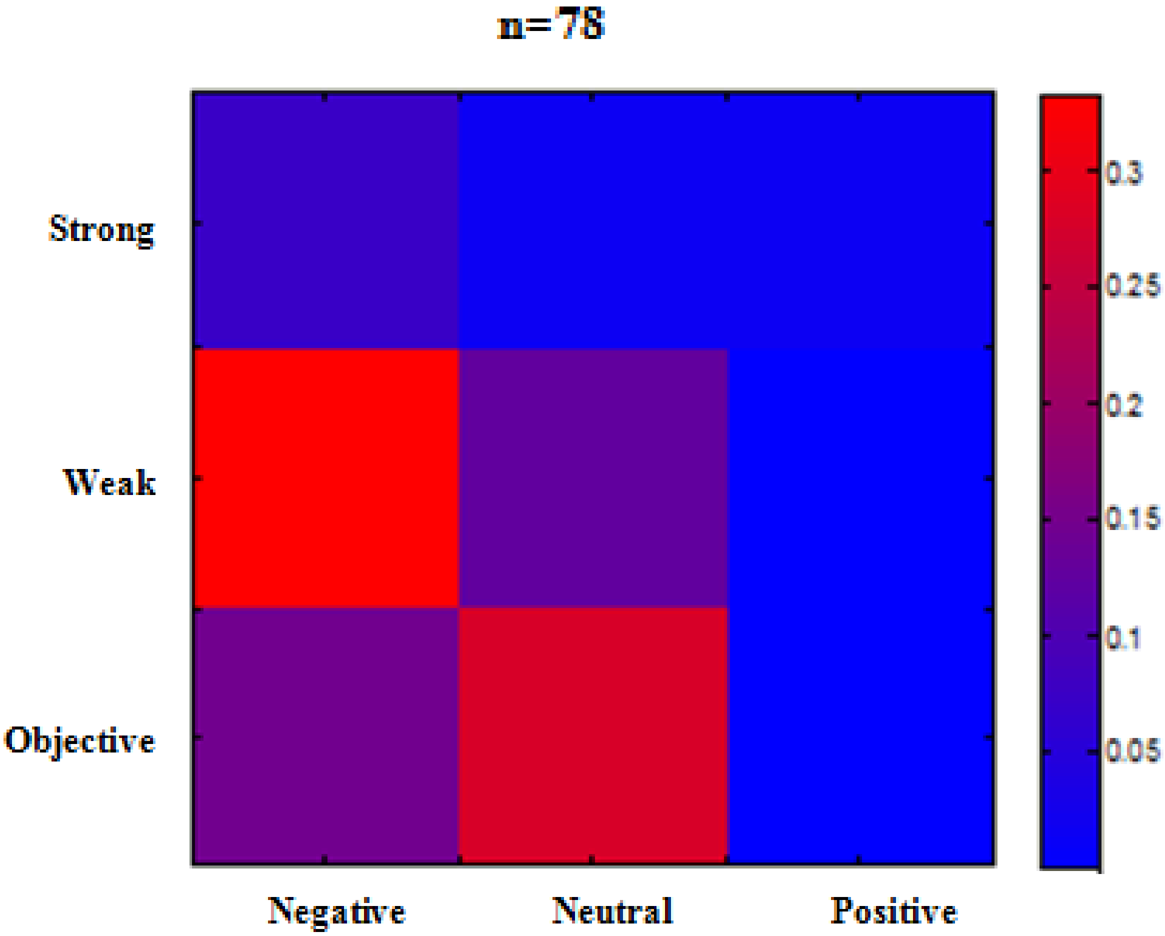


**Figure 9: Colormap of $B_{227}$ ISO Scores ($v_1$)**

For the remainder of the dissertation, I refer to the above visualization as the ISO-mapping model. The ISO-mapping for blogger $B_{227}$ is visualized in Figure 9, above. The first reflection on $B_{227}$ is that the vast majority of their posts fell into objective, neutral to weak, negative scores. Since this single chart only offers a simple summary of the blogger's ISO scores, the chart is also expanded to create a more detailed representation.

An expanded chart better examines how much the colormap differs by using a 9x9 grid instead. From the visualizations, it is easier to pinpoint the opinion and sentiment scores of blogposts on a more detailed scale. This helps to determine any correlations not as easily noticeable in a 3x3 grid.

For the example shown in Figure 10, a comparison of a 3x3 grid and 9x9 grids is displayed for the blogger of interest, $B_{91}$. One observation which was more noticeable for this specific blogger when moving to a 9x9 grid is that as the author presented a more negative tone, the author's opinion became stronger.

Moreover, while a detailed visualization shows the frequency of a blogger's topic sentiment and opinion for an entire collection of blogs at any instance or as a summary of the entire blogger's portfolio, it does not show how the sentiment and opinion of these issues have changed over time in a single chart. A comparison of different time steps side-by-side is shown in Figure 11.

In regards to the timeline for the specific blogger represented in Figure 11, initial ISO scores leaned to objective with negative sentiment. Over time, the blogger's writings began to spread into more negative, subjective regions along with the occasional neutral, objective blogposts. While this chart is helpful in modeling and visualizing the blogger's

ISO scores at the document level, it also serves useful to measure dynamic changes at the sentence-level ISO scores as well with the hope of better understanding the changing degree of opinion as well as positive and negative sentiment within each document.

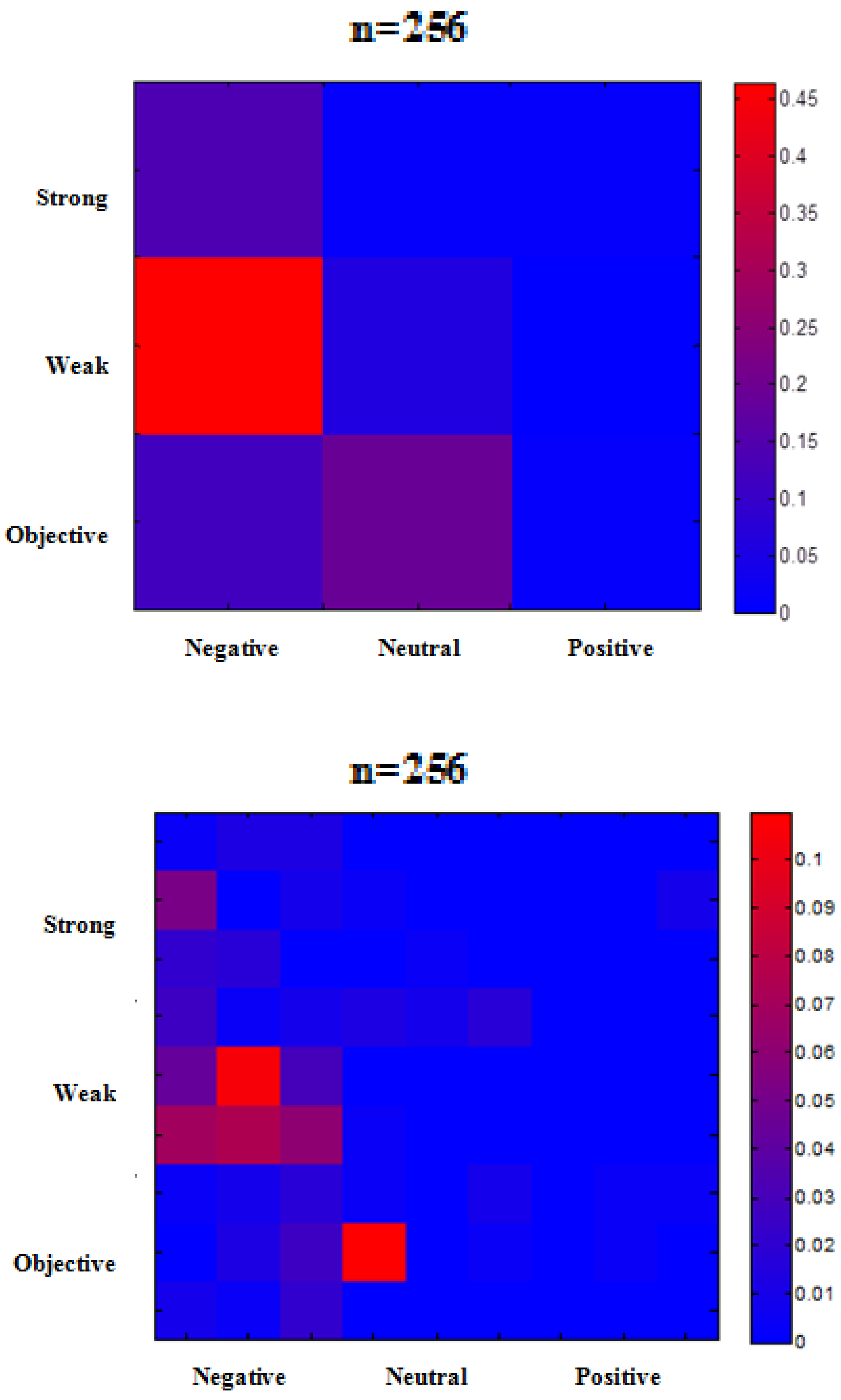


**Figure 10: 3x3 vs. 9x9 Grid Comparison for $B_{91}$ ($v_2$)**

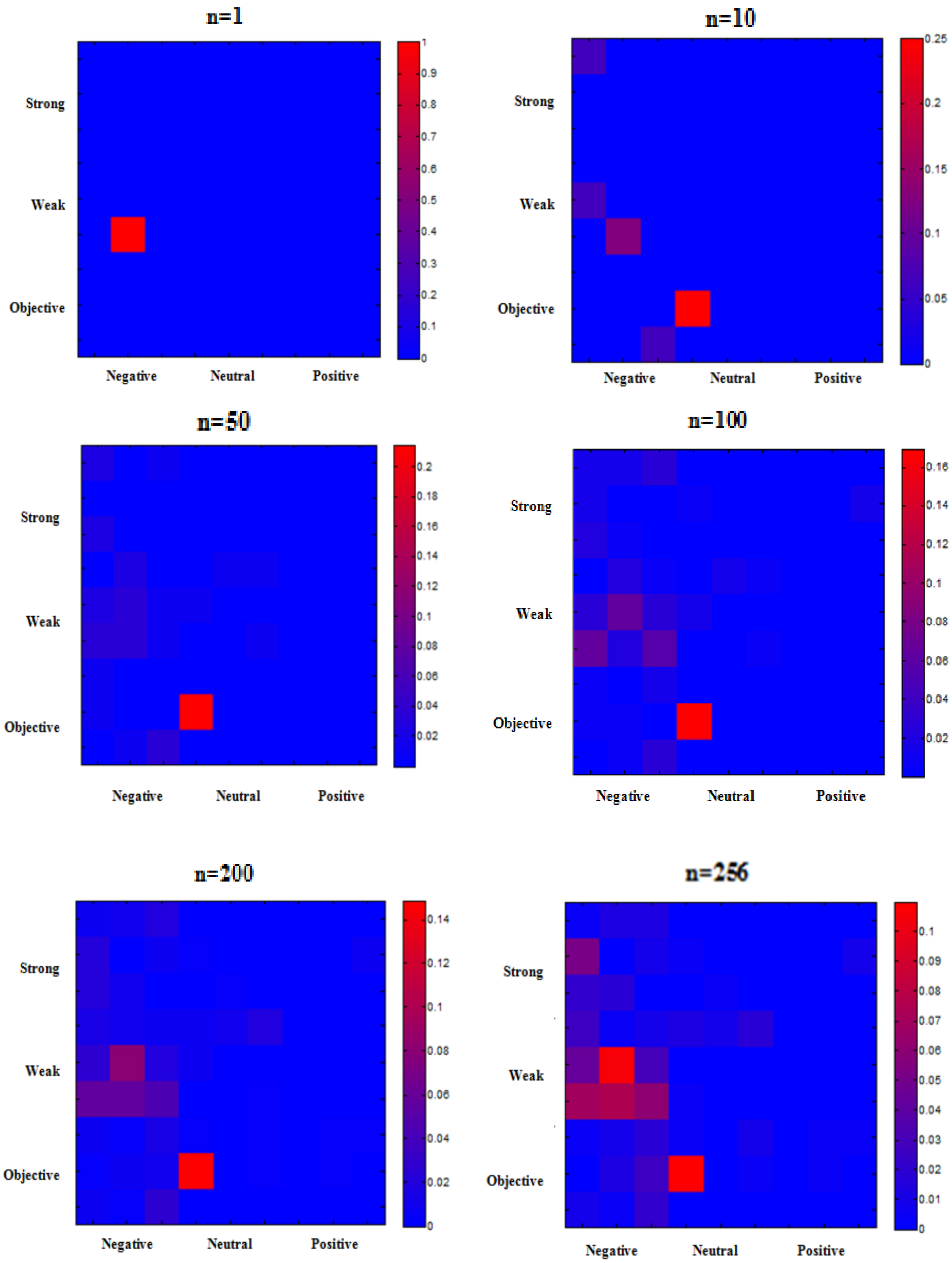


**Figure 11: Time Step Evolution of $B_{91}$ ($v_2$)**

## Sentence-Level ISO Score Visualization

In this section, I examine a model from recent research that has shown to be successful in tracking the evolution of sentiment associated with a variety of topics. As discussed previously in the literature review, the common aspect that occurs for sentiment analysis visualizations is that there only exists an emphasis on tracking the evolution of a single variable; the user or overall community's sentiment associated with the corresponding topic. The goal, similar to the blog-level visualizations, is to incorporate opinion scores over time in the same model and show how each variable independently changes over time.

## Time-Aware ISO (TISO) Model

Another related visualization is a joint topic-sentiment model which tracks and visualizes the evolution of sentiment over time based on Dermouche's Time-Aware Topic-Sentiment (TTS) model [15]. This model considers the two issue viewpoints as the two topics within the model. Additionally, opinion is also incorporated into the topic-sentiment model, thus transforming the existing TTS model into the time-aware ISO (TISO) model. This helps to differentiate between the previous existing model (TTS), and the adapted model integrating both sentiment and opinion (TISO) as I progress throughout the remainder of the dissertation.

It is also important to note that the TTS model incorporates a generative process measuring the evolution of positive and negative sentiment [15]. For the purpose of the TISO model, the ISO scoring model uses the methodology from Chapter 3 followed by graphing the estimated ISO evolution based on the number of sentences in each sentiment and opinion category. A full step-by-step of the TTS model's iterative process as discussed in [15].

Instead of counting the number of documents of each sentiment and opinion, I choose to calculate the running average of the percentage of sentences for each sentiment and opinion combination in each document at the specific time step in a similar way to the blog-level equations for the colormap. Totals are then projected as nine different lines.

$$E_t(s_i, o_j) = \frac{e(s_i, o_j)}{\sum_i \sum_j e(s_i, o_j)} \; X \; 100 \tag{4.2}$$

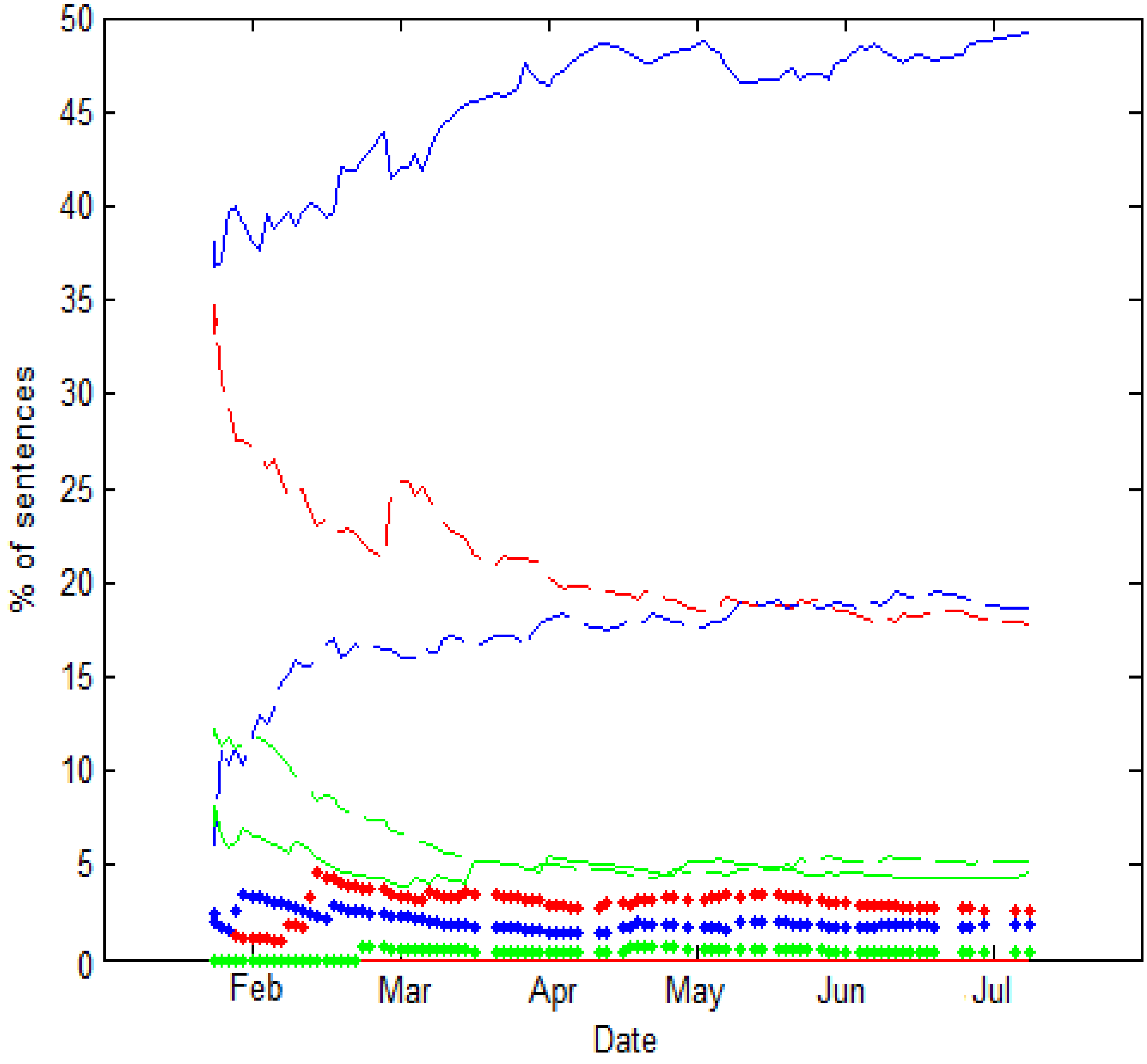


**Figure 12: TISO Visualization for $B_{91}$ ($v_1$)**

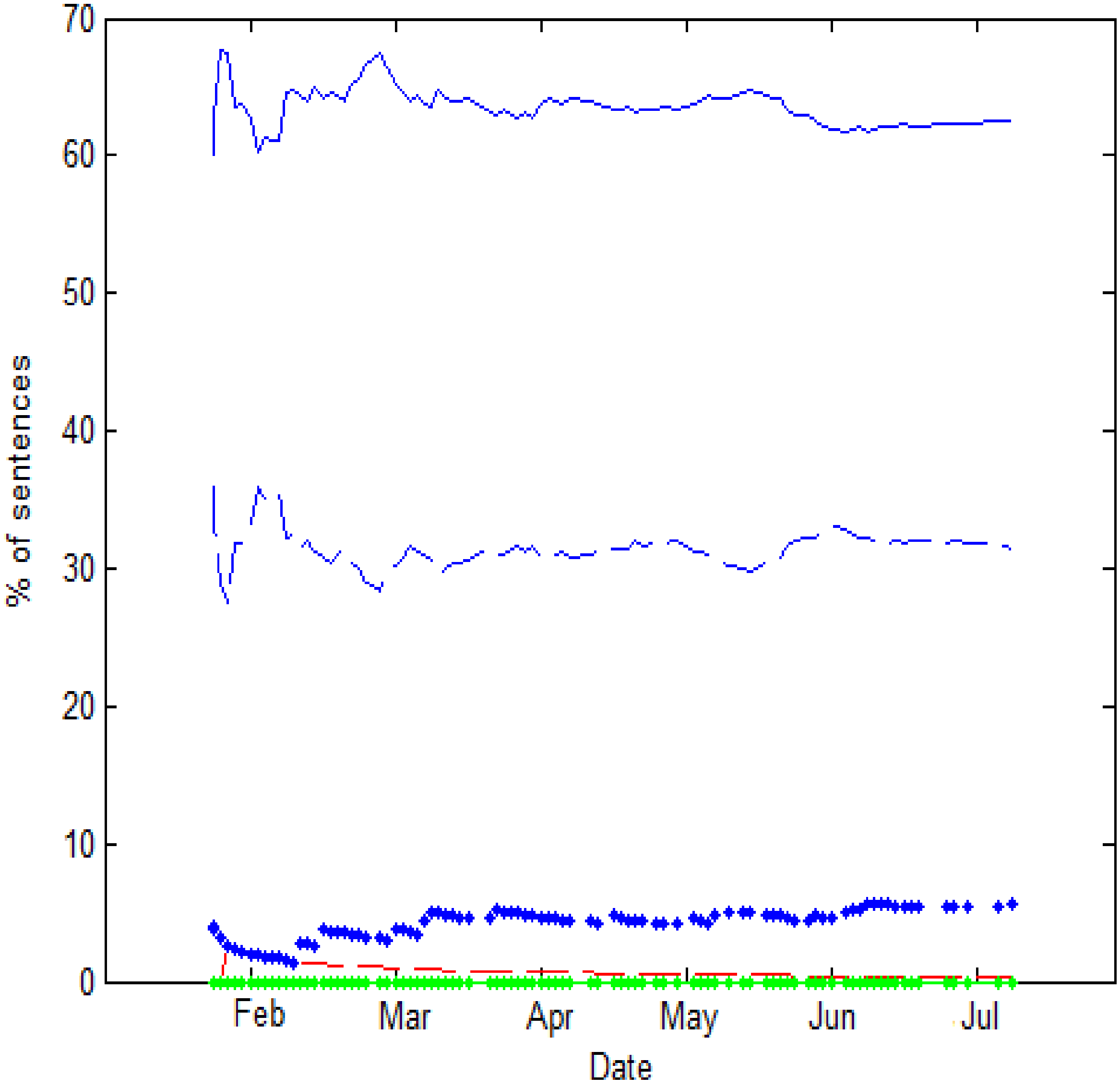


**Figure 13: TISO Visualization for $B_{91}$ ($v_2$)**

Figure 12 and Figure 13 above show the TISO visualization depicted as nine different lines representing red, blue, or green for negative, neutral, and positive sentiment, as well as a solid, dashed, or dotted-line to correspond with opinion (objective, weak, and strong subjective respectively).

As I can see for this specific blogger, initial posts within the blogosphere expressed increased sentiment when discussing issue viewpoint $v_1$, more specifically negative than positive. As for issue viewpoint $v_2$, nearly all sentences held a neutral tone, though the model encountered a fair amount of subjectivity. As the time progressed over six months within the blogosphere, the author's sentiment decreased for $v_1$ while staying neutral for $v_2$. For both issue viewpoints, the blogger's amount of subjectivity expressed increased slightly over time.

For some of the bloggers, having nine lines on a single plot made interpreting changes in subjectivity or sentiment difficult to convey over time, as both sentiment and opinion sometimes fluctuated significantly from post to post. Therefore, opposed to comparing the first visualization with a more complex picture similar to what is done with the blog-level sentiment scores, I elect to compare the current TISO model to a simpler TISO visualization (TISO-lite). In Figure 14 below, there are three solid lines representing sentiments, red/blue/green for negative, neutral, and positive respectively, as well as two dashed lines to display opinions, black/magenta corresponding with objective and subjective opinion.

Viewing the TISO-lite visualization below, I easily see a decrease in sentiment used over time, eventually steadying at around thirty percent on average. On the other hand for opinion, objectivity and subjectivity sentences stabilize near an even 50/50 split. Both of these observations are determined from the original TISO visualization, but they are seen much quicker in the TISO-lite chart.

Additionally, another common observation from the visual results shows that there most likely exists a relationship between the level of sentiment and opinion. At the same time, while scores for both of these variables tended to increase or decrease during the same time increments, the rates at which these variables changed in either direction were not identical to one another. I would like to further examine any connection between sentiment and opinion in future work, but for this research, the main objective is to only show the importance of separately measuring the two variables opposed to scoring them as one combined category.

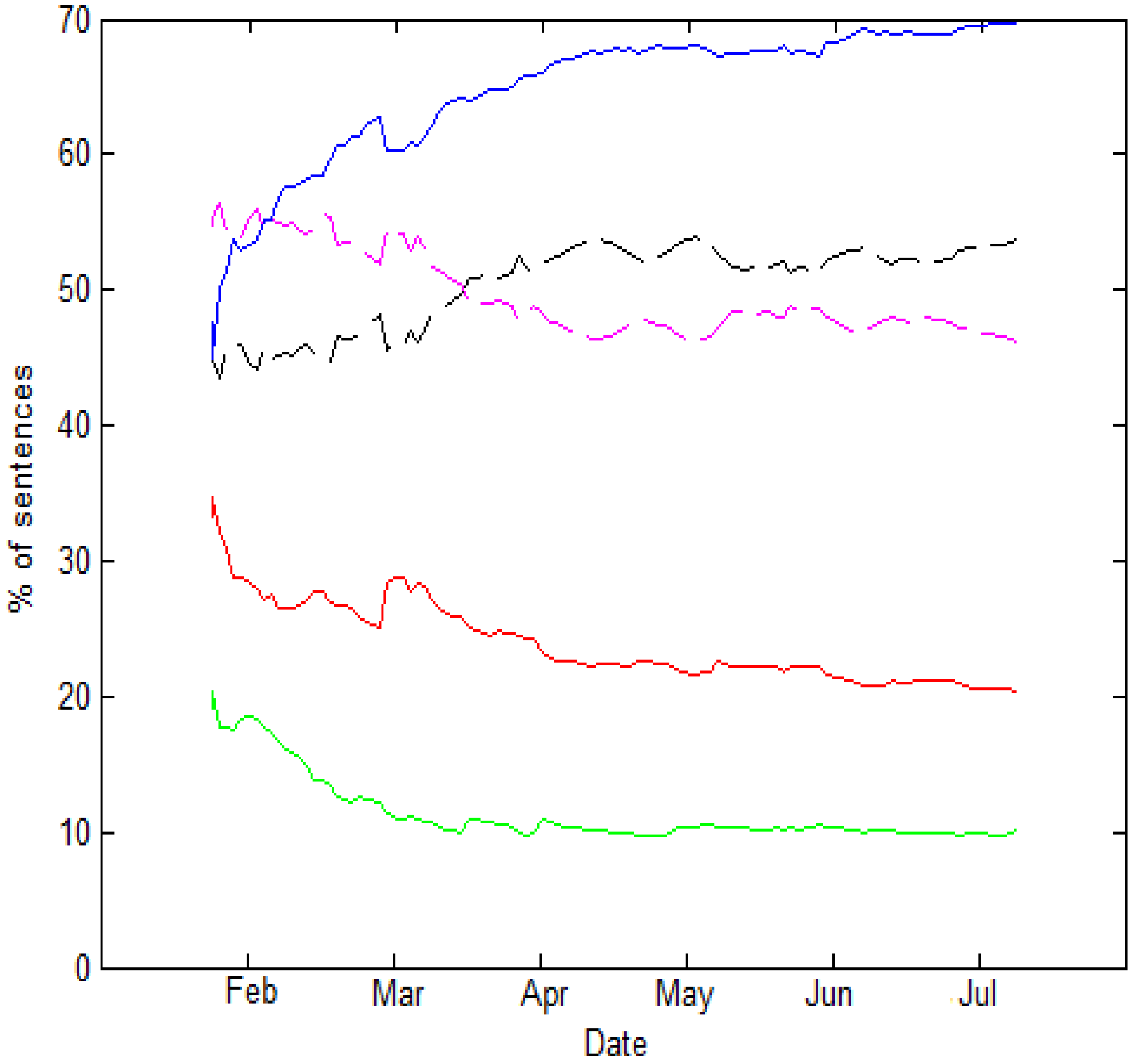


**Figure 14: TISO-lite Visualization for $B_{91}$ ($v_2$)**

## Conclusion

The benefits of utilizing the ISO-mapping visualization include the ability to determine which sentiment-opinion regions that the blogger demonstrates within the writing. However, without animating multiple posts over time, a single graphic cannot show any dynamic changes in sentiment and opinion. This is where using the TISO and TISO-lite models provide more success, by clearly showing changes in sentiment and opinion lines over time to evaluate trends within the date. However, the TISO models make it more difficult to determine specific time steps since it calculates the running average scores. Overall, both models have their own strengths and weaknesses for measuring and showing how frequently the blogger expresses emotion and subjectivity, allowing the ability to provide a useful tool for visualizing the blogger's sentiment and opinion.

Overall in this section, I propose two different models to visually capture the blogger's ISO scores determined from the model in Chapter 3, one for blog-level scores and the other for scores at the sentence-level. While this research only focuses on two visualizations, there may be other models that can be modified to fit sentiment and opinion. I propose future research examining the optimal ways to visualize this problem.

The important takeaway from the results is that I should continue to consider the blogger's sentiment and opinion as independent variables. For this specific case study, the model found frequent correlations between varying changes in sentiment and opinion, but there did exist instances where this was not always the case, showing the importance of tracking the author's sentiment and opinion separately to provide a more detailed picture into the blogger's profile.

Progressing into the next chapter, this model is extended to analyze the second major research problem: measuring the influence of the author's sentiment and opinion and how it impacts neighboring bloggers. ISO scores as well as the changes in those scores for bloggers over time continue to be of value, but now I expand on this effort by examining similarities within different bloggers' ISO scores, with the additional goal of analyzing how closer relationships among connected bloggers can impact either author's level of sentiment and opinion in future posts.

# MODELING THE SPREAD OF OPINION AND SENTIMENT INFLUENCE WITHIN THE BLOGOSPHERE

In the previous two chapters, I modeled how to calculate the blogger's document scores for issue sentiment and opinion and examined how those ISO scores may evolve over time. This chapter expands on those two concepts by analyzing the interactions between neighboring bloggers and how ISO scores for those bloggers evolve over time dependent on the bloggers' proximity to one another. In order to achieve this goal, I start by adding a network model using both existing social network analysis and graph theory techniques to measure the closeness of bloggers to their nearest neighbors. Components from this network model are then used to determine prominent bloggers and the level of influence of other authors, leading to the final step to output a representative model of the evolving sentiment and opinion within the entire blogosphere.

As with the other chapters, I first start by presenting a formal definition of the problem along with descriptions of any new variables. This is followed next by the construction of the network model which shows the proximity of bloggers to one another calculated through the number of followers measuring the frequency of blogger recommendations and followers. In doing so, I propose a new way to measure the level of influence for more prominent bloggers within the network model by incorporating the total counts of all followers, recommendations, as well as comments for each blog. Afterwards, I incorporate a deterministic model of influential spread within the social network by using modified versions of existing epidemic models such that influence is interchangeable to a viral infection within the community. Finally, I review the experimental results from the proposed structure and model and discuss the implications of the findings and what the next potential steps should to expand the model in future work.

## Problem Definition

This section focuses on additional definitions and variables beyond the previous sections to include specific variables and formulas used within the following chapter. First, for this chapter I continue to reference individual bloggers using the same notation, $B_i$. In addition, the focus is now on building a social network connecting the bloggers to one another. Therefore in the network model, each blogger is referred to as a node or vertex. The edge connecting the two vertices, which for this case represents a connection between two of the bloggers, is represented as $B_{ij}$, where the bloggers $B_i$ and $B_j$ denote the connecting node.

Another important factor to consider is that the model records of the order of any two bloggers, because the direction the influence travels is important in the spread of influence. To differentiate between the starting and connecting blogger matters when measuring the influence of individual bloggers on connected nodes, $B_{ij}$ is used to represent the edge itself, but $\vec{B}_{ij}$ signifies the direction of influence on that given edge, $\vec{B}_{ij}$ also be referred to as a directed edge. The distance of the edge is represented as $d(B_{ij})$ and this value reflects how close together those two bloggers are with one another. This distance value is calculated based on multiple factors impacting author influence.

Furthermore, the proposed concept of author influence within the model is based on the total frequency of comments, blog recommendations, and followers in terms of quantity for both the documents as well as for each blogger. The resulting terms and related functions are described in further detail below:

- Comment Counts/ Commenting Individuals: Each blog allows for other members of the blogosphere, along with the original author to comment on the document. Therefore, the set count of comments of $b_m^n$ is denoted as *c(*$b_m^n$*)* as well as the list of all of the other *k* bloggers commenting on the specific blog $C_m^n$ *(*$B_j, \ldots, B_k$*)*. For the purpose of this research, I exclude any comments by the author from the total comment count. Finally, the total comment count between two specific bloggers, $B_i$ comments of $B_j$ blogs and vice versa is denoted *c(*$B_{ij}$*)*.

- Recommendation Count/Recommending Individuals: Each blog allows for the other bloggers to give the option to recommend the blog. For each blog *n* written blogger *m* as represented in Chapter 2, $b_m^n$, there exists a set count of recommendations, *r(*$b_m^n$*)* as well as the list of all of the other *k* bloggers recommending the specific blog $R_m^n$ *(Bj,…,*$B_k$*)*. After calculating recommendations for each blog, the total recommendations count is calculated between two specific bloggers, $B_i$ recommendations of $B_j$ blogs and vice versa, which is represented as *r(*$B_{ij}$*)*, similar to calculating comment counts and individuals.

- Follower Count/Individuals: Likewise, there is the option for bloggers to follow one another. This is done at the blogger level opposed to the individual documents, where a follower score, $f(\vec{B}_{ij})$, represents that blogger $B_j$ is following blogger $B_i$. This score returns a value of 0 or 1 depending if the blogger is following the other. In addition, the follower score also returns a value of 2, which implies that both bloggers are following one another. For example, if $f(\vec{B}_{ij})$ equals 1, then $B_j$ follows $B_i$, but not vice versa. Additionally, the list of followers of blogger $B_i$ is represented similarly to the commenting and recommending individuals and is denoted as the vector $F_i(Bj,\ldots,B_k)$.

As stated in the introduction, the Influential Spread Model (ISM) of resulting ISO scores is directly based on modifications of an epidemic model. Additionally, for other parameters and equations within this model, Discussions of further relevant notations occur after the proposed methodology has been examined further.

## Blogosphere Network Analysis

In this section, I discuss the methodology to add a social network component to the model by investigating the interconnections among authors and their respective blogs within the blogosphere by creating a social network. By examining the active participation within the community over time, I look to address the following questions related to both the roles of the individual bloggers and network connections between them play in regards to the overall schema:

1) Who are the central blogger(s)?

These bloggers are the more popular “A-list” bloggers [26] and are characterized as having a higher frequency count of the three key attributes of all $k$ blogs: comments, followers, and blog recommendations, such that $T(B_m)$ represents the sum of the three variables.

$$T(B_m) = \sum_{i=1}^{k} (c(b_m^n) + r(b_m^n) + f(b_m^n)), \tag{5.1}$$

where $k$ equals the total number of blogs written by $B_m$. More specifically, $k$ represents only on the total number of blogs corresponding to the specific case study.

2) How closely connected are any two bloggers?

The network connections between bloggers show how closely connected they are to one another. For two specific bloggers, person-to-person connections are represented as either a distance separating the two individuals or the degree, which represents the numbers of other bloggers it takes to connect one blogger to the other. These variables are both described in further detail as the overall network structure is discussed.

The bloggers along the connections among bloggers are the two major factors that are needed to create the representative network model of the blogosphere. Within the description of the social network, I refer to the bloggers as nodes or vertices, and the connections between bloggers as edges. Both the nodes and edges are used within the social network to provide a complete description of the overall structure of the blogosphere.

Within the network model, the nodes help to show which bloggers are the most influential or centralized within the blogosphere, while the edges between each node show the number of interactions each blogger has and how close any two bloggers are to one another. The completed network model provides the baseline structure of the blogosphere which is then used to build the influence spread layer on top of the network, thus establishing a methodology representing the distribution of sentiment and opinion between nodes throughout the social network.

### Blogger Nodes

Most importantly for each of the nodes, I look to quantify the importance of the central bloggers representing the influence they have on their nearest neighbors (local connections) and the rest of the bloggers in the community (global connections). To do so, the model utilizes the concept of centrality within a social network analysis model [26]. Recall that the metrics characterize the node's importance are the total count of comments, recommendations and followers of each blog, and $T(B_i)$ equals the sum of these factors for blogger $b_i$. Additionally, since some bloggers have more blogposts than others, the influence scores need to be normalized by the total number of documents, $k$, in which the blogger has posted. Therefore, the normalized total frequency score is as follows:

$$T_N\ (B_i) = T(B_i)\ /\ k, \qquad (5.2)$$

The goal is to determine two scores of influence for each blogger: relative and absolute influence [42]. These scores are then used to determine which bloggers fall more central than others both locally and globally within the blogosphere.

### Blogger Neighborhood

Prior to calculating the above scores, the model also needs to consider which bloggers fall into the same neighborhood as the blogger of interest, $B_i$. While the next portion of the model discusses exact distances and degrees between bloggers, here only the following specific threshold metrics are used to determine when a blogger falls within the same neighborhood. Hence, the blogger $B_j$ falls into the same neighborhood as $B_i$ if it meets at least one of the following three criteria:

*i)* $B_j \in F_i\ (B_o, \ldots, B_m)$

*ii)* $B_j \in C_i^n\ (B_o, \ldots, B_m)$ for at least 10% of *n* blogs

*iii)* $B_j \in R_i^n\ (B_o, \ldots, B_m)$ for at least 10% of *n* blogs

The list of *m* blogger's that meet the above thresholds is denoted as $A_i(B_o, \ldots, B_m)$ to represent the neighborhood around $B_i$ and the total number of bloggers within the neighborhood is denoted as *A* such that the total blogger count equals *m* + 1 to account for the blogger of interest as well.

**Relative Influence**

Once the neighborhood of $B_i$ is computed, the next step is to calculate the two influence scores for each blogger node. The first score to determine is the relative influence which represents the blogger's influence in respect to the overall influence of its corresponding neighborhood. Similar to [42], this model incorporates the appropriate follower and following numbers, but in addition, it also expands on that feature by adding both the recommendation and comment counts as represented in [21].

Therefore the relative influence of the specific blogger $B_i$ within its given neighborhood is defined as the function $I_R(B_i)$, which is then calculated using the following equation.

$$I_R(B_i) = \frac{T_N(B_i)}{\sum_{i=1}^{A}(\mathrm{T}_N(\mathrm{b}_i))} \tag{5.3}$$

While bloggers $B_i$ and $B_j$ may both fall in each other neighborhoods, there are other bloggers that may only fall into only one of blogger's neighborhood. This is why relative influence alone cannot support which of the two bloggers may be more central than the other.

### Absolute Influence

Therefore, the next step is to also calculate the absolute influence, $I_A(B_i)$. To calculate this, the model uses a similar equation as with the relative influence, but absolute influence also incorporates the size of the neighborhood for the specific blogger as well, which is represented as $h_a$. The absolute influence of blogger equals the relative influence multiplied by the neighborhood size as follows:

$$I_A(B_i) = \frac{h_a}{\max(h)} * I_R(B_i) \tag{5.4}$$

Based on analysis of initial results, bloggers with a larger neighborhood and higher relative influence are more likely to be more central blogger holding the most influence in the rest of the community. Table 5 represents a portion of $k_{11,}$ the neighborhood around blogger $B_{11}$ and the associated relative and absolute influence scores.

Once the absolute influences of all nodes are calculated, each blogger is ranked from higher to lowest scores. Throughout the rest of the chapter, I refer to the top ten percent of the bloggers as the "A-list" bloggers, which represent the most central nodes also referred to as the most influential nodes. The "A-list" blogger nodes are used as a starting point for modeling the spread of influence from "A-list" bloggers to the rest of the community, which is discussed following the social network analysis.

In the following section, I continue to build on the social network additions, by measuring the distance that each node has with its neighbors as well as the degree between bloggers and the closest “A-list” bloggers. This allows the creation of a representative picture of the blogosphere, leading to the end objective of discovering what happens to bloggers’ ISO scores at the end state.

**Table 5: Relative and Absolute Influence Scores for $k_{11}$**

| | $T_N\ (B_i)$ | $I_R(B_i)$ | $I_A(B_i)$ |
|---|---|---|---|
| $B_{11}$ | 961 | 0.1380 | 0.0104 |
| $B_{105}$ | 63 | 0.0009 | <0.0001 |
| $B_{119}$ | 38 | 0.0005 | <0.0001 |
| $B_{152}$ | 79 | 0.0011 | <0.0001 |
| $B_{196}$ | 28 | 0.0003 | <0.0001 |
| $B_{248}$ | 93 | 0.0013 | <0.0001 |
| $B_{278}$ | 22 | 0.0003 | <0.0001 |
| $B_{326}$ | 172 | 0.0025 | 0.0002 |
| $B_{395}$ | 160 | 0.0023 | 0.0002 |
| $B_{436}$ | 59 | 0.0008 | <0.0001 |
| $B_{501}$ | 140 | 0.0020 | 0.0002 |
| $B_{562}$ | 79 | 0.0011 | <0.0001 |
| $B_{687}$ | 112 | 0.0016 | 0.0001 |
| $B_{735}$ | 299 | 0.0043 | 0.0003 |

## Blogger Edges

Determining nodes and their level of influence are important is only one of the key components for building the social network structure. The next major step is to then create and measure the distance of connecting edges between pairs of nodes.

The first step of this process is to determine the criteria for creating an edge between two nodes. Once the edge is created, the distance between those two bloggers is measured, while setting a maximum distance when a minimal connection is present. If however there does exist a strong enough connection, then the distance gradually decreased from the maximum value to a distance of zero. Additionally the smallest number of degrees of separation is also calculated between any blogger node and the nearest "A-list" bloggers.

### Existence of Edge

The criteria for determining if an edge exists is very similar to the thresholds for determining if a blogger falls in a given neighborhood. However, the criteria for an edge reduces from the "ten percent" criteria to just a single occurrence in comments or blog recommendations, with the simple reasoning being that an edge connecting two nodes will exist in some form within a social network, no matter how faint, as long as there exists any sort of association between two bloggers whether this is through following, blog recommendations, or comments.

### Distance Calculation

After defining the edges, the next step is to measure the distance of the connecting edges between two nodes. The distance of the edge length represents how close any two of the bloggers are to one another with the basic understanding that if the two bloggers are having more frequent communication with one another, they would have a stronger connection and thus have a shorter distance connecting their nodes. Hence, after examining all possible edges within the social network, the model determines whether the bloggers are scattered evenly through the blogosphere or there instead exists clustering of bloggers where closer groups with very little connections between groups.

To calculate the distance between two bloggers, $B_i$ and $B_j$, the model starts with a baseline distance, $d(B_{ij})$ equal to 1, which represents the starting distance for any connecting edge. From there, the following set of rules is used to reduce the distance of the edge based on how connected two bloggers as shown in Table 6.

**Table 6: Edge Distance Algorithm**

```
1) Set Initial Distance
        a. $d(B_{ij}) = 1$

2) Distance reduction based on comment and blog recommendation counts
        a. Let $cr(B_{ij})$ equal the total occurrences of $B_i$ comments and blog
        recommendations of $B_j$ and vice versa.
        b. $d(B_{ij}) = 1 – [max(cr(B_{ij}) / (n_m + n_n), 0.95)]$

3) Distance reduction based on following count
        a. If $f(\vec{B}_{ij})$ or $f(\vec{B}_{ji}) = 1$, then $d(B_{ij}) = 0.75*d(B_{ij})$
        b. If $f(\vec{B}_{ij}) = 2$, then $d(B_{ij}) = 0.5*d(B_{ij})$
```

Thus by using the above algorithm, a maximum score of one is set for an edge so long as there is a connection between the two bloggers. Additionally, the lowest possible score for an edge between any two nodes would be 0.025, which was manually determined to optimize computing time and not end up with a possible edge distance of zero. For the selected case study, the above algorithm worked well for the creation of edges within the social network, but a further iterative scheme would likely need to be developed to fit a different dataset or more general problem.

### Degrees of Separation

Regarding the edge distances, the shortest distance is calculated between nodes and "A-list" bloggers by looking at the smallest degrees of separation and using the breadth-first-search (BFS) algorithm [50]. The BFS algorithm represents the social network as a tree structure starting at the top for the initial blogger and traversing down the tree until reaching the connected "A-list" blogger.

Interestingly, there were some instances where edge distance based on the degrees of separation between two bloggers was not the shortest overall path. For further analysis of the social network based on edge distances from non-uniform distances in the future, I propose incorporating Dijkstra's algorithm to find the shortest path between the two nodes [8]. However, for the purpose of this research, the degrees of separation provide a clearer picture of how many bloggers fall between the selected node and the more influential nodes.

### Social Network Visualization

Once all node and edge values are calculated, the next step is to visualize the social network structure showing the interactions between the bloggers. Additionally, it also provides a way to generate a picture of the blogosphere by visualizing the prominent bloggers (absolute influence) and relationships among bloggers (edge distances) at different time steps providing the means to implement an influence spread model on top of the social network structure.

For the visualization of the social network, attributes are leveraged from previous research to clearly represent the nodes and edges [52]. The model represents nodes' color to corresponding sentiment scores, with red, green, and grey nodes representing negative, positive, and neutral sentiment. Additionally, edge lengths are directly related to calculated distance measurements from the algorithm and provide more insight on connections over influence. The stored influence values are of more value in the influential spread model discussed next. Future research ideas for expanding the social network analysis include the following:  1) how to incorporate opinion scores into the node color, 2) vary node size based on influence scores and 3) incorporate the directed edges between bloggers.

For this research, I chose to use Gephi network analysis software to create social network visualization. Gephi is open source software used to display social networks in real-time and is effective with larger complex data sets [4]. This software is used to produce the social network, as partially shown in Figure 15, below, providing a representative depiction the overall blogosphere for this specific case study. Also, blogger node identification is removed to increase the clarity of the visualization for this dissertation.

Additionally, statistics from the social network analysis are shown in Table 7 providing the minimum, mean, and maximum of influence and edge values. The social network consisted of 764 nodes and 45,776 edges, with both highly influential nodes to nodes with little or no influence and everything in between.

I notice that the average degree of separation within the social network was significantly lower than expected, due to a few bloggers which had connections with almost all other bloggers. One such blogger, for example, was likely automated, as nearly every story was a summary of daily or weekly events on that topic. Finally, there were many isolated bloggers, some with only one or two posts, that did not have any existing connections with other bloggers, and this is shown through the minimum influences and non-existence of neighborhood size or any degrees of separation.

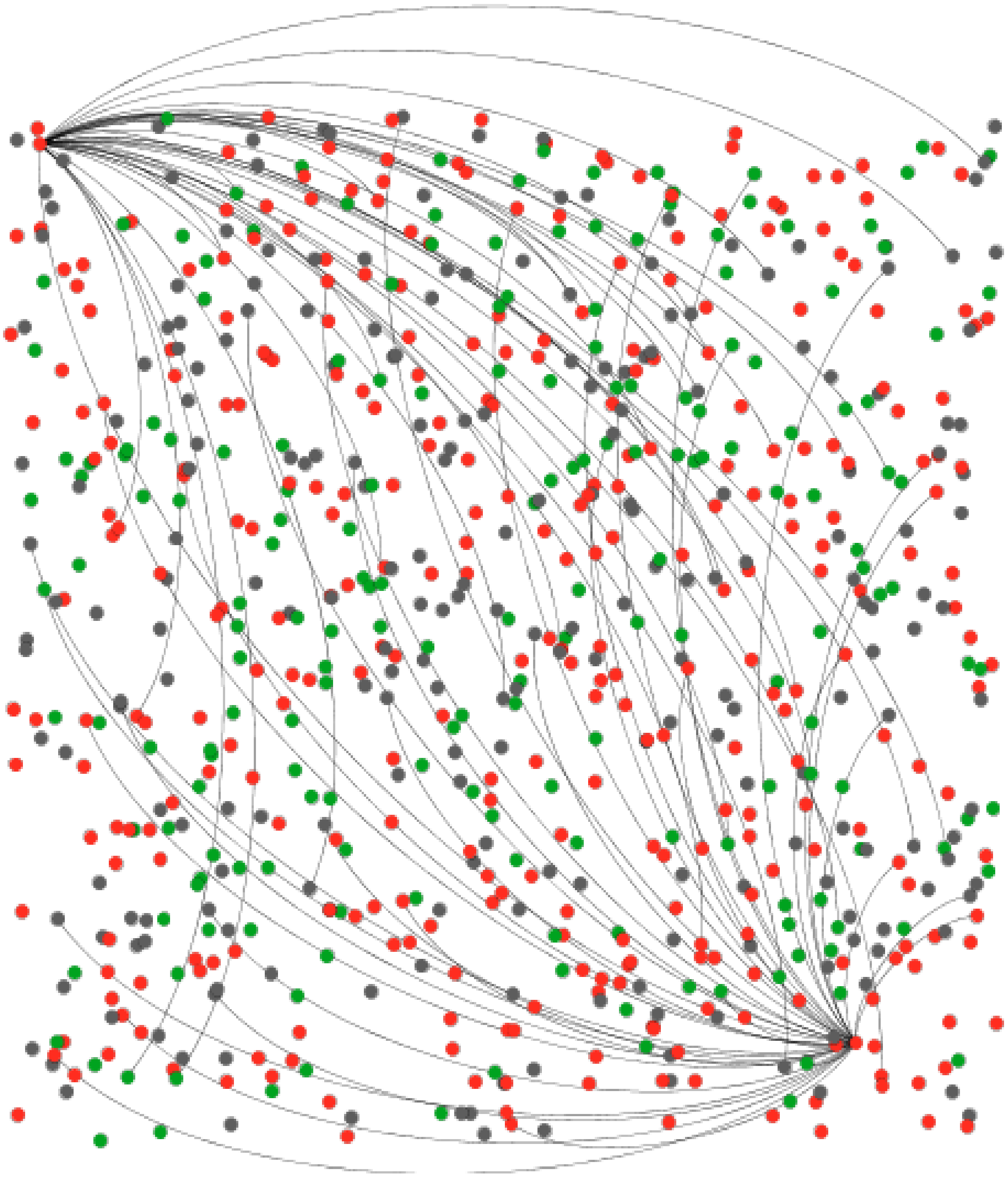


**Figure 15: Blogosphere Network Visualization**

**Table 7: Social Network Statistics (764 bloggers)**

| | $Min$ | $Mean$ | $Max$ |
|---|---|---|---|
| $T_N\ (B_i)$ | 1 | 91.2 | 1002 |
| $I_R(B_i)$ | <0.0001 | 0.0241 | 0.309 |
| $I_A(B_i)$ | <0.0001 | 0.0012 | 0.0112 |
| $h_a$ | N/A | 60.9 | 736 |
| $d(B_{ij})$ | 0.025 | 0.369 | 1 |
| $deg(B_{ij})$ | 1 | 2.5 | N/A |

## Influential Spread Model (ISM)

After developing the social network model to represent the blogosphere, the next step is to implement the Influential Spread Model (ISM). The goal of this additional model is to better understand how both opinions and sentiment change within the blogosphere over time. The key principle of this model is that it emulates how prominent bloggers' influential nodes lead to the spread of influence based on how close two nodes are to one another.

A direct approach of the ISM is that the personal views that one blogger displays in his or her writing directly impacts the closest readers, inflicting them to have similar views in their future posts. This idea can be compared to the spreading of any general information through social interactions, and can be shown using a conventional epidemic model tracking the spread of infectious diseases. The simplest scenario in this model is if a non-infected individual encounters someone that is infected, then there is some fixed probability that the non-infected individual will contract the virus. Based on the social network structure, the infection refers to the personal views expressed by others in their writings [47], so within the ISM model, the type of infection is considered to be the ISO scores of influential nodes.

Throughout the rest of the section, I refer to the terms *infection* and *influence* interchangeably. Furthermore, since there is also a chance for central nodes to influence surrounding bloggers at each time step, the influential impact eventually converges at a some level throughout the entire blogosphere at a specific point in time. Also, since some individuals interact with a close neighborhood of users within the network, those individuals are likely to accept and express similar views as well, based on the concept of herd behavior [47].

### General Epidemiology Model Overview

In this section, I provide an overview of the existing viral models as it relates directly to this research. It is important to better understand the main principles of the generalized epidemic model for interpreting a rapid outbreak of an infectious disease prior to discussing existing and proposed modifications to incorporate both sentiment and opinion.

The main model I explore is based on the original SEIR model [33], which refers to four different components: Susceptible, Exposed, Infected (Influenced), and Recovered. This model is also referred to more simply as the SIR model when removing the exposed state (*E*) from the overall model. Each of the four states within the model is described as follows [25]:

- Susceptible: The host/individual node is susceptible to infection, but no infection (influence) is currently present.
- Exposed: In the early stages of the spread, the host may or may not exhibit obvious signs of the infection (influence).
- Infected/Influenced: Host encounters infectious “influential” individual and becomes infected / influenced.
- Recovered: The host is no longer infectious.

Within the classical SIR model, assumptions are made that individuals are only born into the susceptible class, and if they show signs or catch the disease, they are then moved directly into the infected class, since there is no existence of an exposed state [10], which is modeled by the transmission rate. Hence individuals can never belong to more than one class.

Within the infected state, an individual either becomes deceased or eventually moves to the recovery state depending on the healing or recovery rate. Based on the particular virus, the recovery state describes the individual as either immune from the disease for the rest of their life or in some instances, the individual is only immune for a limited period of time, thus that person would re-enter a susceptible state [10]. A flow diagram is shown in Figure 16. In regards to the idea of sentiment and opinion, there is no existence of immunity. Therefore will be described based off an SIS model with no recovery state, similar to a “flu-like” virus with no immunity.

Furthermore, no bloggers perish within this model thus allowing for the only possibilities to stay in one of only two possible states, susceptible and infected once introduced into the blogosphere model. Although in the future, I may include deaths as in the case of being suspended or deactivated from the blog. The ISM also considers the possibility for spontaneous infection of a susceptible individual within the system as proposed in the SISa model by Hill in 2010 [27], where *a* represents the spontaneous infection rate. The flow diagram of the modified SISa model is shown in Figure 17.

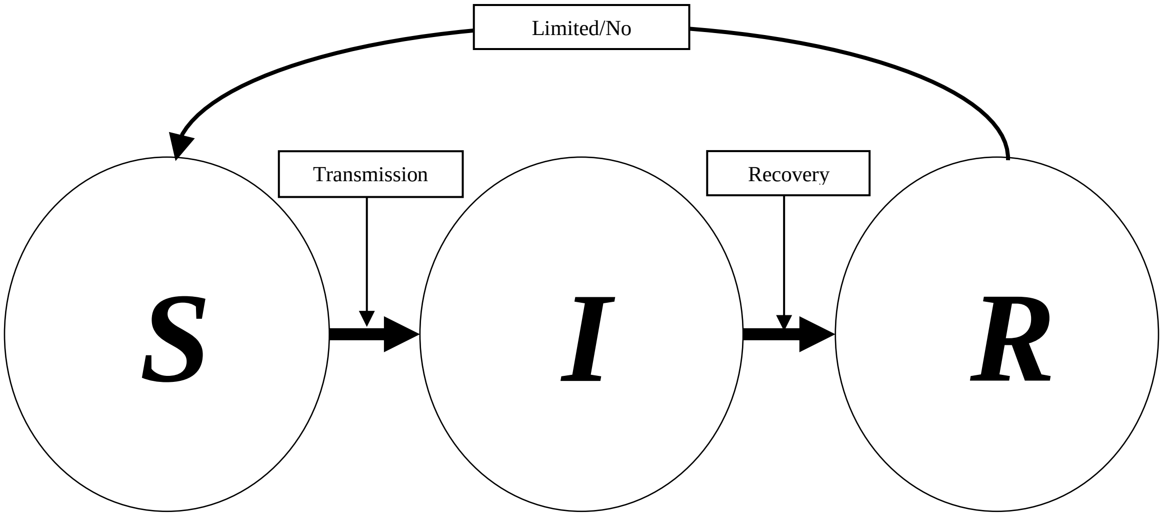


**Figure 16: SIR Model Flow Diagram**

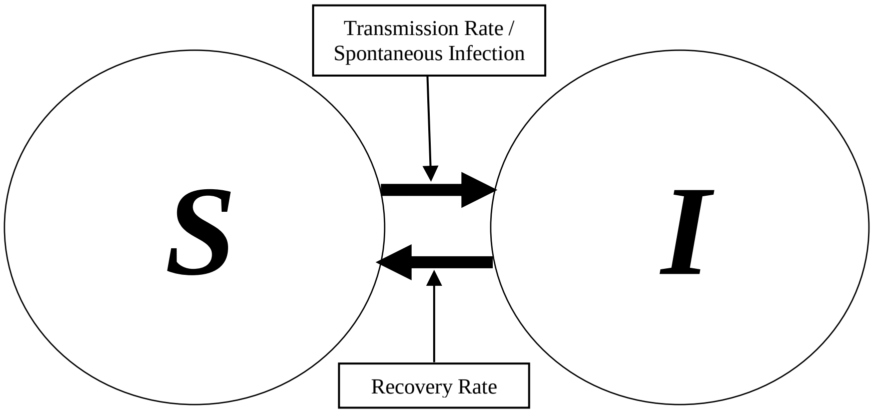


**Figure 17: SISa Model Flow Diagram**

## Incorporating Sentiment and Opinion

There are two major components that need to be incorporated into the ISM for use with sentiment and opinion. First, and more specifically with sentiment, a two-state infection or dual-contagion is used to represent positive and negative scores. This is not necessary for opinion as there is only a single infection state of subjectivity. Secondly, interactions between the two non-competitive viruses (positive vs. negative sentiment) with the addition of a third infection (opinion) are necessary such that a distinct level of competition within the network ranging from no to full competition may now exist between a sentiment and the opinion viruses. Based on the level of competition, these variables can be modeled with either some amount of dependency or complete independence as previously discussed throughout the other components of the dissertation.

### Two-State Infection (Sentiment)

In regards to the two contagions, opinion and sentiment, only opinion fits the general epidemic model, where objectivity represents the opinion of blogger nodes in the susceptible state, and any level of subjectivity exists in infected individuals. For this particular case study, the ISM combines individual nodes where both weak and strong subjectivity are combined into a single infected state. However, in the future, I propose considering only strong subjectivity as the infected state, while weak subjectivity would represent an exposed state.

When examining only sentiment contagion within the social network, there are two unique and independent infected states, based on the positive or negative sentiment. To better understand how to model a two-state infection, ISM utilizes the SOSa-SPSa model introduced by Dodds and Watts in 2014 [16], and while their model labeled the infected states as *O* and *P*, referring to optimistic and pessimistic states respectively, the two infected states refer to positive and negative sentiment within the contagion model. In addition, individuals showing objective sentiment would fall into the susceptible state.

Hence, there are certain key assumptions to follow [16]: when determining which of the three states, susceptible (*S*), positive infected ($I_P$), and negative infected ($I_N$), each node falls into:

- The higher number of positive nodes the susceptible individual is in contact with, then the higher probability that the individual becomes infected with positive sentiment. Similarly, the higher number of negative nodes the susceptible individual is in contact with, then the higher probability that the individual becomes infected with negative sentiment
- The probabilities of rate of recovery back into the susceptible state of objective sentiment from either a positive or negative infection are independent to one another
- Susceptible individuals may also spontaneously become infected with either positive or negative sentiment at certain independent probabilities.

This is the reason to treat the negative and positive sentiment contagion as two mutually exclusive viruses with no competition between one another. Therefore, similar to [16], there will be no transformation in the model directly between infected individuals in $I_P$ and $I_N$.

Additionally, the extension to the ISM model introduces the concept of another contagion based on opinion, where the infected state $I_S$ represents the independent class of individuals infected with subjectivity. However, since the opinion contagion may possibly have some level of competition with either sentiment, it also must be defined as a unique interaction rate between either one of the sentiment contagions along with the opinion contagion. The following section will discuss the interaction between the two interacting viruses within the network.

### Multiple Co-existing Virus Network

Now that a general contagion model with infection states for each virus (opinion and sentiment) is determined, the next step to consider is how opinion and sentiment viruses interact with one another. To do so, ISM follows the ideology that two viruses can still both exist within the same environment when competition exists, as long as it is not full competition [44], which is not the case in this research.

With the addition of the opinion virus within model, there are now three distinct contagions, subjectivity along with positive and negative sentiment, each with its own respective infected state, $I_S$, $I_P$ *and* $I_N$. Although there are now three contagions interacting with one another within the network, there exist as many as five distinct infected states which are as follows: $I_P$ (infected with positive sentiment), $I_N$ (infected with negative sentiment), $I_S$ (infected with subjectivity), $I_{PS}$, (infected with both positive sentiment and subjectivity), and $I_{NS}$ (infected with both negative sentiment and subjectivity). As stated previously, there exists no transmission between individuals in $I_N$

and $I_P$ and this holds true as well for when individuals fall into a dual infection state, $I_{NS}$ and $I_{PS}$.

With the introduction of an opinion contagion, while the transmission rate is calculated independently for the susceptible class, the interaction rate among opinion and sentiment viruses [6] is also critical in determining the spread throughout the network of individuals already infected. If the interaction factor equals zero, then there exists no dual infected classes. However, if the interaction factor is greater than zero, than the probability for infected individuals being infected with another contagion can either decrease or increase depending if the interaction factor is less than or greater than one, respectively.

### Model Formulation

In this section, I formulate the mathematical model of the ISM based on the concepts derived above, giving further details about the variables and assumptions as it relates to the system. Table 8 provides a full list and definitions of the variables used within the model while the transition among states within the flow diagram is shown in Figure 18.

**Table 8: ISM Symbols and Definitions**

| Symbol: | Definition: |
|---|---|
| $\beta_i$ | Transmission Rate: Probability of individual to be influenced from virus based on contact |
| $\gamma_{ij}$ | Interaction Factor: Level of competition between two viruses ($\gamma_{ij}$ = 0 when viruses are mutually exclusive) |
| $\alpha_i$ | Spontaneous Rate: Probability of being influence without any contact |
| $g_i$ | Recovery Rate: Probability individuals recovers from virus |

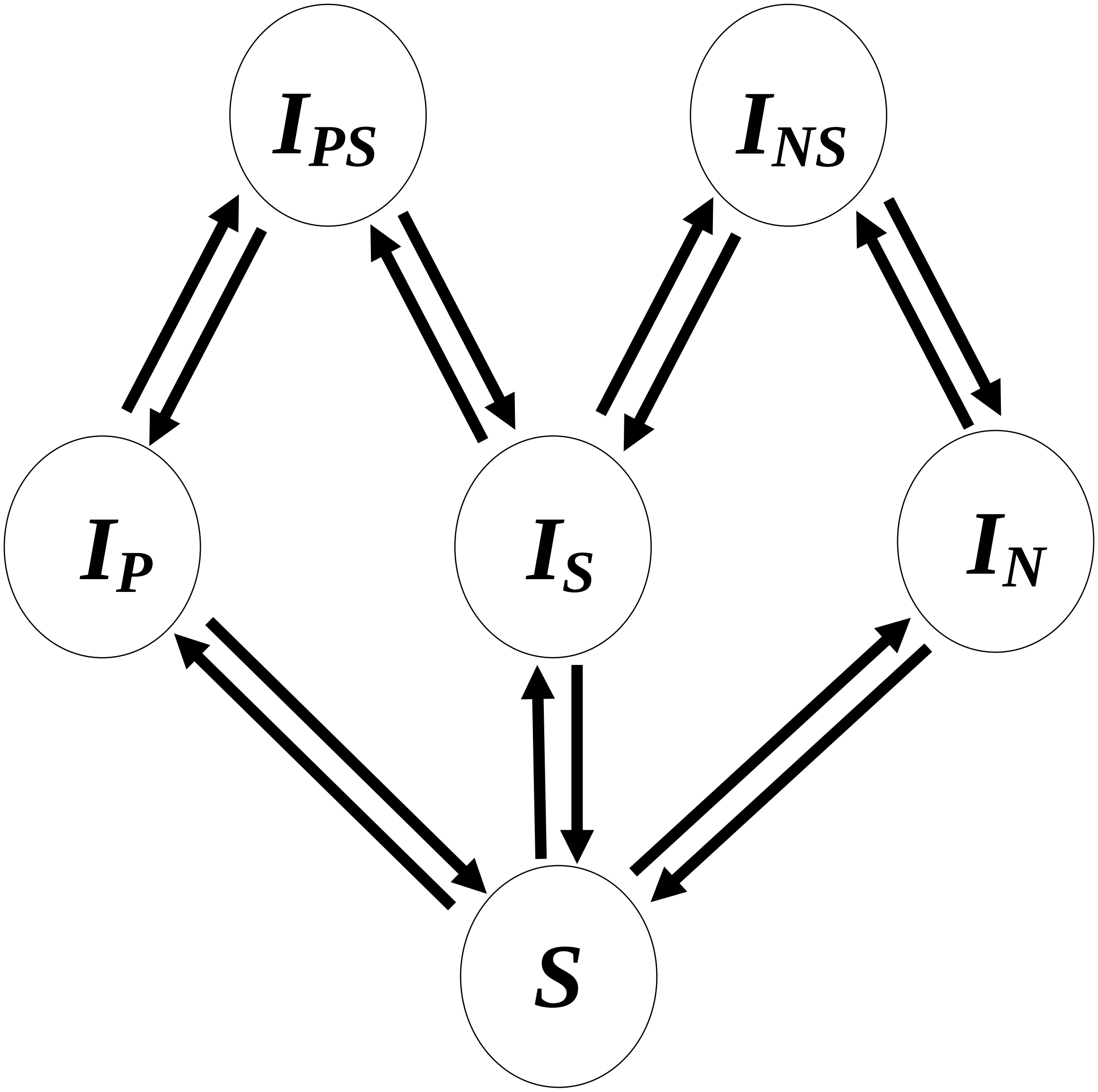


**Figure 18: ISM Model Flow Diagram**

Now that the symbols and states within the ISM are defined, the problem can be represented as the following differential equations where $N$ equals the total number of nodes within the system:

$$N = S + I_P + I_N + I_S + I_{PS} + I_{NS} \tag{5.5}$$

$$\frac{dI_N}{dt} = \beta_N S + g_S(I_{NS}) + \alpha_N S - g_N(I_N) - \beta_S I_N \gamma_{NS} - \alpha_S(I_N) \tag{5.6}$$

$$\frac{dI_P}{dt} = \beta_P S + g_S(I_{PS}) + \alpha_P S - g_P(I_P) - \beta_S I_P \gamma_{PS} - \alpha_S(I_P) \tag{5.7}$$

$$\begin{gathered} \frac{dI_S}{dt} = \beta_S S + g_P(I_{PS}) + g_N(I_{NS}) + \alpha_S S - g_S(I_S) \\ -\beta_P I_S \gamma_{PS} - \beta_N I_S \gamma_{NS} - (\alpha_P + \alpha_N) I_S \end{gathered} \tag{5.8}$$

$$\frac{dI_{PS}}{dt} = \gamma_{PS}(\beta_P I_S + \alpha_P I_S + \beta_S I_P + \alpha_S I_P) - (g_P + g_S) I_{PS} \tag{5.9}$$

$$\frac{dI_{NS}}{dt} = \gamma_{NS}(\beta_N I_S + \alpha_N I_S + \beta_S I_N + \alpha_S I_N) - (g_N + g_S) I_{NS} \tag{5.10}$$

Additionally, by setting the derivatives of each state equal to zero, the equilibrium points are calculated for the differential equations. By making the assumption that $\alpha_P + \beta_P P \neq 0$, $\alpha_N + \beta_N N \neq 0$, and $\alpha_S + \beta_S S \neq 0$, then the equilibrium points for each state can also be calculated.

## Experimental Results

In this section, I start with setup of the experiment followed by model simulations using varying assumptions for the different rates, followed by analyzing any comparisons between the different experimental results.

### Problem Setup

Since the actual solution of the differential equations is difficult to compute [28], the solution of the ISM is determined using a numerical simulation. The first step in the simulation is to set the initial parameters which are dependent on the specific case study. For the initial simulations, parameters are set to fixed values in range with Hill [27] and Dodds and Watts [16] for positive and negative sentiment, along with average influence and ISO scores of the more prominent bloggers. Those parameters within the simulation as follows: $\beta_n = .04$, $\beta_p = .01$, $\beta_s = .06$, $\alpha_n = .12$, $\alpha_p = .03$, $\alpha_s = .16$, $g_p = .14$, $g_n = .10$, and $g_s = .07$. Additional initial values for the six states are determined assuming an average degree of 3 [16], where $S = 313$, $I_P = 324$, $I_N = 308$, $I_S = 309$, $I_{PS} = 291$, $I_{NS} = 304$.

Additionally, I also looked at two different scenarios for $\gamma$. The first simulation assumes the interaction strength between positive or negative sentiment and opinion were equal to another, $\gamma_{ps} = 1$ and $\gamma_{ns} = 1$. The other simulation assumes the existence of a level of dependency between the two sentiments and opinion where $\gamma_{ps} = 5$ and $\gamma_{ns} = 2.5$, where the interaction is much stronger with positive and subjective contagions. Finally, the second simulated scenario is compared to the actual end-state distributions of sentiment and opinion within the blogosphere discussions of this case study discovered from the ISO scores alone.

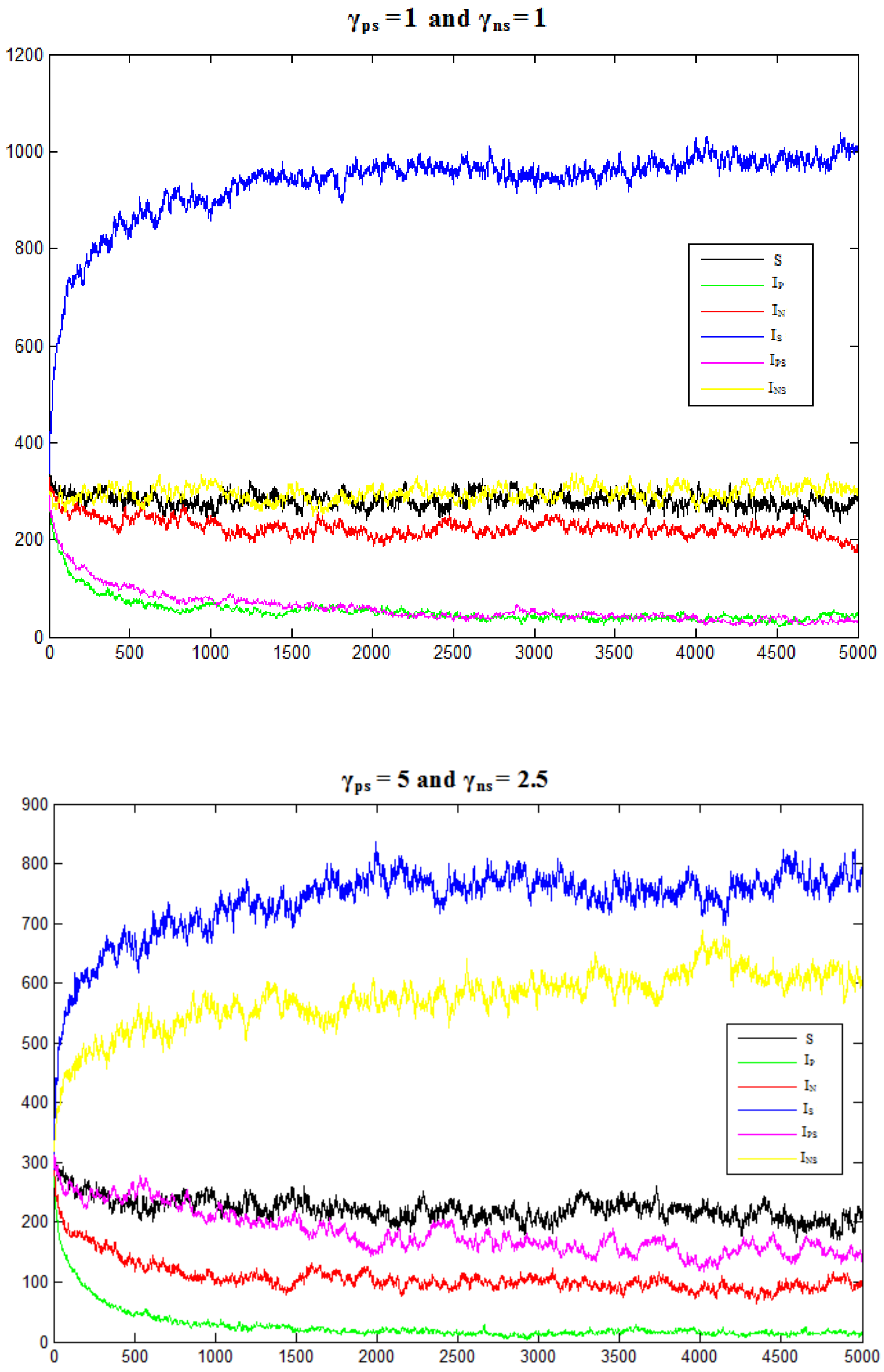
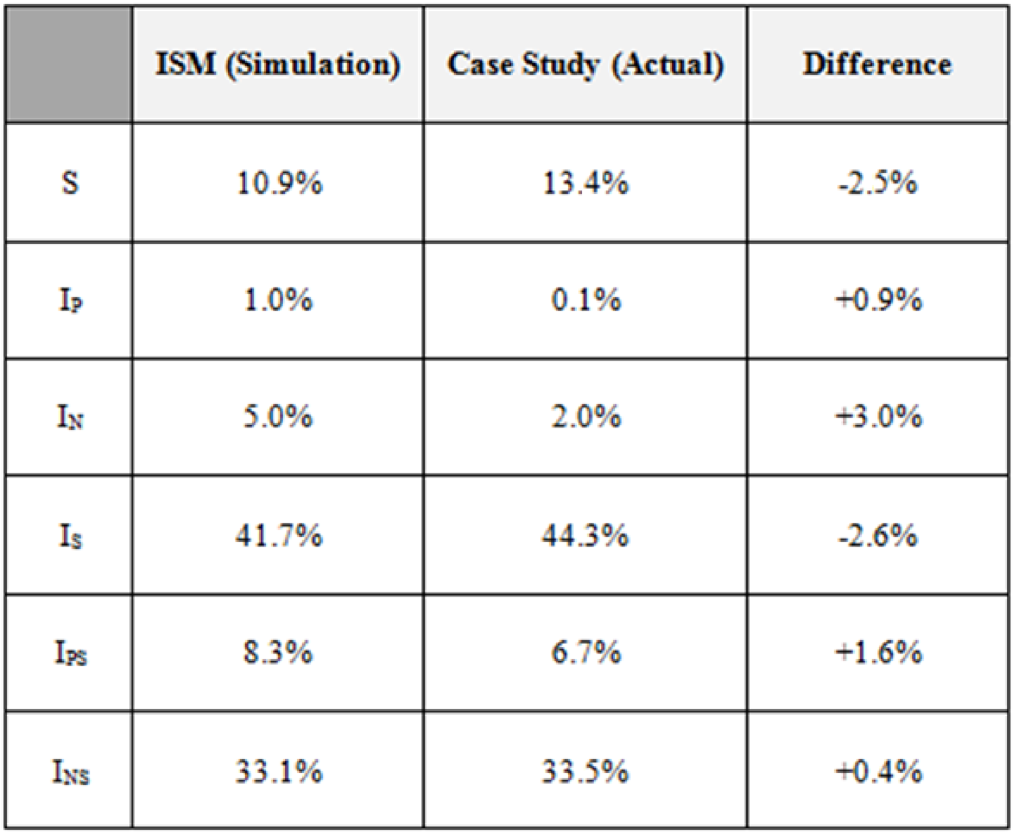


**Figure 19: ISM Simulation Results**

**Table 9: Simulation vs. Case Study (*$v_1$*) Equilibrium Results**



| | ISM (Simulation) | Case Study (Actual) | Difference |
|---|---|---|---|
| S | 10.9% | 13.4% | -2.5% |
| $I_P$ | 1.0% | 0.1% | +0.9% |
| $I_N$ | 5.0% | 2.0% | +3.0% |
| $I_S$ | 41.7% | 44.3% | -2.6% |
| $I_{PS}$ | 8.3% | 6.7% | +1.6% |
| $I_{NS}$ | 33.1% | 33.5% | +0.4% |

## Results

In both scenarios is the coexistence of both sentiment and opinion within the same network, but for the bottom figure, the interaction factor is directly proportional to the values of the two dual contagions, causing a decrease in the values of the single contagion states. Also, for the ISM parameters based on the case study, the probability of negative sentiment transmitting throughout the blogosphere was greater than positive sentiment, and the interaction factor was greater for subjective opinion and negative sentiment than for subjective opinion and negative sentiment. For both simulations, changing the initial values of the different states had no impact on the final equilibrium states.

While assessing the simulation to actual case study results, I discovered that the equilibrium states from the ISM model are very comparable to the final state of the blogosphere using the ISO scores (±3%). The next step in future research is to test the ISM on further issues and determine if using an initial state of ISO scores and the baseline ISM parameters leads to the correct final state of ISO scores. In conclusion, while the ISM simulation did not perfectly fit the actual results of the data from the case study, it is still reasonable for providing an approximate representation of the changes in sentiment and opinion based on blogger influence.

## Conclusion

For this chapter, I propose a method for modeling the spread of influence within the blogosphere using techniques from social networks and infectious disease models. Within the social network, nodes and edges represent bloggers' influence and closeness to others within the community, leading to the determination of clustering neighborhoods as well as the more prominent bloggers. Additionally, the influential spread model represents sentiment and opinion as contagions allowing to measuring the spread of the variables within the blogosphere using a modified epidemic model. As for the experimental results, simulations show how sentiment and opinion impact the entire community as well as provide a moderate representation of sentiment and opinion spreading within the specific case study.

On top of representing the blogosphere through a social network, a key aspect from the ISM is to better understand exact interactions between sentiment and opinion, a new approach compared to interchanging the two variables as one. In the future, I propose further analysis of the ISM to determine any relations between sentiment and opinion both for single issues and across much larger multi-topic networks as well as progress towards predicting past states and initial causes of the influence.

# CONCLUSION AND FUTURE WORK

In the previous chapters of this dissertation, I presented new and modified techniques to answer the two proposed research questions stated below:

1) <u>How can I model a blogger's sentiment and opinion separately for an issue over a specific time period?</u>
2) <u>How do I model a blogger's influence of sentiment and opinion on the entire blogosphere community?</u>

To answer these questions, I first developed the ISO model using a set of algorithms used to independently score and visualize the author's sentiment and opinion expressed within each of their documents. Secondly, I also developed the ISM to model the author's influence of those factors on others within the blogosphere. Throughout the rest of this chapter, I further summarize the main contributions presented throughout this dissertation. This is then followed by a discussion of limitations I came across, as well as extensions to the model for further potential research.

## Summary of Contributions

To address the first problem, I used a methodology for scoring a blog for the level of sentiment and opinion that was introduced, both at the sentence-level and document-level. Each sentence within the document was either characterized as a set of nouns representing issue viewpoints or as a collection of adjectives, verbs, and adverbs to support both the sentiment and opinion features where scores were then calculated and also propagated upwards to calculate the ISO scores for each document. In addition, different models for visualizing the changing scores over a period of time were also discussed. This work was one of the first to determine sentiment and opinion scores within text independently of one another prior to determining any connections between the two variables.

To approach the second research problem, I created two related models. The first model introduced was a social network throughout the blogosphere. This diagram served as a way to determine the more prominent bloggers as well as to measure the spread of influence from them to the rest of the community. To create a realistic representation of the blogosphere's social network, more popular bloggers, denoted as nodes in the network, had a higher influence and larger number of connections, with many being much closer relationships to others, represented as smaller distance edges.

The creation of the social network model then led to the development of the Influential Spread Model (ISM) representing to measure the changes in sentiment in opinion. This model was based on a modified epidemic model with interacting viruses showing how the sentiment and opinion from the more influential nodes spreads throughout the rest of the blogosphere. This work was one of the first to take the concept of sentiment and opinion as different contagions while modeling both spontaneous influences as well as measuring the interactions between those variables.

## Limitations

In this section, I present two key limitations within the methodology. The first limitation is in regards to the initial data collection, in which recommendations, comments, and followers are fixed variables. These values are dependent on the time the data is extracted, and since only one collection was performed at the beginning of the analysis, no changes exist for these values. The other limitation is in regards to sentiment and opinion scoring. When scoring a sentence, I focused on the sentiment and opinion words used to discuss an issue keyword, but these variables did not always directly relate to the author's actual viewpoint (i.e. discussions on others' emotions and opinions).

Using fixed variables impacts both the social network and influential spread model. By not collecting data at multiple instances, relative and absolute influence scores of the blogger nodes and the edges connecting bloggers are fixed as well since blogosphere dynamics are not shown. By using static values, it also impacts the ISM by using the same transmission rate for all iterations. In reality, it would be more beneficial to view any dynamic changes to these variables to better track blogger connection lengths and the spread of influence. However, to perform this method would require significantly more computational steps, but at the same time it would also provide a more accurate picture of the blogosphere.

Additionally, the ISO model scores sentences for sentiment and opinion are based on the frequency of words used in each category. However, what the model does not take into account, is when authors quote or restate what others are saying, which is common in blogs as part of a discussion or rebuttal. While I remove quotes from the posts prior to scoring, if the author is merely paraphrasing others' thoughts, it usually goes unnoticed. Therefore, it is important to understand the limitation that sentiment and opinion scores are not always directly related to the author's viewpoints. This is especially important to note, especially when discussing subjectivity and emotion revolving around such a sensitive issue.

## Future Work

This work provides the basis for other potential areas for future research for extending methodologies and models discussed above. In particular, the next major objectives would be to further measure the dependency of sentiment and opinion to determine both topic-specific and overall general patterns and then to redefine the virus interaction factor ISM contagion model. While some related suggestions for future work have already been discussed in the previous chapters, other additional ideas for future research are discussed below.

The first component was an introductory analysis of independently measuring sentiment and opinion. Sentences were measured to see if sentiment or opinion existed, but as I realized, it was also difficult to determine to which viewpoint each issue referred. A follow up to this research would be to continue adapting existing sentiment analysis techniques such as subject/object identification and aspect-based analysis, all while satisfying the scoring of documents for sentiment and opinion independently prior to measuring potential dependencies.

Furthermore, the current model is only based off the analysis of a single issue. In reality, the author’s opinion and sentiment would likely be present across multiple issues within the blogosphere, and tracking how these factors evolve over time to determine correlations between these different topics. It would be interesting to investigate additional visualization techniques to better understand what interdependencies exist between sentiment and opinion across a single issue as well as a larger multi-topic problem.

The second model (ISM) extended on existing SIR models by including the multiple contagions as well as how authors’ sentiment and opinion influence spread both independently and dependently of one another based on the interaction factor for the viruses. While this research focused on influence directly from connected bloggers within the limited social network, it did not account for external news sources such as current events. Therefore, it could prove to be beneficial to understand the impacts of outside sources on authors’ sentiment and opinion. This would provide more details for how much influence is directly impacted by a current event or outside news story versus direct influence from other bloggers.

Additionally, the ISM only focused on bloggers within a single blogosphere, it would be useful to expand the social network and track the influence between multiple, closely related blogospheres. As stated with the first problem, also tracking influence among multiple topics or issues would be valuable as well. Thus, an interesting next step would be to increase the size of the number of authors and their connections in addition to the number of topics being discussed amongst them.

In conclusion, I would also want to be able to construct a much larger version of both the social network and ISM, all with the bigger objective to better understand which factors have the greatest impacts on an author's expressed level emotion and subjectivity. Continuing down this path would hopefully make it possible to eventually use this model to work backwards and understand what was the initial cause, and whether it was a specific person, event or story, that caused either sentiment or opinion to peak, drop, and cause an influential outbreak throughout the rest of the community.

# REFERENCES

[1] Amiri, Hadi, and Tat-Seng Chua. "Mining sentiment terminology through time." *Proceedings of the 21st Advancing Computing as a Science & Profession (ACM) International Conference on Information and Knowledge Management*. ACM, 2012.

[2] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." *Conference of Language Resources and Evaluation (LREC)*. Vol. 10. 2010.

[3] Banchs, Rafael E. *Text mining with MATLAB®*. Springer Science & Business Media, 2012.

[4] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An open source software for exploring and manipulating networks." *International Conference on Weblogs and Social Media (ICWSM*-8) (2009): 361-362.

[5] Benamara, Farah, Cesarano, Carmine, Picariello, Antonio, Reforgiato, Diego, and Subrahmanian VS. "Sentiment analysis: Adjectives and adverbs are better than adjectives alone." *ICWSM*. 2007.

[6] Beutel, Alex, Prakash, B., Rosenfeld, Roni, and Faloutsos, Christos. "Interacting viruses in networks: Can both survive?." *Proceedings of the 18th ACM Special Interest Group on Knowledge Discover and Data Mining (SIGKDD) International Conference on Knowledge Discovery and Data Mining*. ACM, 2012.

[7] Blei, David M. "Probabilistic issue models." *Communications of the ACM* 55.4 (2012): 77-84.

[8] Bradley, David, "Six degrees of separation." *ScienceBase*. 1 Feb 2008. Web. 10 Dec 2014. <http://www.sciencebase.com/science-blog/six-degees-of-separation.html>.

[9] Bródka, Piotr, Paweél Stawiak, and Przemyséaw Kazienko. "Shortest path discovery in the multi-layered social network." *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*. IEEE, 2011.

[10] Bubniakova, Lenka. *The mathematics of infectious diseases.* Diss. Comenius University, 2007.

[11] Chen, Bi, Zhu, Leilei, Kifer, Daniel, and Lee, Dongwon. "What is an opinion about? Exploring political standpoints using opinion scoring model." *Association for the Advancement of Artificial Intelligence (AAAI)*. 2010.

[12] Chen, Bi. *Topic oriented evolution and sentiment analysis*. Diss. The Pennsylvania State University, 2011.

[13] Chenlo, Jose M., and David E. Losada. "Combining document and sentence scores for blog topic retrieval." *Proceedings of the Spanish Conference on Information Retrieval (CERI 2010)*. 2010.

[14] *DailyKos.* Markos Moulitsas Zúniga, 2002. Web. 15 Aug 2015. <http://www.dailykos.com/special/about>.

[15] Dermouche, Mohamed, Velcin, Julien, Leila, Khouas, and Loudcher, Sabine. "A joint model for topic-sentiment evolution over time." *Data Mining (ICDM), 2014 Institute of Electrical and Electronics Engineers (IEEE) International Conference on*. IEEE, 2014.

[16] Dodds, Peter Sheridan, and Duncan J. Watts. "Universal behavior in a generalized model of contagion." *Physical Review Letters* 92.21 (2004): 218701.

[17] Feinberg, Jonathan. "Wordle." (2014).

[18] Fellbaum, Christiane. *WordNet*. Blackwell Publishing Ltd, 1998.

[19] Fukuhara, Tomohiro, Hiroshi Nakagawa, and Toyoaki Nishida. "Understanding sentiment of people from news articles: Temporal sentiment analysis of social events." *International Conference on Women, Media and Sexuality* (*ICWSM)*. 2007.

[20] Furuse, Osamu, Hiroshima, Nobuaki, Yamada, Setsuo, and Kataoka, Ryoji. "Opinion sentence search engine on open-domain blog." *International Joint Conference on Artificial Intelligence (IJCAI)*. 2007.

[21] Gliwa, Bogdan, Anna Zygmunt, and Jarosław Koźlak. "Analysis of roles and groups in a blogosphere." *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Springer International Publishing, 2013.

[22] *Google Trends*. 2015. Web. 30 Sep 2015. <http://www.google.com/trends>.

[23] Granovetter, Mark. "Threshold models of collective behavior." *American Journal of Sociology* (1978): 1420-1443.

[24] Hammer, Hugo Lewi, Per Erik Solberg, and Lilja Øvrelid. "Sentiment classification of online political discussions: A comparison of a word-based and dependency-based method." Association for Computational Linguistics, 2014.

[25] Haran, Murali. "An introduction to models for disease dynamics." *Spatial Epidemiology, Statistical and Applied Mathematical Sciences Institute (SAMSI)* (2009).

[26] Herring, Susan C., Kkouper, Inna, Paolillo, John, Scheidt, Lois, Tyworth, Michael, Welsch, Peter, Wright, Elijah, and Yu, Ning. "Conversations in the blogosphere: An analysis 'from the bottom up'." *Hawaii International Conference on System Sciences (HICSS), 2005. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE, 2005.

[27] Hill, Alison L., Rand, David G., Nowak, Martin A., and Christakis, Nicholas A.. "Emotions as infectious diseases in a large social network: The SISa model." *Proceedings of the Royal Society of London B: Biological Sciences* 277.1701 (2010): 3827-3835.

[28] Hill, Alison L., Rand, David G., Nowak, Martin A., and Christakis, Nicholas A. "Infectious disease modeling of social contagion in networks." *Public Library of Science (PLOS) Comput Biol* 6.11 (2010): e1000968.

[29] Hohman, Elizabeth Leeds, and David J. Marchette. "A dynamic graph model for analyzing streaming news documents." *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*. IEEE, 2007.

[30] Hui, Peter, and Michelle Gregory. "Quantifying sentiment and influence in blogspaces." *Proceedings of the First Workshop on Social Media Analytics*. ACM, 2010.

[31] Jin, Hua, Zhu, Yatao, Jin, Zhiqiang, and Arora, Sandhya. "Sentiment visualization on tweet stream." *Journal of Software* 9.9 (2014): 2348-2352.

[32] Jo, Yohan, and Alice H. Oh. "Aspect and sentiment unification model for online review analysis." *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining.* ACM, 2011.

[33] Kermack, William O., and Anderson G. McKendrick. "A contribution to the mathematical theory of epidemics." *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 115. No. 772. The Royal Society, 1927.

[34] Lin, Chenghua, and Yulan He. "Joint sentiment/topic model for sentiment analysis." *Proceedings of the 18th ACM conference on Information and Knowledge Management.* ACM, 2009.

[35] Lin, Chenghua, He, Yulan, Everson, Richard, and Ruger, Stefan. "Weakly supervised joint sentiment-topic detection from text." *Knowledge and Data Engineering, IEEE Transactions on* 24.6 (2012): 1134-1145.

[36] Liu, Bing. "Sentiment analysis and subjectivity." *Handbook of Natural Language Processing* 2 (2010): 627-666.

[37] Liu, Zhifeng, Tingting Zhang, and Qiujun Lan. "An extended SISa model for sentiment contagion." *Discrete Dynamics in Nature and Society* 2014 (2014).

[38] Marconi, Guglielmo. "Wireless telegraphic communication." *Nobel Lecture, December* 11 (1909).

[39] McNichol, Tom “TIME.com's first annual blog index.” *The Huffington Post*. 2007. Web. 22 Jun 2015. <http://content.time.com/time/specials/2007/article/0,28804,1725323_1725329,00.html>.

[40] Mei, Qiaozhu, and ChengXiang Zhai. "Discovering evolutionary theme patterns from text: An exploration of temporal text mining." *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* ACM, 2005.

[41] Milgram, Stanley. "The small world problem." *Psychology Today* 2.1 (1967): 60-67.

[42] Mishra, Shubhanshu. *Analysis of social media data to determine positive and negative influential nodes in the network*. Diss. Indian Institute of Technology, Kharagpur West Bengal, India, 2012.

[43] *MLB.com Gameday*. 2010. Web. 7 Oct 2015. <http://mlb.mlb.com/mlb/gameday/y2010/>.

[44] Myers, Seth A., and Jure Leskovec. "Clash of the contagions: Cooperation and competition in information diffusion." Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012.

[45] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and Trends in Information Retrieval* 2.1-2 (2008): 1-135.

[46] Quirk, Randolph, et al. "A comprehensive grammar of the English language." *English Language Teaching (ELT)Journal* 42 (1988): 3.

[47] Sela, Alon, Hila Oved, and Irad Ben-Gal. "Information spread in a connected world." *arXiv preprint arXiv:1406.7538* (2014).

[48] Song, Xiaodan, Chi, Yun, Hino, Koji, and Tseng, Belle. "Identifying opinion leaders in the blogosphere." *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*. ACM, 2007.

[49] Sukute, Karina, "Behavior patterns of the modern opinion leaders in the Latvian market" 2012, Diss. Universität Kassel, Kassel, Germany, 2012.

[50] Sulieman, Dalia, Malek, Maria, Kadima, Hubert, and Laurent, Dominique. "Semantic social breadth-first search and depth-first search recommendation algorithms." *3ième Conférence sur les Modèles et L'analyse des Réseaux: Approches Mathématiques et Informatiques*. 2012.

[51] Wallsten, Kevin. "Political blogs and the bloggers who blog them: Is the political blogosphere and echo chamber." *American Political Science Association's Annual Meeting. Washington, DC September*. 2005.

[52] Wang, Changbo, Zhao, Xiao, Liu, Yuhua, and Zhang, Kang. "SentiView: Sentiment analysis and visualization for internet popular topics." *Human-Machine Systems, IEEE Transactions on* 43.6 (2013): 620-630.

[53] West, Robert, Paskov, Hristo, Leskovec, Jure, and Potts, Christopher. "Exploiting social network structure for person-to-person sentiment analysis." *arXiv preprint arXiv:1409.2450* (2014).

[54] Wiebe, Janyce, and Ellen Riloff. "Creating subjective and objective sentence classifiers from unannotated texts." *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2005. 486-497.

[55] Wiebe, Janyce, Wilson, Theresa, Bruce, Rebecca, Bell, Matthew, and Martin, Melanie. "Learning subjective language." *Computational Linguistics* 30.3 (2004): 277-308.

[56] Wilson, Theresa, Hoffmann, Paul, Somasundara, Swapna, Kessler, Jason, Wiebe, Janyce, Choi, Yejin, Cardie, Claire, Riloff, Ellen, and Patwardhan, Siddharth. "OpinionFinder: A system for subjectivity analysis." *Proceedings of hlt/emnlp on Interactive Demonstrations*. Association for Computational Linguistics, 2005.

[57] Woodly, Deva. "New competencies in democratic communication? Blogs, agenda setting and political participation." *Public Choice* 134.1-2 (2008): 109-123.

[58] Zheng, Minjie, Wu, Chaorong, Liu, Yue, Liao, Xiangwen, and Chen, Guolong. "Topic sentiment trend model: Modeling facets and sentiment dynamics." *Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference on*. Vol. 3. IEEE, 2012.

# BIOGRAPHY

Michael J. Garrity graduated from Mount Vernon High School, Alexandria, VA, in 2006. He then received his Bachelor of Science and Master of Science in Mathematics both from George Mason University in 2010 and 2011 respectively. He has been employed at Modern Technology Solutions, Inc. (MTSI) as a Project Engineer for over six years and is expecting to receive his Doctor of Philosophy in Computational Sciences and Informatics from George Mason University in 2016.