DISCOVERING GEOGRAPHICAL TOPICS

FROM SOCIAL MEDIA

by

ELIZABETH WILLIAMS

JEFF GRAY, CO-COMMITTEE CHAIR
BRANDON DIXON, CO-COMMITTEE CHAIR
JEFFREY CARVER
SUSAN VRBSKY
JASON SENKBEIL

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2017

ABSTRACT

Traditional query-based search engines such as Google are often not able to discover real-time, contextual information such as traffic accidents or severe weather situations. As an alternative, social media can often provide relevant information to a user about important events that are occurring in their environment. However, to obtain this knowledge, a user may be required to wade through a large amount of irrelevant data.

In this dissertation, we describe our research goals for providing relevant contextual information to a user by mining social media. We describe the implementation of our system, GeoContext, which consists of a geotopical clustering system that discovers topics appearing in a social media stream and analyzes where the topics are centered geographically. GeoContext also includes a method for filtering a social media stream by keywords and location coordinates in order to provide more specific topics. In order to find the geographical location of topics, GeoContext must also predict the location of each social media post. However, due to privacy concerns, many social media users do not share their exact geographical coordinates. For this reason, GeoContext includes a technique that predicts locations of posts that are not associated with explicit coordinates, a process called geolocation. Existing research has utilized the content of a post as well as the post author's social media relationships with other users to estimate location. Our research provides a novel approach to geolocation by combining multiple techniques, as well as adding a new technique: estimating location by clustering social media posts of similar topics that are centered in a geographical area.

We evaluate the geotopical clustering portion of GeoContext against a common topic modeling algorithm often used in geotopical clustering, Latent Dirichlet Allocation. We also evaluate the parameters and threshold values implemented within GeoContext. In addition, we evaluate the geolocation portion of GeoContext by collecting geotagged social media posts (posts explicitly tagged with geographical coordinates) and comparing the predicted location from GeoContext against the actual coordinates.

DEDICATION

This dissertation is dedicated to my daughter, Isabella. I love you infinity.

LIST OF ABBREVIATIONS

API	Application Programming Interface
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GCL	GeoContext Locator
GCTC	Geographical Clustering, Topical Clustering
LDA	Latent Dirichlet Allocation
RDF	Resource Description Framework
REST	Representational State Transfer
SPARQL	Sparql Protocol And RDF Query Language
TCGC	Topical Clustering, Geographical Clustering
TF-IDF	Term-Frequency-Inverse Document Frequency

ACKNOWLEDGEMENTS

I would like to thank Dr. Jeff Gray for his guidance throughout my Ph.D. career. His mentorship made this research possible. I would also like to thank my committee members, Dr. Brandon Dixon (co-chair), Dr. Jeffrey Carver, Dr. Susan Vrbsky, and Dr. Jason Senkbeil for their advice and support.

Thank you to the University of Alabama Department of Computer Science and Dr. David Cordes for encouragement through both my undergraduate and graduate time at UA. I also appreciate all the friendship and support from my classmates and colleagues at UA, both during my undergraduate and graduate time.

In addition, I would like to thank Dr. Grace Lewis, Edwin Morris, Ben Bradshaw, and Keegan Williams from Carnegie Mellon University for their assistance and collaboration with performing evaluations.

I would also like to thank the Center for Advanced Public Safety, the University of Alabama National Alumni Association, and the Department of Education GAANN fellowship for providing funding throughout my graduate school career.

Lastly, I would like to thank my parents for being the best baby-sitters, dog-walkers, chauffeurs, chefs, and cheerleaders throughout my graduate school journey.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

With the ever-increasing use of mobile devices, users desire information faster than ever. Although traditional search engines are useful for many queries, they often do not provide relevant information about real-time contextual events. For that reason, users often turn to social media for time-sensitive questions about their environment.

When an emergency situation occurs, nearby citizens often post on social media sites such as Twitter[1] about their experience in the situation. This allows other people to very quickly realize details about the event and know whether friends and family are safe. This action also provides a lasting document of the situation for later investigation (Scapusio, 2017). Law enforcement agencies use social media as tools to investigate crimes and assist in tracking suspects. After the 2013 Boston Marathon explosions, the FBI used social media to broadcast important information to citizens regarding suspects (Jin, Dougherty, Saraf, Cao, & Ramakrishnan, 2013). Police departments are using social media platforms to monitor events such as large protests (Dwoskin, 2016). Social media posts can be used after emergency situations to retrace a timeline of events leading up to the situation (Scapusio, 2017).

Information such as traffic and weather updates are also often posted on social media due to the fact that these updates can offer critical information, such as in the case of a weather emergency, and social media can generally disseminate information more rapidly to users than

[1] http://twitter.com

other media (Sakaki, Okazaki, & Matsuo, 2013). In contrast, because query-based search engines such as Google provide information based on keyword relevance rather than temporal significance, they often provide stale data that is no longer useful with queries such as those about traffic or weather.

In addition, social media is often used for staging social movements. In a poll, nine out of ten Egyptians and Tunisians responded that they used Facebook to communicate and spread awareness of protests (Jin, Dougherty, Saraf, Cao, & Ramakrishnan, 2013). Social media has been used extensively for analysis during the Iran elections, tsunami in Samoa, and earthquake in Haiti (Hong, Ahmed, Gurumurthy, Smola, & Tsioutsiouliklis, 2012). The real-time nature and broad reach of social media is ideal for the posting of time-critical information. These characteristics of social media platforms indicate that modeling data from social media across different locations can reflect the ideas, opinions, and information that is important to people in varied geographical regions. Social media can allow the discovery of the opinions and information of the over 3 billion worldwide Internet users, rather than relying on a limited number of traditional media sources.

1.1. Motivation for Modeling Social Media

Social media regularly provides more detailed information about topics such as current events, emergency situations, traffic, or weather than other outlets. Because social media happens in real-time, it can also provide knowledge more quickly than other media, which in turn helps us make better and faster decisions about current situations.

Due to aspects such as publishing time, traditional media news sources can often be slower at reporting events than social media. Leskovec et al. (Leskovec, Backstrom, & Kleinberg, 2009)

studied the lag times of blogs, which are one type of social media, versus other traditional media sites such as cnn.com and washingtonpost.com. The authors showed that independent media sites and blogs are often very quick at mentioning news, while traditional media websites, although ahead of much of the crowd, lag behind. They also found that there exist a few cases in which news stories began prominently within blogs and only later percolated to mainstream traditional media sites.

A few existing mobile device applications attempt to remedy the information lag by providing users with information based on their context. However, these applications mainly rely on traditional media for news. Currently, none of the state-of-the-art contextual applications (e.g., Google Now[2], Siri[3]) utilize an entire social media stream as a means to source data.

Zhao et al. (Zhao, et al., 2011) compared Twitter usage against traditional media sources, using the New York Times as their case study. They applied Latent Dirichlet Allocation (LDA) to a corpus of New York Times articles and a modified LDA algorithm to a collection of tweets in order to extract topics. The authors then assigned categories to each topic that appeared in both types of media. Their results showed that Twitter and the New York Times differed greatly in their distribution of categories. The topic "Family and life" dominated Twitter, while "Arts," "World," and "Business" were predominant in the New York Times. The authors also found that "Entity-oriented" topics, i.e., topics about celebrities, occur more frequently on Twitter than the New York Times.

[2] https://www.google.com/search/about/
[3] http://www.apple.com/ios/siri/

Social media can provide several unique aspects that other media lack, including, but not limited to:

- *Location specific information* - because many social media platforms incorporate a geographical component, more specific location information can sometimes be extracted from social media posts than traditional media sources. On many of the major platforms, users have the option of including a geotag, or fine-grained location information, with their post.
- *Fast dissemination* - because, unlike traditional media, social media does not require publishing time, and posting typically requires only a few words and the push of a button, social media is often a faster way to push information out to the public. Also, some platforms provide an option to quickly repost information from other users, such as Twitter's retweet[4], which allows information to disseminate among viewers extremely rapidly (Sakaki, Okazaki, & Matsuo, 2013).
- *Opinions of users* - due to the nature of social media, it reflects individual experiences of an event or situation, while traditional media news stories can be viewed as more of a conglomeration of information related to the event. In many cases, traditional news media also attempts and claims to provide impartial coverage of an event. However, sometimes personal opinions are desired regarding a topic for research purposes, such as how people feel about a recently released or upcoming movie. In this case, social media can provide individual reviews.

[4] https://support.twitter.com/articles/20169873

1.2. Topical and Geographical Modeling in Social Media

Despite the large push towards utilizing social media as a tool for analyzing and detecting events, there are still areas in which the existing tools and methods can be improved. Early work in social media research focused on analyzing the characteristics of microblogging services such as Twitter. For example, Java et al. (Java, Song, Finin, & Tseng, 2007) studied the types of information about which Twitter users are posting, as well user's intentions when utilizing Twitter. Dela Rosa et al. (Dela Rosa, Shah, Lin, Gershman, & Frederking, 2011) attempted categorization of tweets into topics using hashtags and compared several different methods of unsupervised and supervised clustering of the tweets, including LDA, K-means clustering, and a Rocchio classifier.

Newer research has focused on applying topic models to streams of posts from social media sources. Topic models are "algorithms that uncover the hidden thematic structure of document collections" (Blei, Topic Modeling). In many cases, approaches have begun to include other aspects of social media into the model, such as temporal and geographical information. Social media analysis has spread to encompass topic modeling (Vosecky, Wai-Ting Leung, & Ng, 2013), event detection (Baldwin, Cook, Han, Harwood, Karunasekera, & Moshtaghi, 2012), bursty topic detection (Diao, Jiang, Zhu, & Lim, 2012), and more. Some methods are being used to track the evolution of topics over time (Yang, Chen, Lyu, & King, 2011). However, because many social media users do not provide fine-grained location information, existing methods are not able to take full advantage of geographical analysis.

In addition, many geographical analysis approaches that utilize clustering algorithms, such as DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), result in clusters based on population rather than relevance. If a cluster appears using a traditional clustering algorithm, the cluster's significance

is only that it is centered at a location with a higher population, not that a larger percentage of people are posting on social media about a certain topic in locations compared to other locations. One of our goals for the research presented in this dissertation was to be able to detect events or topics that people are discussing predominately in one specific location. For that reason, in order to find the relevance of topics to geographical locations, it is important to normalize for population.

The overall goal for our research was to model real-time social media topics based on location and keyword data. Our research method extends the existing models by providing a new and unique method for performing geographical analysis. Many social media platforms do not currently allow querying for specific information by geographical location. For example, a query for the keyword "traffic" on Twitter results in traffic information across the world. Even a query such as "traffic in Los Angeles" results in many non-specific tweets such as "Coffee for the hour of traffic I'm about to sit in by myself" that do not provide much relevant information for users caring about traffic. Also, users must filter through a large amount of information to find relevant results to even simple queries on social media.

The implementation of our research method is called GeoContext, which is able to organize social media information into topics that are geographically centered and contain posts that are topically similar. GeoContext can be used to model how relevant topics within social media are to a specific location. For example, in Chapter 6, we show how GeoContext could detect critical information about airport delays at a specific airport.

GeoContext can also be used to discover opinions about current or political events that are important to users in different geographical locations. GeoContext's analysis server performs social media analysis and investigates how trends on social media relate to various geographical

locations. We chose to use Twitter for GeoContext, and our reasons for this choice are described in Section 2.3.

1.3. Challenges to Geotopical Analysis in Social Media

Performing analysis is required to take advantage of some of the unique aspects of information found on social media. Every second, about 6,000 tweets are published (Twitter Statistics, 2017). 293,000 statuses are updated on Facebook every minute (Top 15 Valuable Facebook Statistics, 2017). Due to this sheer amount of data that exists on online social networks, some type of modeling of the stream of posts is needed to provide insights into the posts.

Because a large number of posts exist, many methods for gleaning information from social media organize posts into topics. Some tweets, like the one shown in Figure 1.1a, do not fit neatly into any topic if the desire is to discover local, regional, or national breaking events due to the fact that this tweet is more personal. Others, like those shown in Figure 1.1b, are redundant due to the fact that they are retweets. In this case, the second tweet does not provide any additional information regarding the topic because it is simply repeating another user's post. Tweets like these represent some of the challenges that need to be considered when choosing algorithms for analysis. By performing topic analysis on a social media stream, clear concepts are revealed and can indicate user interest.

In addition to topic analysis, we decided to focus on geographical analysis. Social media can be a valuable source for information specific to certain locations that may never be relevant enough to be shown by traditional media sources. For example, many people post about non-emergency current events on social media. Watanabe et al. (Watanabe, Ochi, & Onai, 2011)

(a) Thanks friends for keeping me updated.Hahaha
@iamlorainep @aldrichismyname Shalee and Judo

(b) RT @TC_GOP: Listening to Governor Greg Abbott the TCGOP Lincoln Day Dinner
https://t.co/nk9Dp3lTpM/s/tRTO
https://t.co/ff0ti7IiKN

Figure 1.1. Example Tweets (Personal and Retweet)

showed that many users tweet about local events, which they define as "when a certain number of people with a common purpose gather together at the same time and place." We decided to tailor our algorithms to reveal information that is particularly important to certain locations, not just broad topics that are highly popular across an entire social media platform.

In order to perform this type of geographical analysis, the location of the social media posts is needed. Due to privacy concerns, many users choose not to share their locations on social media. On the social networking site Twitter, as few as 0.87% of tweets are geotagged, or associated explicitly with geographical coordinates (Jaiswal, Peng, & Sun, 2013). In our previous research, we found that only 2.63% of tweets we analyzed were tagged with explicit location information (Williams, Gray, & Dixon, 2017). However, some locations can be inferred from the content and metadata of the post. Discovering the locations is vital to having enough geographical information in order to analyze where topics are located.

Given these challenges, our overall research goal is to create a method for modeling geographic topics within a social media stream. Our implementation of this approach, called GeoContext, uses social media to model relevant information about topics such as traffic,

weather, or current events across geographical locations. This overarching goal is comprised of three separate goals: geolocation of tweets, topical analysis, and geographical analysis.

1.3.1. Geolocation of Tweets

Location is often used as a common source for personalization within websites and mobile applications. For example, retailers with mobile apps such as Target often send notifications when a user enters their store. Websites frequently show advertisements based on a visitor's location.

Social media platforms have also provided ways to incorporate their user's location. On many social media platforms, users have the ability to tag their posts with their current location information. Because of the fast dissemination of information on social media, it has been used as a method for detecting natural disasters and gathering information about major events such as the 2013 Boston Marathon explosions (Jin, Dougherty, Saraf, Cao, & Ramakrishnan, 2013). Knowing the location information for these types of events can allow first responders to pinpoint the exact location, as well as the reach, of the event. For example, using a form of geolocation, Sakaki et al. (Sakaki, Okazaki, & Matsuo, 2013) discovered the epicenter of earthquakes as well as the location of the aftershocks. Musaev et al. (Musaev, Wang, Shridhar, & Pu, 2015) analyzed the effects of mudslides through geolocation. In order to perform a full analysis on social media posts to identify major events or discover people's opinions in different geographical areas, location information for the posts is required. As discussed in (Spinsanti, Berlingerio, & Pappalardo, 2013), social media may not always provide locations that are as accurate as GPS, but social media provides public access to location data as well as a social aspect. Also, discovering locations from social media can be useful when traditional location

services such as IP-address geolocation give inconsistent results (Backstrom, Sun, & Marlow, 2010).

Because so many tweets are not associated explicitly with a specific location, GeoContext is unable to take advantage of the location of the tweets to provide location-specific information. Because the ultimate goal of geolocation is predicting all possible locations with 100% accuracy, existing geolocation methods have room for improvement.

The geolocation module within GeoContext, called GeoContext Locator (GCL), can successfully detect many locations within tweets, both broad locations such as cities, states, or large venues such as stadiums, as well as smaller places such as stores, restaurants, or unique destinations that are more difficult to detect. GCL utilizes a novel combination of resources including Dbpedia[5] and Google Maps[6] in order to realize this goal. Dbpedia is a database consisting of structured information extracted from Wikipedia. More specifically, Dbpedia contains information found in the sidebar information boxes on Wikipedia.

GCL combines analysis of the content of the tweet, the location specified on the user's account, and locations of the user's friends and followers to perform geolocation. Previous research has not utilized all of these aspects together to discover a tweet's location. We also present an innovative approach to geolocation by estimating a tweet's location by analyzing the locations of tweets with similar content in real-time.

1.3.2. Topical Analysis

In order to provide topics in social media that are relevant to different users, GeoContext must be able to cluster together social media posts that consist of the same topic. The goal of topic

[5] http://wiki.dbpedia.org/
[6] https://developers.google.com/places/

modeling in the case of social media is to find an effective set of topics that contain posts about a similar or identical event or situation.

In implementing GeoContext's topic modeling module, we wanted to minimize the number of topics that:

1. contain tweets that are too dissimilar in their content
2. should be combined into one topic due to similarity of content

Identifying topics on Twitter using traditional natural language processing techniques can be challenging due to the short allowed length of tweets (140 characters or less). However, the short character limit means that tweets are often limited to a single topic, which makes them a good candidate for topic categorization. Topic modeling algorithms such as LDA (Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003) can provide a solution for analyzing the content of tweets. However, other methods may be needed to improve the classification of tweets beyond what is possible with traditional topic models.

Although many methods choose to remove stop words, many existing methods for topic discovery, such as LDA, treat every word other than the stop words in the tweet equally. This means that topics may have extraneous words that do not contribute to the overall meaning of the topic. On the other hand, GeoContext uses cognitive computing techniques to identify important keywords out of the tweet and cluster the social media posts into topics that contain clearly defined, popular subjects.

GeoContext provides several advantages over existing clustering approaches. First, GeoContext can process social media posts immediately as they are streamed without removing stop words, which are words (e.g., "the" or "a") that are often removed before natural language processing, or stemming terms (i.e., returning terms to their root form, such as transforming the

word “running” into “run”). This method speeds up the processing time of GeoContext as well as reduces the risk of losing the meaning of words.

Second, because of GeoContext’s method of extracting concepts from posts, there is no need for an initial training set. This means that GeoContext is effective as a real-time analysis system, because it can be started immediately on a stream with no prior inputs. Also, GeoContext uses a dynamic number of topics, whereas many implementations of topic modeling algorithms require a fixed topic count defined prior to the modeling process.

1.3.3. Geographical Analysis

The last aspect of GeoContext combines geolocation and topical clustering and determines which topics are important to different geographical regions. To illustrate the importance of discovering geographical topics on social media over traditional searches, we built a query about a traffic incident viewed on Google Maps[7] via Waze[8] (shown in Figure 1.2a). The results from Google include web pages about traffic accidents on the same road that occurred 5 months prior (shown in Figure 1.2b). None of the resulting links are relevant to the traffic in the query, even when the temporal word “today” is included in the query. Because search engines such as Google rely on keyword relevance more than temporal relevance, it is often difficult for users to find relevant information for queries about events such as traffic accidents.

However, as shown in Figure 1.2c, a relevant post from Twitter about the accident was found. The tweet contains more extensive information about the traffic, including the facts that the traffic was caused by fuel on the roadway and that the incident affects northbound traffic (indicated by “NB” in the tweet), as well as the intersection where the traffic occurred. Utilizing

[7] https://maps.google.com
[8] https://www.waze.com

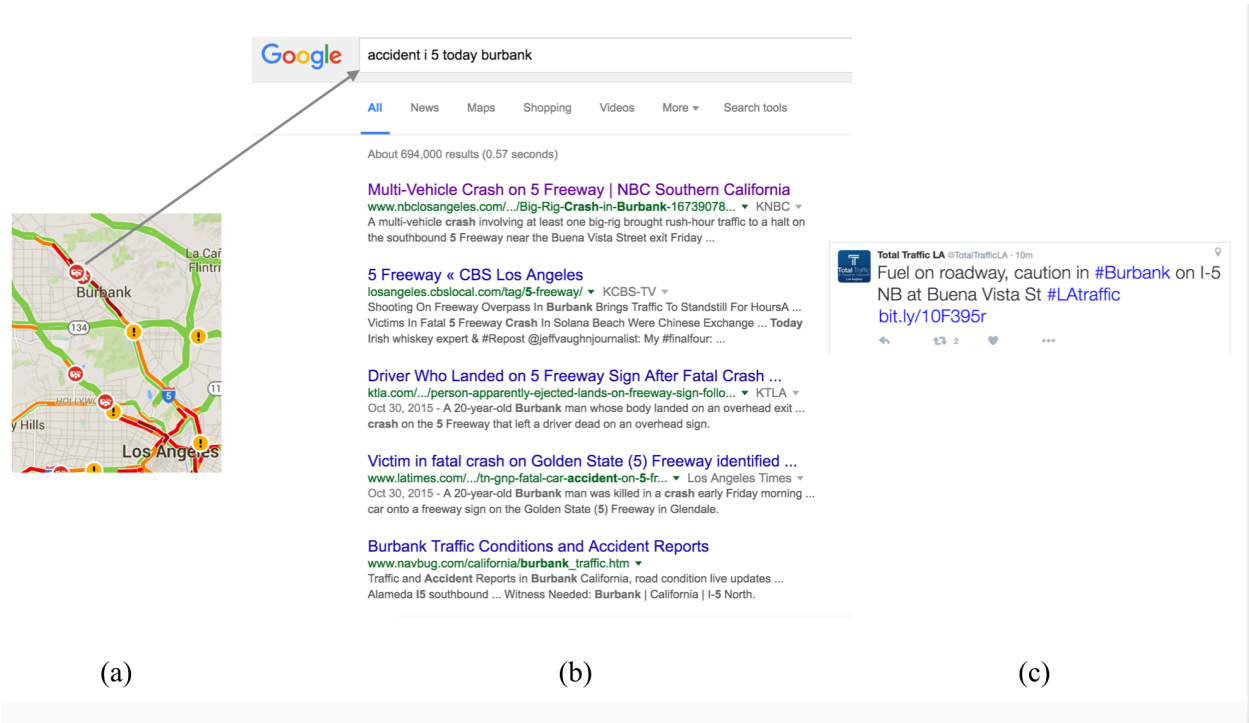

Figure 1.2. Traffic Query Example

the search box on social media platforms such as Twitter does not always result in relevant posts. For the traffic incident, a simple Twitter search query resulted in tweets ranging from several days to several weeks prior to the incident. The geotopical clustering component of GeoContext is able to uncover these types of topics.

1.4. Dissertation Structure

In this chapter, we introduced the problem of modeling social media posts topically and geographically and described the challenges involved with this problem. The improvements that could be made to existing research methods for this problem have also been introduced. We discussed the overall motivations for modeling social media.

The remainder of this dissertation is organized as follows: first, in Chapter 2, we describe background material that introduces key concepts in social media analysis, as well as an

introduction to various social media platforms. In this chapter, we discuss the reason for our choice of using Twitter as our primary research platform. In Chapter 3, we describe an overview of our research algorithm and the process for initializing GeoContext. Chapter 4 outlines the geolocation module, and Chapter 5 describes the geotopical clustering portion of GeoContext. Chapter 6 discusses our evaluation of GeoContext, and Chapter 7 concludes the dissertation.

CHAPTER 2

CONCEPTS AND SOCIAL PLATFORMS

In this chapter, we provide background for social media analysis and introduce some key terms and concepts involved in discovering geographical topics within social media.

2.1. Concepts in Social Media Analysis

Kaplan and Haenlein (Kaplan & Haenlein, 2010) define *social media* as the cross-section of two concepts: Web 2.0 and User Generated Content. They state that Web 2.0 refers to a platform where many users at a time modify content in a collaborative fashion. User Generated Content describes the "various forms of media content that are publicly available and created by end users." With these two Internet terms, the authors define social media as:

> *a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content.*

Some examples of social media platforms are Twitter, Facebook[9], Instagram[10], and YouTube[11]. Given this definition of social media, we can then describe *social media analysis* as the application of algorithms to a social media platform in order to discover insights. As discussed in Chapter 1, we focused on two main areas of analysis in our research: topic analysis and geographical analysis.

[9] http://facebook.com
[10] http://instagram.com
[11] http://youtube.com

2.1.1. Topic Analysis

The first aspect of our research is topical analysis. Given that a document is a single piece of content generated by a user, such as one tweet or one news article, Zhao et al. (Zhao, et al., 2011) defined a *topic* as:

> *a subject discussed in one or more documents. Examples of topics include news events such as "the Haiti earthquake," entities such as "Michael Jackson" and long-standing subjects such as "global warming." Each topic is assumed to be represented by a multinomial distribution of words.*

Given this definition of a topic, *topic detection* is the process of extracting topics from a set of documents, which are some set of terms grouped together. Topic detection can be accomplished in several ways, including document-pivot methods, feature-pivot methods, and probabilistic topic modeling. Feature-pivot methods "group together terms according to their co-occurrence patterns." Probabilistic topic models "treat the problem of topic detection as a probabilistic inference problem." (Petkos, Papadopoulos, Aiello, Skraba, & Kompatsiaris, 2014) LDA is one such example of a topic model that we discuss in this dissertation. Blei et al. (Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003) state that the goal of *topic modeling* is:

> *to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevant judgements.*

Document-pivot methods "group together individual documents according to their similarity" (Petkos, Papadopoulos, Aiello, Skraba, & Kompatsiaris, 2014). In our research method, we

implemented a type of document-pivot method where each document is represented by an individual social media post.

2.1.2. Geographical Analysis

The second aspect of our research is geographical analysis. *Geographical analysis*, or *spatial analysis*, is:

> *the process of examining the locations, attributes, and relationships of features in spatial data through overlay and other analytical techniques in order to address a question or gain useful knowledge.* (ESRI GIS Dictionary)

Many social media platforms, such as Twitter, allow users to incorporate a geographical component to their social media posts through geotagging. A g*eotagged post* refers to a social media post that contains a user's fine-grained latitude and longitude information in its metadata. For tweets that do not contain a geotag, they can go through the process of *geolocation*, which refers to "the process or technique of identifying the geographical location of a person or device by means of digital information processed via the Internet" (Oxford Dictionary).

Due to this inherent geographical aspect included with many social networks, performing some sort of location analysis can provide unique insights. In our implementation, we include a geographical analysis module that uses a unique approach for associating topics with geographical regions.

2.1.3. Geotopical Analysis

Lastly, our research method aims to tie both topic discovery and geographical analysis together in order to extract topics from social media that are centered within a geographical

region. We call this combination of topical and geographical discovery *geotopical analysis* or *geotopical clustering.*

Geotopical analysis adds a location-enabled component to both the individual documents in the topic model, as well as each topic. Yin et al. (Yin, Cao, Han, Zhai, Chengxiang, & Huang, 2011) define a *GPS-associated document* as "a text document associated with a GPS location." They also define a *geographical topic* as "a spatially coherent meaningful theme. In other words, the words that are often close in space are clustered in a topic."

Combining topical and geographical analysis can be accomplished in several ways within social media. Many approaches extended LDA to include a geographical component (Yin, Cao, Han, Zhai, Chengxiang, & Huang, 2011), (Wang, Wang, Xie, & Ma, 2007), (Zhang, Sun, & Zhuge, 2015). In our implementation, we chose to use an adapted approach from the natural language processing community, Term-frequency inverse-document-frequency (TF-IDF) (Sparck Jones, 1972), to associate topics to geographic locations.

2.2. Overview of Latent Dirichlet Allocation

As mentioned previously, LDA is one of the most commonly used topic modeling algorithms. We performed several evaluations, described in Chapter 6, where we compared the results from GeoContext against LDA's results. For that reason, we describe LDA in more detail in this section.

LDA takes as input a corpus, which is a set of documents that contain words not considered by LDA to be in any sorted order. In the algorithm, each document can be a mixture of various topics. The distribution of topics among the documents is assumed to have a Dirichlet prior distribution.

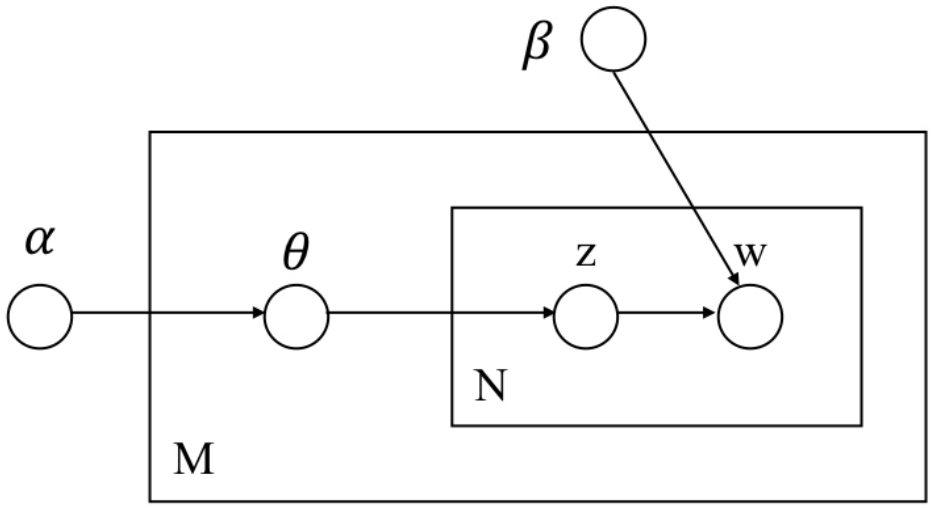

Figure 2.1. Latent Dirichlet Allocation **(adapted from** *(Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003)***)**

LDA produces topics that consist of terms that often appear together in the text of the documents. The number of produced topics is fixed prior to execution of the algorithm. Because LDA assumes that each document is a bag-of-words, where terms within the documents are unordered, all terms are initially treated equally.

LDA uses a sampling method to calculate both the topic distribution over documents and the topic distribution over the terms. A plate notation representation is shown in Figure 2.1. The outer box, denoted M, represents the number of documents, and the inner box, denoted N, represents the number of words within a document. w represents a specific word within a document. z is the topic of the word within the document, and θ represents the topic distribution for a document. The topic distribution is assumed to have a Dirichlet prior, and α and β are the parameters of the Dirichlet prior.

LDA produces two types of output:

1) *Topic distribution over documents*: the mixture of topics found in each document. LDA calculates the percentage of each topic that the document contains.

2) *Topic distribution over words*: the mixture of words within each topic. LDA calculates the percentage each word contributes semantically to each topic.

2.3. Analysis on Various Social Media Platforms

Various forms of social media analysis have been performed on different platforms. Although every platform is distinctive and has a varying set of features, each one has the concept of users and posts. Blogs, which are a somewhat unique type of social media platform due to the generally long length of posts, have been studied within the realm of topic mining (Mei, Liu, & Su, 2006) as well as blog network analysis (Leskovec, McGlohon, Faloutsos, Glance, & Hurst, 2007). Some methods have also applied topic modeling and geographical analysis to online photo collection platforms such as Flickr[12] (Yin, Cao, Han, Zhai, Chengxiang, & Huang, 2011) (Kling, Kunegis, Sizov, & Staab, 2014). Lastly, many approaches have used platforms such as Facebook and Twitter to detect topics and events (Backstrom, Sun, & Marlow, 2010).

2.4. Choice of Twitter in GeoContext

Like most related research (e.g., (Backstrom, Sun, & Marlow, 2010) (Han, Cook, & Baldwin, 2014)) (Cheng, Caverlee, & Lee, 2010), we use Twitter as our platform of choice for analysis. Twitter has several unique features. First, a post on Twitter is referred to as a *tweet.* The user that created and posted the tweet is called the *tweet author.* Authors each have a unique *Twitter*

[12] http://flickr.com

handle, or username. Tweets have a main content as well as many metadata fields that provide more information about the individual tweet and author. An example tweet is shown in Figure 2.1.

We chose to use Twitter specifically for GeoContext for several reasons. First, users on Twitter are limited to short posts of 140 characters. Due to the short size limitations, posts are often limited to one topic, which makes it ideal for clustering tweets into topics. Twitter is often utilized to quickly post news-type updates (Java, Song, Finin, & Tseng, 2007), which allows GeoContext to discover important, relevant information that is intended for a wide audience.

Also, Twitter allows users to connect locations to posts in two ways. Users can attach a geotag directly to the post. This is shown in Figure 2.1 in the "coordinates" field, and includes the exact latitude and longitude of the user with the tweet. Alternatively, Twitter users can tag tweets using Twitter Places[13], which allows a user to tag a post with the name of a location. This tag also includes a bounding box of geographical coordinates around the location. This is shown in Figure 2.1 as the "place" field. Within this field, the "name" field references the name of the place, which is usually a city, state, or major landmark. The "bounding box" field contains the points that form a polygon around the place.

In addition to location information, Twitter allows users to establish relationships with other users using the concept of *following*. If User A follows User B, User A can receive User B's tweets on their Twitter home page. Twitter describes users who User A follows as *friends* and users who follow User A as *followers*. Figure 2.1 shows the "friends_count" and "followers_count" of the example tweet. It is possible to obtain a list of the usernames or Twitter handles of the user's friends or followers by querying the Twitter APIs.

[13] http://dev.twitter.com/overview/api/places

```
{
    “created_at”: “Wed Jan 18 18:49:56 +0000 2017”,
    “id”: 821791752734572500,
    “id_str”: “821791752734572546”,
    “text”: “Yo! I just gotta take a sec to brag about my best
friend. Not only… https://t.co/qAuXIpJZ9t”,
    “source”: “<a href=\”http://instagram.com\”
rel=\”nofollow\”>Instagram</a>“,
    “truncated”: false,
    “in_reply_to_status_id”: null,
    “in_reply_to_status_id_str”: null,
    “in_reply_to_user_id”: null,
    “in_reply_to_user_id_str”: null,
    “in_reply_to_screen_name”: null,
    “user”: {
        “id”: 504937531,
        “id_str”: “504937531”,
        “name”: “Jawwdy”,
        “screen_name”: “jordynnn0720”,
        “location”: null,
        “url”: null,
        “description”: “Probably drinking coffee and planning
my next adventure  Future SLP ”,
        “protected”: false,
        “verified”: false,
        “followers_count”: 457,
        “friends_count”: 524,
        “listed_count”: 6,
        “favourites_count”: 8428,
        “statuses_count”: 13087,
        “created_at”: “Sun Feb 26 20:14:32 +0000 2012”,
        “utc_offset”: -14400,
        “time_zone”: “Atlantic Time (Canada)”,
        “geo_enabled”: true,
        “lang”: “en”,
 “following”: null,
        “follow_request_sent”: null,
        “notifications”: null
    },
    “geo”: {
        “type”: “Point”,
```


Figure 2.2. Example Tweet Metadata

```
        “coordinates”: [
            28.33773485,
            -81.55627962
        ]
    },
“coordinates”: {
        “type”: “Point”,
        “coordinates”: [
            -81.55627962,
            28.33773485
        ]
    },
    “place”: {
        “id”: “4ec01c9dbc693497”,
        “url”:
“https://api.twitter.com/1.1/geo/id/4ec01c9dbc693497.json”,
        “place_type”: “admin”,

“name”: “Florida”,
        “full_name”: “Florida, USA”,
        “country_code”: “US”,
“country”: “United States”,
        “bounding_box”: {
            “type”: “Polygon”,
            “coordinates”: [
                [
                    [
                        -87.634643,
                        24.396308
                    ],
                    [
                        -87.634643,
                        31.001056
                    ],
                    [
                        -79.974307,
                        31.001056
                    ],
                    [
                        -79.974307,
                        24.396308
                    ]
                ]
```

Figure 2.2. (cont.) Example Tweet Metadata

```
]
        },
        “attributes”: {}
    },
    “contributors”: null,
    “is_quote_status”: false,
    “retweet_count”: 0,
    “favorite_count”: 0,
    “entities”: {
        “hashtags”: [],
        “urls”: [
            {
                “url”: “https://t.co/qAuXIpJZ9t”,
                “expanded_url”:
“https://www.instagram.com/p/BPapfGaAJB60ENqRr1PE2-
2stAtwhHCdjsiQRw0/”,
                “display_url”:
“instagram.com/p/BPapfGaAJB60…”,
                “indices”: [
                    68,
                    91
                ]
            }
        ],
        “user_mentions”: [],
        “symbols”: []
    },
    “favorited”: false,
    “retweeted”: false,
    “possibly_sensitive”: false,
    “filter_level”: “low”,
    “lang”: “en”,
    “timestamp_ms”: “1484765396661”
}
```


Figure 2.2. (cont.) Example Tweet Metadata

Twitter provides two sets of APIs that we utilized to perform our analysis. The first is the Twitter Streaming API[14]. The Streaming API provides API users with limited access to the public data streaming through Twitter. The Streaming API has two types: "Firehose," which contains all public Twitter status updates, and "Gardenhose," which provides a small percentage of all public tweets. It is estimated that the "Gardenhose" stream includes 15% of the public Twitter stream (Eisenstein, O'Connor, Smith, & Xing, 2010). Because we do not have access to the "Firehose" stream due to cost constraints, we utilize the "Gardenhose" stream for our implementation. Because the "Gardenhose" stream returns a random sample of all public tweets, we believe that it is still representative of the entire Twitter ecosystem, and the results we obtained will hold for all tweets.

Twitter also provides a REST API, which allows API users to programmatically read and write Twitter data. In our method, we utilized the REST API to obtain lists of a user's friends and followers.

Although we chose to utilize Twitter for GeoContext, it can be adapted easily to use other social networks that have the same characteristics, such as the ability to geotag posts and establish relationships with other users. GeoContext simply processes JSON objects from a social media stream that has content and location information, so any other social network or information provider that attaches geographical coordinates to shared information could be used.

[14] https://dev.twitter.com/streaming/overview

CHAPTER 3

INTRODUCTION TO GEOCONTEXT AND INITIALIZATION

This chapter begins by describing the implementation of GeoContext, our system for discovering geographical topics in social media. We first present an overview of our entire process created to cluster the Twitter stream into topics and perform geographical analysis in Section 3.1. In Section 3.2, we explain how GeoContext is initialized, as well as the method used by GeoContext for filtering the stream by keyword and location.

3.1. Overview of GeoContext

Figures 3.1 and 3.2 show the two pipelines implemented for our research. An example tweet as it moves through the various stages of analysis is shown on the right side of both figures. As described previously in Section 2.1, there are three main steps in the implementation: geolocation, topical clustering, and geographical analysis. We developed two pipelines to determine whether our results varied if topical clustering was performed first, followed by geographical analysis, or geographical analysis was performed first, followed by topical clustering. The two pipelines begin in the same fashion. Results on the differences between the two pipelines are examined in Chapter 6.

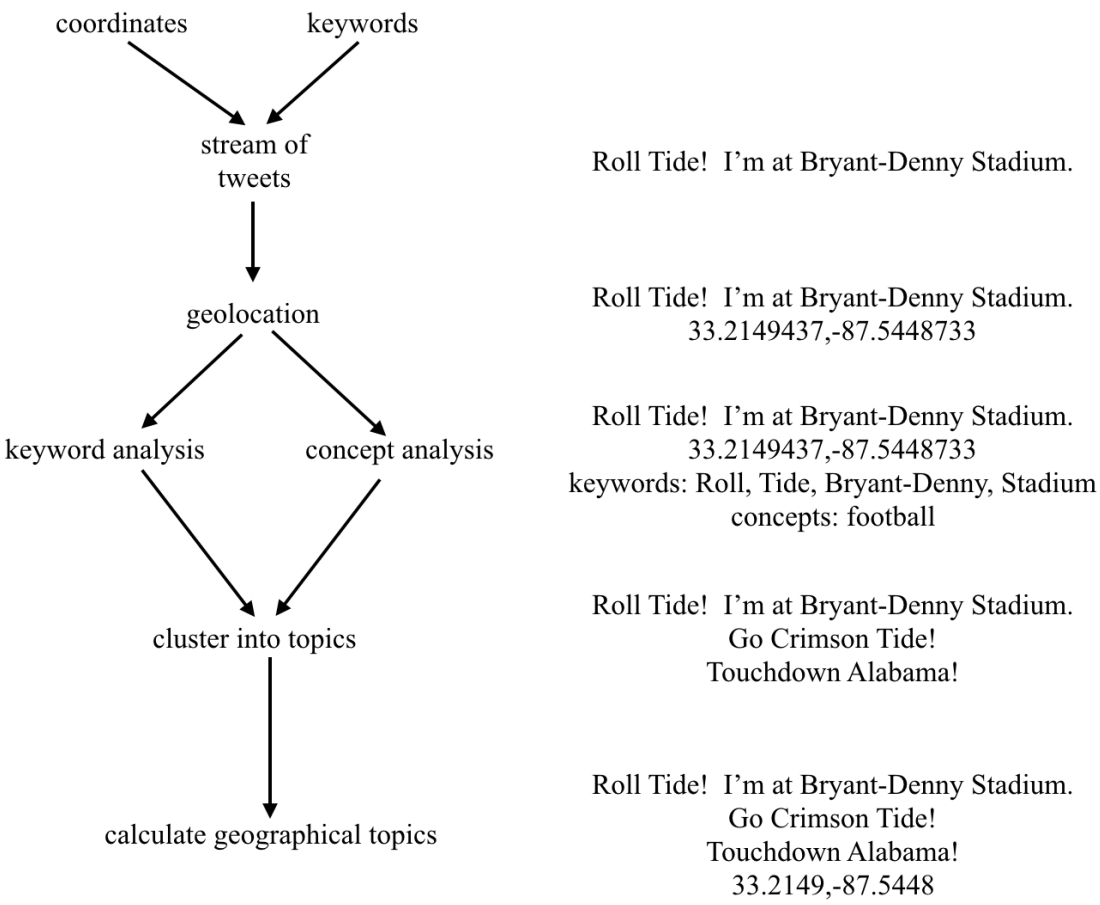

Figure 3.1. Topical Clustering First Pipeline

Both pipelines begin by receiving a stream of tweets as input. The content of an example tweet is shown on the right side of both Figures 3.1 and 3.2 when the stream is first initialized.

The stream can be filtered in two ways if desired. Users of GeoContext have the ability to input location (given in GPS coordinates) or keywords in order to reduce the size of the stream of tweets and reveal topics that are more specific to certain situations. The Twitter Streaming API furnishes the ability to filter the given stream by a set of coordinates or a set of keywords.

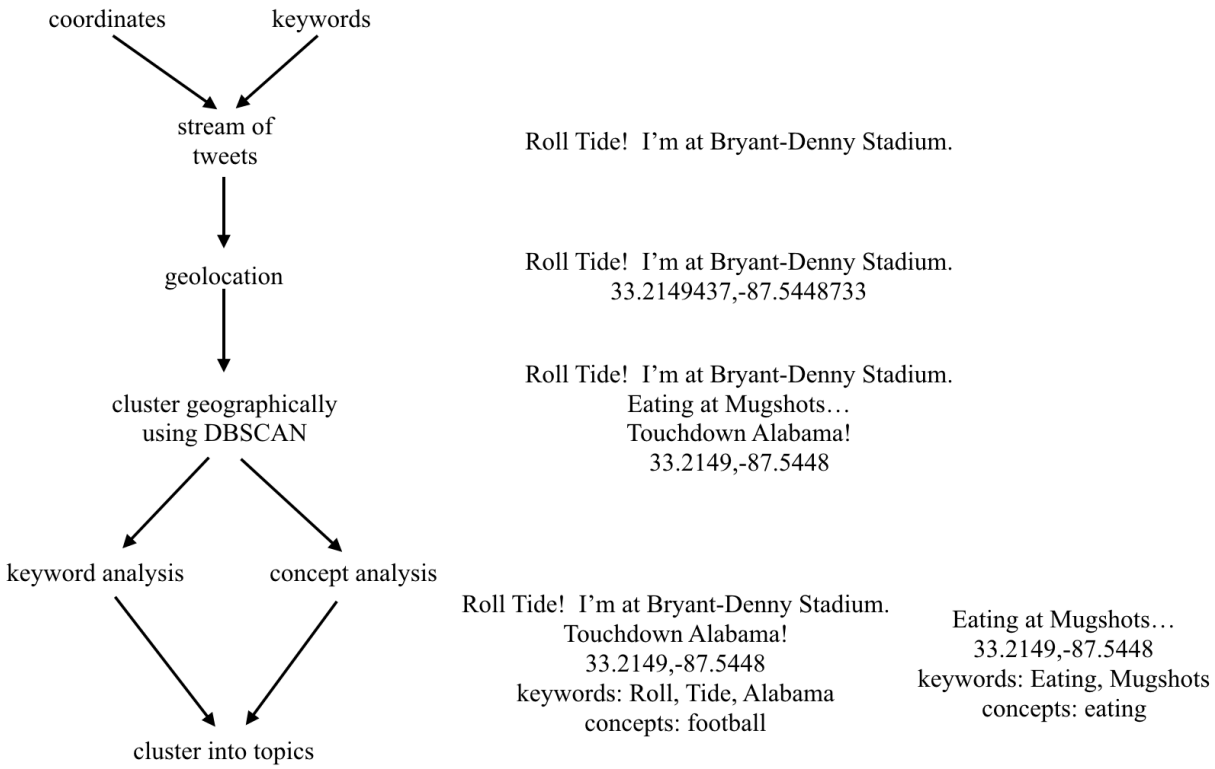

Figure 3.2. Geographical Clustering First Pipeline

If geographical coordinates are given, GeoContext uses the Twitter Streaming API to initiate a stream of tweets that is filtered by the given coordinates. For example, a stream can contain only tweets from a specific city or region so that tweets from one area can be examined in more detail. The Twitter Streaming API will only provide tweets that have been geotagged from this area, so it is possible that non-geotagged tweets have been posted from the specified region that are not returned in the stream.

If keywords are given, GeoContext adds functionality to the Twitter Streaming API by expanding the keywords into a set of all related keywords. Keyword expansion is discussed in Section 3.3. The stream of tweets is then initiated with a filter of the set of keywords. This allows a user to examine only tweets that contain the words "weather," "thunderstorm," and

"hurricane" if desired. If no keywords or coordinates are given, a stream of tweets is initiated through the Twitter Streaming API that returns a sample of all public tweets with no filters or parameters, and the stream is then passed through the geotopical clustering pipeline.

In addition to the optional keyword and geographical coordinates filtering, we chose to filter the stream by English-language tweets only in all of our analyses, because some of the analyses we performed on the tweets in order to achieve clustering are currently available only for English text.

After initializing the stream of tweets, the geolocation step in both pipelines is performed via the GeoContext Locator module. This step uses the text and metadata of tweets to predict locations for any tweets that are not geotagged (i.e., do not have explicit locations attached). At the end of this step, all tweets have geographical coordinates associated with them, as shown in Figures 3.1 and 3.2. The geolocation process is described in Chapter 4.

After the geolocation step, the two pipelines differ. The first pipeline is shown in Figure 3.1. In this pipeline, GeoContext takes the stream of tweets and extracts keywords and concepts from the content of the tweets. Based on the keywords and concepts, it then clusters the tweets along with other related tweets into topics. Lastly, geographical analysis is performed such that the algorithm determines which topic clusters of tweets should be associated to different geographical locations. The geographical analysis is performed using our novel adapted TF-IDF method. We call this pipeline the *topical clustering first pipeline*, and details are described in Section 5.1.

The second pipeline is shown in Figure 3.2. In this pipeline, GeoContext first clusters the stream of tweets geographically by using DBSCAN, a commonly used density-based clustering algorithm. This geographical analysis results in finding clusters of tweets centered in different

locations, as shown in Figure 3.2. This version of the pipeline then takes each individual geographical cluster and separates the tweets within the cluster into different topics. As mentioned in Chapter 1, it is possible that by clustering tweets geographically, this method will result only in geographical clusters of tweets where population is high, such as large cities. However, we wanted to test this hypothesis and compare results from both pipelines to see how much they differed. For that reason, we believe that this pipeline is still a worthwhile endeavor to study. We call this pipeline the *geographical clustering first pipeline*, and details are described in Section 5.2.

3.2. Tools Used to Create GeoContext

To create both pipelines within GeoContext, we used Node.js[15], which is a runtime environment for developing JavaScript web applications. GeoContext needs to connect to several analysis resources, most via REST calls, such as Dbpedia, AlchemyAPI, and the Google Maps API. Multiple calls are needed per tweet in order to geolocate the tweet and perform topical and geographical analysis on the content and metadata of the tweet. Because of the large number of tweets coming through the stream, GeoContext needs the tweets to be processed synchronously in order to maintain near real-time results. For this reason, GeoContext needs to have multiple open HTTP connections at a time. We chose Node.js for our implementation because it has a simple interface for creating HTTP requests and can handle many concurrent connections. On the server side, we utilized the Twit framework[16] for Node.js to initiate a stream of tweets. Twit allows Node.js integration for the Twitter Streaming API, which provides a sample of tweets through an open HTTP connection.

[15] https://nodejs.org/
[16] https://www.npmjs.com/package/twit

```
Elizabeths-MacBook-Pro-6:SocialMedia elizabethwilliams$ node index.js keywords:w
eather,traffic loc:33.2137452,-87.5415876
```

Figure 3.3. Starting GeoContext

GeoContext is set up as three modules: geolocation, topical clustering, and geographical analysis. Output can be displayed and analyzed at any point between the three modules so that each can be evaluated separately.

3.3. Initialization and Keyword and Location Filtering

GeoContext can be started using the command *node*, as shown in Figure 3.3. At initialization time, keywords and geographical coordinates can be accepted as input in order to filter the stream to find topics more specific to different scenarios. This is useful, for example, if a user is interested in getting updated information about a certain topic such as weather. As described by Sakaki et al. (Sakaki, Okazaki, & Matsuo, 2013), users can sometimes tweet about major weather events before official sources can report the events. Any provided keywords are sent through keyword expansion.

The Twitter Streaming API allows filtering by keyword phrases. Any number of keywords can be used, and the API will return tweets that match any of the keyword phrases present. If a keyword phrase contains more than one term separated by spaces, the Twitter API will match tweets containing all of the terms in the phrase, even if the terms are not in order. Punctuation and case are ignored in the tweet matching a keyword phrase.

We are interested in concept matching for GeoContext, rather than the more traditional keyword matching implemented by the Twitter Streaming API. Concept matching allows a user

to input a keyword or keyword phrase, such as *weather*, and GeoContext will track tweets that not only contain the specific word *weather*, but also tweets relevant to the concept *weather*. This might include tweets that contain the terms *rain*, *thunderstorms*, or *hail*.

To implement concept matching, GeoContext expands any keywords the user has provided. We realize this keyword expansion by utilizing cognitive computing techniques. For each keyword in the comma-separated input list, we pass the keyword to the JoBimText distributional semantics framework, described by Biemann and Riedl (Biemann & Riedl, 2013). The JoBimText framework uses a corpus of text such as Wikipedia and analyzes the structure of the text through methods such as a dependency parser. JoBimText extracts pairs of terms from the corpus, a word and another term that describes the context of the word. After obtaining a count throughout the corpus of each pair of terms, which are denoted a Jo (the word) and a Bim (the contextual term), a frequency significance measure is calculated for each unique pair of terms based on whether they frequently are associated with the same words. Terms with the highest significance measure are clustered into sense clusters, such that each word has a sense cluster containing terms that are conceptually related to the word. For example, the word *weather* might have a sense cluster containing the terms *rain*, *thunderstorm*, and *wind*. These clusters can be used to find other words that are similar to a term.

For keyword expansion, we used JoBimText's similarity score feature to calculate terms that are related to each keyword provided to GeoContext. After observation, we consider terms that have a score of at least 70 because terms below the threshold of 70 are less conceptually related. We then pass each similar term, as well as the original keyword term, to the Twitter Streaming API to track.

GeoContext also accepts geographical coordinates as input to allow a user to provide a specific location. GeoContext will pass the coordinates directly to the Twitter stream and begin to receive tweets that are located around the coordinates. Twitter supplies both tweets that are geotagged and tweets that have been tagged with Places whose bounding boxes intersect with the coordinates given. The location query system is useful for discovering topics of tweets that are being discussed in a specific geographical region. For example, events occurring around a city can be discovered using the location query system. In one evaluation discussed in Chapter 6, GeoContext discovered events occurring in two different locations: Tuscaloosa, AL, and New York City, NY.

After initializing the stream with any of the possible parameters, GeoContext sets up a persistent HTTP connection using Twit and begins to receive tweets in the stream. A tweet received from the Twitter stream is returned as a JSON object, as shown in Figure 2.1. As described in Chapter 2, the tweet object consists of the actual content of the tweet and metadata about the tweet, such as geographical coordinates if the tweet is geotagged and a timestamp. The object also contains metadata about the author of the tweet, such as the username, account location, account description, and more. Tweets that are "retweets" (i.e., tweets that are re-published from other users) contain extra metadata. In our approach, we do not consider retweets different from non-retweets; thus, we do not analyze the additional details of retweet metadata.

Once tweets are received, they are then passed through the geolocation step, described in Chapter 4, where GeoContext predicts the location of tweets that do not contain a geotag. In Chapter 5, we describe the geotopical clustering step in the pipeline. In this step, GeoContext

clusters together tweets into topics that represent individual trending concepts. GeoContext then produces resulting topic clusters that can be recommended for certain locations.

CHAPTER 4

GEOLOCATION OF TWEETS

Geolocation, or the prediction of the location of a tweet, is an important step for any geographical analysis performed on social media platforms. Although some social media platforms provide users a way to geotag posts, many users choose not to furnish their location due to privacy concerns.

The current standard for geolocation for many applications is IP-based geolocation, which is fairly accurate at city-level granularity. However, because Internet Service Providers often pool IP-addresses, and IP-addresses are constantly reassigned to users, IP-based geolocation is not always completely exact (Backstrom, Sun, & Marlow, 2010). For this reason, we implemented our own geolocation module in order to predict the locations of Twitter users.

Once the initialization is completed with any needed filters, as described in Chapter 3, the stream of tweets passes through the geolocation step. In order to analyze the location at which different topics in the stream appear, GeoContext needs the location of each individual tweet. As mentioned in Section 1.3, as few as 0.87% of tweets are geotagged, or associated explicitly with geographical coordinates (Jaiswal, Peng, & Sun, 2013). These low statistics indicate that if we only used tweets that are geotagged or tagged with Places, GeoContext would not have enough tweets to glean intelligent topics from the stream. Therefore, we need to mine locations from additional tweets that were not geotagged. To solve this problem, GeoContext includes GeoContext Locator (GCL), which predicts the location of the tweet's author at the time the tweet was posted using the content and metadata of the tweet.

For our approach to geolocation, GCL predicts the current location of a tweet, which corresponds to the current location of the user at the timestamp of the tweet. This differs from several of the existing research approaches that predict the user's home location (i.e., where the user resides) such as (Han, Cook, & Baldwin, 2014) and (Li, Wang, Deng, Wang, & Chang, 2012). GCL is useful for geolocation on a mobile device where the user's current location is often more important than the user's home location. Therefore, if a user tweets while on a trip or out of the region of their home location, GCL will predict the location the tweet was sent, not the location where the user normally resides.

An important detail regarding tweet objects is that although few tweets are geotagged and contain location information attached to the tweet itself, many more users provide account locations, which are locations attached to the user account rather than a specific tweet. Account locations can be freeform text, so they do not follow any convention in terms of representing location, which makes geocoding these locations difficult. Moreover, because account locations can be any text, some users provide a location that is not a true location, such as "Cloud 9." Users also have the option of leaving the account location blank. The three explicit metadata fields and one implicit field that GCL uses for geolocation are displayed in Table 4.1.

Unlike most existing research, GCL combines the content of the tweet, the user account location, relationship graph (i.e., the set of the user's friends and followers), and the extracted topic of the tweet to predict the tweet's location. Tweets collected from the stream are processed through the GCL pipeline, shown in Figure 4.1. In each step, GCL extracts location information from various aspects of the tweet object. GCL then attempts to geocode the raw location

Table 4.1. Tweet Fields Used For Geolocation

User account location	Tuscaloosa, Alabama
Tweet content	Roll Tide! I'm at Bryant-Denny Stadium.
Friends and followers	Tuscaloosa, AL; Nashville; I'm in Northport; Tuscaloosa, y'all
Topic	Alabama football

information, or convert it to geographical coordinates. Because different analysis techniques extract different types of location information better than others, for each field within the tweet, GCL utilizes multiple techniques in order to obtain the most geographical information possible. For example, we found that, when attempting to discover location words in the tweet content, AlchemyAPI and Dbpedia, which are two techniques described in the following subsections, are quite accurate when extracting large, well-known locations, such as cities, sports stadiums, or landmarks. However, these techniques are not able to extract location words such as local restaurants or stores. In order to improve the extraction of location information, we utilized the Google Places API as an additional resource.

For this reason, we combine several techniques for each field. Although some of the techniques can be redundant and result in the same location being extracted multiple times, none of the methods used by GCL are perfect in extracting location information, so multiple techniques are needed in order to provide more coverage and ensure locations are discovered in the text.

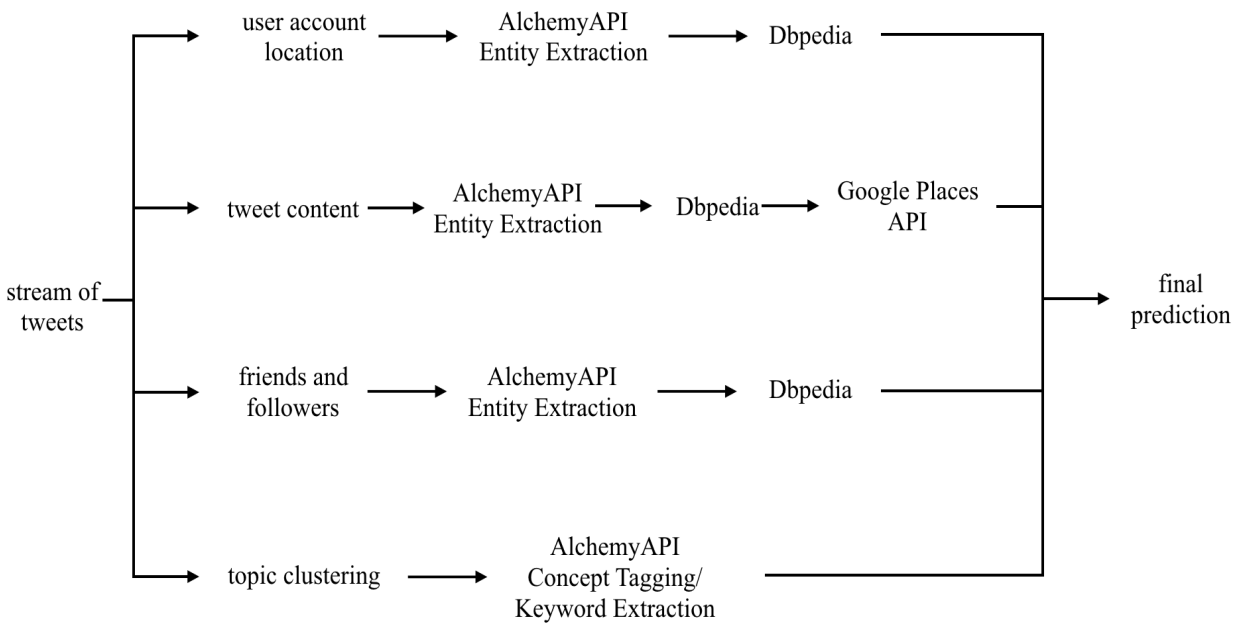

Figure 4.1. GCL Pipeline

Throughout the pipeline, GCL stores an object for each tweet consisting of all estimates of coordinates from each step. For each step, if the geocoding process in that step is successful, the resulting coordinates are stored in the list of predicted locations, along with the step from which they resulted (i.e., "user location" or "content"). At the end of the pipeline, the list of predictions is analyzed, and the most likely set of coordinates is chosen from the list to be the final location prediction for that tweet.

In the following subsections, we describe each step of the GCL pipeline for predicting tweet location.

4.1. Geolocation Using User Location

The first portion of the tweet that GCL checks for location information is the user account location. As described previously, the user account location is a location attached to a user's

account, rather than to each individual tweet. It is possible that a user is not near their home location (i.e., the user is on vacation), so we also utilize other techniques described in the subsequent sections to geolocate a tweet.

This field is shown as "location" within the "user" object in Figure 2.1. On a user's Twitter account, the user can input any freeform text for the user location, or the user has the option to leave the location blank. An example of the user account location is shown in Table 4.1. If the user location is blank, or null, GCL ignores this field and proceeds to the next step, because no location information can be extracted.

Because the user location is freeform text, GCL must be able to extract text that represents locations from the field. GCL performs this extraction in multiple steps using two different analysis techniques: AlchemyAPI and Dbpedia. AlchemyAPI is a set of cognitive computing APIs that include natural language processing techniques. As mentioned in Section 1.3.1, Dbpedia is a database consisting of structured information extracted from Wikipedia. This makes it ideal for discovering entities such as cities or states mentioned in text. Both techniques are explained in detail in the following paragraphs. An example of the user location extraction process is shown in Figure 4.2.

First, GCL utilizes AlchemyAPI's Entity Extraction API[17] to look for text that resembles locations. The Entity Extraction API receives a piece of text as input and returns a ranked list of named entities, such as people, organizations, or locations. As each tweet comes in through the stream, GCL passes the user account location associated with the tweet to the Entity Extraction API and receives back results indicating whether the account location contains any named entities.

[17] http://www.alchemyapi.com/api/entity/textc.html#rtext

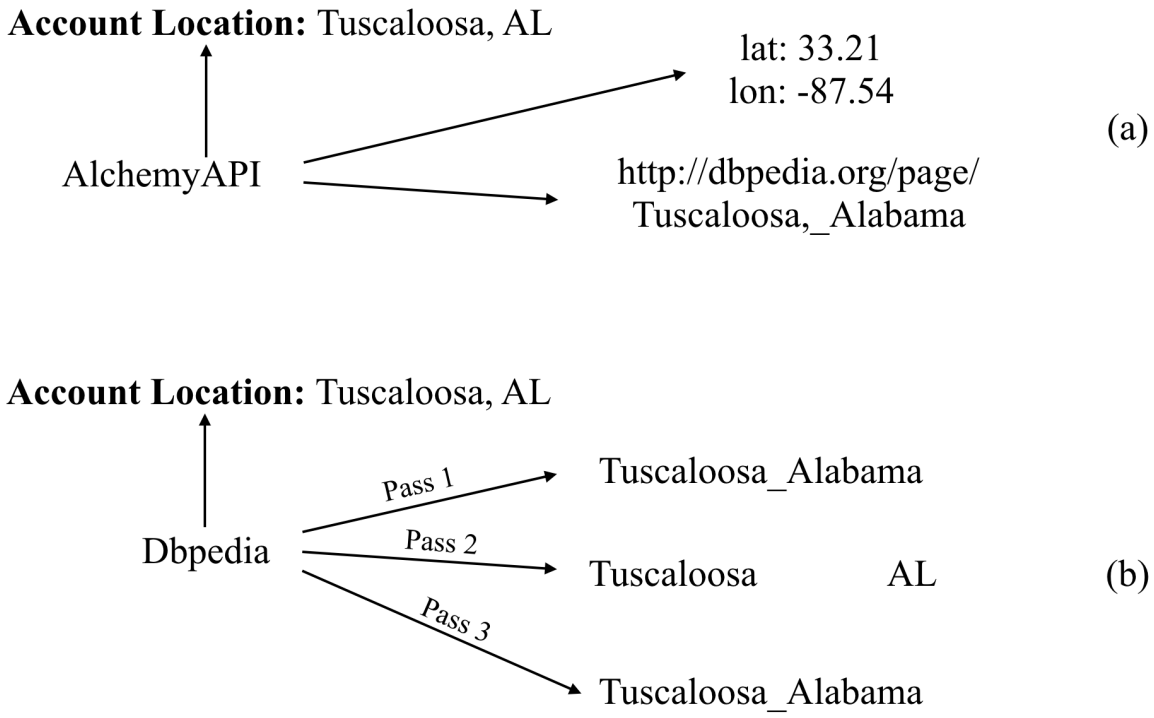

Figure 4.2. User Location Extraction

If results are received, this indicates that the Entity Extraction API has been able to extract some entity information. Each discovered entity is associated with a type, and not all entities found by the API are related to location. For example, the API extracts numerical values and Twitter mentions (i.e., a tweet that contains another user's username, also known as their Twitter handle). Neither of these entities is useful to GCL in determining locations at this time, so we only consider results for entities that contain location information. The types considered by GCL are "City," "Region," "Facility," "StateOrCounty," "Organization," "Company," and "GeographicFeature." "City," "Region," "StateOrCounty," and "GeographicFeature" are clearly types of entities related to the user's location. "Facility" is often related to locations such as

sports stadiums. "Organization" and "Company" can be entities such as sports teams or major corporations, which are often location-specific.

For each entity extracted from the user location whose type matches one of the considered types, GCL attempts to extract geographical coordinates from the results. As shown in Figure 4.2(a), some results simply contain geographical coordinates within the result. In this case, the coordinates are extracted and added to the tweet's list of predicted locations. Some results do not contain geographical coordinates, but contain a link to a Dbpedia entry of the entity. In this case, GCL retrieves the Dbpedia entry for the entity and extracts any coordinates found in the entry. We provide more detail about requesting and extracting coordinates from Dbpedia entries in the next paragraph. Lastly, some results do not contain any disambiguation information, but simply the entity text that was extracted from the user location and the type of the entity. In this case, GCL creates a URL to Dbpedia that is similar to the URLs received from results that contain links to Dbpedia entries by prepending "http://dbpedia.org/resource/" to the entity text. GCL then follows the same procedure as if the results contained a Dbpedia link by sending a request to the entry and extracting any coordinates the entry contains. An example is shown in Figure 4.2. After any possible geographical coordinates are extracted, the coordinates are added to the list of predicted locations. Next, GCL uses a second technique to extract more location data from the user account location field.

We determined through manual observation that the AlchemyAPI Entity Extraction API did not recognize 100% of locations in the stream of tweets. Therefore, we utilized Dbpedia directly as a technique for discovering additional locations in the user account location field.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?s
 WHERE {
  { ?s rdfs:label "Tuscaloosa"@en ; a owl:Thing . }
  UNION
  { ?altName rdfs:label "Tuscaloosa"@en ;
  dbo:wikiPageRedirects ?s . }
 }
```

Figure 4.3. Example SPARQL Query

We created a Dbpedia mirror that is utilized locally using Virtuoso, which is an open-source server for data management[18]. Queries are sent to Dbpedia via a SPARQL (SPARQL Protocol and RDF Query Language) endpoint. SPARQL is a semantic query language for databases that are stored in RDF (Resource Description Framework) format (SPARQL Query Language for RDF, 2017). GCL executes the user location through the Dbpedia extraction step in three passes, shown in Figure 4.2(b). In the first pass, the entire user account location field is used. Because an HTTP request uses some types of punctuation as special characters, punctuation within the tweet can affect the results of the request. For this reason, all punctuation is removed except commas. Spaces are also replaced with underscores.

GCL formulates a query and attempts to retrieve results from Dbpedia. An example SPARQL query is shown in Figure 4.3. If the request succeeds, this indicates that an entity exists within Dbpedia of the user location. Some Dbpedia entries are pointers to another entry; in this case, the entry is disambiguated to another. This occurs especially in cases such as names of cities where there may be more than one result due to multiple cities with the same name. For example, the entry for “Tuscaloosa” disambiguates to “Tuscaloosa,_Alabama.” If the results

[18] https://github.com/openlink/virtuoso-opensource

received from the request indicate that the entity disambiguates to another entry, GCL sends another request to the disambiguated entry's URL.

After results of an entry are received that do not disambiguate, GCL extracts any geographical coordinates that are available from the entry. Any entry that is a location (e.g., city, state, or region) contains coordinates. The coordinates are added to the list of location predictions stored with the tweet.

In the second and third pass, GCL performs the same process of sending a request to Dbpedia and extracting geographical coordinates, but with different text used as the entity. In the second pass, GCL removes all punctuation and splits the user account location text into tokens by spaces and sends each token as the request to Dbpedia. We chose to split the text by spaces, because this allows GCL to discover locations from text such as "Tuscaloosa, y'all," in which extra words are present in the user location that would result in a bad request in the first pass. In the third pass, punctuation is removed, the text is split by spaces, and then every two tokens are concatenated with a space in the middle. This ensures that user locations such as "My office, Beverly Hills" are analyzed properly. "Beverly Hills" can be extracted, while extra words such as "My office" are effectively ignored. We chose not to perform more passes where more than two tokens are concatenated at this time, as it did not seem to affect our results significantly because all fields in a tweet are kept relatively short.

After processing the user account location, GCL passes the tweet and the list of predicted location coordinates to the second step in the pipeline, which is focused on tweet content analysis.

4.2. Geolocation Using Tweet Content

The next step in the GCL pipeline extracts location data from the actual content of the tweet. Twitter limits the tweet text to 140 characters, so each tweet is relatively short. An example of tweet content is shown in Table 4.1. Similar to the user account location, Twitter allows the tweet content to be freeform, and users are able to use any punctuation or character they choose.

For the first part of this step, GCL follows the same procedure when analyzing tweet content as the user account location. First, GCL passes the content of the tweet to AlchemyAPI's Entity Extraction API. As described in Section 4.1, the Entity Extraction API discovers entities present in the text. If the entity is related to a location, GCL is able to geocode the entity name into geographical coordinates, either directly from the Entity Extraction results or indirectly through Dbpedia.

After utilizing the Entity Extraction API, the tweet content is converted to a Dbpedia SPARQL request, as shown in Figure 4.3, and the corresponding Dbpedia entry is requested. An example is shown in Figure 4.4(a). However, many, although not all, users put a valid location in the user account location field, while most words in the tweet content are not words related to location. Therefore, fewer requests sent from the tweet content analysis succeed than from the user account location analysis.

As with the user account location, several passes are performed over the content. Each pass is shown in Figure 4.4(b). First, the entire tweet with punctuation removed is requested as a Dbpedia entity. In the second pass, GCL removes punctuation and tokenizes the tweet content by splitting it by spaces. Each token is requested as a Dbpedia entity. In the third pass, GCL removes punctuation, splits the content text by space, and concatenates each two consecutive tokens together with a space in between.

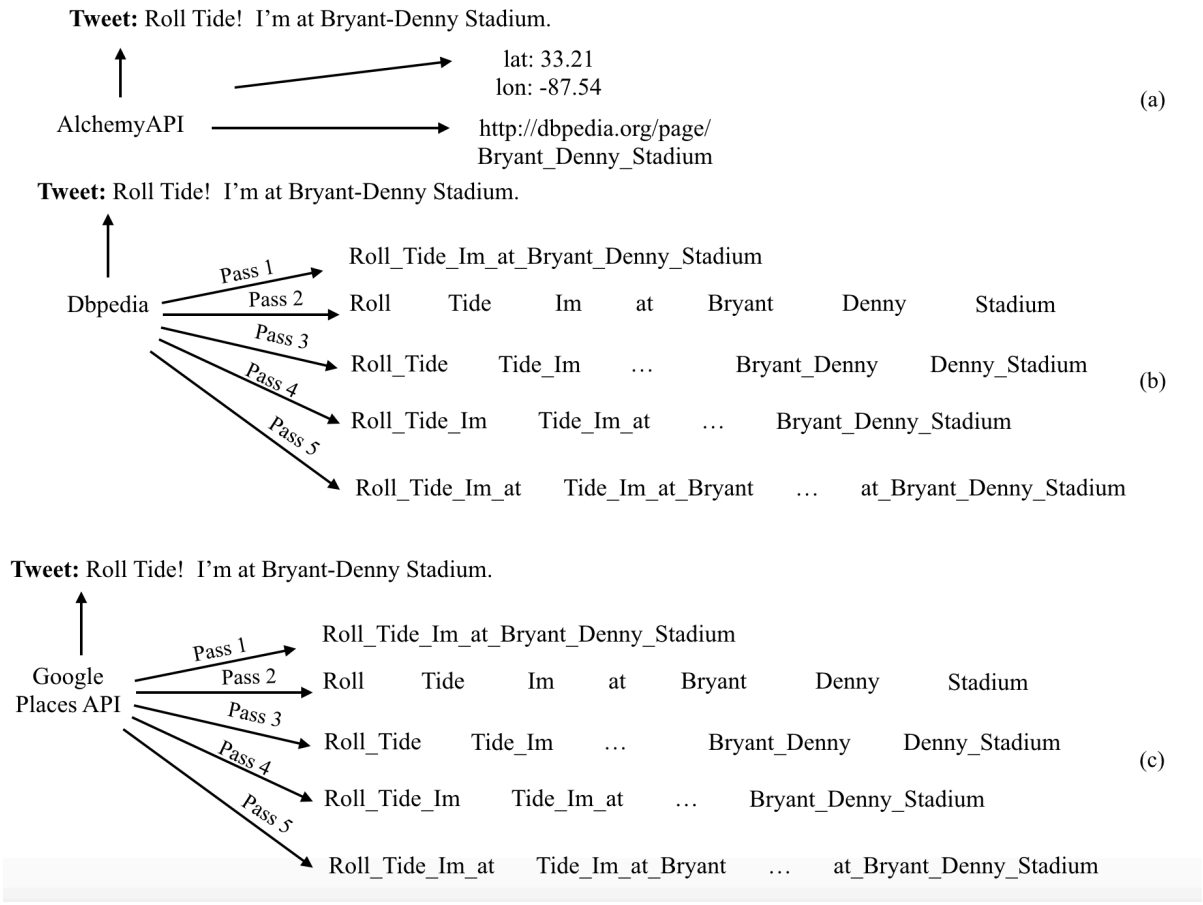

Figure 4.4. Content Extraction

Because the content of tweets is generally longer than the user account location, we decided to implement more passes over the content. In the fourth and fifth passes, GCL concatenates every three and four tokens together. Each token combination is requested as a Dbpedia entity. After each request, if the request is valid and the results contain geographical coordinates, the coordinates are added to the list of possible location predictions for the tweet.

AlchemyAPI and Dbpedia are both able to extract broad, well-known locations that would appear in Wikipedia, such as cities, state parks, and sports stadiums, among others. However, neither method is particularly effective at recognizing local venues such as restaurants or stores that are mentioned in the content of the tweet. As a last technique for extracting these additional

locations, GCL utilizes the Google Places API, part of the Google Maps API, to search for names of locations.

GCL uses the same method for the Google Places API as for Dbpedia. The process, which is very similar to the process for Dbpedia, is shown in Figure 4.4(c). The content of the tweet has punctuation removed and is split by spaces into tokens. Five passes are completed by GCL: in the first pass, the entire tweet text is sent in the request; in the second pass, individual tokens are sent; in the third pass, every two consecutive tokens are sent; in the fourth pass, every three consecutive tokens are sent; and in the fifth pass, every four consecutive tokens are sent. Each request is sent to the Google Places API and a JSON object of results is retrieved.

If a resulting location is found, then the results include latitude and longitude data, which is added to the list of predictions. The Google Maps API tends to err on the side of quantity over quality; it produces many results, although many of the produced results are not accurate. The tweet and list of predictions are then passed to the third step in the GCL pipeline: discovering location information through friends and followers.

4.3. Geolocation Using Friends and Followers

The third step in the GCL pipeline for predicting tweet location analyzes the friends and followers of the tweet author. Within the Twitter social media platform, users can follow each other, which means they may receive updates (depending on Twitter's home page selection algorithm) on their Twitter home screen when a followed user posts a tweet. A user's friends and followers form a relationship graph, which can be analyzed for geolocation purposes.

Some existing geolocation approaches (described in Section 4.6) utilize the relationship graph of a user to predict location. Like these approaches, we also analyze the relationship graph of each tweet author. McGee et al. (McGee, Caverlee, & Cheng, 2013) discovered peaks in the distribution of friends and followers around the area where the user lives. This result shows that although Twitter allows users from all over the world to communicate, users tend to have a collection of friends and followers near their same location. The friends and followers of a user are valuable for finding the user's home location, especially in the case where the user does not provide a user account location or the user account location is not an actual place. However, the friends and followers location is not always accurate, especially in the case where a user is on a trip or away from their home location. It is for this reason that we combine multiple techniques for predicting location. We describe GCL's process for determining the final predicted location in Section 4.5.

GCL is able to collect the friends and followers of each user through the REST APIs provided by Twitter. The REST APIs allow the lookup of friends and followers based on a Twitter ID. An object representing the friend or follower, respectively, is returned when requested via the API. The object contains metadata about the friend or follower, including username (also known as a user's Twitter handle), description, and user account location. Following existing approaches, we utilize the user account location for each friend or follower when calculating a predicted location.

After collecting the user account locations for friends and followers of the tweet author, we follow a two-step process that is similar to the user location and tweet content steps in the GCL pipeline. GCL first passes each user location of each friend and follower to the AlchemyAPI Entity Extraction API. If a result is received that indicates a location has been found in the user

account location, GCL parses the result and adds the location to a list of friends and followers locations. In the second step, GCL utilizes Dbpedia to extract location information. GCL removes punctuation from each user location of each friend and follower and then converts the full user location to a Dbpedia URL. GCL also splits the user locations by spaces and converts each token to a Dbpedia URL. Requests are then sent to each Dbpedia URL. Similar to the process described in Section 4.1, GCL collects any geographical coordinates contained by the Dbpedia entity received from the request result. These coordinates are also added to the list of friends and followers locations.

After coordinates are extracted from all friends and followers user locations, GCL needs to analyze the list of coordinates to determine a reasonable prediction for the location of the original tweet. Existing approaches (Backstrom, Sun, & Marlow, 2010) (Li, Wang, Deng, Wang, & Chang, 2012) (McGee, Caverlee, & Cheng, 2013) have mainly used probabilistic models to determine which friends and followers might be in the same geographical region as the user. We employ a different technique for analysis of the friends and followers location. GCL takes the entire list of friends and followers' locations from both steps (i.e., the AlchemyAPI step and the Dbpedia step) and clusters all locations using DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), in order to discover where the majority of the user's friends and followers are located. We chose to use the density-clustering package for Node.js[19] to implement the DBSCAN clustering step.

There are two parameters to the DBSCAN algorithm: the minimum number of points to form a cluster and the cluster radius. We chose 2 for the minimum number of points because some

[19] https://www.npmjs.com/package/density-clustering

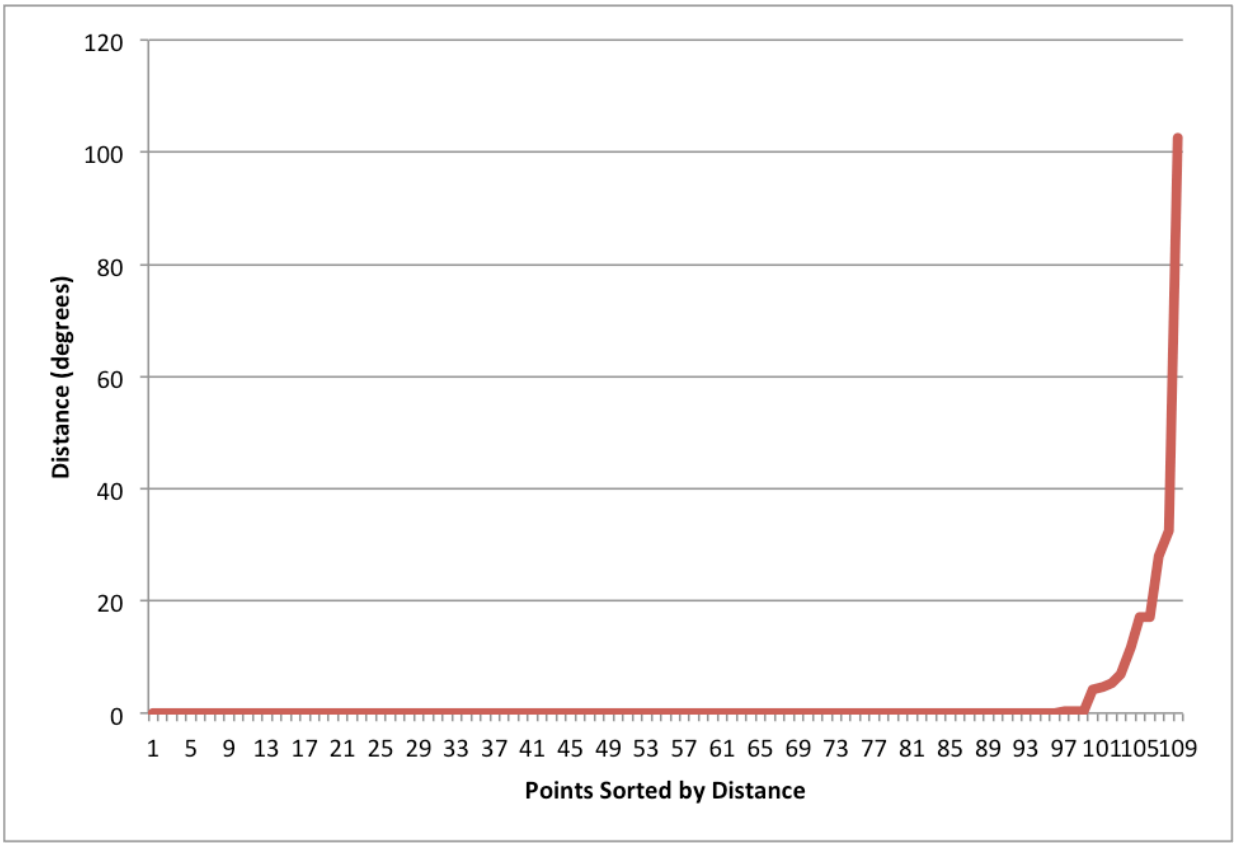

Figure 4.5. K-distance Graph

Twitter accounts have few friends or followers, yet it is still beneficial to obtain some result for clustering. For example, for a user with 4 friends, if 2 of the friends live in the same location, GCL is able to make a prediction that the user lives near the cluster of 2 friends.

For the cluster radius parameter, we first tried utilizing the standard technique for estimating the value of the cluster radius, which is calculated by finding the K-Nearest Neighbors for each point and then creating a k-distance graph from the resulting distances. The standard technique is then to choose the radius parameter as the point at which the k-distance graph bends sharply. The K-Nearest Neighbors algorithm is a machine learning algorithm used to find the closest (in terms of distance) k points to each point in the set. The k-distance graph plots each point with the distance to some given nearest neighbor point. We used the K-Nearest Neighbors

implementation from scikit-learn[20] to calculate the k = minPts = 2 nearest neighbor for each of the friends and followers geographical coordinates. We created a k-distance graph where the x-axis represents points sorted by increasing distance, and the y-axis represents the distance to the 2nd nearest neighbor. A sample plot is shown in Figure 4.5. We followed the recommended estimation algorithm and chose the point where the graph bends sharply, which is 0.

Because a large number of Twitter users select a city name as their user location, and many Twitter users have at least several friends and/or followers in their home city, a majority of the friends and followers coordinates lists contain repeating coordinates from that city. This results in most of the k-distance graphs looking like Figure 3, where the distance to the 2nd nearest neighbor is 0 for a large number of the points in the friends and followers list.

This standard technique for parameter estimation works well for continuous data, but friends and followers coordinates are discrete (using a single latitude-longitude for each city). Because of the unique nature of the locations present in the Twitter relationship graph, this technique results in a cluster radius of 0 for a majority of tweets. A parameter of 0 for the cluster radius eliminates some valid clusters being formed. Due to the list of friends and followers coordinates being discrete, we chose to augment the radius with a value of 0.5. We compared the average error distances for the cluster radius values of 0.2, 0.5, and 1, and 0.2 resulted in the lowest average error.

[20] http://scikit-learn.org/stable/modules/neighbors.html

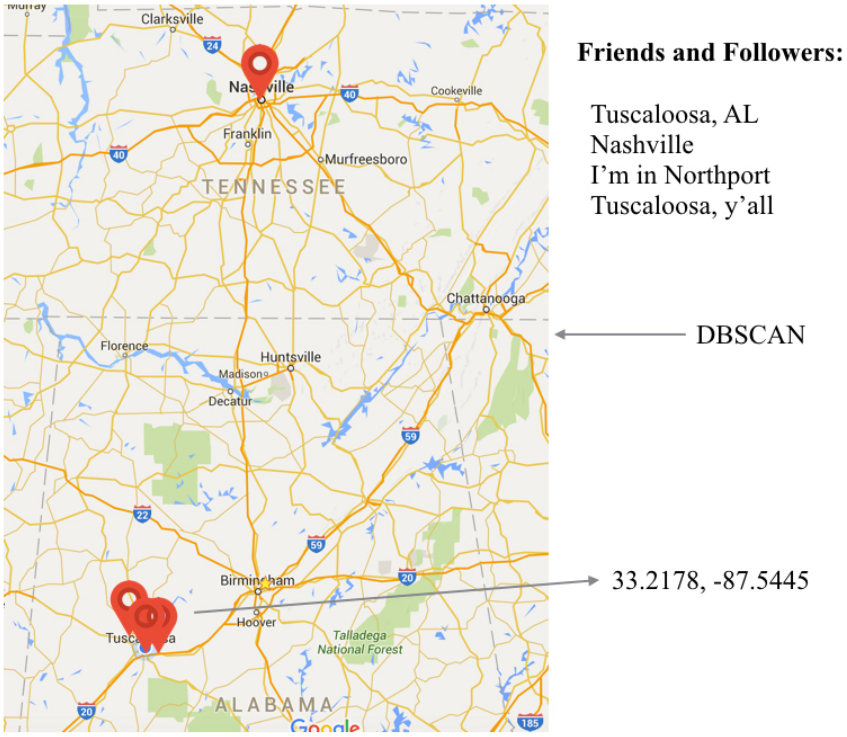

Figure 4.6. DBSCAN Clustering Process

The DBSCAN algorithm clusters points that are densely packed together and considers points in low-density regions to be outliers. Because the DBSCAN algorithm is able to ignore outliers by not including them in a cluster, GCL is able to ignore single locations that are far away (e.g., many people have a friend who lives in another city) and focus on friends and followers locations that are clustered together geographically. We use the density-clustering package[21] available for Node.js for our implementation. Figure 4.6 shows a graphical representation of the clustering process for friends and followers locations.

After DBSCAN clusters the locations, GCL looks for the largest cluster. We chose to pick the largest cluster because this cluster represents where most of the user's friends and followers

[21] https://www.npmjs.com/package/density-clustering

are located. GCL chooses the midpoint of the cluster as the estimated location of the tweet. The midpoint coordinates are added to the list of location predictions for the tweet.

4.4. Geolocation Using Tweet Topic

In addition to using a combination of content-based geolocation and relationship-based geolocation, we chose to use topic-based geolocation as a possible prediction source for tweet location. In work described in Chapter 5, we extracted the topic of tweets, or concepts that the tweet is discussing. We clustered the tweets by topic and discovered where the clusters were located geographically. In this way, we could understand how topics that users are tweeting about differ from place to place.

GCL takes advantage of the topical clustering algorithm and is able to predict the location of some tweets based on their topic. In order to extract the topic of a tweet, GCL utilizes the AlchemyAPI Concept Tagging[22] and Keyword Extraction APIs[23]. Both APIs provide results that contain ranked topics pertaining to the tweet. Table 4.1 shows the extracted topic for the example tweet, which is discussing Alabama football.

After topics are extracted from the tweet, GCL clusters the tweet along with existing tweets. The clustering process is described in Section 5.1. If the tweet is matched to a cluster of tweets with a similar topic, GCL then determines whether that topic cluster is centered in a geographic region or whether it is distributed evenly. If the topic cluster is centered in one region, it can be predicted that the tweet's location is also within that geographic region. In this case, coordinates of the geographic region are placed into the list of possible predicted locations for the tweet. The tweet is then passed to the final step in the GCL pipeline.

[22] http://www.alchemyapi.com/api/concept-tagging
[23] http://www.alchemyapi.com/api/keyword-extraction

4.5. Choosing a Final Prediction

In the final step of the GCL pipeline, the list of estimated locations from the previous four steps is analyzed and a final predicted location is chosen. First, if any of the estimated locations are close to each other, they are clustered together using the same DBSCAN algorithm as used by GCL in the friends and followers clustering step. We evaluated several different parameters for the cluster radius for this step, including 0.1, 0.3, 0.5, and 0.7 degrees, and we found that the accuracy of the final predicted location did not depend on the value of the cluster radius. If a cluster is found, this indicates that at least two techniques produced predicted locations that were close together. GCL takes the average of the geographical coordinates in the cluster with the largest number of coordinates and considers the average location as the final predicted location. If no clusters are found, this indicates that either there is only one predicted location, or the predicted locations are not within 0.5 degrees of each other.

In the case where no clusters are found, GCL selects the most likely estimated location. In order to calculate a final prediction out of the list of estimated locations, we conducted two different experiments to determine which methods produced accurate locations most consistently.

In the first experiment, we analyzed 139 unique geotagged tweets and considered which techniques produced a result within 30 km of the tweet's actual location. The results are displayed in Table 4.2. The right-most column in Table 4.2 displays the percentage of the technique that was accurate out of the number of tweets that had an accurate result. This shows how much each particular method contributes to the overall accuracy.

Table 4.2. Results From First Experiment

Technique	**Number of Accurate Results**	**Percentage of Accurate Results**
Friends and followers	43	57.75%
User account location - AlchemyAPI	19	26.76%
User account location – Dbpedia with one token	14	19.72%
User account location – Dbpedia with two tokens	13	18.31%
Content – AlchemyAPI	20	28.17%
Content – Dbpedia with one token	4	5.63%
Content – Dbpedia with two tokens	4	5.63%
Topic	22	30.99%

As shown, a tweet's friends and followers produce the correct location with the highest percentage. This is not surprising, because almost every Twitter account has at least one friend or follower, and the average number of followers is 208 (Roberts, 2012). In contrast, not every user has an account location or mentions a location within their content. The related topic method is the next highest percentage of accuracy, and following that is the user account location and the tweet content. The results from this experiment are used to predict the final location.

We determined that 139 tweets were not enough to consider a full evaluation, so we decided to repeat the experiment with more geotagged tweets. In the second experiment, we extracted location information from 5,000 geotagged tweets and determined, for each tweet, which

technique produced the closest location to the actual location of the tweet. The results are shown in Table 4.3. The left-most column and middle column contain the name of each technique along with a brief description. The right-most column contains the number of results where each technique was the closest to the actual location of the tweet.

In this experiment, we found that the technique of geolocation by topic did not produce any locations that were more accurate than other techniques. We believe this is due to the fact that there are too many tweets with varying topics in the dataset. A dataset with a smaller number of topics, such as a collection of tweets for a specific domain, may benefit from topic geolocation. To this end, we present this result for reference for future research in geolocation.

Following the results of this experiment, if no cluster is found via DBSCAN, GCL selects the location produced by the friends and followers step, if it exists. If it does not exist, GCL selects the location produced by the google three step, progressing through the other techniques listed in Table 4.3.

4.6. Related Work

Existing research in the area of geolocation has been focused in mainly two areas: 1) geolocation based on the content of the social media post, and 2) geolocation based on the relationships of the user with other users on the social media network. We first discuss research based on the content of the post.

Table 4.3. Number of Accurate Predictions Per Technique

Step	**Technique**	**Number of Most Accurate Predictions**
Friends and followers	Midpoint of cluster of friends' and followers' locations	1761
Google three	Google Places API with three consecutive tokens from content	672
Google two	Google Places API with two consecutive tokens from content	609
Alchemy user location	AlchemyAPI Entity Extraction of user location	551
Google four	Google Places API with four consecutive tokens from content	520
Dbpedia user location	Dbpedia of entire user location and tokenized user location	440
Alchemy content	AlchemyAPI Entity Extraction of content	226
Dbpedia content two	Dbpedia of entire content and two consecutive tokens of content	96
Dbpedia user location two	Dbpedia of two consecutive tokens of user location	76
Dbpedia content three	Dbpedia of three consecutive tokens of content	35
Dbpedia content four	Dbpedia of four consecutive tokens of content	8
Google	Google Places API with entire content and tokenized content	1
Topic	Topic of content	0

4.6.1. Content-Based Geolocation

Several approaches are based on comparing a tweet to previous tweets with known locations to discover similarities between the tweets. Tweets that are determined to be similar can be inferred to have similar locations. A significant number of geotagged tweets occur as a result of the user having other location-based social networks, such as Foursquare, that send automatic geotagged messages to their Twitter account. Watanabe et al. (Watanabe, Ochi, & Onai, 2011) created a database from tweets that were posted via Foursquare. They were then able to use the database to look up place names in non-geotagged tweets and predict the location of the non-geotagged tweets. Ikawa et al.'s (Ikawa, Enoki, & Tatsubori, 2012) approach for predicting user locations involves extracting keywords from tweets in a training set. Keywords are then extracted from the test set tweets, and the keywords are compared to those in the training set. Cosine similarity is computed between the keywords, and the location associated with the keyword set in the training set is estimated as the location of the tweet in the test set. These approaches are similar to our method for determining location via tweet content topic. However, Watanabe et al's (Watanabe, Ochi, & Onai, 2011) method is only able to geolocate tweets that have place names in the text, such as the name of a restaurant, while ours is not limited to only location names.

Several approaches predict locations within a grid cell rather than as geographical coordinates. Wing and Baldridge (Wing & Baldridge, 2011) ran their geolocation algorithm on Wikipedia documents, rather than a more traditional type of social media platform such as Twitter. However, like the previously described related work, the authors also utilize the content of the document to predict a location of the text. Their approach divides the Earth into varying sized cells and predicts a cell for each document. Their model calculates the distribution of

words over different locations and compares the word distribution of each document to the word distribution of each geographic cell, eventually choosing the cell with the highest similarity. Baldwin et al. (Baldwin, Cook, Han, Harwood, Karunasekera, & Moshtaghi, 2012) also predict the location of the author of each post within a grid cell on a map. Their approach utilized a naive Bayes classifier to approach the problem of geolocation. They split each Twitter post into tokens and consider each token as a feature in the classifier.

Some approaches utilized existing web services or other APIs in order to perform geolocation. Jaiswal et al. (Jaiswal, Peng, & Sun, 2013) utilized a named-entity extraction module, ANNIE, to extract possible locations from the content of Twitter posts. The locations were then mapped to geographical coordinates (a process called geocoding) using the geonames.org web service. In this approach, the authors take into account temporal information present in the tweet content. For example, if the word "tomorrow" is present in the tweet, the location mentioned in the tweet will be predicted to occur one day after the timestamp of the tweet. Baucom et al. (Baucom, Sanjari, Liu, & Chen, 2013) discussed their approach for analyzing how Twitter models the real world through the example of social media discussions about a basketball game. In order to perform the analysis, the authors geolocate tweets by passing the location associated with each user's account (not the location associated with each tweet) through the Google Maps geocoordinates API.

Several approaches involved creating models to calculate the distribution of text over geographical areas. Han et al. (Han, Cook, & Baldwin, 2014) experimented with several algorithms for location prediction, including a generative Naive Bayes model and KL divergence (Kullback & Leibler, 1951). The authors attempted to predict the "home location" of the user associated with each tweet and assumed that the users remain in the same location throughout the

dataset. Hong et al. (Hong, Ahmed, Gurumurthy, Smola, & Tsioutsiouliklis, 2012) outlined their approach for defining a model that describes the global distribution of topics. The geographical location portion of their model is a collapsed Gibbs sampler (Gemen & Gemen, 1984) for locations. Yuan et al. (Yuan, Cong, Zhao, Ma, & Sun, 2015) described their model, EW^4 that uses a generative process to model tweets along with their day, time, words, and location. The model is able to predict user location by incorporating the temporal aspect of the tweet. It utilizes both location identifiers (e.g., text descriptions of the location) and geographic coordinates in the model to better predict location. Cheng et al. (Cheng, Caverlee, & Lee, 2010) also modeled the distribution of words over locations to discover "local words," or words that are used more frequently in a localized region. They extracted words that are used frequently in one point and whose usage drops off rapidly around that central point. Tweets containing "local words" are predicted to be in the locations where "local words" occur.

Although these existing approaches all geolocated users and tweets based on the content of the tweet, no existing methods utilized Wikipedia entries or Google Maps as we did. We evaluated our method against several of these approaches in Chapter 6, and found that using both techniques was able to improve the accuracy of GCL above the existing approaches.

We next discuss the research based on the relationship graph of the user within the social network.

4.6.2. Relationship-Based Geolocation

Backstrom et al. (Backstrom, Sun, & Marlow, 2010) described their approach for predicting location based on the relationships of the user on a social network. When analyzing the relationship graph of a social network user, the mean or median location of the user's friends

may not be accurate. For example, if the user has one friend living far away, that "outlier" friend will inaccurately influence the user's location. The authors constructed a probabilistic model that determines the likelihood of a given location being the actual location of the user. The model is based on the probability that each of the user's friends would have a friend living in that location. To compute the location prediction, the model computes the likelihood that each friend's location is the user's location. Similarly, we account for these "outlier" friends by clustering by density using DBSCAN.

McGee et al. (McGee, Caverlee, & Cheng, 2013) extended Backstrom et al.'s approach of using relationships to predict location by including tie strength, or a measure of how much two users interact, in their prediction. Unlike previous work on relationship-based geolocation, this approach does not treat friends equally. The authors construct a model consisting of a tree classifier and a maximum likelihood estimator to predict location.

Like Han et al. (Han, Cook, & Baldwin, 2014) described in the content-based geolocation section previously, Li et al. (Li, Wang, Deng, Wang, & Chang, 2012) were interested in predicting users' home locations. However, they utilize the relationship graph of the user rather than the content of the tweet. Their model can analyze the likelihood that the user is in various locations of his friends based on the probability that an edge between the user and the friend exists without the user living in the friend's location. The model can also take into account the "influence scope" of the user. For example, a celebrity Twitter user is more likely to have followers in distant locations.

Similar to our approach, Rout et al. (Rout, Bontcheva, Preotiuc-Pietro, & Cohn, 2013) utilized Dbpedia as a resource for looking up location information. They used regular expressions to extract locations from the user account locations of each user's friends. Like Han

et al., the authors also take into account the probability of the existence of an edge between the user and a friend depending on the population size of the friend's city. Their model also considers whether an edge in the relationship graph is unidirectional or bidirectional (i.e., whether the friendship is reciprocated).

The existing approaches described in this section are the state of the art in geolocation research. Our pipeline for predicting location differs from these approaches in several ways. GCL combines both aspects of geolocation research, content-based geolocation and relationship-based geolocation, in order to gather location information from multiple sources within the social media post. Also, we introduce topic-based geolocation, which geolocates tweets with other tweets of the same topic.

Our approach differs from several existing methods, especially in terms of relationship-based geolocation. Several algorithms described in Section 4.6.2 utilized techniques such as analysis of tie strength, or the amount two users interact online. We believe that although GCL performed well in evaluations, described in Chapter 6, incorporating a measure of how much users connect and treating friends and followers with different measures could likely improve GCL even further. We plan to study this phenomenon in future work.

CHAPTER 5

GEOTOPICAL CLUSTERING

Geotopical clustering allows topics to emerge from a stream of tweets and to be associated with where those topics are located. Within geotopical analysis, topical analysis can cluster tweets into well-formed concepts, where all tweets in a cluster are talking about the same idea or event. The geographical analysis portion then adds a location-enabled component to the concepts. Once the geolocation step is complete, tweets are associated with locations, fulfilling the concept of GPS-enabled documents. Each document is represented by a tweet.

We decided to implement two types of geotopical clustering. The first type has topical clustering performed first, followed by geographical clustering. The second type has geographical clustering performed first, followed by topical clustering. Previous research has not compared the two approaches. In this chapter, we describe both clustering types in detail. Following, in Chapter 6, we perform evaluations to compare the techniques.

GeoContext provides a new technique for geotopical clustering, which involves clustering tweets into topics for a user's specific area and performing geographical analysis on worldwide tweets in order to find relevant information. After performing the geolocation step, in order to provide a recommendation of relevant information to users in different geographic regions, we need to determine which topics mined from the Twitter stream appear in various locations.

Our method provides several advantages over existing clustering approaches. First, GeoContext can process tweets immediately as they are streamed without removing stop words,

(i.e., words such as “the” or “a” that are often removed before natural language processing) or any terms needing to be stemmed (i.e., returning terms to their root form). Also, because of GeoContext’s method of extracting concepts from tweets, there is no need for an initial training set.

GeoContext’s topical clustering approach differs slightly from traditional topic modeling approaches. Where topic modeling techniques extract terms from a collection of documents that represent individual topics, topical clustering techniques cluster documents together that pertain to individual topics. Both methods are related in the sense that they involve determining what documents or terms should be considered to have similar content. However, we wanted to highlight this slight distinction between the two.

In the following subsections, we describe the topical clustering first technique, followed by the geographical clustering first technique.

5.1. Topical Clustering First, Geographical Clustering Second

In the initial implementation, we decided to perform topical clustering first. An overview of this method is shown in Figure 5.1. We refer to this implementation as TCGC.

After passing through the geolocation step, tweets are analyzed and clustered in real-time. The clustering process is performed synchronously, which is important due to the volume of tweets. We first describe the topical clustering step, which uses keyword and concept extraction to match tweets together with similar topics.

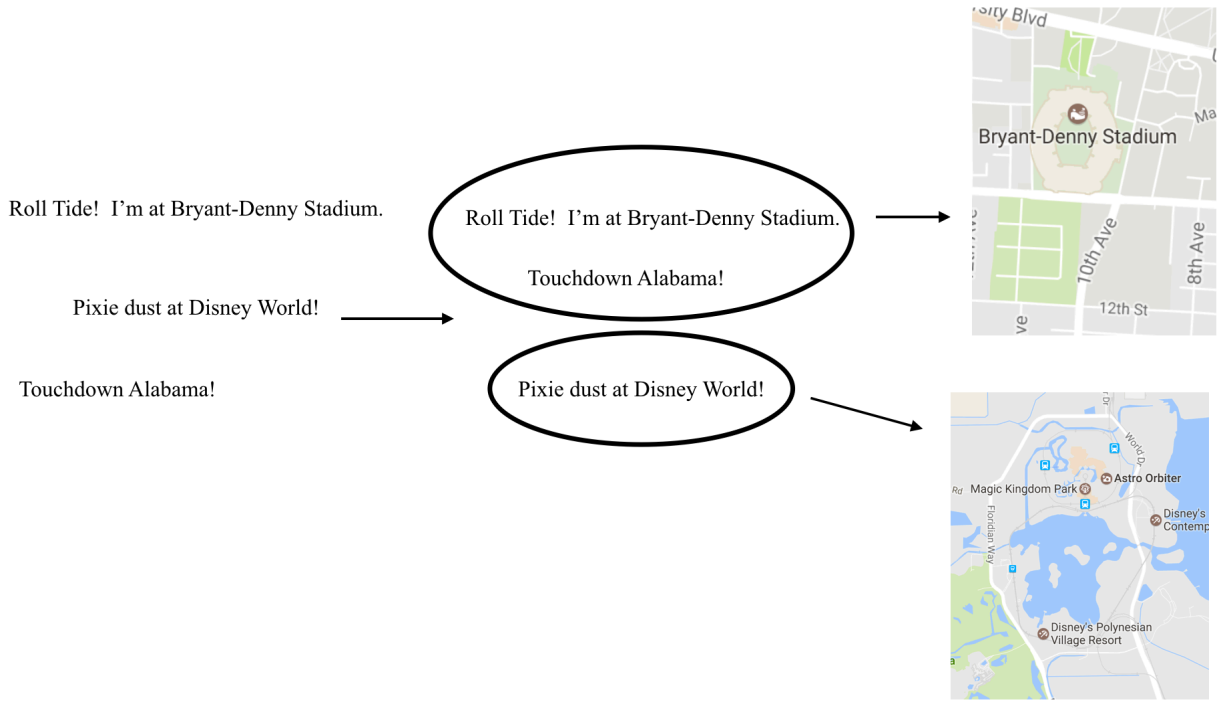

Figure 5.1. TCGC Implementation

5.1.1. Topical Clustering

After passing the tweet through the geolocation step, we begin the topical clustering step. A common method for clustering text into topics is to use LDA (Blei, Ng, & Jordan, Latent Dirichlet allocation, 2003), which is described in Section 2.3. As mentioned previously, LDA is a topic model that takes in a corpus, or body, of text separated into documents. Each document contains words in no sorted order. The model produces a selection of topics, which are collections of words found in the documents. The topics are based on which words appear together most often. LDA can also determine what percentage of each document is composed of each topic, as well as a percentage of how each word influences each topic. A graphical representation of LDA is shown in Figure 2.1.

However, LDA requires the number of topics to be determined beforehand, which is impractical to calculate for a real-time, global system. Also, LDA only considers words that appear directly in the text, which limits the algorithm's ability to detect any underlying meaning of the text. To address this problem, we implemented GeoContext to add new topics dynamically as they appear in the Twitter stream, and prune topics if they are not tweeted about enough.

We utilize AlchemyAPI's concept tagging and keyword extraction APIs to extract topics from each tweet. The topics returned from the concept tagging API are not simply terms extracted directly from the tweet, but are concepts of the tweet. For example, a tweet that contains song lyrics could result in concepts of the recording artist or the year the song was published. Keywords returned from the keyword extraction API mine important words directly from the tweet. We elect to use both the concept tagging and keyword extraction APIs because, although there is sometimes overlap between concepts and keywords extracted from a tweet, both provide useful information about the content of the tweet. Example keywords and concepts are shown for tweets in Figure 5.2.

After concepts and keywords are extracted from the tweet, GeoContext clusters the tweet along with existing tweets into topic clusters. To determine which topic cluster the tweet should be matched with, GeoContext needs to determine which topic cluster contains the same topics as the tweet. Tweets in the same cluster are discussing the same topic, whether it is an event, popular celebrity, or news subject. GeoContext calculates a similarity score between the tweet and each topic cluster. The tweet is placed into the topic cluster with the highest similarity score to the tweet.

The similarity score represents the similarity between two tweets. It is calculated based on whether the tweets contain the same hashtag, as well as the ranked concepts and keywords extracted from the tweets. Our calculation of the similarity score between two tweets is shown in Formula 1.

In the similarity score algorithm, GeoContext first checks whether the two tweets have any hashtags in common. Hashtags are tags that can indicate a user is posting about a specific topic. A hashtag, which is simply a string preceded by the # symbol, can provide important metadata about the topic of the tweet. Hashtags can also aid in search on Twitter and allow users to easily join in a conversation about a topic (She & Chen, 2014), which makes them ideal as a tool for discovering topics. We also chose to compare hashtags because they can often express a popular topic (Cui, Zhang, Liu, Ma, & Zhang, 2012).

If the current tweet contains hashtags that match hashtags present in another topic, the similarity score is assigned a value of 1 and the tweet is added to that topic. Hashtags that are not exactly the same, but refer to the same event, often end up in the same topic due to appearing together in tweets. Users can include multiple hashtags in each tweet related to the same topic.

$$tweetSimilarity = \max\left(hashtagsMatch(t1, t2), \prod_{a=0}^{b} avg(relevanceScore(t1_a), relevanceScore(t2_a))\right) \quad \text{(Formula 1)}$$

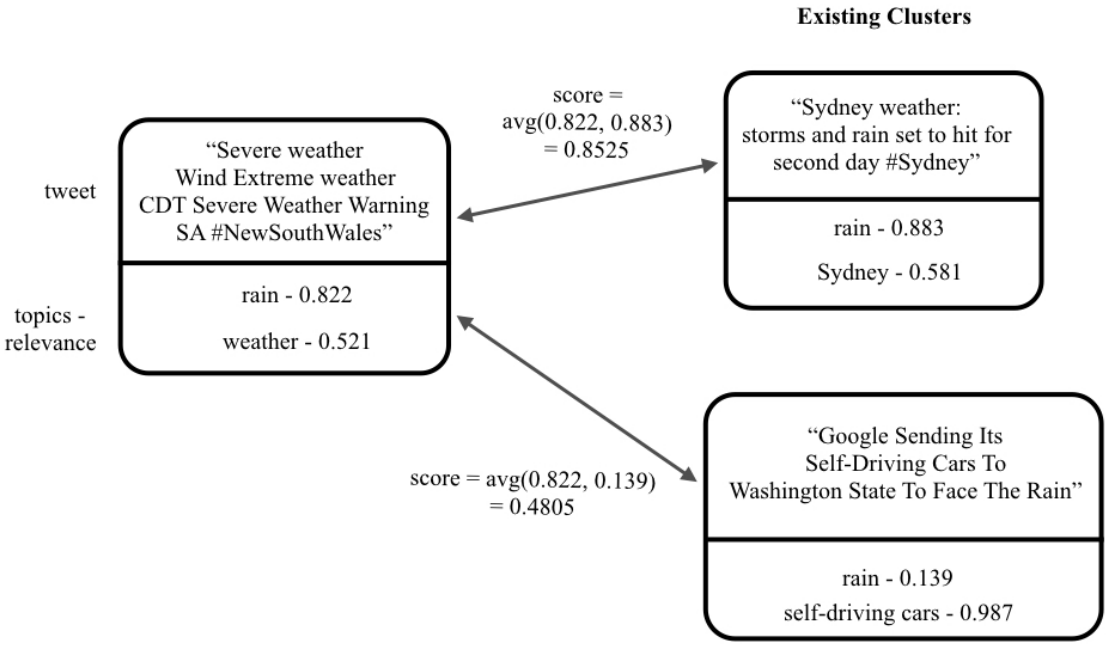

Figure 5.2. Calculating Similarity Scores Between Tweets

If no hashtags are matched, we then compare the concepts and keywords returned from AlchemyAPI of the two tweets. Each concept and keyword is associated with a relevance score from the concept tagging and keyword extraction APIs. The relevance score is a percentage that indicates how much each concept or keyword influences the tweet. As shown in Formula 1, GeoContext computes the similarity score by multiplying together the average of the relevance scores of all keywords or concepts that match between the two tweets. Given that *t1* and *t2* are the two tweets being compared, and *b* is the number of concepts that match between *t1* and *t2,* $t1_a$ and $t2_a$ are the a^{th} concepts that match between *t1* and *t2*. The relevance scores of $t1_a$ and $t2_a$ are averaged, and the averages of the relevance scores of the *b* matching concepts are multiplied.

$$clusterSimilarity = \frac{\sum_{m=0}^{n} tweetSimilarity(t, t_m)}{n} \quad \text{(Formula 2)}$$

This way, tweets are matched only with topic clusters that contain only similar topics with high relevancy scores, rather than matching the secondary topics of the tweet with low relevancy scores.

For example, Figure 5.2 shows a tweet on the left side being compared with two different tweets on the right side. The tweet on the left clearly has a main, or primary, topic of "weather." In the same vein, the top tweet on the right also has a primary topic of "weather." However, the bottom tweet on the right contains the word "rain," but is mainly about "self-driving cars." We consider the primary topic for this tweet to be "self-driving cars," while the secondary topic is "weather." For our research purposes, we desire tweets to be clustered based on their primary topics.

To illustrate the calculations of the similarity score using the relevance score, Figure 5.2 shows the tweet on the left side with a term "rain" with a relevancy score of 0.822. The two tweets on the right side both contain the same extracted term, so the average of the relevance scores for the term "rain" for each tweet is computed as the similarity score. The similarity score of the left tweet with the top right tweet is 0.8525. Because the term "rain" for the bottom-right tweet has a much lower relevancy score of 0.139, the similarity score between the left tweet with the bottom-right tweet is 0.4805. If a simple keyword matching algorithm were used, these two tweets might end up in the same topic cluster even though their primary topics are different. By taking the relevancy score into account, we avoid matching these two tweets into the same topic cluster.

```
Input: currentTweet, topicClusters
Output: The tweet will be added to the topic cluster with the highest similarity
score = 1
scores = []
for each cluster in topicClusters do
    for each clusterTweet in cluster do
        currentTweetTopics = getConceptsandKeywords(currentTweet)
        clusterTweetTopics = getConceptsandKeywords(clusterTweet)
        for each currentTopic in currentTweetTopics do
            for each clusterTopic in clusterTweetTopics do
                if currentTopic == clusterTopic then
                    score = score * average(relevance(currentTopic), relevance(clusterTopic))
                end
            end
        end
        scores.push(score)
    end
    if average(scores) > 0.5 then
        add currentTweet to cluster
        break
    end
end
```

Algorithm 5.1. Clustering Tweets

In order to place a tweet into a cluster, the similarity score is calculated between the tweet and every other tweet in every existing topic cluster. Our calculation for a topic cluster's similarity score is shown in Formula 2. Given that t is the current tweet, t_m is the m^{th} tweet in the topic cluster, and n is the number of tweets in the topic cluster, the topic cluster's similarity score to the current tweet is the average of the similarity scores between the tweet and all tweets in the topic cluster. We utilize this method for calculating the similarity between tweets rather than existing topic modeling approaches such as LDA because our method is able to compare tweets based on their underlying meaning and topic, rather than treating each word in the tweet equally.

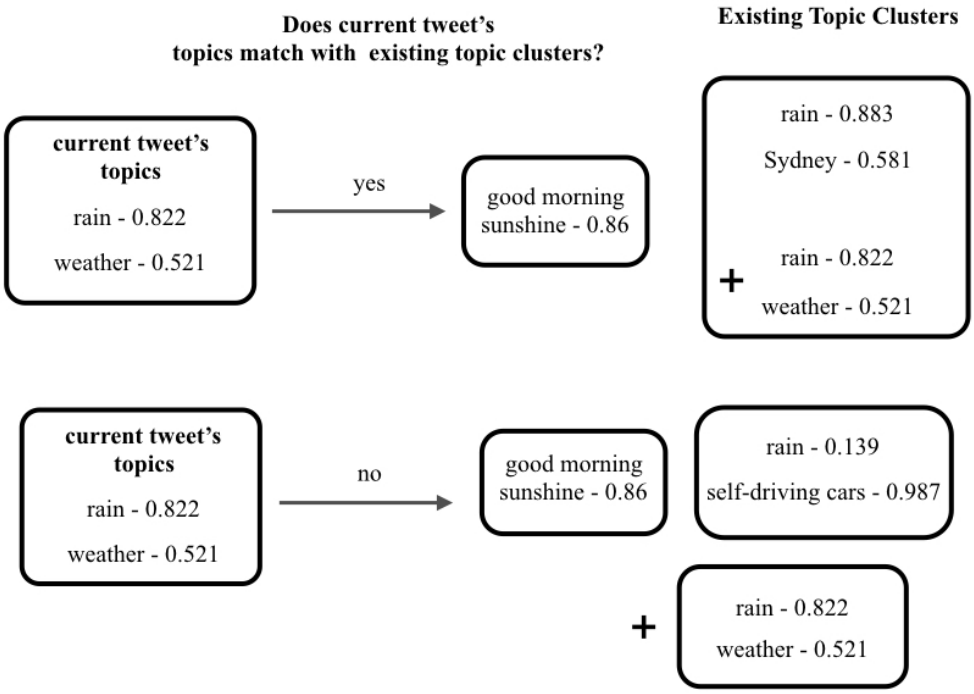

Figure 5.3. Clustering Tweets Into Topics

If a tweet matches with a topic cluster of other tweets, it is added to that cluster. This situation is shown at the top of Figure 5.3.

There are many cases in which a tweet may not belong in any existing cluster. A tweet's topics may not be related enough to tweets that have already been processed and clustered. This situation is shown at the bottom of Figure 5.3. In this case, a new cluster will be created and the tweet will be placed into the new cluster.

In order to determine whether the concepts and keywords of the tweet have below average similarity scores with the topics of the existing clusters, we use a threshold value for the

similarity score. Relevancy scores for the concepts and keywords range from 0 to 1. Therefore, our calculated similarity score will range from 0 to 1. We chose 0.50 as our threshold similarity score, because this value represents topics that are of average similarity. We conducted an empirical study to determine the accuracy of this threshold value. Results from this study are presented in Chapter 6.

If a tweet's topics do not contain similarity scores above 0.50 with any existing topic cluster, a new cluster will be created and the tweet will be placed into the new cluster. If there is at least one existing topic cluster whose similarity score to the current tweet is above 0.5, the tweet will be added to that cluster. Pseudocode for our algorithm for clustering tweets is shown in Algorithm 5.1.

Because our method has to compare each new tweet to every existing tweet in the system, it is easy to see that the time complexity of the algorithm can increase exponentially within a short period of time. However, GeoContext solves this problem in two ways. The first way is by synchronously performing the clustering process for new tweets.

The second way GeoContext runs in near real-time is by pruning topic clusters. While it is running, every fifteen minutes, the topic clusters are pruned. Pruning involves looking at each topic cluster and determining if it has become "stale." We define a "stale" cluster as one that has not had any new tweets added in 24 hours. If a cluster becomes "stale," it is deleted from the collection of total topic clusters. If stale clusters are not removed, the storage and analysis of so many tweets can greatly affect performance of GeoContext. The length of time between pruning sessions and the length of time between the last tweet added to a cluster and the cluster becoming stale are both threshold values that are evaluated in Chapter 6.

Although pruning allows GeoContext to run in real-time, in future work, we would like to examine ways to improve the performance time. We plan to determine whether a method exists that would not require the comparison of every new tweet to every existing tweet. We also plan to determine whether a different storage solution for tweets would improve the performance of GeoContext.

5.1.2. Geographical Clustering

After the topical clustering step, geographical clustering can be performed on the existing topic clusters that have been calculated. An example of how topic clusters containing tweets are associated to locations is shown in Figure 5.1. The goal of this process is to determine, for each topic cluster, whether the tweets are clustered in one or more geographic location or spread out across a larger geographic area, for instance, the entire United States. Using this goal, we can recommend topics to users that are more specific to their location.

To perform geographical clustering, we adapted the TF-IDF algorithm for our process. TF-IDF stands for Term Frequency-Inverse Document Frequency, and is a statistic that determines how important, or meaningful, a word is to a document within a corpus (Sparck Jones, 1972). The statistic eliminates words that are not unique within the corpus, so are not meaningful to a document. For example, if we consider the corpus of all Wikipedia pages of all universities, the word "university" probably occurs many times. However, the word does not add much meaning or differentiation to any particular university Wikipedia page. This example is shown in Figure 5.4.

TF-IDF calculates the meaningfulness of a word by determining that the word "university" occurs many times across all documents, indicating that they are common across all text and is

The University of Alabama (Alabama or UA) is a public research university located in Tuscaloosa, Alabama, United States, and the flagship of the University of Alabama System. Founded in 1820, UA is the oldest[4] and largest of the public universities in Alabama. UA offers programs of study in 13 academic divisions leading to bachelor's, master's, Education Specialist, and doctoral degrees. The only publicly supported law school in the state is at UA. Other academic programs unavailable elsewhere in Alabama include doctoral programs in anthropology, communication and information sciences, metallurgical engineering, music, Romance languages, and social work.

Auburn University (AU or Auburn) is a public research university in Auburn, Alabama, United States. With more than 22,000 undergraduate students and a total of more than 28,000 students and 1,200 faculty members, it is one of the state's largest universities [11] as well as one of two public flagship universities in the state.[12][13] [14]

The University of Georgia,[6] founded in 1785, also referred to as UGA or simply Georgia, is an American public Land-grant, Regional Sun Grant, National Sea Grant, and National Space Grant research university. Its primary location is a 762-acre (3.08 km2) campus adjacent to the college town of Athens, Georgia, approximately an hour's drive from the global city of Atlanta. It is a flagship university[7] that is ranked tied for 18th overall among all public national universities in the 2017 U.S. News & World Report rankings.[8] The university is classified in the highest ranking, "R-1: Doctoral Universities – Highest Research Activity", with the Carnegie Classification of Institutions of Higher Education classifying the student body as "More Selective," its most selective admissions category.[9] The university has been labeled one of the "Public Ivies," a publicly funded university considered to provide a quality of education comparable to those of the Ivy League.

Figure 5.4. TF-IDF Example

not particularly important to one piece of text. In our case, using an adapted version of TF-IDF, we can discover whether a location occurs commonly throughout all topic clusters of tweets, indicating that the location simply has a higher population, or whether it is occurring more within a specific topic cluster, indicating that that topic cluster is important to that location.

GeoContext considers each topic cluster to be a document and each geographic location of each tweet in that topic cluster to be a word. In this way, GeoContext can sense whether a location is clustered more within a certain topic, or whether it occurs commonly throughout all topics. We decided to use this algorithm rather than a more traditional clustering algorithm such as K-means (MacQueen, 1967) or DBSCAN because tweets follow a population distribution. More tweets are posted in locations where the population is higher, and thus using a traditional clustering algorithm would simply cluster tweets around population centers. We are interested in

```
Input: topicClusters
Output: The topic clusters will be associated with any relevant geographical locations

for each cluster in topicClusters do
    if concept has more than 3 tweets  // otherwise cluster is not large enough to produce
                                                relevant result
        for each tweet in cluster do
            matchingLocationsInCluster = 0
            for each othertweet in cluster do
                if isClose(tweet, othertweet)
                    matchingLocationsInCluster += 1
                endif
            end
            termFrequency = matchingLocationsInCluster / totalLocationsInCluster

            clustersWithMatchingLocation = 0
            for each cluster in topicClusters do
                for each othertweet in cluster do
                    if isClose(tweet, othertweet)
                        clustersWithMatchingLocation += 1
                        break
                end
            end
            inverseDocumentFrequency = log(totalTopics / clustersWithMatchingLocation)

            tfidf = termFrequency * inverseDocumentFrequency

            if tfidf > 0.8 then
                recommend location for topiccluster
            endif
        end
    endif
end
```

Algorithm 5.2. Adapted TF-IDF

$$tf(t,d) = f(t,d) \quad \text{(Formula 3)}$$

$$idf(t,D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \quad \text{(Formula 4)}$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad \text{(Formula 5)}$$

also discovering topic clusters around geographic locations with varying population densities. Using adapted TF-IDF allows us to discover topics that are influencing a certain geographic area, even if the population density of the area is small.

The TF-IDF statistic is computed by the formula shown in Formula 5. Term frequency is shown in Formula 3 and is calculated as the number of times a word *t* appears in a document *d*. Following our assumptions outlined previously, we calculate term frequency as the number of times a certain geographic location appears in a topic cluster. Inverse document frequency is shown in Formula 4 and is calculated as the logarithmically scaled fraction of the number of documents in the corpus *N* divided by the number of documents *d* in the corpus *D* that contain the word *t*. We calculate inverse document frequency by dividing the total number of topic clusters by the number of topic clusters that contain the geographic location and taking the logarithm. The TF-IDF statistic is then determined by taking the product of the term frequency and the inverse document frequency, as shown in Formula 5.

Pseudocode for our implementation of the adapted TF-IDF algorithm to associate locations to topic clusters is shown in Algorithm 5.2. We calculate the TF-IDF for each geographic location in each topic cluster, and if the result is above a threshold value, we can infer that the geographic location occurs more often in that topic cluster than other topic clusters. We chose 0.2 as our threshold value, because after inspection of the results, this value represents locations that occur

Figure 5.5. GCTC Implementation

several times in one topic cluster, but few times in all other topic clusters. Therefore, if a location has a TF-IDF value of higher than 0.2, we recommend that topic cluster for the geographic location. Using our adapted TF-IDF statistic allows us to possibly discover multiple geographic locations that are important for a topic cluster, meaning that people are tweeting about a topic clustered in multiple locations.

5.2. Geographical Clustering First, Topical Clustering Second

In the second implementation, we perform geographical clustering first. An overview of this method is shown in Figure 5.5. We refer to this implementation as GCTC. In this implementation, GeoContext takes in a stream of tweets, clusters tweets by their geographical locations, and then breaks each geographical cluster into topics.

Similar to the TCGC implementation where tweets were clustered by topic first, with GCTC, we first perform geolocation on the stream of tweets and then begin the clustering process. In the GCTC implementation, geographical clustering is performed first. Because there are no topic clusters yet, we cannot perform geographical analysis using adapted TF-IDF as with the

TCGC implementation. Instead, we chose to utilize DBSCAN to cluster tweets into different locations.

After the tweets are passed through the geolocation step, GeoContext saves their geographical coordinates. When the list of coordinates reaches a multiple of 1000, we cluster the list of coordinates. Our parameters to the DBSCAN algorithm are 0.5 for the cluster radius and 5 for the minimum number of points to form a cluster. We chose 0.5 for the radius because a radius of 0.5 degrees is approximately the size of an average city, so DBSCAN will cluster tweets within cities. We chose 5 for the minimum number of geographical coordinates to place in a cluster because a location with less than 5 tweets does not generally have enough tweets to successfully cluster topically in the next step of the pipeline.

The DBSCAN algorithm clusters points that are densely packed together and considers points in low-density regions to be outliers. This process differs from K-means and other clustering algorithms that cluster points based on closeness to a mean point. We chose DBSCAN because GeoContext requires tweets to be clustered by the density of the tweets in various geographical areas so that it can discover areas where tweets are occurring the most. DBSCAN returns a list of clusters and lists of all points within those clusters. It also analyzes which points do not belong within a cluster (classified as "noise") and returns a list of those points.

After the geographical clusters are formed through DBSCAN, we utilize the same topical clustering system as described previously for the TCGC implementation. For each geographical cluster, we create topic clusters using the same topical clustering implementation described for TCGC. Each tweet's concepts and keywords are compared against tweets in existing topic

clusters, and if the similarity scores indicate that the tweets have similar topics, they are clustered together. Otherwise, the tweet is placed in a new topic cluster within the geographical cluster.

We present the results of our evaluation of both methods of geotopical clustering in Chapter 6.

5.3. Related Work

Research in the area of geotopical clustering has typically been in the area of creating geographical models of topics that appear in a stream. Many systems use LDA as a base technique. Also, most of the research described in this section utilizes Twitter as the predominant social media platform.

Kim et al. (Kim, Lee, & Kyeong, 2013) detected "hot topics" from Twitter posts by normalizing high frequency words over time. This approach allowed words with a frequency that dramatically increased in a short period of time, such as words related to holidays or major events, to appear. They also used a Louvain community detection algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) to discover in which states topics were being tweeted. Their method, however, has a drawback in that some topics may be suppressed if the topics contain mostly high frequency words.

Yin et al. (Yin, Cao, Han, Zhai, Chengxiang, & Huang, 2011) introduced Latent Geographical Topic Analysis, or LGTA, which is their extension of LDA to take geographical information into account within a corpus of text. Rather than cluster text by document as in traditional LDA, the LGTA algorithm uses a textual corpus clustered by region to derive topics from text. LGTA discovers topics that are grouped together by geographical region. There are several limitations to their method; namely, the fact that the number of desired geographical

topics must be known beforehand. Also, parameters to the algorithm must be estimated prior to the algorithm, making it inefficient for use on a real-time system, because parameters may need to be re-estimated often.

Zhang et al. (Zhang, Sun, & Zhuge, 2015) described their system for clustering text by topic and geographical location. Their approach is similar to Yin et al., in that they use LDA to discover topics in the corpus, a collection of unordered textual documents. They also separate the corpus by region. The authors combine LDA with DBSCAN to produce six different topic and geographical clustering algorithms. With all algorithms, however, the number of topics and clusters must still be set beforehand. These parameters may be difficult to determine for a large-scale, real-time system.

Other research has expanded beyond LDA. Vosecky et al. (Vosecky, Wai-Ting Leung, & Ng, 2013) introduced their Multi-Faceted Topic Model, which incorporates all facets of information present in tweets, including people, location, and organization entities and a time element. Hong et al. (Hong, Fei, & Yang, 2013) built their Content Model based on Binomial Logistic Regression. The Content Model extracts content from tweets by expanding the URLs found in many tweets. The Content Model also takes into account the number of retweets.

Son et al. (Son, Noh, Song, & Park, 2012) described their method, called Probabilistic Explicit Semantic Analysis (PESA), which compares locations for the purpose of location recommendation. The authors represent each space as a set of topics gleaned from Wikipedia. Like our work, they attempt to calculate a semantic distance between topics associated with a location in order to determine which locations are similar. However, they are not applying this work to social media and mining topics from user posts.

Gao et al. (Gao, Cao, He, & Li, 2013) adapted the K-means clustering algorithm to cluster tweets with other tweets of the same topic. They applied TF-IDF to textually cluster the tweets. Like our research implementation, they also used an adapted TD-IDF method; however, the authors did not include a geographical component with that portion of their research. Rather, they applied a novel pattern mining algorithm in order to perform geographical analysis on the tweets.

Sakaki et al. (Sakaki, Okazaki, & Matsuo, 2013) presented their approach for detecting earthquakes and other major events by analyzing a real-time stream of tweets. The authors use a classifier to determine if a user is tweeting about an event happening in real-time or whether the tweet is not referring to a major event or is irrelevant. This work differs from our work in that Sakaki et al. are detecting only pre-defined events of a large scale by filtering by keywords related to the event. Their approach also will not detect multiple events occurring in different locations simultaneously, while GeoContext is able to detect multiple events of any type at differing locations automatically.

Hong et al. (Hong, Ahmed, Gurumurthy, Smola, & Tsioutsiouliklis, 2012) modeled a stream of tweets across geographical locations. Through their model, they are able to predict a location of a user given the topics of the tweet and a user's location history. Although they are mining topics from each tweet, similar to our work, their system does not attempt to model events as they happen across Twitter.

Musaev et al. (Musaev, Wang, Shridhar, & Pu, 2015) presented their approach for classifying textual documents within social media by resolving the meaning of ambiguous words. For example, they disambiguate the term "landslide" to the type of weather phenomenon, a

Fleetwood Mac song, or a mudslide cocktail. In our approach, we disambiguate words by associating them with the other words in the content of the tweet.

CHAPTER 6

EVALUATION

In order to determine how well GeoContext is able to perform, we performed several evaluations. We first evaluated the geolocation module separately and tested how accurately it was able to predict locations for tweets. We also performed several evaluations solely on the topical clustering module in order to determine how well it was able to cluster tweets into topics. We compared the topical clustering module against LDA. Lastly, we tested the geographical clustering module to determine whether it could associate locations with topic clusters.

In this section, we describe all evaluations we completed of both the geolocation and geotopical clustering modules within GeoContext. For all evaluations, the GeoContext pipeline is set up as a Node.js server. We executed all evaluations on a MacBook Pro 2.5 GHz with 16 GB RAM.

6.1. Evaluation of Geolocation Module

We performed two empirical evaluations to test the accuracy of GCL, the geolocation component of GeoContext, in predicting locations of tweets. In this section, we describe our experimental setup and results for both evaluations.

Because GCL has a step that analyzes a Twitter user's friends and followers to predict location, lists of all user's friends and followers were needed for both evaluations. These lists are not included in the tweet objects that are obtained from the Twitter Streaming API. Rather, they

need to be fetched from the Twitter REST API. Because most Twitter REST API calls are restricted to the retrieval of 15 users' friends and followers list per 15 minutes, including friends and followers fetching, we decided to pre-fetch tweets along with each tweet's friends and followers list.

6.1.1. First Evaluation of Geolocation Module

The first evaluation was performed early in our research. This testing was done prior to adding in utilization of the Google Places API to GCL, as well as prior to improving the final step of choosing a single predicted location. In our first evaluation, we collected 409 total tweets, all of which were geotagged. We used geotagged tweets so that we had a baseline of known accurate locations for all tweets. We ran all 409 tweets through GCL, including their user locations and friends and followers.

In this evaluation, we analyzed two types of results from GCL. First, we wanted to determine the accuracy of any of the techniques within GCL. We looked at all techniques within GCL (i.e., user location, tweet content, and friends and followers) and considered whether any of the techniques were able to produce an accurate predicted location. In this portion of the evaluation, we were effectively testing all steps in GCL except for the last (i.e., choosing a final predicted location). We checked the user's actual location (from the geotag) against the list of all estimated locations for each tweet from each technique. If any of the estimated locations are correct, we considered that tweet to have an accurate location prediction. For example, for a tweet, if the friends and followers step produced an accurate location prediction, even if it was not chosen to be the final prediction, we considered this to be an accurate prediction.

Because this evaluation was very preliminary, we only considered an "accurate" location prediction to be within 30 km for this portion of the evaluation. We examined GCL's accuracy in more detail in the evaluation described in Section 6.1.2. GCL was able to gather the correct location within 30 km 51.83% of the time within one of the methods described in the pipeline. 212 tweets had a correct predicted location within at least one of its methods. This shows that for city- level accuracy, GCL is able to produce an accurate location within at least one method about half of the time.

For the second portion of the evaluation, we examined the final step of GCL: choosing a final predicted location out of the list of estimated location coordinates. This is obviously a more real-world analysis than the first, because in a real system, one final location would usually need to be chosen, rather than several estimates of a user's location.

For the second portion of the evaluation, we analyzed the final predicted location from GCL. In order to calculate a final prediction out of the list of estimated location, we ran an experiment to evaluate which methods were most likely to produce an accurate predicted location. This experiment was run prior to the testing described in Section 4.5. We ran 139 unique geotagged tweets through GCL and considered which techniques produced a result within 30 km of the tweet's actual location. The results are displayed in Table 6.1.

The rightmost column in Table 6.1 displays the percentage of the technique that was accurate out of the number of tweets that had an accurate result. This shows how much each particular method contributes to the overall accuracy. As shown, a tweet's friends and followers produce the correct location with the highest percentage. This is not surprising, because almost every Twitter account has at least one friend or follower, and many have quite a few. In contrast, not every user has an account location or mentions a location within their content. The related topic

Table 6.1. Results From First Geolocation Evaluation

Technique	**Number of Accurate Results**	**Percentage of Accurate Results**
Friends and followers	43	57.75%
User account location - AlchemyAPI	19	26.76%
User account location – Dbpedia with one token	14	19.72%
User account location – Dbpedia with two tokens	13	18.31%
Content – AlchemyAPI	20	28.17%
Content – Dbpedia with one token	4	5.63%
Content – Dbpedia with two tokens	4	5.63%
Topic	22	30.99%

method is the next highest percentage of accuracy, and following that is the user account location and the tweet content. The results from this experiment were used to predict the final location.

We then ran the same 409 tweets from the first experiment through GCL and evaluated the final predicted location against the geotagged location. The overall prediction accuracy of the final prediction choice (i.e., comparing the final prediction by GCL to the geotagged location for

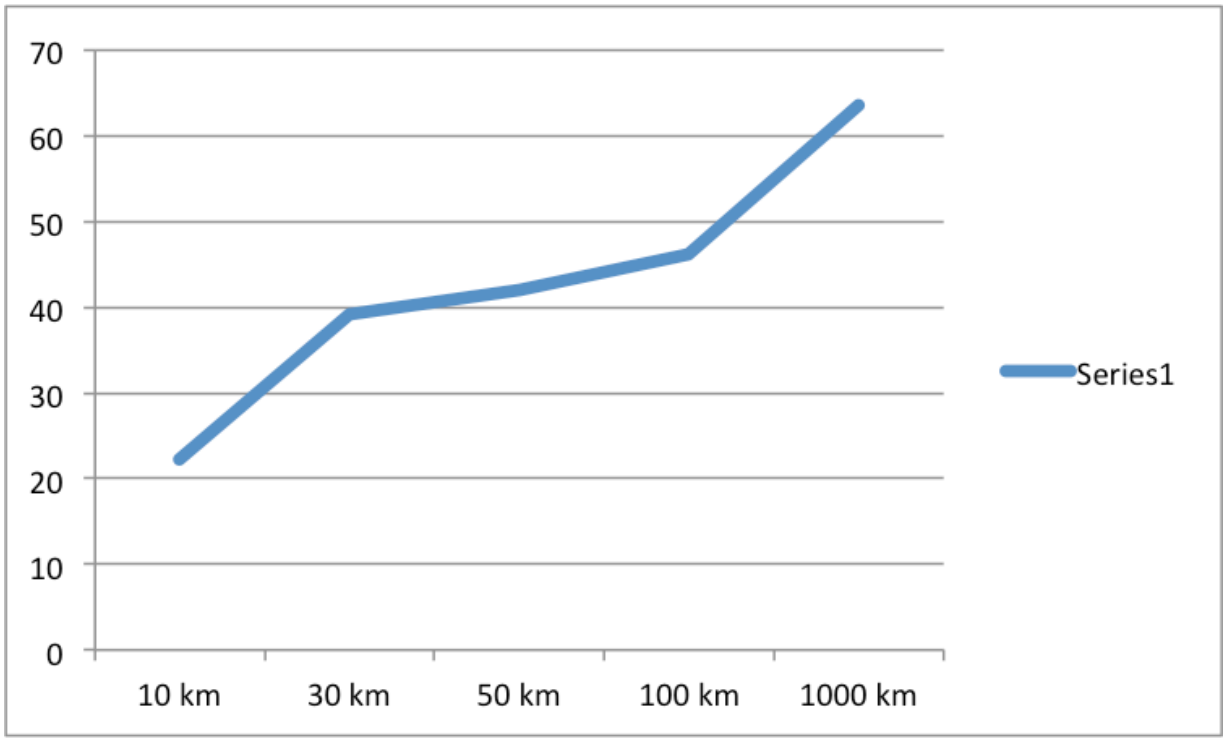

Figure 6.1. Accuracy of GCL in First Evaluation

each tweet) was 39.12% within 30 km, with 160 tweets having an accurate location prediction.

Using the friends and followers geolocation method contributed to an accurate prediction 65.09% of the time. User account location was the next most accurate method, contributing to the final prediction 60.38% of the time. Topic geolocation contributed to the correct location 18.87% of the time. This shows that topic geolocation can be an effective method, when combined with other existing methods, for predicting social media location.

Figure 6.1 shows the accuracy of GCL across multiple distances. As displayed, 22.25% of total tweets analyzed were accurate within 10 kilometers. 42.05% of tweets were accurate within 50 kilometers. These percentages reflect the distance of the final predicted location with the actual geotagged tweet location.

Our results show that GCL is fairly accurate in predicting a location for a social media post. Although some existing approaches resulted in a higher accuracy, these methods were predicting the home location of a user. This is arguably an easier assumption to make, because users spend more time at their home location, so friends and followers plus the user account location can often produce a correct result for the home location. GCL attempts to predict the tweet's current location, however, which we argue is a more useful approach when performing social media analysis. If a user is on a trip away from their home location and is tweeting about their current location, the content of that tweet should be correlated with the current location, not the home location where the information may not be relevant.

6.1.2. Second Evaluation of Geolocation Module

We determined that the first evaluation of GCL was severely lacking in the number of tweets. To address this problem, we performed a second, more extensive evaluation. For this testing, we streamed and collected 24,221 geotagged tweets from the Twitter Streaming API using twit, a node.js library for retrieving a Twitter stream. Friends and followers from each tweet were collected from the Twitter REST API also using twit. The tweets were collected in May 2016. We filtered the tweets by English-language only because some of the analyses performed by GCL are available only for English. 15,462 unique users are represented in the dataset. We did not limit the geographical location of the collected tweets, but because all tweets are English-language, the USA, United Kingdom, Canada, and Australia are the most common countries represented.

As with the first evaluation, we decided to use only geotagged tweets so that we could analyze effectively whether the location predicted by GCL was the user's actual location at the

time of the tweet. This could produce a bias, because it is possible that the content of geotagged tweets contains more location information than non-geotagged tweets. However, because there is no way to know the exact geographical coordinates of non-geotagged tweets, using geotagged tweets is the only possible method for truly analyzing whether GCL can predict accurate locations.

After tweets were collected, we streamed the tweets as JSON objects through GCL. For each tweet, GCL analyzed the accuracy of the final predicted location at various distances. The results from this experiment are described in Section 4.5.

For this evaluation, we compared the final predicted location against the actual geographical coordinates of the geotagged tweets in our dataset. Figure 6.2 shows the accuracy of GCL at various error tolerances. We define accuracy as the percentage of tweets whose final predicted location is the same as the geotagged location, within the various error tolerances shown in Figure 6.2.

Figure 6.2 also shows the comparison of GCL with several existing geolocation approaches (Ikawa, Enoki, & Tatsubori, 2012) (Han, Cook, & Baldwin, 2014) (Cheng, Caverlee, & Lee, 2010). As shown, GCL shows improvement over Ikawa et al. and Han et al.'s methods and compares within 1% of Cheng et al.'s method. Although 160 km is quite a large distance for many applications using geolocation, such as event detection systems, we are unable to compare GCL's accuracy to Cheng et al.'s at a lower threshold because they did not evaluate a lower error distance.

There are two main reasons that contribute to the results. Some tweets do not have any location information that can be extracted. For example, the tweet "@MaichardCLBRQ @aldenrichards02 @mainedcm cn't watch it" does not have any location data in the content.

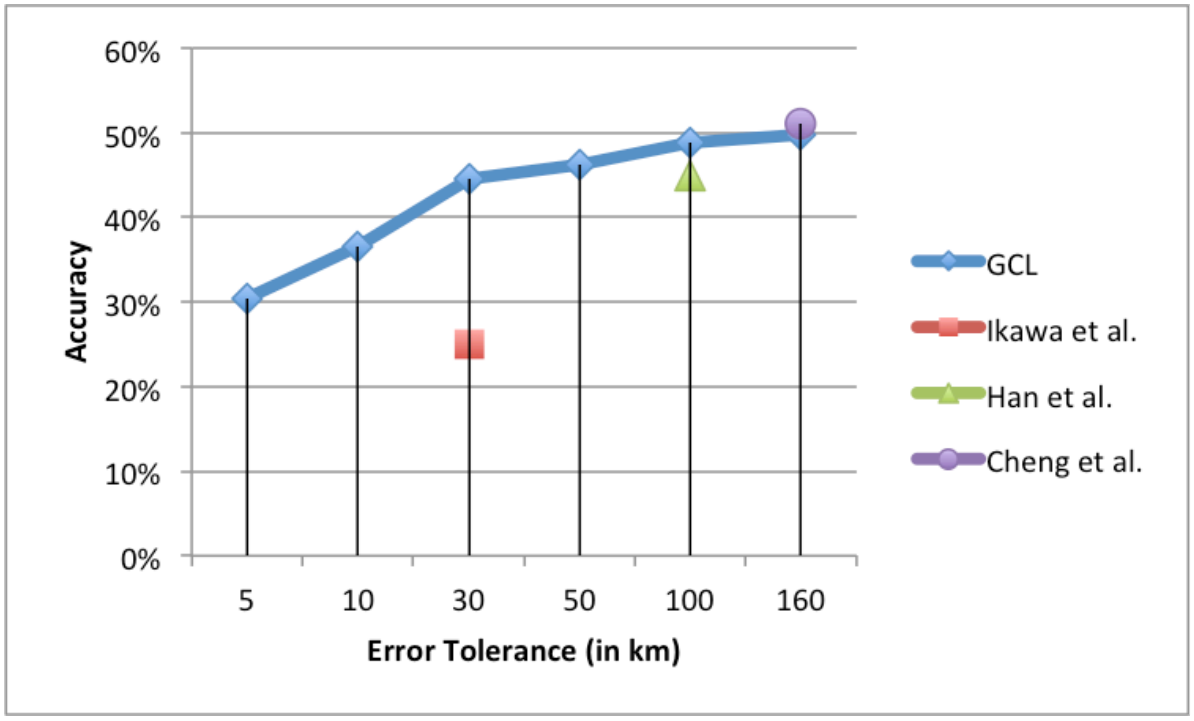

Figure 6.2. Accuracy of GCL in Second Evaluation

The tweet author also does not have a user location or any friends or followers. In this case, there exists no information within the tweet object that could be used to predict a location.

The second case is where an accurate location is extracted by a technique, but it is not chosen to be the final predicted location. We determined how many tweets had a technique produce a predicted location accurate within 30 km and compared this value to the number of tweets with a final predicted location accurate within 30 km. We found that approximately 17% of all tweets had extracted location information accurate within 30 km, but the final predicted location was larger than 30 km. Although GCL already shows significant improvement over existing methods, this value suggests that GCL offers additional promise in terms of geolocation due to the fact that it is able to extract even more accurate location information.

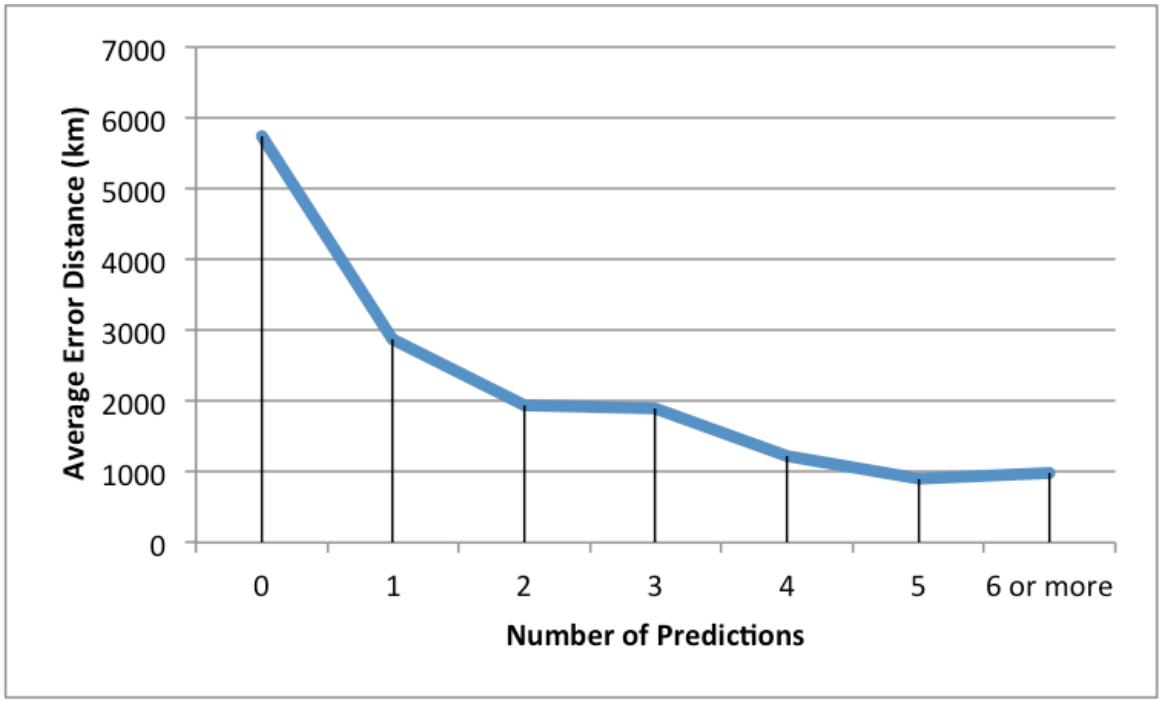

Figure 6.3. Average Distances Per Number of Predictions

Some of the existing techniques analyzed different types of datasets than our evaluation of GCL, which could be recognized as a threat to validity. However, because our analysis of GCL used a dataset without any type of filter except for English-language, we argue that this dataset is the most difficult to geolocate, yet it also represents the most real-world set of tweets. Ikawa et al. (Ikawa, Enoki, & Tatsubori, 2012) excluded tweets that were replies or retweets. GCL was able to geolocate retweets, which are copies of tweets from other users, so if they contain location information, it is likely not relevant to the author of the retweet, but to the original author. This characteristic makes retweets more difficult to geolocate than non-retweets.

The average distance from the final prediction to the actual geotagged location of all tweets in the dataset is 2561 km, which is quite large. However, as shown in Figure 6.3, the average distance depends on the amount of location information found in the tweet. In order to enhance the usability of GCL, we decided to include every tweet in our dataset, even if the tweet does not

contain usable location information. For the calculation of the average error distance of all tweets, GCL considers every tweet to have a default location of (0,0). Thus, tweets with no extracted predictions, such as tweets with no user location, no location data in the content, and a small number of friends, have the default location as the final prediction and contribute greatly to the high average error distance. If we exclude those tweets, the average error distance is 2295.94 km.

6.2. Evaluation of Geotopical Clustering Module

We performed several evaluations of the GeoContext geotopical clustering system. In several evaluations, we compare the results of GeoContext's topical clustering system against LDA, described extensively in Section 2.3. We describe all evaluations in the following subsections.

6.2.1. Evaluation I

In the first evaluation of the geotopical clustering module, we analyze the topic clusters provided by both implementations of GeoContext and compare the clusters to those produced by LDA, which is commonly used in other geotopical clustering research. In order to effectively evaluate the same tweets in our system and LDA, we used a set of streamed tweets from August 2015 for both techniques. We streamed the tweets through GeoContext's topical clustering first implementation and then ran the same tweets through the geographical clustering first implementation and LDA.

With both the TCGC implementation and the GCTC implementation, we discovered several topic clusters whose topics were trending on Twitter. The topics with the most number of tweets

Table 6.2. TCGC Results

Topic Num	Num Tweets	Extracted Topic	Example Concepts	Recommended Location
Topic 1	95	Celebrity Big Brother (UK TV show)	#CBB, TV, celeb housemates	London, UK
Topic 2	30	London, National Burger Day	#London, #NationalBurgerDay	London, UK
Topic 3	20	Celebrity Big Brother (UK TV show)	#cbb,	London, UK
Topic 4	18	Astrology	#Pisces, #Leo	Berlin, Germany
Topic 5	17	News	#news, #Iran	London, UK
Topic 6	16	Jhalak Dikhhla Jaa (Indian TV show)	#InjusticeToVivianDsena	New Delhi, India
Topic 7	13	MSG2 trailer release (movie)	#MSG2TrailerLaunch, Gurmeetramrahim	New Delhi, India
Topic 8	13	Market research/Business	Profit, business, forecast	London, UK
Topic 9	12	Leila de Lima (Philippine Secretary of State)	#DeLimaBringTheTruth	Manila, Philippines
Topic 10	11	School	#tipsforyear7s	London, UK

Table 6.3. GCTC Results

Topic Num	Num Tweets	Extracted Topic	Example Concepts	Geographical Cluster
Topic 1	14	Celebrity Big Brother (UK TV show)	#cbb, TV, celeb housemates	London, UK
Topic 2	76	Job Advertisements	#job	Washington, DC, USA
Topic 3	81	Job Advertisements	#job	Boston, USA
Topic 4	60	Celebrity Big Brother (UK TV show)	#CBB, TV	Manchester, UK
Topic 5	11	Celebrity Big Brother (UK TV show)	#CBB, TV	Sheffield, UK
Topic 6	14	Job Advertisements	#job	Los Angeles, USA
Topic 7	57	Celebrity Big Brother (UK TV show)	#CBB, TV	London, UK
Topic 8	16	Cameron Dallas	#followmecam	Brasilia, Brazil
Topic 9	18	Job Advertisements	#job	Los Angeles, USA
Topic 10	13	Job Advertisements	#job	Chicago, USA

consisted of tweets advertising job openings, followed by two topics consisting of tweets talking about 5 Seconds of Summer and Justin Bieber (popular musicians), respectively, and a topic consisting of tweets talking about football in Europe. As might be expected, however, these topic clusters are spread across large geographic areas, not clustered around one or several locations. Using our adapted TF-IDF algorithm, we extracted topic cluster recommendations for various locations using both the TCGC and GCTC approaches. We extracted the ten most populous topic clusters that were recommended to various locations and present these results for the TCGC approach in Table 6.2. We perform the same extraction for the GCTC approach and present the results in Table 6.3.

The TCGC results clearly show that different topics are important to different geographical areas. We converted geographical coordinates to city names, and the cities for which the topics are recommended are displayed in the rightmost column. By extracting topics relevant to different locations rather than simply the most popular topics, we can filter out topics that users may not care about. For example, in this evaluation GeoContext recommended tweets about a UK TV show to users in London, while tweets about an Indian TV show were recommended to users in New Delhi. Topics 1 and 3 both consist of tweets talking about a UK TV show. This is due to the fact that users were using different hashtags for the same TV show, and GeoContext was not able to recognize that the hashtags were related. However, both topics were recommended to users in the same location, London.

The results for the GCTC implementation are shown in Table 6.3. These results clearly indicate what topics are the most popular in different areas. We found that the geographical clusters were very well-defined and did not contain outliers. The topic clusters within each geographical cluster were also well-defined and contained tweets that were all closely related

Table 6.4. LDA Results (No Clustering)

Topic Num	20 Topics	50 Topics	100 Topics
Topic 1	I'm love srt don't amp good it's day people time follow great lol can't make today happy back you're work	I'm love srt don't amp good it's day people time follow great lol can't make today happy back you're work	I'm love srt don't good amp it's day people time follow great can't lol make today back you're work life
Topic 2	B***qualityrt greg kidding wwe ya'll families where's romance brick slowly cools noooo ipostpraksrt longtime punishment thee fleek honour receiver #arsenal	Hack laughed trips innovation #india treat continues Monmouth bliss ari berne #porn alispagnola dish milestone ers malibuselfies usc exhibition rap	# week makes #followmecam win happy years give morning awesome heart excited car real_liam_payne link nice cool talking football past
Topic 3	Happiness #nyc discount childhood recounts oxford confusing struggles elmasritrt quizthe favorites medicine cos punch chili would've trap horror broadcaster jared_carrabis	Engagement navy joey lingerie vibes ties slick peoples plastic cuffing snd bend nollywood konstantinos vid satan adwords bare brand-new threatened	Miserable aidancmorenort niece auction they'd berahino liam's shelter creative weeknd's #opportunity whut thee checking supplier programme pete teachers actively louis
Topic 4	Kills smiling #defiance potus how's supplemental coat deserves spain's fifthharmonyrt turned cycling s*** stressful subway harsh burnley suspension maya tuition	Acc deal money stupidest #perfect noooo edsa turnt vevort reporters golfer tart Oklahoma hopes tonite chin intro byrt dramatic faze	Standard hrt mentions ignores allinallbeautyrt eamaddennfl kpop greedy dummies bacon ruby sporting purse dudes ruining walsh platform tyler cultural wwe
Topic 5	Madison court ordered similar mode facing bus h*** details islands failing habits celebs hoverboard diff insta complain colin push recording	Undercover California Scottish follower hopeful lawsuit corners carry gabeturner backyard degree uptown students mail basically hpa onion Leverkusen bedrooms preferably	Tag tebow impossible hug insane thatsabinegirl #advertising bullet s beys complaining gateway unitekcollege split reds you're enjoying sight silver tickets murderer

conceptually. The largest topic clusters over all geographic clusters are displayed in Table 6.3. These topic clusters align with the largest topic clusters found in the TCGC implementation. However, although many of the largest topic clusters extracted are similar over different areas, we were also able to extract more location-specific events, such as topics about baseball games. As with the TCGC implementation, by recommending topics that are trending in different geographical areas, GeoContext can provide more relevant information to users in those areas, rather than topics that are important in other areas of the world.

We used Mallet[24] to run LDA on the same set of tweets with varying number of topics. First, we ran LDA on the set of tweets with no prior clustering or filtering. These LDA results are shown in Table 6.4. Using GeoContext's geotopical clustering algorithm, we discovered about 50 topic clusters that are of significant size (more than 10 tweets), so we ran LDA with 20, 50, and 100 topics. Due to space constraints, we display only the 5 highest-weighted topics from LDA in Table 6.4 for each run. As is evident from the results, LDA produces topics that are much less defined than GeoContext. We believe this is due to the fact that, although LDA removes stop words, many other words such as "I'm," "cool," and "nice" are not removed. These terms are common, therefore they show up within the produced topics, but they do not add significant meaning to a tweet.

Next, we clustered the tweets using DBSCAN prior to running them through LDA. Each geographical cluster was considered a document for input to LDA. The results are shown in Table 6.5. The topics from this approach are more defined that the non-clustered results. For example, topic 2 for 20 topics, topic 3 for 50 topics, and topic 4 for 100 topics all contains terms regarding popular musicians. However, the resulting topics are still much less defined compared

[24] http://mallet.cs.umass.edu/

Table 6.5. LDA Results (Clustering)

Topic Num	20 Topics	50 Topics	100 Topics
Topic 1	I'm love it's don't good amp time people day great lol can't today back #job work life happy night make	I'm love don't it's amp time day good great can't lol today back work you're night life that's make people	I'm love don't it's amp day people great can't today back you're life happy that's I've live home watch year
Topic 2	Follow love camerondallas harry_styles justinbieber #followmecam sos real_liam_payne niallofficial day hey happy cam carterreynolds make Louis_tomlinson you're luke_brooks smile nashgrier	Video people good free check hashtag youtube photo follow god world hope music person years happy news heart city hours	Good time video man feel free youtube big photo music person make years feeling friends top full times news real
Topic 3	Sosfamily tha nowplaying sound Denmark stories icemoon active break yep edit #dkshame staff Australia success task split hii japan officer	Love follow camerondallas harry_styles justinbieber sos #followmecam day real_liam_payne happy make cam niallofficial Louis_tomlinson hey carterreynolds smile nashgrier birthday mtv	#job lol work s*** game we're time good school hate job latest girl click hot play weekend #hiring talk high
Topic 4	Seattle road imam fancy hero jeans portrait manhattan French led #nfl brick skills wedding education state #autocar busty falls #seattle	Posted photo facebook storm silence psychologicai ignore don hero morning #aldubgettingcloser crochet notice grand rosymcmichael cherrycrush hub tablecloth values girlideas	Love follow harry_styles camerondallas justinbieber sos #followmecam real_liam_payne make niallofficial Louis_tomlinson happy day cam carterreynolds smile hey birthday photo nashgrier
Topic 5	Blessed sets drivers shopping empty legend farm lies longer ooh tuition Puerto pair earth solo leader deal studios expecting raining	Wind temperature rain humidity hpa kit ops barometer rising dry grow it's sold flying challending theory wsw Erika drugs wishing	Commercial gear cars campaign playoffs topic bulls*** #art delays jonahmarais hes ignoring hiring smiles freeze techcrunch lane countdown overheard turkey

Table 6.6. Keyword Query Results

Topic Num	Keyword Extraction	No Keyword Extraction
Topic 1	Transportation jobs	#traffic
Topic 2	Travel	Johor Causeway traffic
Topic 3	Road closures/accidents	Manila traffic
Topic 4	Items for sale	Portland road closure
Topic 5	UK Football	#driverdiaries

to topics discovered from GeoContext. Terms related to jobs and hiring are spread over several topics. Also, many topics that were discovered with GeoContext are missing with LDA. For example, there are no topics that include Celebrity Big Brother, which was one of the most popular topics discovered by GeoContext. Interestingly, for all 6 LDA runs, the highest-weighted topic is very similar in each run. This topic consists mainly of common terms used in tweets.

We also analyzed our results from the GeoContext keyword query system. We set the keyword parameter as "traffic" in the keyword query system. The keyword query system expands the term "traffic" to other terms such as "congestion," "travel," and "transportation." We also initialized a stream of tweets without the keyword expansion. We extracted concepts and keywords from 2000 tweets streamed from Twitter from streams both with keyword expansion and without keyword expansion. The five topic clusters with the most tweets are displayed in Table 6.6 for both streams. As shown, the topics in the stream with keyword expansion were able to discover multiple topics related to all types of traffic, shown in topics 1 through 4. Topic 5 was discovered due to users tweeting about a football team being in the "top

Table 6.7. Location Query Results

Topic Num	Num Tweets	Extracted Topic	Example Concepts
Topic 1	153	hiring	#job, hiring
Topic 2	18	weather	rain, weather forecasting
Topic 3	16	WWE NXT	Sasha Banks, #NXTtakeover
Topic 4	10	WWE NXT Brooklyn	#NXTtakeoverbrooklyn, WWE
Topic 5	7	University of Alabama	#rolltide, MDB

flight," and "flight" was a term to which "traffic" was expanded. Although the results from the keyword expansion system discovered some tweets not completely related to traffic, the results from the system with no keyword expansion are much less defined. For example, Topic 1 includes any tweets with the hashtag #traffic, which included tweets from road construction to driving website traffic.

Finally, we analyzed our results from the GeoContext location query system. We set the coordinate parameters as the geographical coordinates of the University of Alabama. We extracted concepts and keywords from 6096 tweets streamed from Twitter. The top 5 most populous topic clusters are displayed in Table 6.7. Topics 1, 3, and 4 are similar to some of the most populous topical clusters discovered from the geotopical clustering system. However, perhaps the most interesting result from the location query system is topic cluster #5, which consists of tweets talking about the marching band preview night, a local event occurring on the

campus of the University of Alabama. This event was not highly publicized even on the University of Alabama calendar, showing that GeoContext can be a useful tool for clustering what people are tweeting about in an area and discovering new topics that may not be able to be discovered elsewhere.

6.2.2. Evaluation II

In the second evaluation, we compared both approaches of the system (i.e., topical clustering first and geographical clustering first) along with LDA, which is commonly used in other geotopical clustering research. To perform this evaluation, we analyzed the topic clusters provided by GeoContext and compared the clusters to those produced by LDA. In order to effectively evaluate the same tweets in both implementations of GeoContext and LDA, we used a set of streamed tweets from February 2016 for both methods in this particular evaluation. From this dataset, 362,419 total tweets were used.

We performed the GeoContext evaluation with several different parameters. First, we set the keyword query system to track the keywords *weather* and *traffic*. Next, we tracked two different locations: Tuscaloosa, AL and New York City. We chose these locations to compare cities of very different populations. Lastly, we left the keyword and coordinate filters empty and discovered geographical topics within the entire tweet stream.

For each set of parameters, we display the five most relevant and populous results from GeoContext for both the TCGC and GCTC implementations. Due to the large amount of tweets and metadata within the topic clusters, we chose to display selected tweets from each topic, separated by semicolons, within each cell. For the keyword tables and no filter table, each cell also contains the geographical location to which that topic is recommended. We provide both

Table 6.8. “Traffic” Keyword Results

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
TCGC	Accident on E2 Lebuhraya Utara Selatan #kltu still delaying traffic 28m more than usual; traffic is slow from the 9th mile toll; left lane is being obstructed (3.083,101.65 - Kuala Lumpur, Malaysia)	Disabled vehicle, right lane blocked in #Hollywood on I-95 NB near Sheridan St (25.787,-80.224 - Miami, FL, USA)	Accident, right lane blocked in #Dallas on 35E SB at Lp12 Walton Walker; merge, stop and go traffic (32.866,-96.896 - Dallas, TX, USA)	#M65 Delays Near J8 eastbound caused by congestion (51.5,-0.116 - London, UK)	Traffic Update: As of 12:11 PM, Light to moderate traffic; Garcia-Xavierville, Aurora #mmda (14.583,120.966 - Manila, Phillipines)
GCTC	Stopped traffic in #Pinson on Hwy 75 NB, delay of 1 min; Slow traffic in #Alabaster on Cahaba Valley Road SB between Highway 52 and US 31, delay of 3 mins (33.586,-86.697 - near Birmingham, AL, USA)	Closed due to accident in #Harrisburg on I-81 NB between Linglestown-Paxtonia and Manada Hill; Stop and go traffic in #Harrisburg on I-81 between the 83 split (40.239,-76.934 - Harrisburg, PA, USA)	Accident in #Exton on Rt-100 SB before Pottstown Pike; right lane blocked in #Exton (39.95,-75.166 - Philadelphia, PA, USA)	#LagosMarathon ; I can’t imagine how a state that deals with the worst kind of traffic decided to hold #Lagos Marathon and block roads (6.453,3.395 - Lagos, Nigeria)	Accident in #UDistrict on I-5 NB near 45th St. #traffic; right lane blocked in #UDistrict (47.661,-122.322 - Seattle, WA, USA)
LDA	man crotcheskill closed today skip waze drivers yyc stop sanmobkandar peteboyle local blog lights air police presence promotion alexa hwy	jam road stay pledging manager til wardens facebook kpk city exit business delays usual walk don show helped	seo website jeff pond time glass nearby bridge content yyc documents highway trafficsocial driving exchange n asa officer tips details	amp social work lane marketing reactive league update lagos team leads jams left fans visit event bspodnetwork lekki totaltrafficbhm safety	supporting alert congestion controllers state massive start beat info send captures signal adds media week east projects find kltu council

the coordinates in latitude/longitude format, as well as a text form of the location. The location filter tables do not include a location, because all tweets and topics are already centered around a location. For the LDA executions, the five most relevant topics are displayed.

Table 6.8 shows topics obtained using the keyword "traffic." GeoContext was able to extract traffic information for several cities across the world. The major benefit to GeoContext's traffic analysis over other mediums that provide traffic information is the level of detail. For example, in the TCGC implementation, Topic 1 includes the exact mile number at which the traffic begins, and Topics 1, 2, and 3 contain the lane closures. Topics discovered from GeoContext also contain the reason for the traffic. Topics 1, 2, and 3 include traffic information due to an accident or disabled vehicle, while traffic in Topic 4 is due to congestion. Also, some topics contain delay information. For example, in the GCTC implementation, Topic 1 includes the amount of time the delay is expected to take. This kind of detailed traffic information is useful for drivers who can make a more informed decision about whether to take a detour or a different route. Topic 4 in the GCTC implementation is not directly related to road closures or car accidents, but because it is still a traffic-related topic, we do not consider it to be out of scope. As shown, LDA is able to extract terms related to traffic and cars, but the resulting topics leave out much of the information that is important to users, such as the delay time, road names, and lane closures, because these terms are not as common within the tweets as more general traffic terms such as "traffic," "driving," or "hwy."

GeoContext is able to cluster together social media posts to find trending topics that are better refined and focused than topics found with traditional topic models such as LDA. Topics discovered with LDA tended to have many terms unrelated to the overall topic, while topics discovered with GeoContext contained terms that were more relevant.

Table 6.9 “Weather” Keyword Results

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
TCGC	Gale warning issued; Small Craft Advisory (38.883,-77.016 - Washington, D.C., USA)	More snow! Might as well jump in face first; Here’s the #Ottawa forecast (45.420,-75.69 - Ottawa, Canada)	How the army rescued a hero from #Siachen snow; Lance Naik Manamanthappa Koppad’s condition deteriorating (28.6,77.2 - New Delhi, India)	Beautiful peak at Spring today in #Seattle; Temperature 59degrees few clouds (47.609,-122.333 - Seattle, WA, USA)	Storm-battered Cruise Ship Returning to Homeport; Weather Forecast Wasn’t ‘Anything Near’ What Happened (38.883,-77.016 - Washington, D.C., USA)
GCTC	Special Weather Statement issued February 09 by NWS (38.883,-77.016 - Washington, D.C., USA)	Let’s now have a squizz at the west coast: Perth’s weather for the week; More snow for some before Wednesday morning (-35.3,149.116 - Canberra, Australia)	56.7F - Humidity: 38% - Wind: 7.6mph (29.875,-92.218 - near New Orleans, LA, USA)	Winter Weather Advisory issued by NWS; #WxPA until February 10 at 1:00 AM EST (38.883,-77.016 - Washington, D.C., USA)	Hazardous Weather Outlook (HWO) #WXMeteorology (38.883,-77.016 - Washington, D.C., USA)
LDA	weather channel updates cold wind pressure est wednesday km/h closed issued tomorrow precip temperature hpa bad beach makes coast perfect	weather updates channel today humidity warm rain temp schools inclement conditions nws love hate hum cloudy don rising hot summer	snow forecast day advisory tonight good west fair light pluto special steady live home feel work effect cancelled	winter mph current degree school days news back bar chill statement alert clear feb dew week stay party forecast	due amp visibility https county morning sunny inches reputation yourgoddesssss current storm giveaway coastal change stop nature man falling thing

Table 6.9 shows topics obtained using the keyword "weather." GeoContext was able to extract weather information for several cities across the world. Like the traffic keyword evaluation, when compared to LDA, the topics discovered using GeoContext contain a much greater level of detail. The topics found from the LDA evaluation mostly contain general terms related to weather, which is not particularly useful to a user desiring weather information for a specific location. TCGC discovered two topics that do not contain information related to weather conditions: Topics 3 and 5. However, the topics contain tweets about current events related to weather, so we do not consider those to be outliers. With topics obtained from GCTC, we noticed that many topic clusters that contained weather information for the United States were centered around Washington, D.C. This is due to the fact that the National Weather Service is located in Washington, D.C.; thus, the tweets regarding weather for the entire United States were centered there. In future work, we plan to analyze the content of the topic cluster to determine which location for which the topic cluster is relevant.

Table 6.10 shows topics obtained using the location set as Tuscaloosa, AL, USA. The 50th NFL Super Bowl was occurring during the period of evaluation, so topics related to the game were discovered. Topics related specifically to Tuscaloosa were also discovered, such as a popular advertised Valentine's Day party that was retweeted multiple times. A topic related to nearby weather is displayed in the results for both TCGC and GCTC, showing that filtering by location can be useful in discovering local events. Lastly, the TCGC implementation was able to find a topic containing tweets related to a University of Alabama (located in Tuscaloosa, AL) basketball game, which again shows that GeoContext is able to discover events relevant to people in different geographical areas.

Table 6.10. Location Tuscaloosa Evaluation

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
TCGC	Ripley SW Limestone Co. Rain Today 0.00in #alwx; What's the latest on snow for North Alabama Monday nite?; A super sunset on Super Bowl night #valleywx #SuperBowl; Stewart takes flight to get the Panthers on the board	Cam Newton wins NFL Most Valuable Player Award! #SB50; I believe I am becoming a #CamNewton Fan; Denver Carolina Official prediction	Pre-Valentines Day Party Hosted by the Ques + Kappas Saturday @ Museum Bar	There's a s*** storm coming clintons way and I would not want to face some of these guys in a debate; I do believe what Jeb Bush said about being a good commander-in-chief; #GOPDebate Rubio takes a flamethrower to the Democrats in answering the abortion question	I was watching a basketball game; AlabamaMBB leads 57-41 #RollTideBasketball; Glad to see #PeytonManning get another superbowl win
GCTC	Ripley SW Limestone Co. Temp 58.8 Wind:5.4mph Steady Rain Today 0.30in. #alwx #valleywx	#SB50 Good game so far; These are the best two teams playing for #SB50	If nothing else, I do believe what Jeb Bush said about being a good commander-in-chief; Rubio takes a flamethrower to the Democrats in answering the abortion question		
LDA	super bowl good broncos game great happy time party tonight feel bar npre-valentines nastyvalentine kappas miss bad damn congrats	don't panthers amp back life superbowl hate didn s*** football big f*** real money wonderful school taking count morning today	love peyton win today man manning make newton saturday night ques museum season birthday minute play tweet tomorrow lmao work	cam day hope baby bama home watching give wanna aint song run making playing ready wouldn't state hair sbvote hard	lol people alabama year ajcib thought guess made makes national time winning rubio smh photo wait video perfect bout forgot

In this case, the GCTC implementation only discovered three events total. It was not able to create a topic cluster containing information about the basketball game, unlike the TCGC implementation. However, the topics discovered by GCTC were well-defined and contained tweets all related to one concept. TCGC was able to discover more topics, but the topic containing tweets about the University of Alabama basketball game also contained some tweets related to the Super Bowl. This may have occurred due to both concepts being related to sporting events. The LDA run resulted in topics that were very mixed conceptually, with terms related to the Super Bowl, the local party, and the debate located in the same topics. LDA also was not able to discover topics that contained any terms related to the local weather, while both TCGC and GCTC were able to identify weather topics.

Table 6.11 shows topics obtained using the location set as New York City, NY, USA. GeoContext was able to discover several topics relevant to people located in New York City, such as Topics 3 and 5, which include tweets about the Rangers and Knicks, sports teams located in New York. GeoContext is able to extract the fact that people in New York City would care more about these sports teams than people in other locations. Similar to the Tuscaloosa filter results, GeoContext also discovered topics related to the political debate. The GCTC implementation resulted in a topic related to New York Fashion Week, which is a large event located in New York City. Topic 5 in the GCTC implementation contains tweets that are not highly related on the surface, but after manual examination, the tweets in this topic cluster are mainly related to tourism and photography.

Table 6.11. Location New York City Evaluation

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
TCGC	These three powerful #entrepreneurs are part of my #BlackHistoryMonth plugs; So honored to join this celebration of art, activism, & progress! #BlackHistoryMonth	Another #losewithcruz supporter diverting from issue; Woke up from catnap to sad news about @WWEDanielBryan	The #NYRangers are forced to play a TEAM game without 2 of their best; Good evening at the hockey #cheapseats #nyr	Of course racist Republicans cause the #FlintWaterCrisis then blame #PresidentObama for poisoning the water; Watching @HillaryClinton appearing like a leader	Wow! #NYKnicks Head Coach Derek Fisher was just fired.; Kurt Rambis takes over as interim HC for the #Knicks
GCTC	Off-ramp reopened in #EastRutherford on Rt-3 EB at Service Rd.; stopped traffic back to 61st St, delay of 19 mins	#snowing but not sticking\#nyc; My niece and nephew from the Philippines having a ball in the snow	Home sweet #giuliettanewyork #nyfw; New York Fashion Week here we come!	Still running high off #birthday #weekend fumes; It was a happy Birthday to this guy!	What a view #newyorkcity; #worldtradecenter #lower manhattan; #instagramNYC; #nyc @ Winter Village At Bryant Park Ice Rink
LDA	love people york back today nyc night man f*** thebachelor show tonight god school miss real guys big gonna	amp lol time day s*** raw hope lmao hate home week omg thing free b**** play finally friend hell mom	don good happy life make great year feel game bad stop girl morning things give doesn't guy n**** knicks makes	work birthday f***ing made a** didn't watch damn wanna live black amazing true team girls baby super won yeah start	olivia don't follow white weekend heart rangers art move feels wtf clinton news d*** calling rose tap facebook matt

With the location query evaluations, there is the least amount of difference in topics between the TCGC implementation and GCTC implementation. This is likely due to the fact that, because the stream of tweets is coming from the same geographical area, the geographical clustering portion of GeoContext is less relevant. Because the topical clustering portions are the same between the two implementations, the topics are similar. However, because the stream still passes through the geographical clustering step, and there is a slight variation in topics, we include results from both implementations in this discussion.

An interesting observation is the difference shown in political topics between the two locations. During the evaluation period, the stream captured a Republican debate. Unsurprisingly, the topic concerning the debate from the Tuscaloosa, AL, stream is much more favorable towards the candidates than the topic concerning the debate from the New York City stream. This reveals another possible use case for GeoContext as a tracker for how people feel about current events such as political affairs.

Finally, Table 6.12 shows topics obtained using the stream with no keyword or location filters. In this evaluation, GeoContext was able to discover any type of topic from any location. Several of the topics clearly display how protests or debates are geographically located. For example, Topic 2 from TCGC and Topic 1 from GCTC contain tweets about a political event occurring in India regarding a university. The topics show that there are many tweets calling for the shutdown of the university. Also, Topic 5 that was discovered from GCTC contains tweets protesting against Monsanto, an agriculture company that has had a significant role in creating genetically modified food. The company has been linked on Twitter to the recent Zika virus. Both of these examples show that GeoContext can extract important topics from a social media stream that shows the political and social leanings of people in different geographical locations.

Table 6.12. No Filter Evaluation

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
TCGC	Got my face contoured and hair slicked back for an upcoming #TryGuys video (3.425,-76.528 - Cali, Colombia)	Why should taxpayers pay for such an institute that's nurturing anti-national professors and student?; It's high time to take action against JNU #ShutDownJNU (28.6,77.2 - New Delhi, India)	Exxon knew about Climate Change Almost 40 Years Ago #science (3.425,-76.528 - Cali, Colombia)	#IfIHadTrumps Money I would give artists money so they can continue making work that can change our culture; Something even money just can't buy (3.425,-76.528 - Cali, Colombia)	#Ascendant #MediumCoeli for Berlin, DE for now; #Libra #Astrology (52.516,13.383 - Berlin, Germany)
GCTC	Why should taxpayers pay for such an institute that's nurturing anti-national professors and students? #ShutDownJNU; It's high time to take action against JNU (28.6,77.2 - New Delhi, India)	What is love for Maine Mendoza?; AldubAcronyms : Win or lose, ADN got your back! #VoteMaineFPP (14.583,120.966 - Manila, Philippines)	#BestFanArmy Directioners #iHeartAwards; ; #VoteDirectioners UK (-34.603,-58.381 - Buenos Aires, Argentina)	Exxon Knew about Climate Change Almost 40 Years Ago (3.425,-76.528 - Cali, Colombia)	MarchAgainstM StopMonsanto Thanks for raising awareness about #TPP, just shared your tweet (3.425,-76.528 - Cali, Colombia)
LDA	love amp don weather day good life today video happy https follow free make win girls s*** man black	people kca channel updates girl live watch real feel things made give hope ready stop wanna vote guys world	time lol year night youtube god show home baby kanyewest years valentine aldub money read friend stay february fun	back f*** didn't amazing doesn't zaynmalik talk long true white listen talking movie lil iphone room water sexy mainedcm	great person damn family mind sex checked post tickets hit friday hear sleep text job leave periscope games bae

As displayed in Table 6.12, LDA was not able to discover any of the topics found by GeoContext. This is due to the fact that LDA is unable to filter out terms that are not necessarily stop words, but not related to the overall concept of the tweet. Because of this, the topics resulting from LDA contain terms such as "good," "didn't," "today," and "love," which do not add meaning to a topic cluster.

Overall, more topics were discovered by the TCGC implementation. However, the topics discovered by the GCTC implementation were more defined and contained tweets very related to one specific topic. This shows that both implementations could be useful in discovering geographical topics in a social media stream.

We conclude this evaluation by mentioning some of GeoContext's statistics. GeoContext works completely in real-time. GeoContext receives approximately 18 tweets from the Twitter stream per second. Out of those, it is able to geolocate between 1 and 4 tweets per second. This limitation is not due to the processing time, but rather to the percentage of tweets from the stream that contain location information. GeoContext then extracts topics and keywords and adds the geolocated tweets to a topic cluster within an average of 1 second. These statistics show that GeoContext is implemented as a truly real-time system that can be used to extract geographical topics that are relevant to various users in a real-world system.

6.2.3. Evaluation III

In the third evaluation, we tested the accuracy of the resulting clusters from GeoContext with various configurations. Specifically, we tested four different threshold values present with GeoContext: the similarity score threshold, the time value between prunings of the topic clusters,

Table 6.13. Varied Experimental Values

Threshold Value	Possible Values
Similarity Score	0.2, 0.4, 0.6, 0.8
Pruning Time	15 min, 30 min, 12 hrs, 24 hrs
Stale Cluster Time	24 hrs, 48 hrs
TF-IDF Threshold	0.2, 0.4, 0.6, 0.8

the time threshold at which a topic cluster is considered "stale," and the adapted TF-IDF threshold value.

When GeoContext performs the topical clustering step, as described in Section 5.1.1, it calculates a similarity score between two tweets by taking the average of the relevance scores of any matching concepts or keywords between the tweets. Tweets that have a score over a threshold value will be clustered together into the same topic. This threshold value is the first that we evaluate.

Also, when GeoContext performs geographical analysis on the topics produced from the topical clustering step (as described in Section 5.1.2) in order to determine whether each topic is centered at a location or spread across a larger region, GeoContext uses TF-IDF, a statistic used in natural language processing that shows how important or meaningful a term is to a document (Sparck Jones, 1972). GeoContext calculates an adapted version of the TF-IDF statistic in order to determine how important a location is to a topic cluster. If the TF-IDF value for a location

and a topic cluster is above a threshold, then the topic cluster is considered to be centered at that location. This threshold value is the second that we evaluate.

After a period of time, the collection of topic clusters is pruned to remove any clusters that have not had tweets added recently, or "stale" clusters, as described in Section 5.1.1. If "stale" clusters are not removed, the storage and analysis of so many tweets can greatly affect performance. The length of time between pruning sessions is the third threshold value evaluated. The length of time between the last tweet added to a cluster and the cluster becoming "stale" is the fourth threshold value evaluated.

In this empirical evaluation, we tested GeoContext in order to determine the accuracy of these threshold values used in the clustering process. For this evaluation, we collected a dataset of 14,817 tweets throughout April 2016. We clustered the tweets using GeoContext with 48 different configurations and then tested the TF-IDF threshold with 4 different values separately. Due to rate limits imposed by AlchemyAPI, we obtained the concepts and keywords for all tweets, as well as geolocated the tweets, prior to running the experiment. All tweets were then clustered using GeoContext.

We first present the evaluation of the topical clusters produced by GeoContext. We evaluated 48 different configurations of the three threshold values that affect the topic clusters: the similarity score value, the pruning time value, and the stale cluster time value. Table 6.13 shows the possible values for each of these threshold values. Table 6.14 shows the resulting five largest topic clusters for each configuration. The concepts and keywords that matched within the similarity score calculations are shown for each topic cluster. Because the matching concepts and keywords are the factor that makes tweets within the cluster similar, we believe that they give an accurate representation of the overall topic of the cluster.

Table 6.14. Topic Clusters With Various Configurations

Configuration (Sim. Score, Pruning time, Stale time)	*1*	*2*	*3*	*4*	*5*
0.2, 15 min, 24 hrs.	Weather, climate change, people	English-language films, American films, 1990s music groups	Apple Inc., bid	Friends, time	Thanks
0.4, 15 min, 24 hrs.	People	English-language films, American films	Friends, time	Retweets	Thanks
0.6, 15 min, 24 hrs.	People	English-language films	Time	Retweets	Thanks
0.8, 15 min, 24 hrs.	English-language films	People	Lol	Retweets	Thanks
0.2, 30 min, 24 hrs.	Weather, people	Bid, dress	Guys, thanks, mom	English-language films, American films	Thermodynamics, time
0.4, 30 min, 24 hrs.	Things, people	English-language films	Friends, time	I'm, retweets	Bid
0.6, 30 min, 24 hrs.	People	English-language films	Time	Retweets	Thanks
0.8, 30 min, 24 hrs.	English-language films	Retweets	People	Lol	Thanks
0.2, 12 hrs, 24 hrs.	Weather, people	Bid	English-language films, American films	Thermodynamics, time	Internet slang, I'm, guys, retweets
0.4, 12 hrs, 24 hrs.	People	English-language films	Time	Life, retweets	Retweets
0.6, 12 hrs, 24 hrs.	People	English-language films	Time	Life, retweets	Retweets
0.8, 12 hrs, 24 hrs.	English-language films	Retweets	People	Lol	Thanks
0.2, 12 hrs, 24 hrs.	Weather, people	Bid, dress	English-language films	Thermodynamics, time	Internet slang, retweets
0.4, 12 hrs, 24 hrs.	Things, people	English-language films	I'm, retweets	Bid	Life, photography
0.6, 12 hrs, 24 hrs.	People	English-language films	Time	Retweets	Thanks
0.8, 12 hrs, 24 hrs.	English-language films	Retweets	People	Lol	Thanks

Table 6.14. (cont.) Topic Clusters With Various Configurations

0.2, 15 min, 48 hrs.	Bed, thermodynamics, people	Cause, heart, English-language films	Bid	Life, retweets	Time
0.4, 15 min, 48 hrs.	People	English-language films	Retweets	Time	Thanks
0.6, 15 min, 48 hrs.	People	English-language films	Retweets	Time	Thanks
0.8, 15 min, 48 hrs.	English-language films	Retweets	People	Lol	Thanks
0.2, 30 min, 48 hrs.	Bed, thermodynamics, people	Cause, English-language films	Bid	Life	Time
0.4, 30 min, 48 hrs.	People	English-language films	Time	Life, retweets	Retweets
0.6, 30 min, 48 hrs.	People	English-language films	Retweets	Time	Thanks
0.8, 30 min, 48 hrs.	English-language films	Retweets	People	Lol	Thanks
0.2, 12 hrs, 48 hrs.	Bed, thermodynamics, people	Cause, heart, English-language films	Bid	Life	Time
0.4, 12 hrs, 48 hrs.	People	English-language films	Time	Life, retweets	Retweets
0.6, 12 hrs, 48 hrs.	People	English-language films	Retweets	Time	Thanks
0.8, 12 hrs, 48 hrs.	English-language films	Retweets	People	Lol	Thanks
0.2, 24 hrs, 48 hrs.	Bed, thermodynamics, people	Cause, heart, English-language films	Bid	Life	Time
0.4, 24 hrs, 48 hrs.	People	English-language films	Time	Life, retweets	Retweets
0.6, 24 hrs, 48 hrs.	People	English-language films	Retweets	Time	Thanks
0.8, 24 hrs, 48 hrs.	English-language films	Retweets	People	Lol	Thanks

Similarity score: Because relevance scores from the Alchemy API Concept Tagging and Keyword Extraction APIs range from (exclusive) 0 to 1, the similarity score value also ranges from (exclusive) 0 to 1. We decided to choose sample values of 0.2, 0.4, 0.6, and 0.8 so that the range is covered in evaluation.

As seen in Table 6.14, it is clear that the lower the similarity score, the broader the concepts within the topic cluster. This is not surprising, due to the fact that more keywords and concepts contribute to the similarity score of the tweets if there is a lower similarity score threshold. With a lower threshold value, several of the topic clusters contain more than one topic that is not related. For example, the (0.2, 15 min., 24 hr.) topic cluster contains tweets about both weather and people. These tweets are separated into two distinct clusters with the higher similarity score threshold values.

Interestingly, many of the five largest topic clusters for each similarity score value are the same or very similar, even as the pruning time value varies. This correlation suggests that the similarity score value and the stale cluster value are the strongest in influencing the topic clusters.

Pruning Time: We chose the values 15 minutes, 30 minutes, 12 hours, and 24 hours for the time between pruning sessions. We believe that waiting longer than 24 hours will keep too many old topics in the system, since many trends in Twitter are fairly short-lived. Also, there are often so many topic clusters after 24 hours that if old ones are not removed, so many tweets are analyzed within the topical clustering step that performance is affected.

As displayed in Table 6.14, there exists basically no discernable difference between the clusters produced by the various pruning time values, while holding the similarity score threshold value and the stale cluster threshold value constant. This indicates that the pruning

time value does not have a discernible effect on the topic clusters. Therefore, the configuration of the pruning time threshold can be determined by any other means desired.

Stale Cluster Time: We chose the values 24 hours and 48 hours for the time threshold at which a topic cluster becomes "stale." This value represents the time between the addition of the last tweet to the cluster and the time at which the cluster becomes "stale" and should be removed. We believe that a value shorter than 24 hours would result in topic clusters being removed while they are still relevant, because trends on Twitter tend to occur over at least one day.

There is only a slight difference between the topic clusters produced by the 24 hour and 48 hour values. Unexpectedly, there are a few topics that appeared with the 24 hour value that did not appear in the 48 hour value clusters. For example, "Internet slang," "photography," and "guys" were all matching concepts or keywords that appeared in topic clusters with the 24 hour value. Prior to the experiment, we expected the 48 hour value clusters to have more range in topics because more clusters are kept, because the clusters are allowed to be older. However, the additional concepts found in the 24 hour value clusters may be overshadowed by the larger group of clusters with the 48 hour value.

Overall, the evaluation shows that the similarity score between tweets should be higher in order to produce topic clusters that consist of one topic each. Also, the time at which topics are pruned does not have any discernible effect on the topic clusters. Lastly, the time at which a topic cluster becomes stale produces more concepts within topic clusters with a lower value.

Adapted TF-IDF Threshold: We also evaluated the adapted TF-IDF statistic threshold value. This is the value at which a topic cluster is considered to be centered at a geographical location.

Table 6.15. Topic Clusters With Recommended Locations

Topic	TFIDF Value	Topic Cluster	Location
Aftermath of Brussels attack	0.2	RT @WPXI: Local prayer vigil held for victims of terrorist attacks in Brussels, Pakistan : s: t.co , Modi leads attack on Nuke terror at global summit, warns of state actors working with terrorists : , RT @USATODAY: Brussels Airport partially opens 12 days after terror attack : via @usatoday , Brussels Airport Partially Reopens 12 Days After Terror Attack: The Brussels airport is expected to restart fl...	38.88333333333333,-77.01666666666667
Spam tweets	0.4	RT @jedydynysem: Selfies you weren't meant to see Shhh!!, RT @jedydynysem: Selfies you weren't meant to see Shhh!!RT @jedydynysem: Selfies you weren't meant to see Shhh!!, RT @jedydynysem: Selfies you weren't meant to see Shhh!!, RT @jedydynysem: Selfies you weren't meant to see Shhh!!, RT @jedydynysem: Selfies you weren't meant to see Shhh!!	38.88333333333333,-77.01666666666667
None	0.6	None	None
None	0.8	None	None

The geographical analysis is a unique aspect of GeoContext, in that it can reveal topics that are specific and important to a geographical location. In this dataset of tweets taken from April 2016, there were tweets containing information about events occurring after the terrorist attack in Brussels, Belgium, in March 2016. We were specifically interested in whether GeoContext could discover these tweets as a topic. This type of information can reveal the opinions of people in different locations about a large worldwide event. Discovering these tweets as a topic can also show that GeoContext is able to consolidate tweets about a certain topic into one cluster.

Grouping the tweets can assist anyone performing social media analysis about the event, from news agencies to individuals.

Table 6.15 displays the largest resulting topic cluster that has a recommended location for each adapted TF-IDF threshold value. This means that GeoContext considers the topic cluster to be centered at that geographical location. As shown, with an adapted TF-IDF value of 0.2, a topic cluster consisting of tweets about the aftermath of the Brussels attack was revealed. The recommended location for this topic cluster was Washington, D.C., which is not surprising as the event is related to national security and therefore the topic contains many tweets from news agencies and government programs located in Washington, D.C.

Also shown in Table 6.15, the adapted TF-IDF value of 0.4 was not able to reveal the Brussels attack topic. Rather, this value resulted in topics that contained spam tweets. We believe that the prevalence of spam topics with this value was due to the fact that spam tweets generally come from a similar location, and the spam accounts simply post retweets from each other. Due to the high volume of tweets being retweeted by the spam account, a larger topic cluster was created, and because the tweets all come from the same location, GeoContext considered the topic to be centered at that location.

Lastly, as displayed in Table 6.15, the threshold values of 0.6 and 0.8 do not result in any topic clusters being centered at any location. These threshold values are simply too high for any geographical locations to be discovered as meaningful to a topic cluster.

Overall, it is clear from this evaluation that the adapted TF-IDF threshold value that is able to produce topic clusters such as the Brussels attack aftermath that are geographically centered is 0.2. It is clear that a higher value adapted TF-IDF value requires almost all tweets within the topic cluster to be at one specific location, rather than more slightly spread out.

Because GeoContext is intended as a way to provide contextual information about temporal events, it runs in real-time. As mentioned previously, the Gardenhose variety of the Twitter stream is estimated to provide about 15% of the public Twitter stream, which equates to approximately 18 tweets per second. In our evaluation, since the tweets were pre-geolocated and concepts and keywords were pre-extracted, we were able to determine the fastest possible time that GeoContext is able to calculate similarity scores and cluster tweets. The clustering process occurs at an average rate of 350 tweets per second. Because the clustering process occurs at a faster rate than the rate at which GeoContext can receive tweet objects from Twitter, it is clear that GeoContext is able to work in real-time and process tweets as they come in.

6.2.4. Evaluation IV

In this evaluation, we performed a second evaluation of both GeoContext and LDA as methods for discovering topics within a social media stream. We utilized two datasets for evaluation. The first dataset from (Aiello, et al., 2013) contains tweets from the 2012 United States presidential elections. The second dataset from (Zou, Fekri, & McLaughlin, 2015) contains tweets that are categorized into various rumors and truths. These datasets were chosen as a representation of topics likely to be of interest to users.

For both datasets, we compared results from GeoContext and LDA against ground truth topics that are included with both datasets. The ground truth topics for the first dataset are keywords and headlines that were extracted from mainstream media reports about the elections. The ground truth topics for the second dataset are tweets clustered into the rumor and truth topics.

Table 6.16. Example Elections Dataset Topics

bernie sanders win wins won call called calling projecting project projects projection hold held senate senator vermont vt
sc carolina romney mittromney mitt wins call projects called held won calling project projection win projecting hold
ma massachusetts wins call projects called held won calling project projection win projecting hold elizabeth warren
barackobama barack obama best come yet

6.2.4.1. Elections Dataset

The Elections dataset consists of tweets from the November 11, 2012, U.S. presidential election. The entire set of tweets is partitioned into timeslots. Each ground truth topic extracted based on media reports is assigned to one time slot. A time slot can have more than one ground truth topic. 64 ground truth topics are present in the dataset.

The dataset consists of 524,886 tweets. The tweets are broken into 26 individual timeslots, where each timeslot is ten minutes long. Example topics are shown in Table 6.16. Topics include the re-election of Barack Obama and his running mate, Joe Biden, over nominee Mitt Romney. Later timeslots contain topics indicating Obama's victory speech. The dataset also include elections to the United States Senate and House of Representatives, as well as some state governors.

Consistent with (Aiello, et al., 2013), we first calculated topics using LDA and GeoContext for each timeslot of the dataset. The number of topics calculated by LDA was 10 for each timeslot. Topics discovered using both LDA and GeoContext for some sample timeslots are

shown in Table 6.17.

We then calculated three metrics for the evaluation of the discovered topics: topic recall, term precision, and term recall.

Definitions of these metrics are:

1) *topic recall*: topic recall is the total number of topics detected out of the ground truth topics. A topic is considered to be detected if all terms in the ground truth topic are present in the detected set of keywords.

2) *term precision*: for a detected topic and some matching ground truth topic, term precision is the number of correctly detected terms in a topic out of the total number of terms in the detected topic.

3) *term recall*: for a detected topic and some matching ground truth topic, term recall is the number of correctly detected terms in a topic out of the total number of terms in the ground truth topic.

The total topic recall, term precision, and term recall is computed by taking the microaverage of the individual topic recall, term precision, and term recall for each timeslot. The total values for each of the three metrics are shown in Table 6.18. We also show the topic recall, term precision, and term recall across all timeslots for both LDA and GeoContext in Figures 6.4, 6.5, and 6.6.

A limitation exists with this dataset in the evaluation that results from the method of calculating ground truth topics, as described in (Aiello, et al., 2013). Because the ground truth topics were not extracted directly from the dataset tweets, but rather from news stories that described the timeslots, it is not guaranteed that tweets exist with the terms in the ground truth

Table 6.17. Sample Discovered Topics Over Timeslots

	Timeslot 0	Timeslot 9	Timeslot 12	Timeslot 18	Timeslot 25
LDA	cnn election luck night obama good win results early lead	wins wisconsin ryan michigan amp rom- ney obama ohio win home	mccaskill wins akin projects romney electoral obama votes state mitt	elected obama congratulations black voted term win years president america	barack election speech victory obama gt chicago live president supporters
GeoContext	#romney,surprises, Indiana, #Romney, #Obama, #USelection,indiana, kentucky, #obama, #kentucky, #ROMNEY, #Obama2012, RACE, Vermont, #ElectionDay2012, Kentucky, #romney, #romney #romney #romney, votes	vote, america, RT, Romney, Obama, BBCNewsUS, #Florida, LIVE, t.co, rt, romney, #election2012, florida, Florida waiting, line, election, #obama, guy, Ohio, Michigan, auto, job, votes, #FLORIDA #election2012, Vote, country, #Obama, VOTE, #stayinline, #OBAMA, #obama2012, #OBAMA #FORWARD, stay, obama	line, #TEAMOBAMA #Obama2012 #Forward, FLORIDA, #stayinline, #Obama2012, #Election2012, #stayinline,votes LINE, polls,	presidentWashingtonstructure, Obama, Dems, GOP, Senate, change, House, t.co, you., heart, #Election2012, White House, RT, chance, leader, #election2012, Retweet, #FourMoreYears, NBC, election obama, power President	president, #obama, President,#Obama, campaign headquarters, 5L0Y7c4H,Chicago,speech, chicago, t.co MT73sKxJ, stage, Chicago, live, way, long voting lines, issues, CNNelection, chicago, RT, VP MartinSchulz, EU, USA, America, Congrats, admiration, respect, speeches, ChelseaMFine Art, ObamaWon

Table 6.18. Total Metric Results

	Topic Recall	**Term Precision**	**Term Recall**
LDA	0.312	0.31	0.539
Geo-Context	0.562	0.468	0.675

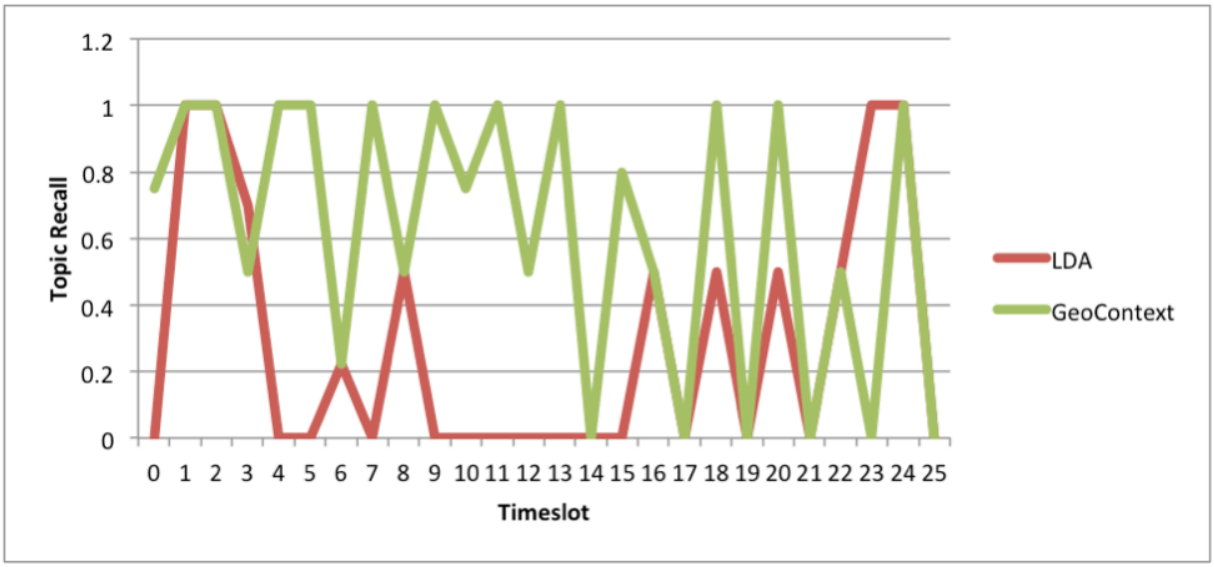

Figure 6.4. Topic Recall

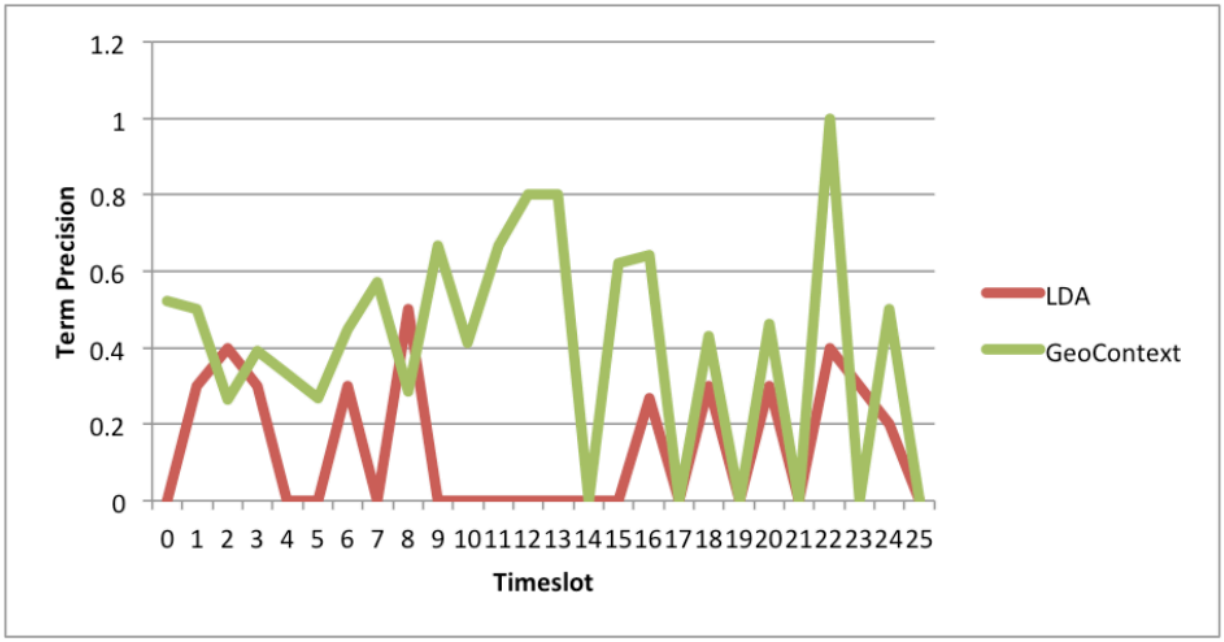

Figure 6.5. Term Precision

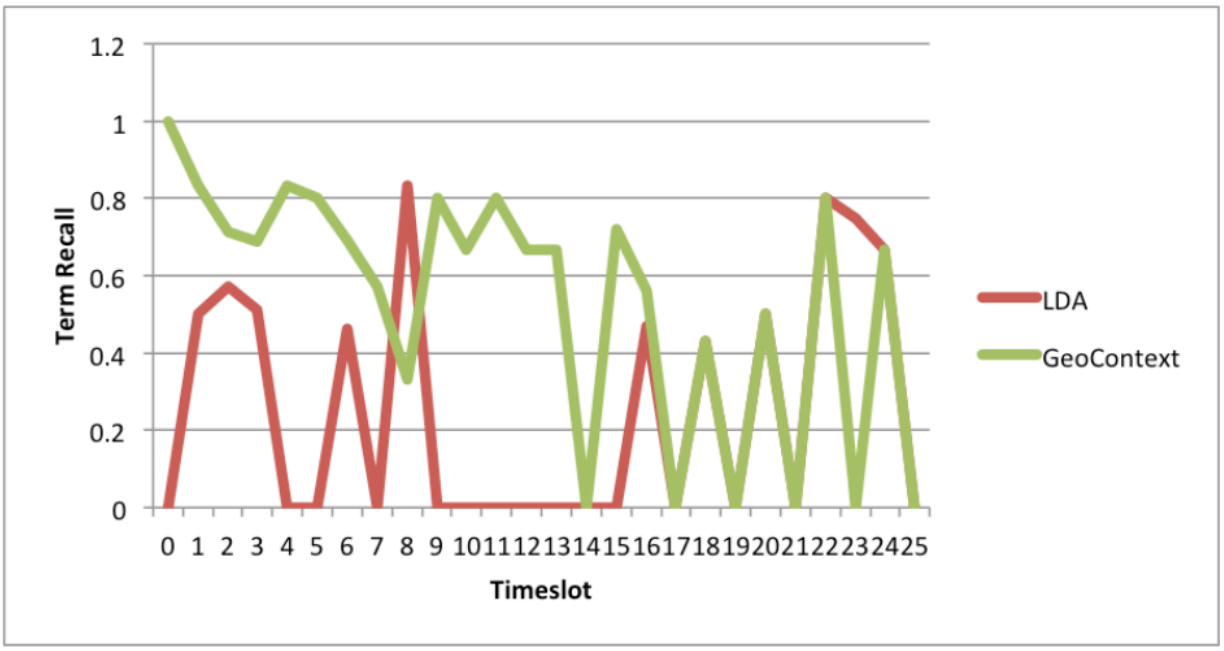

Figure 6.6. Term Recall

topics. For example, a ground truth topic in timeslot 0 is "bernie, sanders, senator, senate, vermont, vt, wins, call, projects, called, held, won, calling, project, projection, win, projecting, hold." However, there exists only one tweet from timeslot 0 in the dataset that contains the term "bernie" and none that contain the term "sanders." Furthermore, the tweet that contains the term "bernie" does not contain the term "senator," so it is highly improbable that these terms would occur in the same topic produced by any topic discovery method, making this ground truth topic impossible to reproduce.

6.2.4.2. Rumors and Truths Dataset

The Rumors and Truths dataset consists of sets of tweets broken up into various topics. Some topics are popular rumors that circulated throughout Twitter, while others are topics that are true. For this dataset, we decided to evaluate how well the topics discovered by LDA and GeoContext matched the dataset topics, which are human-labeled.

The dataset consists of 41,952 individual tweets. The tweets are categorized into 152 distinct topics. Example truth topics are "Mold inside Capri Sun drinks," "Storm hitting Bay Area," and "iPod classic discontinued." Example rumor topics are "NASA warns of six day blackout," "Actor Macaulay Culkin found dead," and "Malia Obama is pregnant."

Table 6.19 shows five example topics from the dataset, LDA, and GeoContext. For each topic, the leftmost column shows a description of the topic and an example tweet from the ground truth dataset. The table also indicates whether the topic is a rumor or a truth. The middle column contains the corresponding topics found by LDA, and the rightmost column contains the corresponding topics found by GeoContext from the topic.

For this dataset, LDA produced topics in several instances that were mixtures of more than

Table 6.19. Rumors and Truths Evaluation

Dataset Topics	**LDA Topics**	**GeoContext Topics**
Truth: David Ryall died "@eonline: RIP David Ryall. The Harry Potter actor has died at age 79."	ryall; david; died; actor; harry; potter; stop; customers; overweight; serving	david ryall, elphias doge, #harrypotter star, peace, harry potter, outnumbered, David Ryall, excellent actor, good films, tv progs, #harrypotter, harry potter, actor, age, t.co
Truth: North Korea Sony attack "North Ko- rea AINT PLAYING! RT @necolebitchie Sony Hackers Threaten 9/11-Type Attack On Theaters (cont)"	1:north; korea; sony; internet; outage; hack; attack; service; restored; report 2:north; ko- rea; time; capsule; sony; internet; outage; paul; attack; boston	north korea, pr win, Sony investigators, at- tack probe, North Korea, links, source, Sony attack, supporters, attack probe-source, sus- pect, denial, sony pictures, house intel, hacks, sony, information, internet outage, Internet services, outage, restoration, night, nkorea outage, online uncertainties, micro- scopic corner, case study, internet, Internet outage, dispute, U.S., experts, i4u news, NKorea outage, Internet ha, AP, LONDON, wifi password, South Korea, Dec, access, torontostar, tit, tat, fingers, Sony movie, Internet service, tensions, hack, internet out- ages, attack, t.co, t.co qijAQlHmA0, #dyn- research #lesleywroughton, dispu, apparent attack, web outage, Web outage, t.co klrSf- FKqFX, Internet
Truth: West Virginia train derailed "Fireball erupts into sky as derailment sends tanker into river: A CSW train derailed, pouring crude oil into a..."	1:train; west; derailed; oil; virginia; crude; freight; carrying; fire; derailment 2: train; oil; derailed; virginia; west; crude; carrying; news; fire; freight	Train Derailment, freight train, explosion th, Oil Spill, crude oil, West Virginia, Monday
Rumor: Bobby Shmurda stabbed "Bobby Shmurda stabbed to death in prison ladies and gents. Guess you could say he was alive about a week ago #toosoon"	1:shmurda; stabbed; death; bobby; jail; fake; cell; mate; news; rikers 2:bobby; shmurda; stabbed; death; jail; killed; rice; tamir; cell; mate	jail, bobby shmurda, death l0l, death, prison, cell mate, jail tho, yea, man, jail, way
Rumor: Obama lowers drinking age to 18 "Effective 6/4/2015 President Obama Signs Amendment To Lower The Legal Drinking Age To 18"	obama; age; drinking; lower; legal; lower- ing; june; law; lowered; signed	legal drinking age, obama, obama signs amendment, president, obama bout

Table 6.20. Tweet Precision and Tweet Recall

	Tweet Precision	**Tweet Recall**
Keywords Only	0.927	0.209
Keywords and Concepts	0.814	0.170

one ground truth topic. For example, the topmost topic in Table 6.19 contains terms about both the actor David Ryall's death as well as a rumor about the restaurant chain McDonald's stopping service for overweight customers. Also, LDA produced more than one topic for several of the ground truth topics. These instances are indicated by numbering in the table.

Because the Rumors and Truths dataset does not include ground truth topics, but rather ground truth tweet clusters, we did not compute the topic recall, term precision, and term recall metrics. Instead, we calculated the tweet precision and tweet recall produced by GeoContext for this dataset. Because the Rumors and Truths dataset consists of clusters of tweets as ground truth, it is well-suited for these metrics. GeoContext can be used for clustering tweets in addition to discovering topics, so it can be useful to determine how well the clusters are formed.

We define the tweet precision and tweet recall as follows:

1) *Tweet precision:* the percentage of correctly clustered tweets out of all tweets in a cluster. We calculated the total tweet precision as the average of the tweet precision for each cluster of tweets.

2) *Tweet recall:* the percentage of correctly clustered tweets out of the total number of tweets in the ground truth cluster. As with tweet precision, we calculated the total tweet recall as the average of the tweet recall for each cluster of tweets.

We noticed during manual evaluations that the concepts extracted by AlchemyAPI's Concept Tagging API was not always accurate in describing the topic of the tweet. Because of this observation, we decided to calculate the metrics both with and without GeoContext's concepts. The total tweet precision and tweet recall over all tweet clusters are shown in Table 6.20.

As displayed in the table, both metrics are higher for GeoContext using keywords only. Also, the tweet precision is high, indicating that tweets are correctly clustered together. However, the tweet recall is somewhat low, indicating that GeoContext splits the ground truth clusters apart into multiple clusters. In future work, we plan to investigate GeoContext's algorithm to determine the cause of this splitting.

6.2.4.3. Discussion of Results

The results from this evaluation process clearly show the benefits of using keyword relevance over traditional topic modeling approaches for topic discovery within social media. As shown in Table 6.18, GeoContext was able to identify more ground truth topics than LDA. The term precision and topic recall metrics were also higher for GeoContext than LDA, showing that GeoContext was able to create more topics that

have more related terms than LDA and more topics that contain the terms from the ground truth topics. The high values for these metrics indicate that GeoContext is able to better discover individual events in a social media stream that do not contain mixed topics.

GeoContext contains a drawback in that it can create a dynamic number of topics, which can affect processing time and readability for users if the number of topics is too large. To address this issue, GeoContext also includes a pruning module, which prunes topics that have not had any new tweets added in a certain amount of time. However, because we wanted to evaluate GeoContext's results directly against LDA's results, which does not consider time, we elected not to use this module.

We conclude this evaluation by mentioning some of GeoContext's statistics. GeoContext works completely in real-time. GeoContext receives approximately 18 tweets from the Twitter stream per second. Out of those, it is able to geolocate between 1 and 4 tweets per second. This limitation is not due to the processing time, but rather to the percentage of tweets from the stream that contain location information. GeoContext then extracts topics and keywords and adds the geolocated tweets to a topic cluster within an average of 1 second. These statistics show that GeoContext is implemented as a truly real-time system that can be used to extract geographical topics that are relevant to various users in a real-world system.

CHAPTER 7

CONCLUSION AND FUTURE WORK

Although search engines such as Google are useful for many static queries, they are not always useful for finding real-time information. Instead, social media can be a valuable resource for discovering specific information about particular situations such as traffic or weather scenarios. Because social media posts can be updated immediately and disseminated quickly, social media can be more useful for uncovering data about time-critical situations than other media. In the following sections, we conclude the dissertation and present future work.

7.1. Conclusion

In Chapter 1, we introduced the problem of social media analysis and why social media can provide insights into events not found on any other media. In Chapter 2, we provided background to social media analysis and introduced terms associated with our research. Chapter 3 provided an overview of our algorithm for analyzing social media through topical and geographical clustering, as well as the parameters and initialization of GeoContext, our implementation of our research in geotopical clustering and analysis. In order to perform geographical clustering, tweets need to be associated with locations.

In Chapter 4, we described GeoContext Locator (GCL), our method for predicting locations of tweets that are not already geotagged. GCL utilizes several different techniques for geolocation and combines those methods in an intelligent manner. It is able to geolocate 39.12%

of tweets when run on our test set. GCL improves on existing geolocation approaches by using both friends and followers, as well as content and topical clustering as methods.

In Chapter 5, we described the topical clustering and geographical analysis portions of our implementation of GeoContext, which is a novel method for discovering relevant contextual topics in a social media stream. We implemented two versions of GeoContext's geotopical clustering module. In the first version, topical clustering is performed first, followed by geographical clustering. In the second version, geographical clustering is performed first, followed by topical clustering.

GeoContext is able to discover topics that are unique to various locations and recommend topics of interest for users in those locations. We also implemented a system for GeoContext to filter the stream by keywords and location coordinates in order to produce a more specific set of topics appearing in the social media stream. The keyword query system uses cognitive computing techniques to expand keywords into collections of keywords that represent a context. The location query system provides clusters of tweets around specific locations.

Finally, we outlined all evaluations that were performed on GeoContext and GCL in Chapter 6. We evaluated all resulting topics extracted from a stream of tweets, and GeoContext was able to discover more defined topics than LDA, an algorithm commonly used in topical clustering implementations.

7.2. Future Work

We outlined three main challenges to geotopical clustering in social media in Section 1.3 (i.e., geolocation, topical clustering, and geographical analysis) and expanded upon these challenges in Chapter 2. We addressed these challenges with our research. However, there are

still several main areas in which we plan to pursue further investigation. We have broken these areas down into future work within GCL, topical analysis, geographical analysis, and performance time of GCL.

In addition to these areas, we plan to refine some of the ways users can interact with GeoContext. Specifically, we plan to create a more visual representation of GeoContext, so that users can view where topics are centered on a map. Because topics in Twitter are dynamic (Diao, Jiang, Zhu, & Lim, 2012), the visual representation can include animations that will show how topics can change in locations over time.

We also plan to examine how to identify tweets posted by ground users, which are users that are physically present in an area where an event occurs. The ground users can give us the most accurate and fast source of information regarding the event. Ground users can also be a source of verification or rejection of inferences made about the event. For example, rumors can spread quickly about topics, especially those that are high-profile. Ground users can indicate whether the rumors surrounding an event are accurate or not.

Lastly, we plan to perform more empirical evaluations that utilize other metrics, such as the F1 score (Van Rijsbergen, 1979), which combines precision and recall into a single metric. We believe that utilizing other metrics will enable us to more effectively study the results of GeoContext.

7.2.1. Future Work: Geolocation

The first challenge addressed by GeoContext is to provide a geolocation module that is able to predict the geographical coordinates of a tweet using the tweet content, user location, friends and followers' locations, and topic.

GCL utilizes several different techniques for geolocation and combines those methods to form a final prediction of the location of a tweet. GCL improves on existing geolocation approaches by utilizing unique ways of extracting location information by querying Dbpedia and the Google Places API.

In the future, we plan to improve GCL in several ways. First, we plan to treat friends and followers with different measures by incorporating a measure of how much users connect. For example, McGee et al. (McGee, Caverlee, & Cheng, Location Prediction in Social Media Based on Tie Strength, 2013) calculated the tie strength between two Twitter users, which determined how much the users communicated. They then used this measure to geolocate the users, because they determined that users who communicate more frequently often live closer together. GCL currently treats all friends and followers the same and does not take into account the level of communication between users. In the future, we plan to explore the use of a similar metric in order to improve the friends and followers technique step within GCL.

Second, we plan to improve the selection strategy for the final predicted location. Although the algorithm used by GCL in this experiment was able to produce an accurate final prediction for about 30% of all tweets within 5 km, there exist instances in which the final prediction location was inaccurate, even though one or more techniques produced an accurate location estimate, due to the selection strategy used by GCL to choose a final prediction. In this situation, an accurate location is extracted by a technique, but it is not chosen to be the final predicted location. As mentioned in Chapter 6, approximately 17% of all tweets had extracted location information accurate within 30 km, but the final predicted location was larger than 30 km. Because GCL extracts a very large amount of location information from a single tweet, the final prediction step can be improved to choose between all of the location prediction results produced

from the various techniques. In the future, we will study how to more intelligently decide between the results to choose the final prediction.

Third, we plan to improve the topic technique within GCL. We believe that the technique has potential, but needs further refining. We also believe that using all techniques in combination with the topic as probabilistic measures could further improve the accuracy of GCL.

7.2.2. Future Work: Topical Clustering

The second challenge is to cluster tweets into representative topics using keyword and concept analysis. Topic discovery in a social media stream can be an invaluable tool for identifying major events around the world. Using social media to gather information can allow us to utilize the opinions and data of millions of people, rather than only a few traditional media outlets.

There are several areas of topical clustering we plan to focus on in future research. First, we plan to utilize sentiment analysis of individual tweets in future work to determine the overall sentiment of a tweet cluster (Maynard, Dupplaw, & Hare, 2013). This is useful in use cases such as the political analysis mentioned in Chapter 6. In that case, sentiment analysis can provide an indication of the feelings of different geographical areas about different topics or candidates without needing to perform manual observations.

Second, we plan to improve GeoContext's algorithm for creating topic clusters. As shown in Section 6.2.4.2, on some occasions, topics can consist of two different unrelated topics, where they should be separated into two topics. We plan to investigate the cause of our algorithm merging two topics together.

Also, we have found that there are situations when more than one topic cluster exists that

consists of the same topic. In this case, the multiple topic clusters should be merged. We also plan to investigate why this type of split is occurring. In future research, we will refine the topical clustering algorithm to discover more distinct topics.

7.2.3. Future Work: Geographical Analysis

The third and last challenge is to perform geographical analysis. The goal of this step is to analyze where the topic clusters are centered geographically using our adapted TF-IDF algorithm.

Geographical analysis is an invaluable, yet relatively unexplored area of research for social media topic discovery. Unearthing whether a topic is spread across a large region or centralized to a smaller location can provide much insight to different opinions and information on social media.

In future work, we plan to further refine the adapted TF-IDF algorithm to better understand where topics are geographically located. Although the evaluations presented in Chapter 6 showed that GeoContext's geographical analysis module is able to successfully associate locations with topics in a majority of cases, there is a situation where the algorithm can be improved. Through our evaluations, we found that sometimes tweets are geotagged in a certain location, but their content is about another location. For example, as described in Section 6.2.2, some weather tweets had a geotag of Washington, D.C., USA, but the tweets were regarding weather in other states. This was due to the National Weather Service being located in Washington, D.C. In the future, we plan to determine whether we should take this type of phenomenon into account in our geolocation module, and analyze the content of the topic cluster to determine the location for which the topic cluster is relevant.

7.2.4. Future Work: Performance Time

Lastly, we would like to examine ways to improve the performance time of GeoContext. We plan to determine whether a method exists for pruning tweets that would not require the comparison of every new tweet to every existing tweet. We believe that it may be possible to take an "average" of the keywords and concepts of a topic cluster and compare an incoming tweet to the "average," rather than every tweet in the topic cluster.

We also plan to determine whether a different storage solution for tweets would improve the performance of GeoContext. As the number of tweets increases over time, a larger and faster database storage system may be required in order for GeoContext to remain in near real-time status. We plan to investigate various solutions to this storage problem and determine which is the best fit for GeoContext.

REFERENCES

Aiello, L., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., et al. (2013). Sensing Trending Topics in Twitter. *IEEE Transactions on Multimedia, 15* (6), 1268-1282.

Backstrom, L., Sun, E., & Marlow, C. (2010). Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. *19th International Conference on World Wide Web*, (pp. 61-70). Raleigh, NC.

Baldwin, T., Cook, P., Han, B., Harwood, A., Karunasekera, S., & Moshtaghi, M. (2012). A support platform for event detection using social intelligence. *Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 69-72). Avignon, France.

Baucom, E., Sanjari, A., Liu, X., & Chen, M. (2013). Mirroring the Real World in Social Media: Twitter, Geolocation, and Sentiment Analysis. *International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, (pp. 61-67). San Francisco, CA.

Biemann, C., & Riedl, M. (2013). Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *Journal of Language Modelling, 1* (1), 55-95.

Blei, D. (n.d.). *Topic Modeling*. Retrieved from https://www.cs.princeton.edu/~blei/topicmodeling.html

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research , 3*, 993-1022.

Blondel, V., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics, 2008* (10), P10008.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information Processing Systems*, (pp. 288-296). Vancouver, BC.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, (pp. 759-768). Toronto, Ontario.

Cui, A., Zhang, M., Liu, Y., Ma, S., & Zhang, K. (2012). Discover Breaking Events with Popular Hashtags in Twitter. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, (pp. 1794-1798). Maui, HI.

Dela Rosa, K., Shah, R., Lin, B., Gershman, A., & Frederking, R. (2011). Topical Clustering of Tweets. *Social Web Search and Mining.* Beijing, 8 pages.

Dey, A., Abowd, G., & Salber, D. (2001). A Conceptual Framework and a Toolkit For Supporting the Rapid Prototyping of Context-Aware Applications. *Human-Computer Interaction, 16* (2), 97-166.

Diao, Q., Jiang, J., Zhu, F., & Lim, E.-P. (2012). Finding Bursty Topics from Microblogs. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, (pp. 536-544). Jeju, Republic of Korea.

Dwoskin, E. (2016, November). *Police Are Spending Millions to Monitor the Social Media of Protesters and Suspects*. Retrieved from https://www.washingtonpost.com/news/the-switch/wp/2016/11/18/police-are-spending-millions-to-monitor-the-social-media-of-protesters-and-suspects/?utm_term=.b1400833df7b

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical Variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (pp. 1277-1287). Stroudsburg, PA.

ESRI GIS Dictionary. (n.d.). Retrieved 2017, from http://support.esri.com/other-resources/gis-dictionary/term/spatial%20analysis

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm For Discovering Clusters in Large Spatial Databases With Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (pp. 226-231). Portland, OR.

Gao, X., Cao, J., He, Q., & Li, J. (2013). A Novel Method for Geographical Social Event Detection in Social Media. *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, (pp. 305-308). Huangshan, China.

Gemen, S., & Gemen, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6* (6), 721-741.

Han, B., Cook, P., & Baldwin, T. (2014). Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research, 49* (1), 451-500.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A., & Tsioutsiouliklis, K. (2012). Discovering Geographical Topics in the Twitter Stream. *Proceedings of the 21st International Conference on World Wide Web*, (pp. 769-778). Lyon, France.

Hong, Y., Fei, Y., & Yang, J. (2013). Exploiting Topic Tracking in Real-Time Tweet Streams. *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, (pp. 31-38). San Francisco, CA.

Ikawa, Y., Enoki, M., & Tatsubori, M. (2012). Location Inference Using Microblog Messages. *21st International Conference on World Wide Web*, (pp. 687-690). Lyon, France.

Jaiswal, A., Peng, W., & Sun, T. (2013). Predicting Time-sensitive User Locations from Social Media. *Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (pp. 870-877). Niagara Falls, Ontario.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. *Proceedings of the 9th WebDKK and 1st SNA-KDD workshop on Web Mining and Social Network Analysis*, (pp. 56-65). San Jose, CA.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological Modeling of News and Rumors on Twitter. *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, (pp. 8:1-8:9). Chicago, IL.

Kaplan, A., & Haenlein, M. (2010). Users of the World, Unite! The Challenges and Opportunities of Social Media. *Business Horizons, 53* (1), 59-68.

Kim, H.-G., Lee, S., & Kyeong, S. (2013). Discovering Hot Topics using Twitter Streaming Data. *Proceedings of the 2013 IEEE/ACM Internationall Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, (pp. 1215-1220). Niagara, Ontario, Canada.

Kling, C. C., Kunegis, J., Sizov, S., & Staab, S. (2014). Detecting Non-Gaussian Geographical Topics in Tagged Photo Collections. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, (pp. 603-612). New York, NY.

Krikorian, R. (2013). *New Tweets Per Second Record*. Retrieved from Twitter Blog: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

Kullback, S., & Leibler, R. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics, 22* (1), 79-86.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the Dynamics of the News Cycle. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 497-506). Paris, France.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., & Hurst, M. (2007). Cascading Behavior in Large Blog Graphs. *Proceedings of the Seventh SIAM International Conference on Data Mining*, (pp. 551-556). Minneapolis, MN.

Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C.-C. (2012). Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. *18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1023-1031). Beijing, China.

MacQueen, J. (1967). Some Methods for Classificiation and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, (pp. 281-297). Berkeley, CA.

McGee, J., Caverlee, J., & Cheng, Z. (2013). Location Prediction in Social Media Based on Tie Strength. *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*, (pp. 459-468). San Francisco, CA.

Maynard, D., Dupplaw, D., & Hare, J. (2013). Multimodal Sentiment Analysis of Social Media. *BCS SGAI Workshop on Social Media Analytics,* (pp. 44-55). Cambridge, UK.

Mei, Q., Liu, C., & Su, H. (2006). A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. *Proceedings of the 15th International Conference on World Wide Web*, (pp. 533-542). Edinburgh, Scotland.

Musaev, A., Wang, D., Shridhar, S., & Pu, C. (2015). Fast Text Classification Using Randomized Explicit Semantic Analysis. *2015 IEEE International Conference on Information Reuse and Integration*, (pp. 364-371). San Francisco, CA.

Oxford Dictionary. (n.d.). Retrieved 2017, from https://en.oxforddictionaries.com/definition/geolocation

Petkos, G., Papadopoulos, S., Aiello, L., Skraba, R., & Kompatsiaris, Y. (2014). A Soft Frequent Pattern Mining Approach for Textual Topic Detection. *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics*, (p. 25). Thessaloniki, Greece.

Roberts, J. (2012, 10 10). *Typical Twitter User Is a Young Woman With an iPhone & 208 Followers*. Retrieved from GigaOm: https://gigaom.com/2012/10/10/the-typical-twitter-user-is-a-young-woman-with-an-iphone-and-208-followers/

Rout, D., Bontcheva, K., Preotiuc-Pietro, D., & Cohn, T. (2013). Where's @wally?: A Classification Approach to Geolocating Users Based on Their Social Ties. *24th ACM Conference on Hypertext and Social Media*, (pp. 11-20). Paris, France.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development. *IEEE Transactions on Knowledge and Data Engineering, 25* (4), 919-931.

Scapusio, D. (2017). *Fort Lauderdale Airport Shooting: On Social Media, People Reflect on the Close Call They Had*. Retrieved from http://www.wftv.com/news/trending-now/fort-lauderdale-airport-shooting-on-social-media-people-reflect-on-the-close-call-they-had/482283650

She, J., & Chen, L. (2014). TOMOHA: TOpic MOdel-based HAshtag Recommendation on Twitter. *Proceedings of the 23rd International Conference on World Wide Web*, (pp. 371-372). Seoul, Korea.

Son, J.-W., Noh, Y.-S., Song, H.-J., & Park, S.-B. (2012). Location Comparison Through Geographical Topics. *Proceedings of 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, (pp. 311-318). Macau.

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation, 28*, 11-21.

SPARQL Query Language for RDF. (n.d.). Retrieved from https://www.w3.org/TR/rdf-sparql-query/

Spinsanti, L., Berlingerio, M., & Pappalardo, L. (2013). Mobility and Geosocial Networks. In *Mobility Data - Modeling, Management, and Understanding* (pp. 315-333). New York, NY: Cambridge University Press.

Top 15 Valuable Facebook Statistics. (2017). Retrieved from https://zephoria.com/top-15-valuable-facebook-statistics/.

Twitter Statistics. (2017). Retrieved from http://www.internetlivestats.com/twitter-statistics/

Van Rijsbergen, C.J. (1979). Information Retrieval.

Vosecky, J., Wai-Ting Leung, K., & Ng, W. (2013). Dynamic Multi-Faceted Topic Discovery in Twitter. *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*, (pp. 879-884). San Francisco, CA.

Wang, C., Wang, J., Xie, X., & Ma, W.-Y. (2007). Mining Geographic Knowledge Using Location Aware Topic Model. *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, (pp. 65-70). Lisbon, Portugal.

Watanabe, K., Ochi, M., & Onai, R. (2011). Jasmine: A Real-Time Local-Event Detection System Based on Geolocation Information Propagated to Microblogs. *20th ACM International Conference on Information and Knowledge Management*, (pp. 2541-2544). Glasgow, Scotland.

Williams, E., Gray, J., & Dixon, B. (2016). Evaluating GeoContext: A System for Creating Geographical Topics from a Social Media Stream. *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, (pp. 1-7). New York, NY.

Williams, E., Gray, J., & Dixon, B. (2016). Mobile Context Recommendations from Social Media through Geotopical Clustering. *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, (pp. 1-7). New York, NY.

Williams, E., Gray, J., & Dixon, B. (2017). Improving Geolocation of Social Media Posts. *Pervasive and Mobile Computing (accepted, in publication)*.

Wing, B. P., & Baldridge, J. (2011). Simple Supervised Document Geolocation With Geodesic Grids. *49th Annual Meeting of the Association for Computational Linguistics*, (pp. 955-964). Portland, OR.

Yang, H., Chen, S., Lyu, M., & King, I. (2011). Location-Based Topic Evolution. *Proceedings of the 1st International Workshop on Mobile Location-Based Service*, (pp. 89-98). Beijing.

Yin, Z., Cao, L., Han, J., Zhai, Chengxiang, & Huang, T. (2011). Geographical Topic Discovery and Comparison. *Proceedings of the 20th International Conference on World Wide Web (WWW)*, (pp. 247-256). Hyderabad, India.

Yuan, Q., Cong, G., Zhao, K., Ma, Z., & Sun, A. (2015). Who, Where, When, and What: A Nonparametric Bayesian Approach to Context-aware Recommendation and Search for Twitter Users. *ACM Transactions on Information Systems, 33* (1), 2:1-2:33.

Zhang, L., Sun, X., & Zhuge, H. (2015). Topic Discovery of Clusters from Documents with Geographical Location. *Concurrency and Computation: Practice and Experience, 27* (15), 4015-4038.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, (pp. 338-349). Dublin, Ireland.

Zou, J., Fekri, F., & McLaughlin, S. (2015). Mining Streaming Tweets for Real-Time Event Credibility Prediction in Twitter. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* , (pp. 1586-1589). Paris, France.