ESTIMATING TOPONYM CONTENT OF SOCIAL MEDIA DATA OF THE NORTHERN TRIANGLE

by

Molly Phillips
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geoinformatics and Geospatial Intelligence

Committee:

____________________________________	Dr. Arie Croitoru, Thesis Chair
____________________________________	Dr. Andrew Crooks, Committee Member
____________________________________	Dr. Matthew Rice, Committee Member
____________________________________	Dr. Dieter Pfoser, Department Chairperson
____________________________________	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
____________________________________	Dr. Ali Andalibi, Interim Dean, College of Science
Date: ______________________________	Fall Semester 2019 George Mason University Fairfax, VA

Estimating Toponym Content of Social Media Data of the Northern Triangle

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Molly Phillips
Bachelor of Science
United States Air Force Academy, 2018

Director: Arie Croitoru, Associate Professor
George Mason University

Fall Semester 2019
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my loving family; I would not be here without all of your support.

ACKNOWLEDGEMENTS

I cannot believe my time here in the GGS department is already coming to a close. I would like to take this opportunity to thank my advisor Dr. Croitoru for all of his guidance, support, and flexibility in every step of my short time here at George Mason University. I would also like to thank my committee members, Dr. Andrew Crooks and Dr. Matthew Rice, for their valuable suggestions. A special thank you to Melanie Swartz for her unmatched patience and valuable insight throughout my analysis process. I would be remiss if I did not mention the United States Air Force Academy Geospatial Science Department which gave me this amazing opportunity to be studying here right now. I am so excited to take my new knowledge back to USAFA and teach in the Geospatial Science Department. Lastly, I would like to thank my family for their constant love, support, and encouragement to finish strong.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Application Programming Interface	API
Central American Regional Security Initiative	CARSI
Comma Separated Value	CSV
Department of Homeland Security	DHS
Department of State	DoS
Gross Domestic Product	GDP
International Narcotics and Law Enforcement Affairs	INL
International Organization for Migration	IOM
Linked Open Data	LOD
Named Entity Recognition	NER
Tab Separated Value	TSV
United States	US
United States Agency for International Development	USAID

ABSTRACT

ESTIMATING TOPONYM CONTENT OF SOCIAL MEDIA DATA OF THE NORTHERN TRIANGLE

Molly Phillips, M.S.

George Mason University, 2019

Thesis Director: Dr. Arie Croitoru

Twitter is a microblogging social media platform where users post tweets. Despite the 280-character limit, Twitter data can be harvested and analyzed in order to gain valuable information. While geolocated tweets give in-depth location information, they comprise only a small percentage of tweets. This thesis uses a Twitter dataset collected on Northern Triangle based cartel keywords, and a bounding box of the world. The Northern Triangle, known for its reputation of drug and gang violence, is the area of Central America consisting of Guatemala, El Salvador, and Honduras. Violence stemming from this region has been known to migrate north through Mexico and into the United States. This thesis aims to examine the presence of toponyms on Twitter, their resolution, the types of user accounts who tweet toponyms, and how toponym usage changes over time. In order to examine these toponym-related issues, and using the Northern Triangle region as a case study, 15.3 million tweets related to the Northern

Triangle were collected over a period of one year were processed and analyzed. The data processing included two primary steps, namely data enrichment using a Named Entity Recognition (NER) tool, and data analysis in which the enriched data was examined to explore key trends in toponym prevalence across space, time, and user characteristics. Results show that roughly 1 in 4 tweets contains a toponym, a country was 10 times more likely to be mentioned than a city, and the most prolific users were individuals. This work is a novel application of geolocation to a new social media dataset.

1. INTRODUCTION

The Northern Triangle is the area of Central America made up of El Salvador, Honduras, and Guatemala shown in Figure 1. In the 1980’s, these Northern Triangle countries experienced devastating civil wars that left them in a state in which violence would remain largely unchecked [1]. To this day, violence has driven asylum seekers from El Salvador, Guatemala, and Honduras north into Mexico and the United States. Figure 2 depicts the major migration routes that the asylum seekers have taken. In addition to asylum seekers, cartels and gangs originating in the Northern Triangle have also been known to spread northward [34]. Not only have they brought violence with them, but they have also introduced drugs into the United States through the southern border [2]. U.S. officials report that 90 percent of documented cocaine that flows into the United States passes through the Northern Triangle region [2]. Therefore, it comes as no surprise that the Northern Triangle represents a major region of interest to the US Department of Homeland Security, whose primary goals are to ensure border security and deny the entry of these illegal drugs.

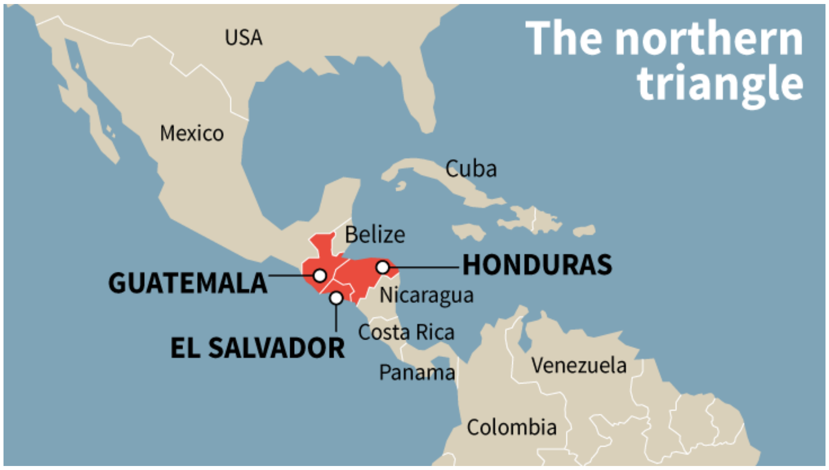


Figure 1 Northern Triangle Countries [34]

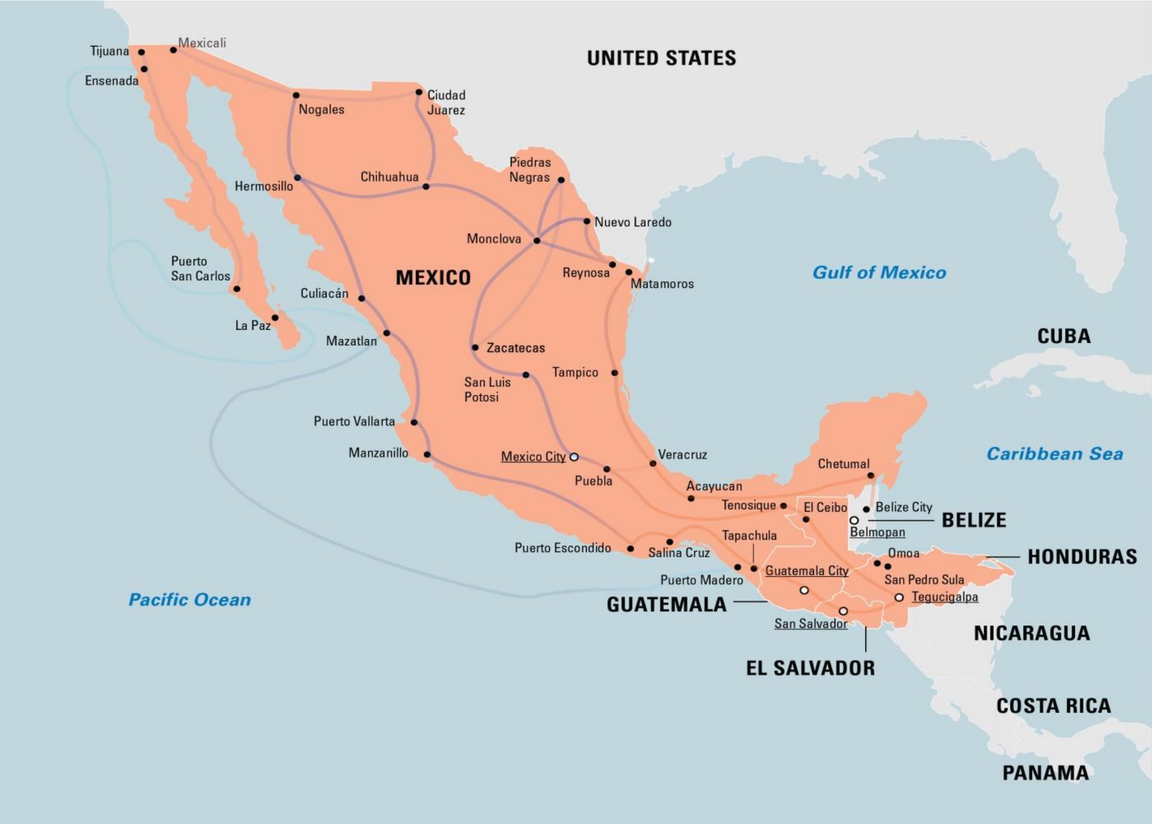


Figure 2 Migration Routes from Northern Triangle Countries [35]

The US Department of Homeland Security (DHS) has a mission to, “safeguard the American people, our homeland and our values” and was born after the September 11th, 2001 terrorist attacks on the United States [41]. In order to help the US Department of Homeland Security monitor these dangerous cartels and gangs from the Northern Triangle, social media can serve as both a cost and time efficient means of providing accessible information. Social media enables users to share personal and valuable information such as their interests, behaviors, sentiments towards topics, relationships statuses, and their location. The emergence of social media participation and information sharing has brought a new model of geospatial information contribution [10]. Out of all of the social media platforms, Twitter, considered a microblogging platform due to its tweet limits of 280 characters, is especially useful for tracking geolocations. In fact, significant research has been dedicated into examining the value of geolocations attached to tweets [12, 13, 14, 15]. However, there is a gap in research when it comes to the presence of toponyms in tweets. Understanding how toponyms appear in tweets is important because it can help scholars provide intelligence to officials to catch cartels. Driven by the US Department of Homeland Security’s interest in the North Triangle and the benefits from social media analysis, this study seeks to investigate the presence of toponyms in tweets.

Toponyms, or place names, provide crucial information about a place such as location along with historical and cultural context. Toponyms are important because they refer to a specific geographic longitude/latitude footprint [21]. They also help a reader understand the location context of a document. This is especially true in the context of microblogs due to the shortened length of the posts. The identification of toponyms in tweets is an area that received little scholarly attention, but could prove to be useful in location and content analysis when paired with data from a reliable gazetteer.

Gazetteers are a dictionary of toponyms. The primary purpose of a gazetteer is to assign precise latitudinal and longitudinal coordinates to a given place. Once toponyms have been located in tweets, it will be important to run the toponyms through a reliable gazetteer in order to attach a precise location to the place mentioned. Doing so enables multiple toponyms within a single tweet, or a group of tweets, to be compared by distance and resolution.

However, selecting a gazetteer proves to be challenging undertaking. There is a plethora of digital gazetteers available online, most of which are region specific. Being region specific has its advantages as these gazetteers are particularly useful in disambiguating places. For example, in the GeoNames global Gazetteer, a well-known non-region-specific gazetteer, there are over 87,000 unique entries for "Washington" [24]. This absurdly large amount of entries would be unworkable. On the contrary, a regionally specific gazetteer has less overlap in the reoccurrence of a toponym, thereby making it easier to attach the correct coordinates to the places mentioned. Unfortunately, due to the extensive time and resources required to create a region-specific gazetteer, there is currently no gazetteer encompassing the Northern Triangle region. Thus, this presents a challenge that this study overcomes by using geocrowdsourced information from Wikipedia and DBpedia.

My research will be start by identifying named entities, specifically toponyms, present in tweets in our Twitter dataset. The task of mapping toponyms to locations is challenging due to insufficient Northern Triangle gazetteers and a large degree of ambiguity in the short text of the tweets, but is necessary in order to learn more about the Northern Triangle countries and their cartel presence. This information is of interest to the Department of Homeland Security because analyzing the social media data to learn about the Northern Triangle and its cartels provide reliable information and are a cost and time savings. It will also better equip the US with more background knowledge of the Northern Triangle and its cartels in order to best handle the current situation.

While the toponym recognition and resolution will be a challenge, overcoming them will provide an abundance of information on the dangerous Northern Triangle. This will be useful in order to gain an understanding of the origins of the drug and gang related cartel violence and better allow the Department of Homeland Security to protect the US against the threats these cartels present.

Before explaining the methodology (Chapter 4) and results of the experiment in (Chapter 5), I will first give a brief history of gang violence and drug crime in the Northern Triangle and a literature review of the current state-of-the-art analysis methods in (Chapter 2). It is important to understand the current drug and crime dilemma in the Northern Triangle and its impact on those countries in addition to the United States in order to understand why this data is important to analyze in a time cost manner.

2. LITERATURE REVIEW

This chapter starts with summarizing the history and current state of the Northern Triangle and the US involvement with Northern Triangle countries (Sections 2.1-2.4). It then proceeds to discuss the current state-of-the-art methods for named entity recognition (Section 2.9) and social media analysis (Section 2.5) to include discourse on digital gazetteers and the use of DBpedia as a gazetteer (Section 2.8). These topics are essential to understanding the important concepts associated with the analysis methods for this study.

2.1 Crime in the Northern Triangle

The Northern Triangle, composed of the Central American countries of Guatemala, Honduras, and El Salvador has been ranked as one of the deadliest regions in the world. According to The Igarape Institute, a Brazilian based think tank, reports that out of these countries, El Salvador had the highest homicide rate in 2017, reaching sixty homicides per one-hundred thousand citizens [4]. Honduras ranked as the fourth highest while Guatemala ranked as fourteenth highest with homicide rates of forty-two per one-hundred thousand and twenty-six per one hundred thousand citizens respectively [4]. These three countries contain the cities that made up the top 5 of the highest homicide rate cities in 2016 [4].

Not only do these countries demonstrate high levels of violence, but they also rank within the poorest countries located in the Western Hemisphere. In 2018, all three countries ranked in the bottom quartile for gross domestic product (GDP) per capita among Latin American states [1]. About 60 percent of Hondurans and Guatemalans live below their countries' national poverty lines. Statistics like these have illuminated the reasons why roughly 265,000 of the 30 million Northern Triangle inhabitants have emigrated in recent years. One should note that this number represents a yearly amount of emigrant instead of a total amount. Data has suggested that this number will most likely double in 2019 [1]. While trends and statistics such as these have not exactly bolstered the reputations of these countries, they have caught the attention of departments who want to offer help such as the United States Department of State (DoS) and the United States Agency for International Development (USAID).

2.2 US Policy on Crime in the Northern Triangle

The DoS is attempting to address the many issues in the Northern Triangle by providing more than $2.6 billion in foreign assistance to Central American countries in fiscal years 2015-2018 through their Bureau of International Narcotics and Law Enforcement Affairs (INL) [5]. The INL's mission is to limit transnational crime and illegal drugs flowing from the Northern Triangle to the United States by helping local law enforcement and its partners enforce the rule of law as well as developing and managing policies and programs. One of these programs is the Central American Regional Security Initiative (CARSI), which responds to threats within the region and is designed to stop the flow of narcotics, weapons, and cash generated by drug sales [6].

2.3 United States Aid to the Northern Triangle Countries

Due to the violence created by transnational crime and illegal drugs, many citizens of the Northern Triangle have emigrated to neighboring countries and the U.S., which has subsequentially sparked great USAID focus in the region. Between 2015 and 2017, USAID contributed $2.5 million to the International Organization for Migration (IOM) program titled “Northern Triangle Information Management Initiative,” which manages, collects, analyzes, and shares migration information to help the populations in those three countries [7]. Not only does the program aid in the migration of vulnerable citizens, but it also obtains information and reports on displacement, human mobility, and repatriation. In addition to the Northern Triangle Information Management Initiative, the USAID also contributed $16.8 million to IOM’s Return and Reintegration in the Northern Triangle Program, which aids returning migrants such as unaccompanied children and families [7]. It is clear that the United States has poured massive amounts of aid and resources into this region, but why?

2.4 Importance of the Northern Triangle to the United States

The simple answer is that it has inflated immigration into the United States to an unmanageable degree. According to the Inter-American Development Bank’s Plan of the Alliance for Prosperity in the Northern Triangle, it states nearly 10% of the population in the three countries has decided to leave and an estimated 265,000 Central Americans migrate to the United States each year [8]. In addition, the report states 50,303 children from Northern Triangle countries were detained at the United States border from January 1st through August 31, 2014 [8]. The United Nations’ Department of Economic and

Social Affairs reported that out of the approximately 2.9 million people migrating from the Northern Triangle in 2017, over 80% of them had an end destination for the United States.

Unfortunately, these migrants are an important component in the financial growth of transnational crime organizations originating from the Northern Triangle. The RAND Corporation published a report for the Department of Homeland Security's Science and Technology Directorate in April of 2018 which states the smuggling of unlawful migrants from the Northern Triangle generated between $200 million and $2.3 billion for drug trafficking transnational crime organizations in 2017 [9]. In addition to covering the direct methods of how these groups receive money, RAND has also stated that these organizations could be collecting anywhere from $30 million to $180 million in taxes for migrants passing through their territory [9]. This additional income significantly aids in illegal drug trafficking from the region to the United States. For this reason, it proves of interest to this study to find out whether Twitter data contains enough specific toponyms to be used as a reliable method of locating and identifying the most dangerous locations in the Northern Triangle, which could then be used to track gang-related and drug-related crimes stemming from this region.

2.5 Social Media Analysis

As the accessibility to the Internet across the globe is growing, many people are gaining access to social media at a rapidly increasing rate. Social media allows users to share personal and valuable information such as their interests, behaviors, sentiments towards certain topics, relationships statuses, and most important to this study, their location.

There are a multitude of different social media sites including Facebook, Instagram, and Twitter. Different countries have different social media sites that are most popular in the region. For example, VK is the Facebook equivalent in Russia. Studies that focus on targeted ads and their use are best fit for Facebook and Instagram due to the ad frequency on those platforms [42]. Out of the many forms of social media, microblogging has quickly emerged as one of the most popular choices for users. Exemplifying this trend towards using a microblogging platform is Twitter. Twitter was founded in 2006, and has gained significant popularity over the years with 319 million active monthly users reported in 2016 [11]. This platform allows users to post content, known as tweets, with up to 280 characters. The emergence of social media participation and information sharing has brought a new model of geospatial information contribution [10].

Specifically, Twitter has sparked scholarly interest in analyzing current events as its open, free, and large public contributions represent a seemingly-endless source of free data to mine through the Application Programming Interfaces (API). Previously, researchers have used Twitter to analyze communication and coordination after Hurricane Katrina hit, coordinate hostage situations during the terrorist attacks in Mumbai, and even help organize anti-government protests in Iran and Moldova [10]. Twitter provides researchers with diverse opportunities to develop techniques and algorithms to investigate user behavior, track patterns of user activity, and monitor social interactions [10]. For these reasons, Twitter is the most appropriate microblogging platform for this study.

2.6 Location Information on Twitter

Social media data has been used widely for detecting geographic events ranging from influenza epidemic to earthquakes [12,13]. Notably, social networking sites like Twitter provide invaluable information to track illicit drug patterns by learning how illicit drugs spread from one place to another [14]. Crucial data from social media feeds include spatial, temporal, and social information. Users frequently comment or post about an event happening in or affecting their location, or refer to locations that represent momentary social hotspots [13]. Twitter data can include three types of geographical data: precisely geolocated tweets which are registered when a user shares the location at the time of tweet submission, account location which is based on the home location provided by user at the time of account creation, and the presence of one or more toponyms within the text potion of the tweet [15]. In June 2019, Twitter removed the automatic precise geolocation of tweets [15]. While toponyms are the most challenging to extract and identify, they can be useful as they contain the location the tweet is referring to, rather than just where the user posting is located. The extraction of toponyms from the tweets will get also around the issue of the precise geolocations being removed from tweets.

2.7 Toponym Resolution

Toponym resolution is a term coined by Jochen Leidner in 2004 to describe the mapping from a place name in a prose text to an extensional representation of a location in a to which the place name refers [21]. It is the grounding of place names to their physical location and represents an important problem when analyzing the Twitter data in the Northern Triangle Region. Place names are highly ambiguous with some having hundreds of possible geographic referents [20]. Another common problem is the use of colloquial, or informal, place names such as “the village market” or the “entertainment district” in a city. Scott McDermott found that a heuristic approach can assist in the identification and disambiguation of colloquial place names [44].

One method of overcoming the resolution problem is to use indirect supervision to create large amounts of training data from links and annotations in Wikipedia. DeLozier et al. describes this method in detail including error analysis methods in their paper *Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles* [20]. First, place names, location information, and coordinates are extracted from the Wikipedia pages provided by the NER tools. Then, this location information is compared to the context of the original block of text to disambiguate the location and provide better toponym resolution.

2.8 Gazetteers

A gazetteer is defined as geospatial dictionaries of geographic names consisting of the core components of a name, a location, and a type [27]. Gazetteers are used to attach coordinates to a known toponym through geoparsing. There are two main types of gazetteers: trigger gazetteers which contain key words that indicate possible presence of an entity, and entity gazetteers which contain entities that are usually proper nouns [26]. While there are a multitude of digital gazetteers online, most were often created for a specific purpose and likely for a specific geographic region. With a regionally specific gazetteer, there is less overlap in the reoccurrence of a toponym which makes it easier to attach the correct coordinates to the places mentioned. Due to the extensive time and resources required to create a regional gazetteer, the cost associated, and the instability and violence in the region, there is currently not a gazetteer specific to the Northern Triangle region [43]. Therefore, it is challenging to find a complete gazetteer online to use for geoparsing [23]. While each of the following gazetteers described in Table 1 have some broad information on the Northern Triangle countries, none of them are complete with refined city details for the Northern Triangle countries.

Table 1 Digital Gazetteers

Gazetteer	**Description**
Geonames	Created by Yahoo! and contains over 25 million global geographical names [24]
Getty Gazetteer	Approximately 1.3 million entries and is used to improve the access to information on art, architecture, and material culture [23]
Falling Rain	Global gazetteer with subsections for individual countries that are then divided by region. The coordinates are not precise and the API is incomplete and is not being developed anymore. [25]

Creating gazetteer for an area is challenging due to the vast resources that go into obtaining the toponyms, their specific coordinates, and their type. It becomes even more challenging to do for a violent and dangerous underdeveloped area like the Northern Triangle.

2.9 Named Entity Recognition Tools

There are off-the-shelf Named Entity Recognition (NER) tools that are used to analyze text for a multitude of different named entities and label them with categories based on the entity type. For example, the entity type could be people, places, or part of speech. However, there are challenges in recognizing entities in text due to varying languages, toponym ambiguity, and limited availability to context of the toponyms due to the short length of a tweet [22]. Using NER on Twitter data poses additional challenges due to slang terms, use of hashtags, and the generally unstructured nature of the text. However, there are NER tools that specialize in microblog, or tweet-style analysis. There are a handful of state-of-the-art NER tools on the market today, of which TextRazor serves as the most beneficial NER tool for this study.

TextRazor was a startup that began in London in 2011. It is a web service that analyzes text in 12 different languages which are English, Chinese, Dutch, French, German, Italian, Japanese, Polish, Portuguese, Russian, Spanish and Swedish [19]. Text analysis in TextRazor includes named entity recognition, disambiguation, confidence and relevance scores, and many other types of analysis. Furthermore, it also provides links to external resources such as Wikipedia for disambiguation [17]. TextRazor uses the hybrid approach of combining machine learning and rule-based methods for analysis [17]. Additionally, TextRazor is known for its advantages with named entity linking, the process of tagging an entity while simultaneously disambiguating it. After comparing many different NER tools on a dataset of tweets, Derczynski et al. found TextRazor to have the highest precision of named entity linking with an accuracy of 65% [30].

2.10 Geoparsing Methods

Once a NER tool has identified the toponyms in Twitter data, the next step is to analyze the different geoparsing methods and decide which method to use. Geoparsing is valuable in many real-world applications such as social media event analysis, emergency response, and much more [12, 13, 14, 15]. Unfortunately, it is still regarded as a challenge due to varying languages in text, toponym ambiguity, and limited availability to context of the toponyms [22]. In addition to these issues, publicly available gazetteers are incomplete or built for only a specific region as mentioned above [23]. This is another challenge that must be overcome in order to precisely locate the toponyms in the Twitter data.

Leidner and Lieberman describe that there are three basic families of methods that are currently in use to recognize toponyms in text. These methods are gazetteer lookup based, rule based, and machine learning based. The gazetteer lookup method consists of using a predefined gazetteer, or dictionary of place names, to compare to the text word by word or even character by character until a match is found. The rule-based method is defined as, “A set of symbolic rules in a domain-specific language encodes a decision procedure that permits an interpreter to decide whether a word is a toponym or not” [16]. Lastly, the machine learning based method consists of sliding window being moved over the text, and at each position a set of properties known as features are computed [16]. Leidner and Lieberman describe that, “Features may comprise checks for strings, length computations, capitalization, and the like, and are frequently Boolean tests. Based on a training corpus containing gold data, feature configurations that are most highly correlated with toponyms are extracted” [16].

2.11 DBpedia as a Gazetteer

DBpedia is a project aiming to extract structured content from the information created in the Wikipedia project [28]. The results from TextRazor provide the DBpedia and Freebase tags for the entities recognized in the tweets. In spite of possessing pertinent information, a limitation of Freebase is that it has not received an upgrade since 2013, meaning that one could consider some of its data dated and not as accurate as current systems. To make up for this, this study employs DBpedia, which contrary to Freebase, is created from information contained in Wikipedia, meaning that it is constantly being updated with the most current information [28].

DBpedia is the hub of the Linked Open Data cloud. Linked Open Data (LOD) refers to linked data which "can be freely used, modified, and shared by anyone for any purpose" [31]. The purpose of the LOD cloud is to publish datasets online and interlink them. Consequently, DBpedia contains many links to other datasets in the cloud such as Freebase, GeoNames, CIA World Factbook, DBLP, and Project Gutenberg [31]. Not only has DBpedia been used extensively in the research community for the Semantic Web, but it also has been noted as a reliable system in commercial settings [31]. For example, companies such as the BBC [32] and the New York Times [33] use DBpedia to organize their content.

Below are the DBpedia definitions of the different resolutions of toponyms analyzed in this study and their hierarchy:

Place- “Immobile things or locations; any concentration of population” [29].

Populated Place- “As defined by the United States Geological Survey, a populated place is a place or area with clustered or scattered buildings and a permanent human population (city, settlement, town, or village) referenced with geographic coordinates” [29].

Country- “A geopolitical area–often synonymous with a sovereign state; a region that is identified as a distinct entity in political geography” [29].

Settlement- “A community where people live” [29].

City- “A relatively large and permanent settlement, particularly a large urban settlement” [29].

2.12 Synthesis

This literature review started by giving a brief history of gang violence and drug crime in the Northern Triangle (Section 2.1), discusses the importance of the Northern Triangle to the United States (Section 2.4), reviews the emergence of social media location analysis (Section 2.5-2.6), discusses digital gazetteers and the role they play in geoparsing (Sections 2.8 and 2.10), names the current state-of-the-art named entity recognition tools (Section 2.9), identifies the challenge of toponym resolution (Section 2.7), and assesses the reliability of DBpedia as a gazetteer (Section 2.11).

The current literature supports the idea of using social media data in order to analyze an area or current event. In addition, it demonstrates that although challenging, toponyms can be extracted from Twitter data and used for further analysis. However, there is a gap in the current literature when it comes to a gazetteer of the Northern

Triangle countries and the evaluation of the extent of toponym usage in Twitter data. My research seeks to fill this gap by analyzing toponym usage in Twitter, specifically focusing on who uses them and the frequency to which they are used.

3. PROBLEM STATEMENT AND RESEARCH QUESTIONS

After introducing and evaluating literature that outlines the problems in the Northern Triangle, its importance to the United States, and analyzing different tools used for NER and toponym resolution, one can now formulate a task in order to state what goals this thesis aims to address. The literature review covered gazetteers, toponym recognition, grounding, toponym resolution, and place name disambiguation, but there is a gap in the literature on the analysis of the extent of toponym usage in tweets. Part of this gap stems from the challenges in identifying toponyms using NER. This is likely due to the short length of a tweet and the lack of context the they provide to the NER tools.

Another issue that arises when dealing with global Twitter data is the issue of processing multiple languages. Some challenges include different spellings of the same name, and a common word in one language being a toponym in another language.

3.1 Research Questions

The central argument of this thesis is that if the toponyms can be identified in tweets, grounded to a specific location, and are prevalent on Twitter, then they can be used to better analyze and evaluate the area surrounding the locations mentioned on Twitter. In doing so, the goal would be able to provide the US Department of Homeland Security with information on the locations in the Northern Triangle that are tied to drug, gang, and cartel violence. The following research questions can be derived from the above problem:

1. To what extent do people use toponyms in tweets?
    a. What is the resolution of these toponyms?
2. Which types of users are most likely to use toponyms?
3. How does toponym usage change over time?

It is important to address these questions in order to understand the extent of toponym usage within the Northern Triangle Twitter dataset. The results can be used to narrow the search timeframe and speed up the analysis process for those looking for the presence of toponyms in tweets. In addition, this information can be used in the future to focus search efforts by pinpointing specific user types who are known to tweet toponyms at an above-average rate. The results could also indicate the hot spots mentioned in the Northern Triangle dataset and possibly even depict the routes that cartels take from the Northern Triangle countries into the United States based on the location and prevalence of toponyms being used. Lastly, the results could demonstrate when these cartels are moving.

In addition to attempting to uncover this specific information from the Northern Triangle dataset, the purpose of this work also serves to guide future studies in where to begin when analyzing Twitter data for a toponym-related problem. This includes the specific types of users to target, the time of day on which to concentrate, and the overall resolution of the problems that can gain insight from analyzing toponym content in Twitter data.

4. METHODOLOGY

In order to evaluate the toponym presence and resolution in the Northern Triangle Twitter dataset, data needed to be collected from Twitter. After collecting the data, this study follows a systematic methodology shown in Figure 3 in order to better understand the dataset and synthesize the results into a manageable product which one can analyze.

First, an examination of the literature (Chapter 2) was crucial in determining the current state-of-the-art tools and best evaluation techniques to tackle the Twitter dataset. After identifying TextRazor as the named entity recognition tool of choice for this study, the text component of each tweet was run through it. The output from TextRazor included a category specifically containing locations that were identified in the text. It also provided a link to the Wikipedia page for each location. This link is vital because attached to it are the latitude and longitude of the specific locations along with other important information detailing location resolution.

Once the coordinates of each of the toponyms were obtained, evaluation of the resolution of each of these locations began. Some locations were specific while others were a whole city, state, or even country. Once the toponym resolution was determined from each tweet, the most prevalent toponyms in the dataset were analyzed. Then, the data was evaluated based on both the time the tweet was posted and the type of user that was tweeting toponyms. Ultimately, this analysis aims to explore patterns and details regarding which type of users, whether news agencies, government, businesses etc. are tweeting toponyms and when they are most commonly tweeted. Figure 3 shows the workflow diagram applied to process the data. Python 3 in Jupyter Notebook was used for the coding and analysis performed in this thesis.

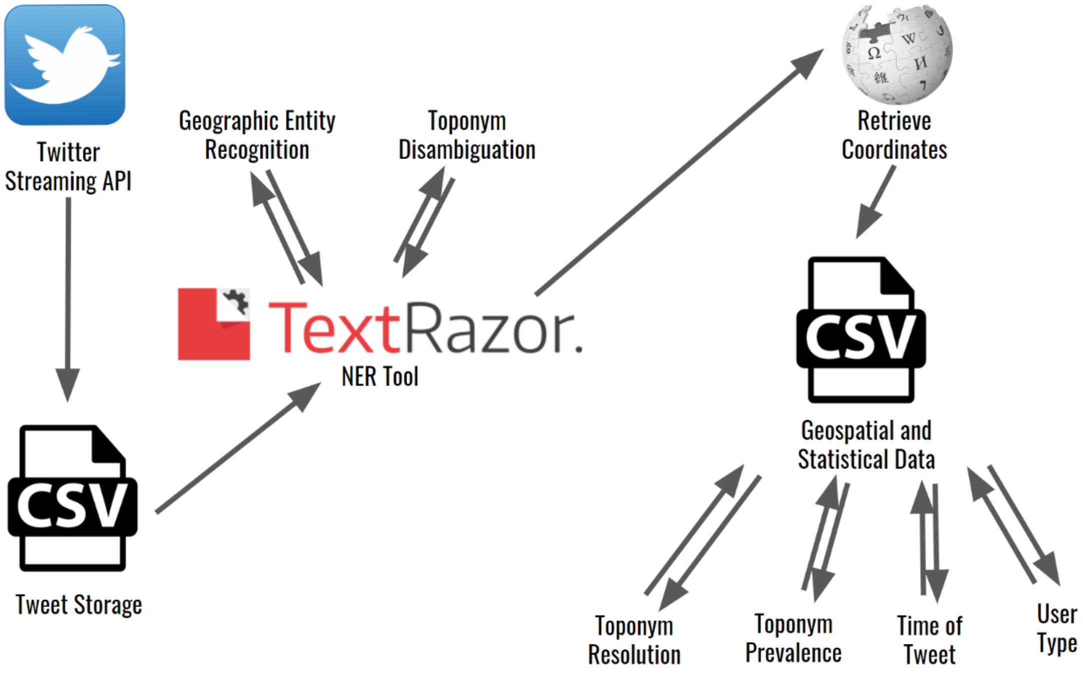


Figure 3 Workflow Diagram

4.1 Data

This study employs a dataset consisting of a collection of tweets from Twitter. Table 2 shows the Northern Triangle based gang and cartel keywords used to collect the tweets for this study. The geographical boundary for the collection was set to the entire world. Tweets were collected by the Center for Geospatial and Open-Source Intelligence using the GeoSocial Gauge System developed by Croitoru et al. [36].

Table 2 Cartel Related Keywords Used for Harvesting Data

Keywords
sinaloa
cartelde_sinalo
cartel
cartels
loszetas
zetas
carteldelgolfo
ms-13
ms13

The dataset contained 15,360,463 tweets collected from August 1, 2018 until August 31, 2019. The tweets were collected using the union of cartel related keywords and a bounding box of the entire world. Table 3 shows a summary of the collected data by month to include the total number of tweets, unique tweets, retweets, precisely geolocated tweets and their percentage out of the total dataset, along with the number of unique authors per month. The totals listed on the bottom of the table are a sum for every column except for unique authors. Instead, this total represents the number of unique authors in the whole dataset due to the presence of some authors, who tweet across multiple months.

Table 3 Summary of Dataset

Date	Total Tweets	Unique Tweets	Retweets	Precisely Geolocated Tweets	Unique Authors
08/2018	1042223	216087	826136	2462 (0.24%)	577545
09/2018	779721	177238	602483	2021 (0.26%)	446923
10/2018	742434	173883	568551	1811 (0.24%)	438712
11/2019	842783	187353	655430	2036 (0.24%)	511434
12/2018	950147	203745	746402	2121 (0.22%)	533185
01/2019	1720213	313942	1406271	3551 (0.21%)	735885
02/2019	1438420	265869	1172551	3087 (0.21%)	654353
03/2019	1215725	250045	965680	2538 (0.21%)	610205
04/2019	1698785	288089	1410696	3322 (0.20%)	784565
05/2019	1291183	241185	1049998	3161 (0.24%)	636357
06/2019	1257943	250574	1007369	2255 (0.18%)	589462
07/2019	1332607	231382	1101225	2259 (0.17%)	661144
08/2019	1048279	220803	827476	2029 (0.19%)	543803
TOTAL	**15360463**	**3020195**	**12340268**	**32653 (0.21%)**	**4517315**

Once the data is collected from Twitter, it can be exported in various formats. A tab-separated values (TSV) file was selected for our study due to the stability it provides as compared to a comma-separated values (CSV). Figure 4 shows the total number of tweets collected per day for the entire 396 days, or 13 months, contained in the dataset.

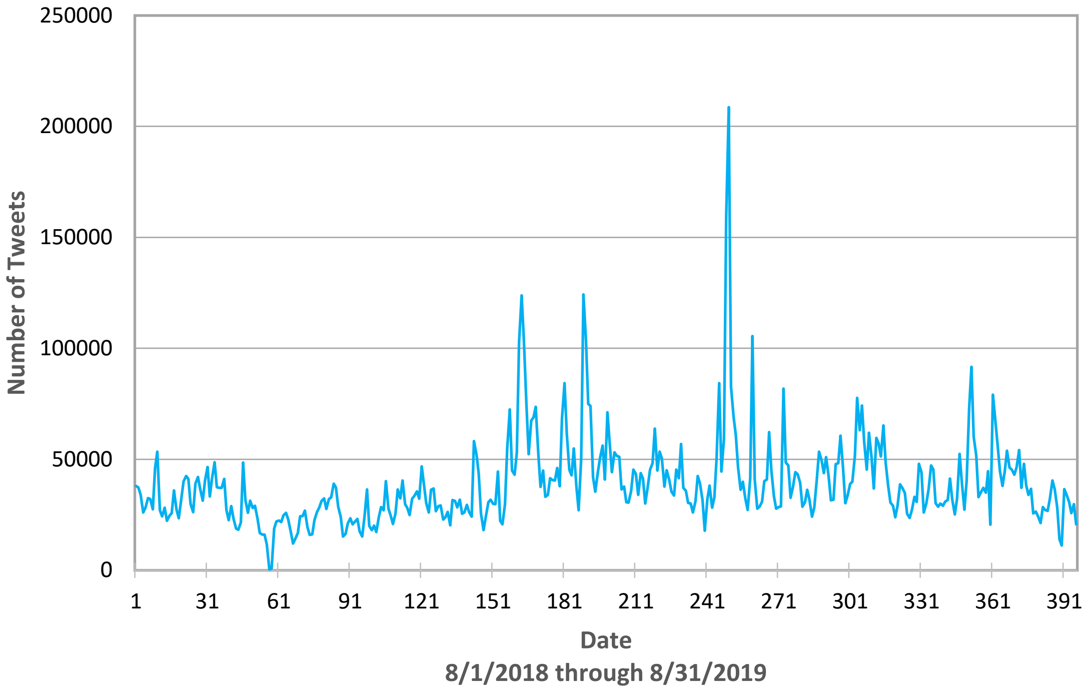


Figure 4 Total Number of Tweets Collected Per Day

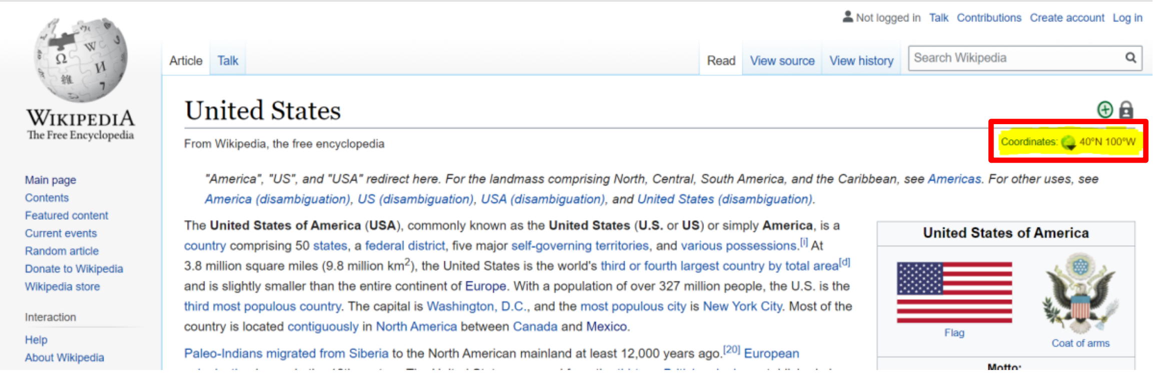
4.2 Named Entity Recognition

TextRazor does not require any preprocessing of the text, automatically detects 142 different languages, and can perform named entity recognition and disambiguation in English, Chinese, Dutch, French, German, Italian, Japanese, Polish, Portuguese, Russian, Spanish, and Swedish [19]. These traits in conjunction with the DBpedia results provided made TextRazor the ideal NER tool for the dataset. The first step in the analysis is to run the text portion of the tweets collected through the TextRazor API in order to identify the toponyms mentioned in the tweets.

In order to collect this information, the selected method included using batch processing to string 700-1000 tweets together into one call to the TextRazor API. This was done in order to maximize the use of the 200 KB limit on the size of one document that the TextRazor API could handle. Once each document had passed through TextRazor and toponyms had been identified, the results were broken back out and merged back to the TSV file that contained the original data from each tweet. The TextRazor results were then inspected using TextRazor entity details extracted from various web sources, including Wikipedia, DBpedia and Wikidata [19]. This step showed that TextRazor tagged all identified toponyms as an entity type of “Place” and provided the link to the location on Wikipedia.

4.3 Geoparsing

Once the toponyms were identified in the tweets, had been tagged as a “Place”, and TextRazor had provided the link to Wikipedia, the next step included the use of Wikipedia API [18] in order to scrape the coordinates from the top right corner of the Wikipedia page. This step was particularly necessary because Wikipedia webpages can have coordinates listed anywhere throughout the webpage, in the info box on the right side of the page, or in the top right corner of the page. Therefore, the failure to scrub this information could have resulted in superfluous coordinates throughout the remaining steps. The coordinates in the top right corner, highlighted in Figure 5, were used for consistency and accuracy [18]. It is important to note that these coordinates from Wikipedia mark the centroid of a place if the place is described by a polygon.



United States

From Wikipedia, the free encyclopedia

"America", "US", and "USA" redirect here. For the landmass comprising North, Central, South America, and the Caribbean, see Americas. For other uses, see America (disambiguation), US (disambiguation), USA (disambiguation), and United States (disambiguation).

The **United States of America** (**USA**), commonly known as the **United States** (**U.S.** or **US**) or simply **America**, is a country comprising 50 states, a federal district, five major self-governing territories, and various possessions.[i] At 3.8 million square miles (9.8 million km^2), the United States is the world's third or fourth largest country by total area[d] and is slightly smaller than the entire continent of Europe. With a population of over 327 million people, the U.S. is the third most populous country. The capital is Washington, D.C., and the most populous city is New York City. Most of the country is located contiguously in North America between Canada and Mexico.

Paleo-Indians migrated from Siberia to the North American mainland at least 12,000 years ago.[20] European

United States of America

Figure 5 Location of Coordinates on Wikipedia [37]

Once the toponyms were run through the Wikipedia API and the coordinates were found, the coordinates and the toponyms were then merged back to the TSV file with the original information from each tweet. In doing so, the analysis of the toponym content within the dataset could begin.

4.4 Toponym Statistic Analysis

After the tweets had been run through both the TextRazor and Wikipedia APIs and the information had been merged back to the tweet dataset, the next step was to calculate toponym statistics in order to answer the research questions. Pandas data frames and Jupyter Notebook were used with Python 3.0 in order to create columns of counts of the varying resolutions of the toponyms. Once these columns were added to the data frames, the complete dataset was exported as a CSV file for analysis. Depending on the size of the CSV file, it was either opened in Excel or again in Pandas in order to view the data and perform the analysis.

5. RESULTS AND ANALYSIS

Upon completion of the data analysis and production of toponym counts along with their resolution, the next logical step was to create graphics that best visualize the data in order to answer the research questions. This includes the analysis of the resolution of the toponyms present within tweets, the types of users who use toponyms in tweets, and the time in which toponyms are being used.

5.1 DBpedia Definitions of Toponym Resolutions

DBpedia is a project aiming to extract structured content from the information created in the Wikipedia project [28]. The results from TextRazor provided the DBpedia tags for the toponyms located in the tweets. Chapter 2.11 gives the definitions of the different resolutions of toponyms analyzed in this study and Figure 6 shows their hierarchy.

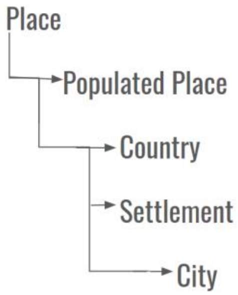


Figure 6 Hierarchy of Toponym tags from DBpedia

5.2 Research Question 1

The first research question provides a broad look into the context of the use of toponyms in Twitter and the results help guide our analysis further. Research question 1 states: To what extent do people use toponyms in tweets? What is the resolution of these toponyms? In order to answer these questions, this study employed a histogram with the count of toponyms in a single tweet for the full dataset shown in Figure 7, the unique tweets shown in Figure 8, and the retweets shown in Figure 9. The histograms for the unique tweets and retweets are further broken up by the specific resolutions of the toponyms present. It is important to note that the y-axis is a base 10 log scale for all of the histograms below. This is due to the nature of the data; it is most common for a tweet to have 1 toponym than for it to have 10 or more toponyms.

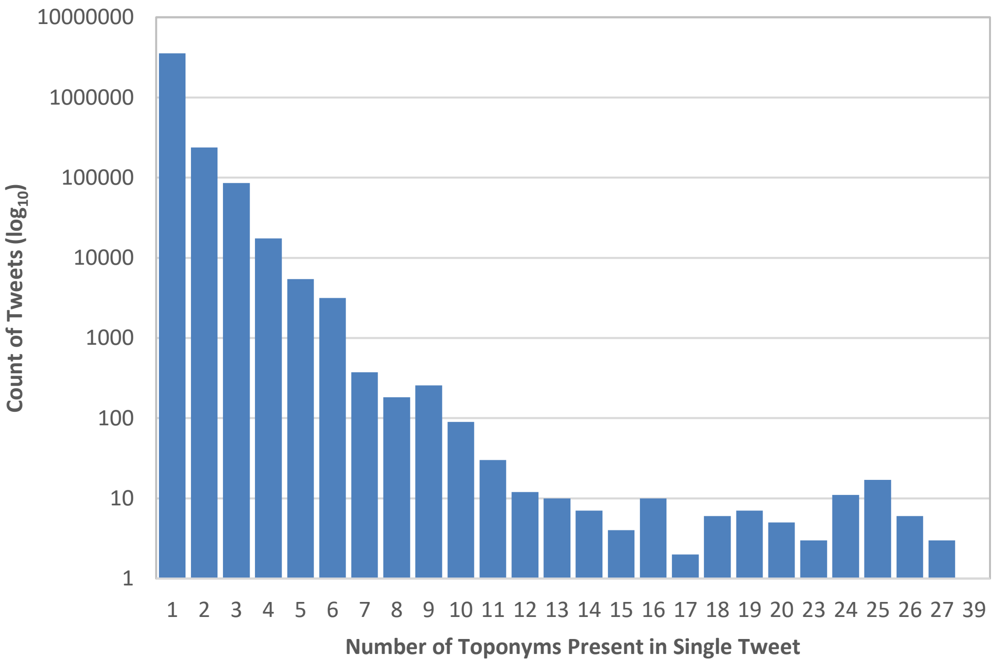


Figure 7 Toponym Counts for the Full Dataset (y-axis is shown in logarithmic scale)

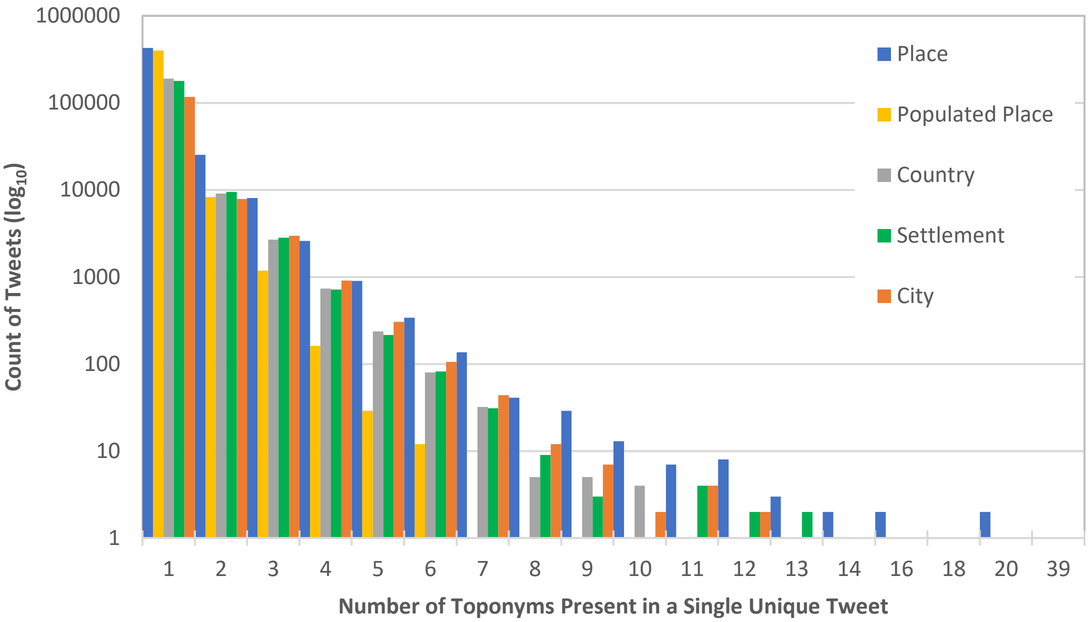


Figure 8 Histogram of Toponym Mentions in Unique Tweets (y-axis is shown in logarithmic scale)

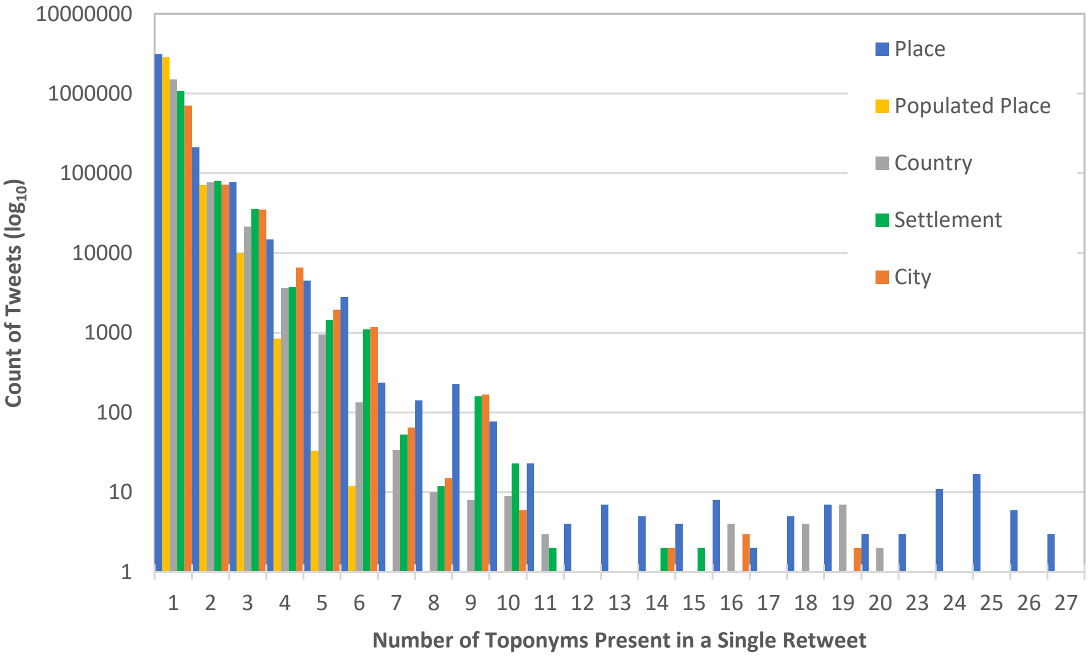
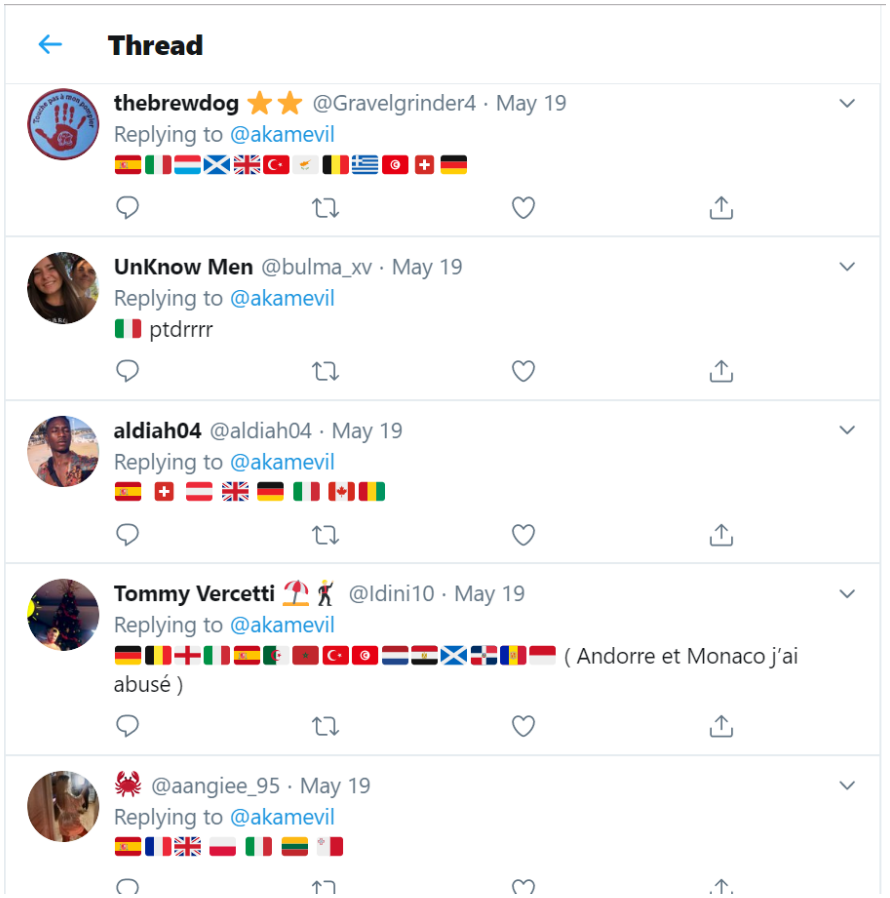


Figure 9 Histogram of Toponym Mentions in Retweets (y-axis is shown in logarithmic scale)

It is also important to note that TextRazor turned all country flag emojis used in tweets into the corresponding two letter country code. As a result, this caused all flag emojis used in the dataset to be returned as a place with the resolution of country. A manual check revealed that country flags in tweets were the cause for most of the tweets that have 12 or more places mentioned in them. Figure 10 depicts a thread that was captured in the dataset. The first tweet said to “Quote the different country flags you have visited.”



Thread

thebrewdog ⭐⭐ @Gravelgrinder4 · May 19
Replying to @akamevil
🇪🇸🇮🇹🇱🇺🏴󠁧󠁢󠁳󠁣󠁴󠁿🇬🇧🇹🇷🇨🇾🇧🇪🇬🇷🇹🇳🇨🇭🇩🇪

UnKnow Men @bulma_xv · May 19
Replying to @akamevil
🇮🇹 ptdrrrr

aldiah04 @aldiah04 · May 19
Replying to @akamevil
🇪🇸 🇨🇭 🇦🇹 🇬🇧 🇩🇪 🇮🇹 🇨🇦🇬🇳

Tommy Vercetti ⛱️🕺 @Idini10 · May 19
Replying to @akamevil
🇩🇪🇧🇪🏴󠁧󠁢󠁥󠁮󠁧󠁿🇮🇹🇪🇸🇩🇿🇲🇦🇹🇷🇹🇳🇳🇱🇪🇬🏴󠁧󠁢󠁳󠁣󠁴󠁿🇩🇴🇦🇩🇲🇨 (Andorre et Monaco j’ai abusé)

🦀 @aangiee_95 · May 19
Replying to @akamevil
🇪🇸🇫🇷🇬🇧 🇵🇱 🇮🇹 🇱🇹 🇲🇹

Figure 10 Thread of Tweets with Country Flags [38]

As ascertained, the most common type of toponyms were locations tagged as place followed closely by populated place, country, settlement, and finally city. This is expected because it follows the hierarchy DBpedia uses in the resolution of the tags of types of place. Furthermore, this trend is consistent through the dataset whether the tweet was unique or a retweet. It should also be noted that terms like “United Nations” and “Hell” were tagged as a place in the dataset, although they did not have coordinates attached to them due to the nature of the place.

After analyzing the above histograms, scatterplots of the frequency of the countries and cities mentioned in the unique tweets and retweets were produced. The country index created in order of prevalence in the dataset can be viewed in Appendix A-Country Index.

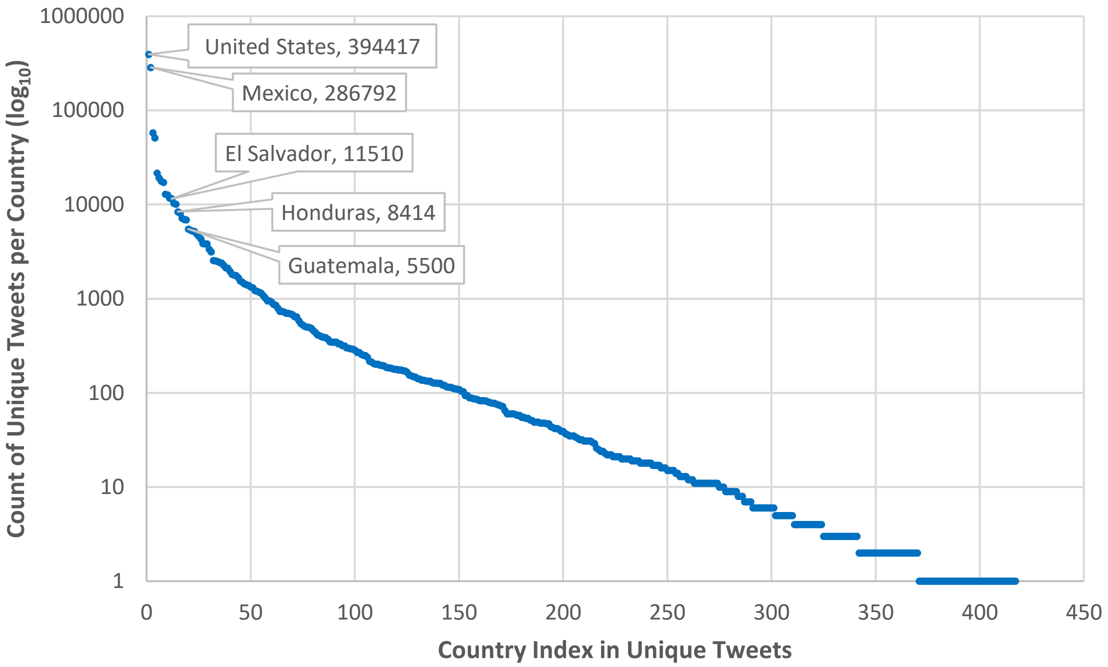


Figure 11 Country Frequency in Unique Tweets (y-axis is shown in logarithmic scale)

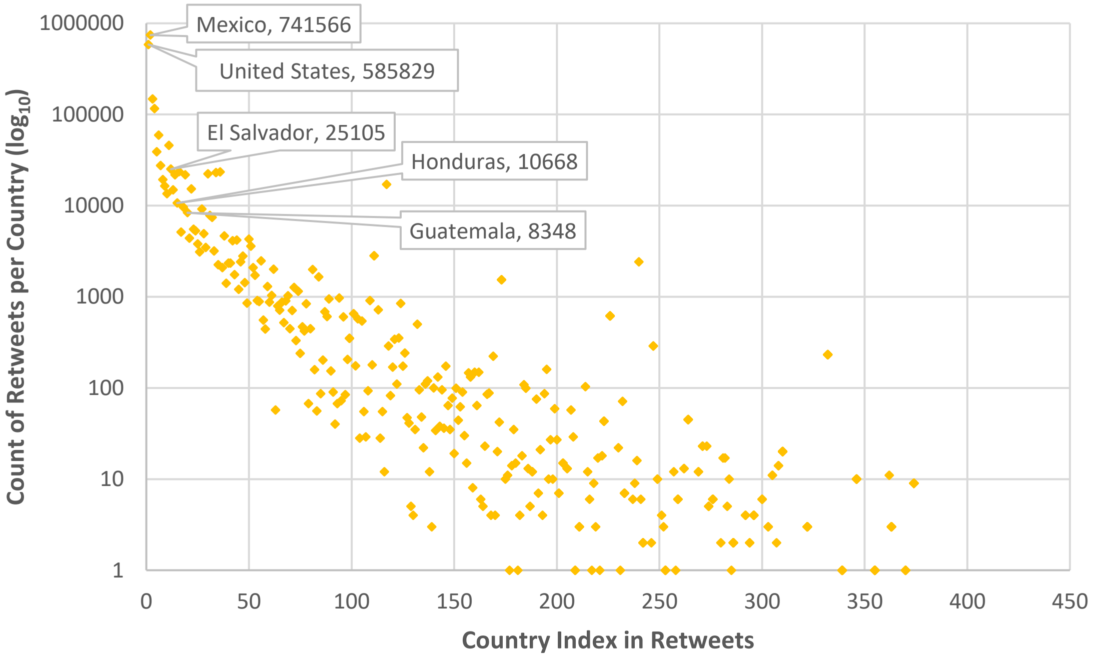


Figure 12 Country Frequency in Retweets (y-axis is shown in logarithmic scale)

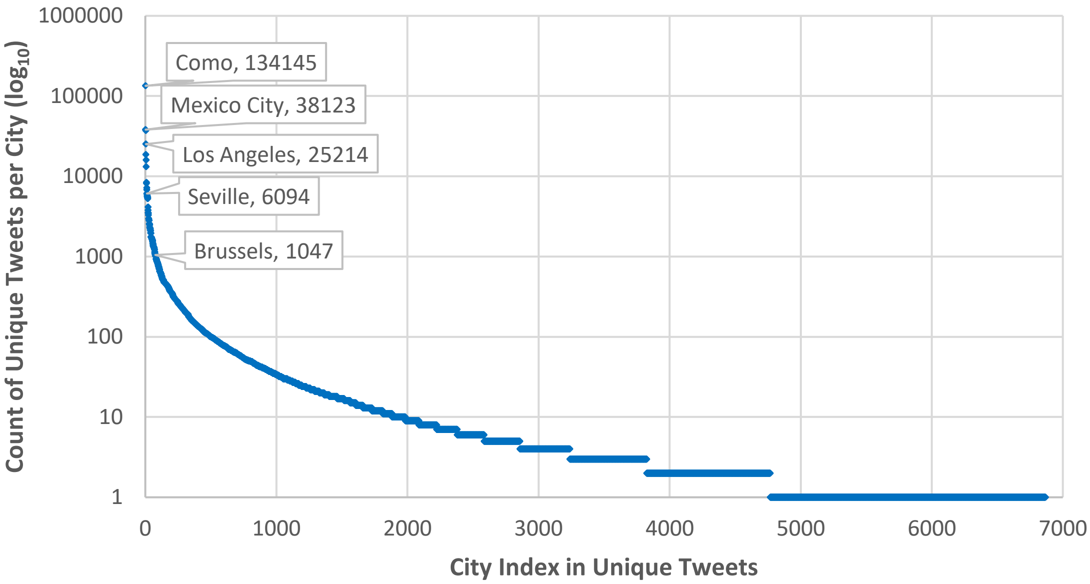


Figure 13 City Frequency in Unique Tweets (y-axis is shown in logarithmic scale)

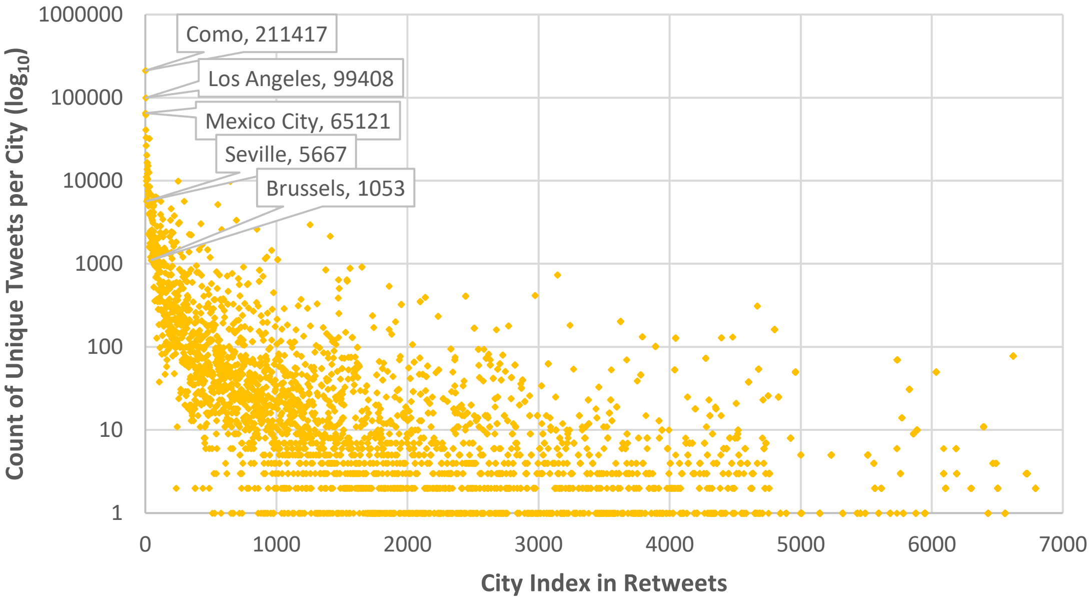


Figure 14 City Frequency in Retweets (y-axis is shown in logarithmic scale)

It is important to note that the order of cities on the x-axis in Figure 13 and Figure 14 stayed consistent just as they did for the order of the countries on the x-axis in Figure 11 and Figure 12. As before, these scatterplots also have a y-axis that is log base 10 due to the large volume of tweets.

Figure 11 and Figure 12 reveal that the United States was the most prevalent country mentioned in unique tweets with 394,417 tweets mentioning the US, while Mexico was most prevalent country in the retweets with 741,566 mentions. All three of the Northern Triangle countries were in the top twenty countries mentioned in both unique tweets and retweets.

Figure 13 and Figure 14 show the top cities mentioned in unique tweets and in retweets. Como, Italy, a small town in the Lombardy region, was tagged as the most mentioned city in both unique tweets and retweets. This demonstrates one of the limitations of using an NER tool on microblog posts. Due to the lack of context in the short length of the tweets, Como was tagged as a location in Italy. “Como” is also Spanish for “How.” When “Como” was found in tweets, it was likely the beginning of a question in Spanish, but due to the capital “C,” TextRazor tagged it as a location and provided the link to the Wikipedia page for Como, Italy. Other notable cities at the top of both the unique tweets and retweets were Mexico City, Mexico and Los Angeles, United States.

In addition, a series of choropleth maps were created depicting the number of times each country was mentioned in the full dataset shown in Figure 15, the unique tweets in Figure 16, the retweets in Figure 17, and in the precisely geolocated tweets shown Figure 18. The scale for each of the choropleth maps above is based on the log base 10 of the counts of each country. This was done in order to bridge the gap between countries like the United States with close to 1 million mentions across the dataset and countries like Bhutan that only had 1 mention.

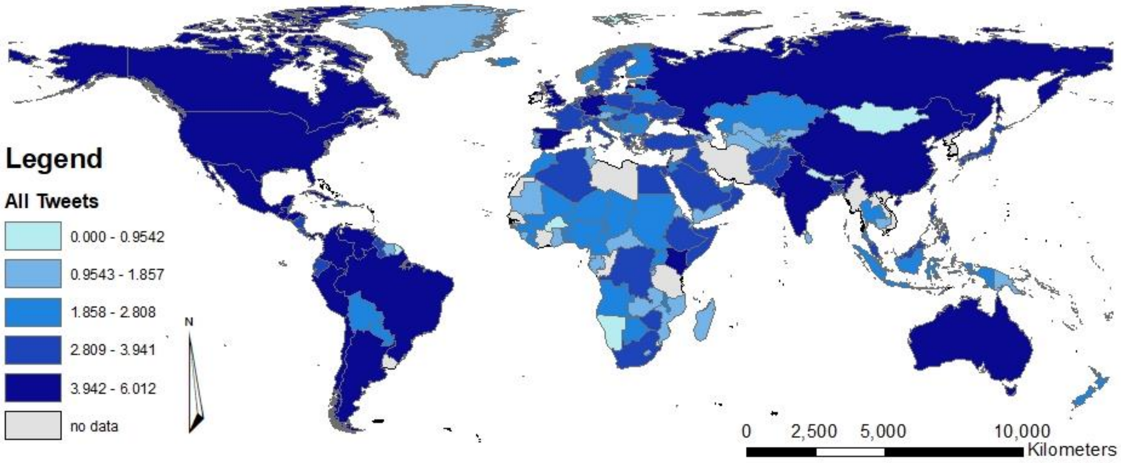


Figure 15 Choropleth Map of Country Mentions in Full Dataset (Legend is shown in logarithmic scale)

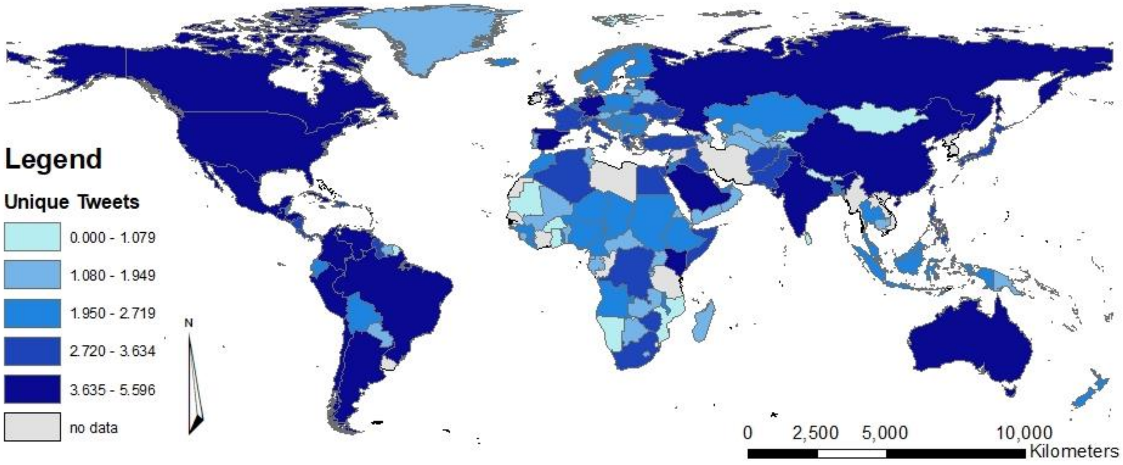


Figure 16 Choropleth Map of Country Mentions in Unique Tweets (Legend is shown in logarithmic scale)

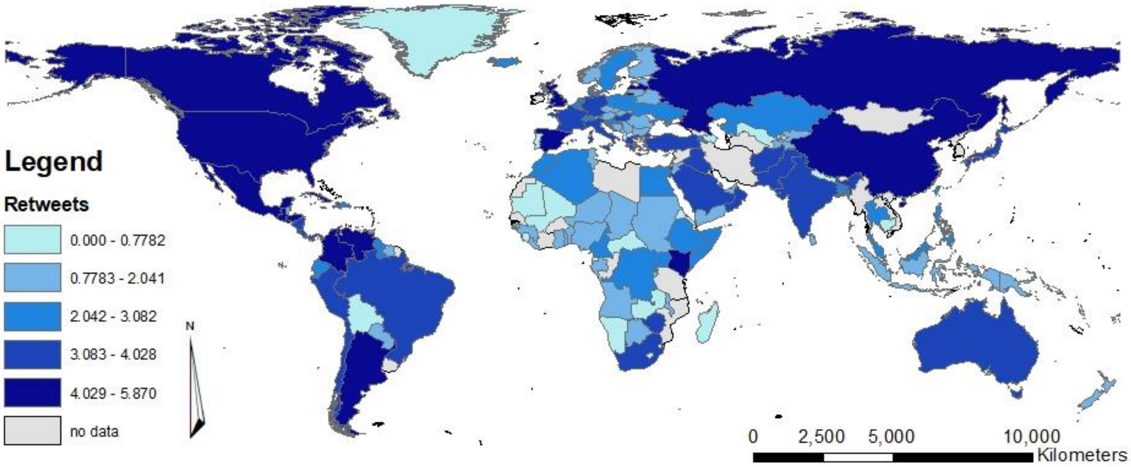


Figure 17 Choropleth Map of Country Mentions in Retweets (Legend is shown in logarithmic scale)

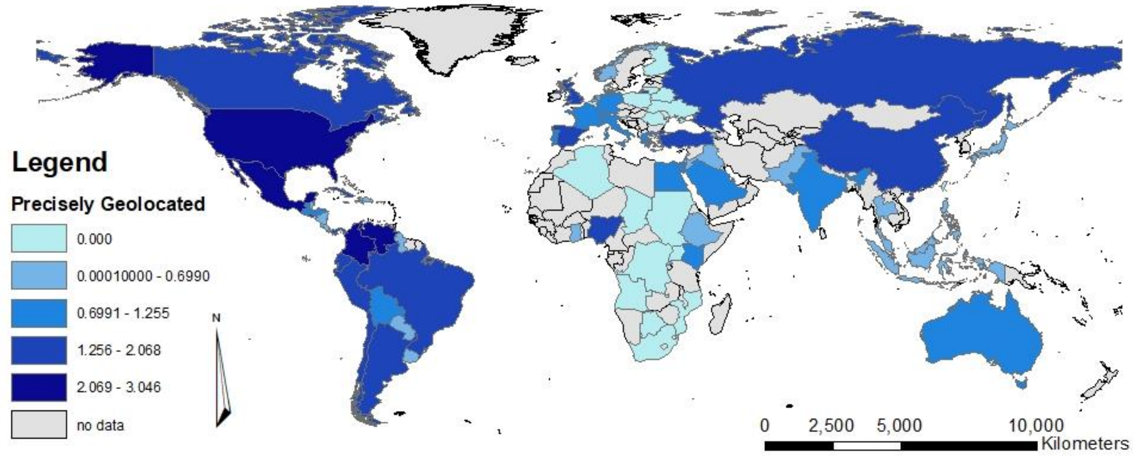


Figure 18 Choropleth Map of Country Mentions in Precisely Geolocated Tweets (Legend is shown in logarithmic scale)

All three Northern Triangle countries along with the United States, Mexico and surrounding Central American countries were in the most mentioned group in all four of the maps above. This suggests that the dataset contained valuable information of people tweeting about the dangerous Northern Triangle and the surrounding countries.

In addition to the maps of country mentions, a series of heat maps were created depicting the number of times each city was mentioned in the full dataset shown in Figure 19, the unique tweets in Figure 20, and the retweets in Figure 21. The scale for each of the heat maps is the same and shows the total number of times each city was mentioned.

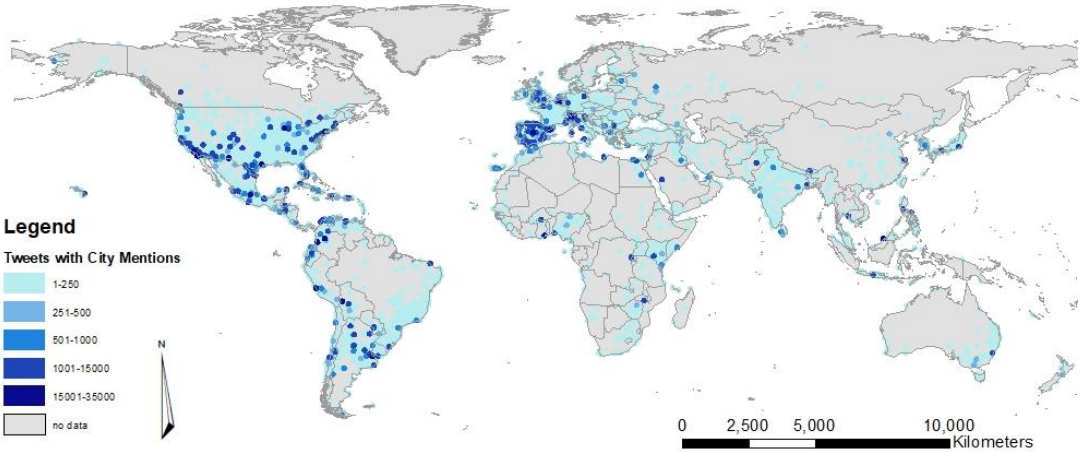


Figure 19 Heat Map of City Mentions in All Tweets (Legend is count of tweets)

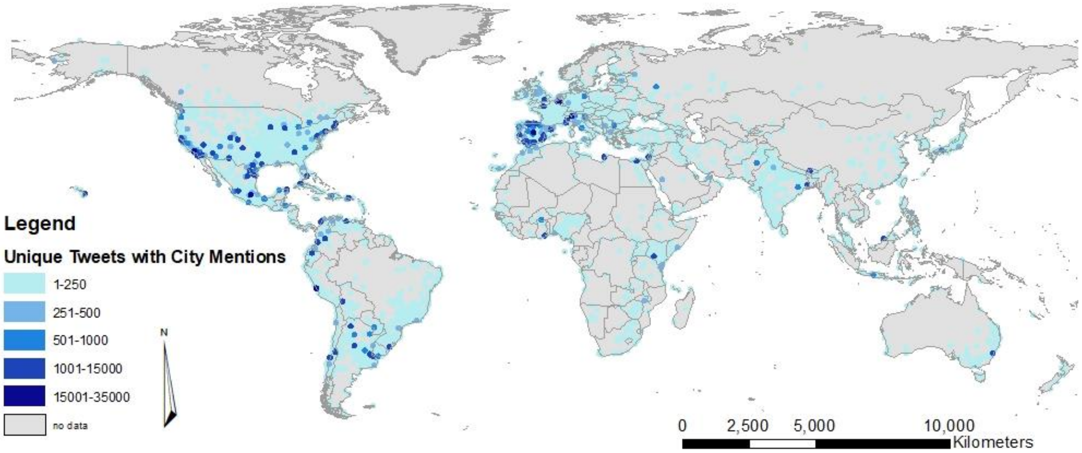


Figure 20 Heat Map of City Mentions in Unique Tweets (Legend is count of tweets)

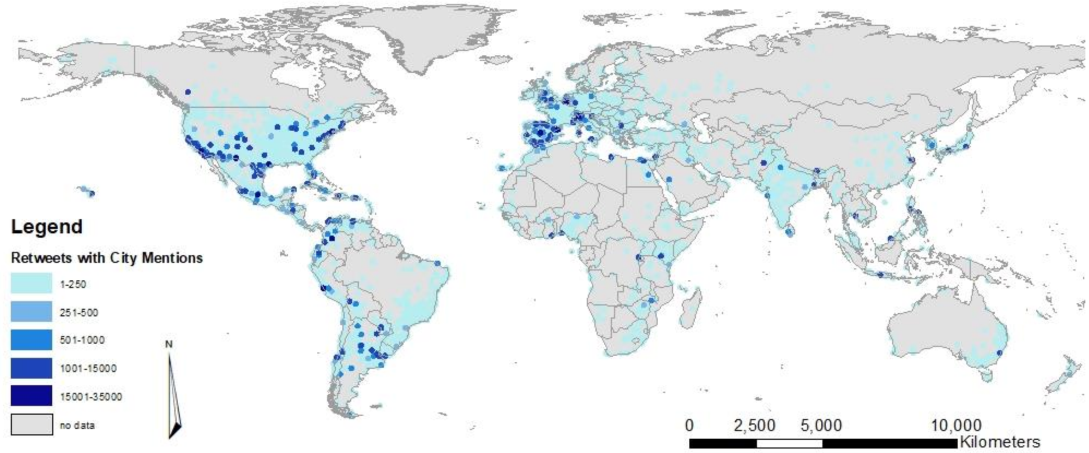


Figure 21 Heat Map of City Mentions in Retweets (Legend is count of tweets)

Figure 22 shows a zoomed in view of the Northern Triangle countries and their cities mentioned. The cities are shown with a graduated symbol size based on the total count of mentions for each city.

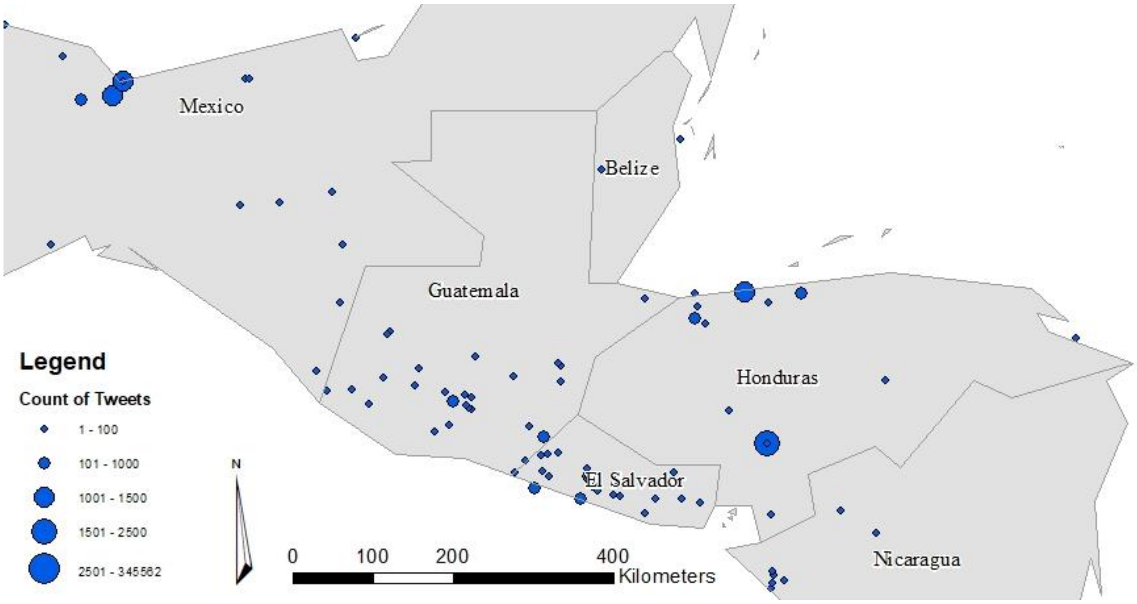


Figure 22 Map of Northern Triangle Country Toponyms

Error! Reference source not found. shows a detailed breakdown of the total mentions, mentions in unique tweets, and in retweets of the Northern Triangle countries and their cities. The table shows all cities from the three Northern Triangle countries with 50 or more mentions in the dataset.

Table 4 Northern Triangle Toponym Counts

Northern Triangle Toponyms	Total Tweets	Unique Tweets	Retweets
El Salvador	57498	11726	45772
Guatemala	28613	6901	21712
Honduras	31912	10123	21789
Tegucigalpa	2389	165	2224
Tela	1411	615	796
San Salvador	624	343	281
Mejicanos	390	177	213
Antigua Guatemala	361	8	353
San Pedro Sula	281	216	65
La Ceiba	131	79	52
El Adelanto	130	2	128
La Libertad, La Libertad	111	56	55
Acajutla	110	33	77
Comayagua	93	73	20
San Ignacio, Belize	72	39	33
Quetzaltenango	68	31	37
Tapachula	67	66	1
Izalco	67	12	55
Ilopango	61	31	30
Choluteca, Choluteca	57	51	6
Chichicastenango	57	14	43
Guatemala City	50	29	21
Mixco	50	19	31

The final analysis that was performed in order to answer the first research question was the calculation of the distance between the location of the toponym mentioned in the tweet to the location of the user at the time of the tweet. For this analysis, only precisely geolocated tweets were used. Figure 23 and Figure 24 show the resulting histograms.

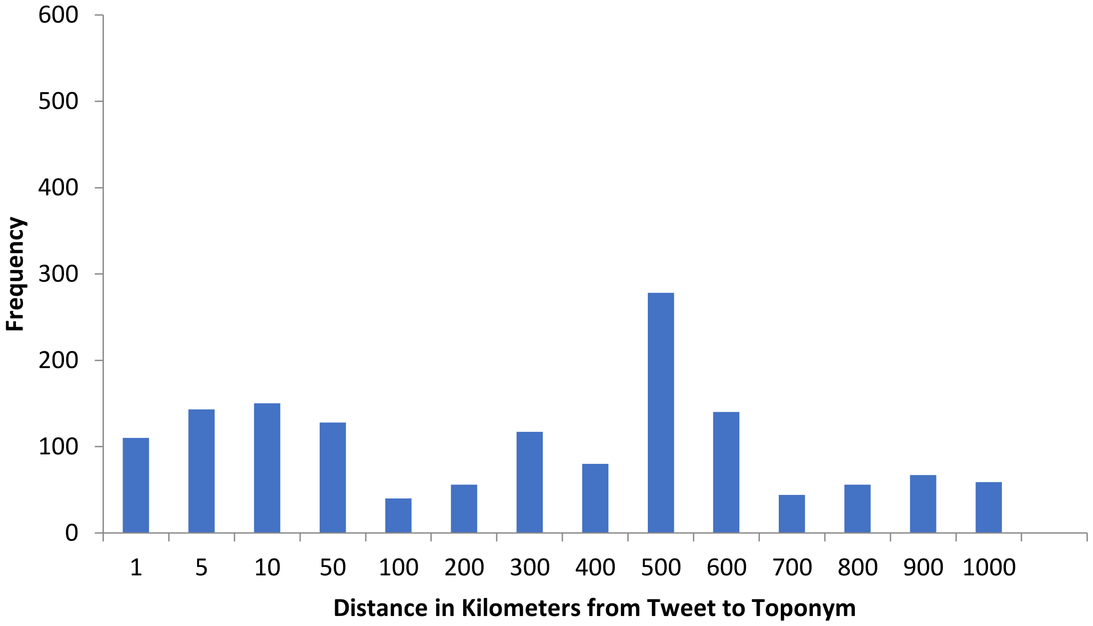


Figure 23 Distance up to 1000 km from a Geotagged Tweet to a Toponym Mentioned in a Tweet

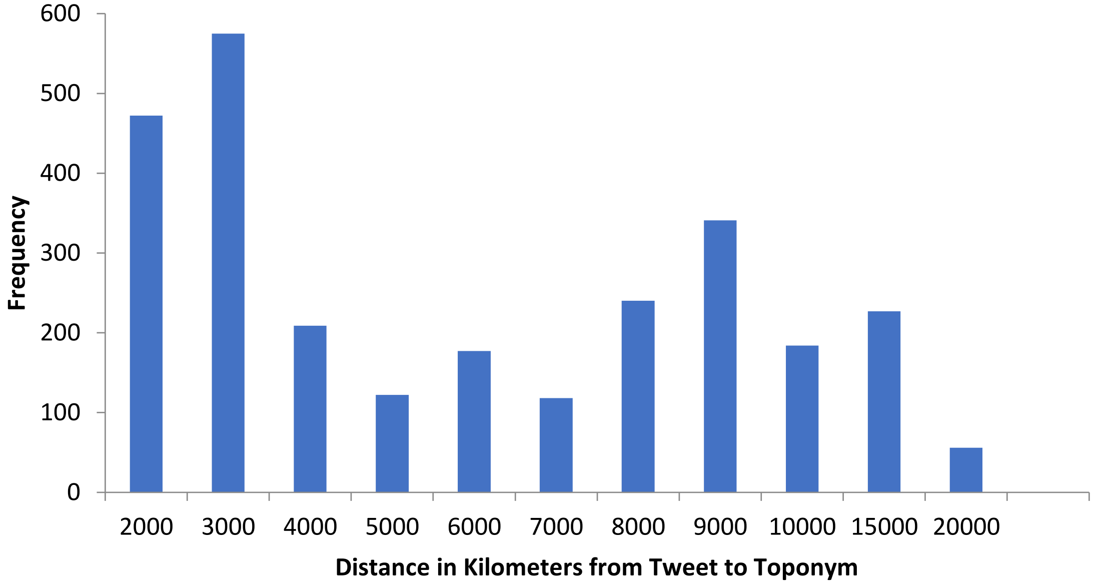


Figure 24 Distance more than 1000 km from a Geotagged Tweet to a Toponym Mentioned in a Tweet

These histograms were split into two different distances. The first distance was deemed to be any distance from 0 to 1,000 kilometers, and the second was any distance 2,000 kilometers or further away from the user. Based on these definitions, there were 1,468 users who included a toponym in their tweet regarding a place that was 1,000 kilometers or less away from where they were tweeting and 2,721 users who mentioned a location that was 2,000 kilometers or further away from them.

5.3 Research Question 2

An examination of the second research questions provides insight into which types of users include toponyms in their tweets. Research question 2 states: Which types of users are most likely to use toponyms? In order to answer this question, this study examined the most prolific users of toponyms throughout the dataset.

Figure 25 shows the total number of unique toponyms tweeted by users across the entire 13-month timeframe. The data suggests that most users use about 35 different toponyms over the span of 13 months. This is significant because examining unique toponyms as opposed to the total count of toponyms will help better evaluate the interconnectedness of the places which users referenced.

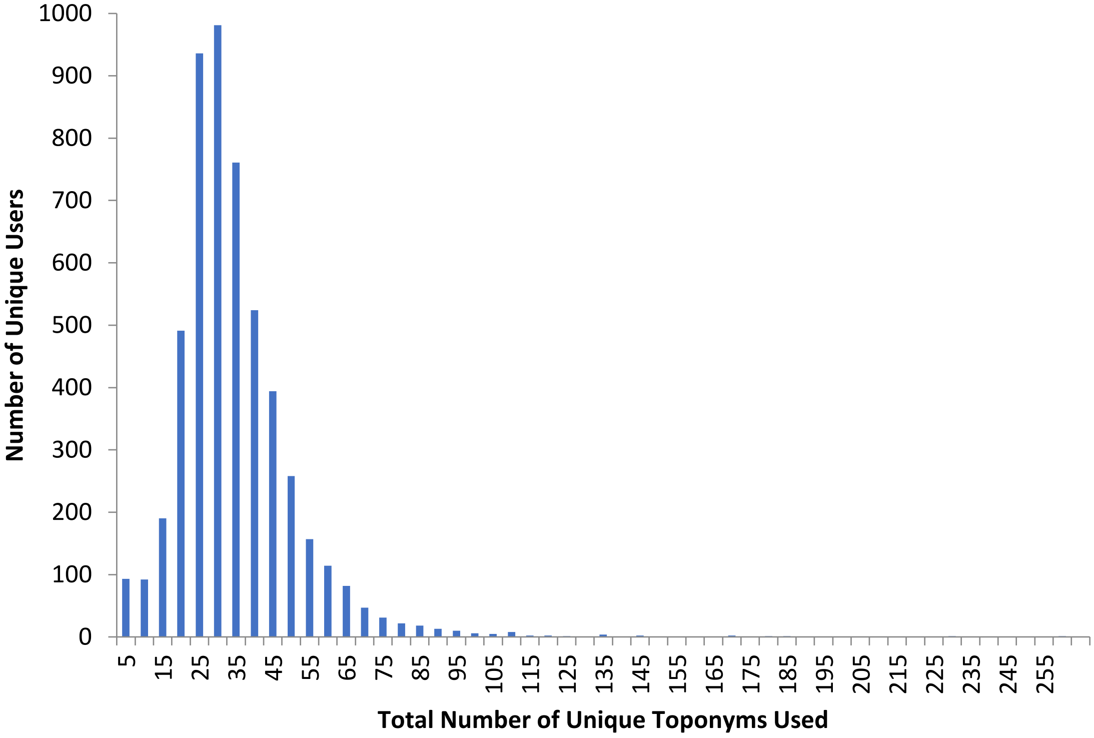


Figure 25 Unique Accounts According to Unique Toponym Count

After analyzing this chart, the top 100 user accounts based on their total unique toponyms were manually inspected. This resulted in Table 5.

Table 5 Top 100 Users Account Types based on Unique Toponym Count

User Type	Count
Individual	80
Reporter	12
News	3
Agency	4
Religious Organization	1

These user types were based on the information in user biographies, their profile and header pictures, their follower and following counts, and their most recent tweets. 85 of the top 100 authors profiles were in Spanish and 15 profiles were in English.

In addition to the unique users, the top ten languages in tweets that have one or more toponym present were identified. This is shown in Table 6. Figure 26 and Figure 27 illustrate the distribution of languages for tweets that had 1 or more toponyms. As predicted based on the keywords used to collect the data, 97% of the dataset was in either Spanish or English. Figure 27 shows the breakdown of the remaining 3% of languages.

Table 6 Top 10 Languages of Tweets with 1 or More Toponyms

Language	Number of Tweets with 1 or more Toponyms
English	1680539 (58.52%)
Spanish	1110591 (38.67%)
Portuguese	22297 (0.78%)
French	20185 (0.70%)
Catalan	15583 (0.54%)
Italian	3966 (0.14%)
German	3902 (0.14%)
Japanese	2348 (0.08%)
Dutch	2011 (0.07%)
Turkish	1521 (0.05%)
Other	8924 (0.31%)

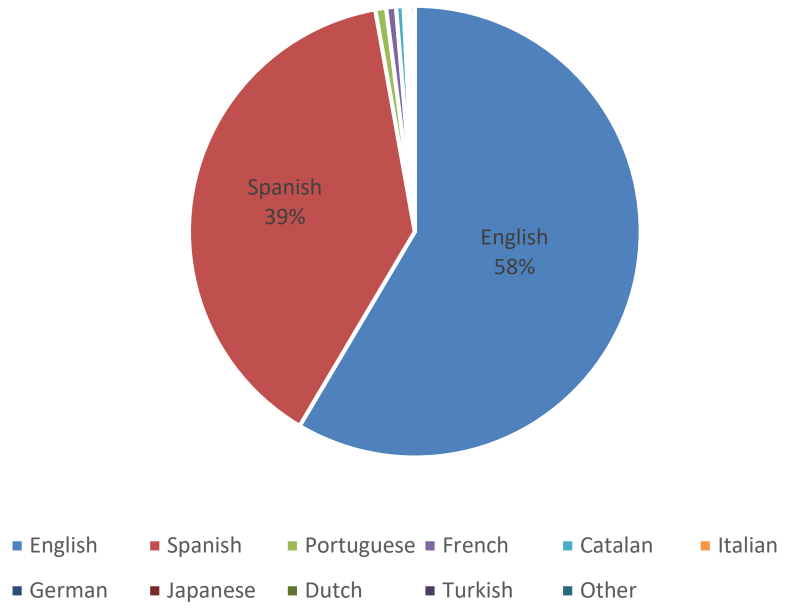


Figure 26 Language Distribution of Tweets that had 1 or More Toponyms

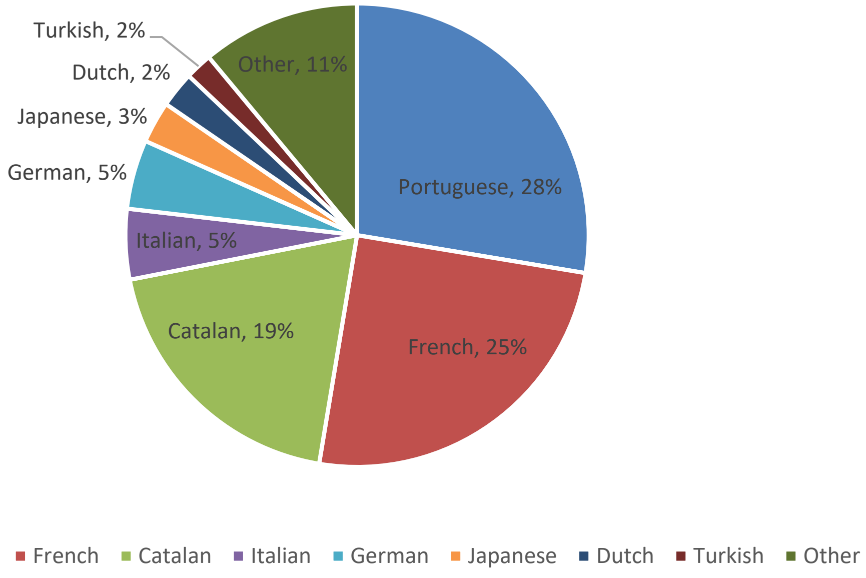


Figure 27 Language Distribution of Tweets with 1 or More Toponyms without English and Spanish

5.4 Research Question 3

The final research question will give a broad look into the times that toponyms are used on Twitter and the results will help guide our analysis further. Research question 3 states: How does toponym usage change over time? In order to answer this final research question, we calculated the counts for the total number of toponyms used per day across the 396-day dataset.

Figure 28 shows the total number of tweets, unique tweets and retweets combined, per day compared to the number of tweets with one or more toponyms per day. Then, Figure 29 shows the daily percent of tweets with 1 or more toponym out of the total number of tweets.

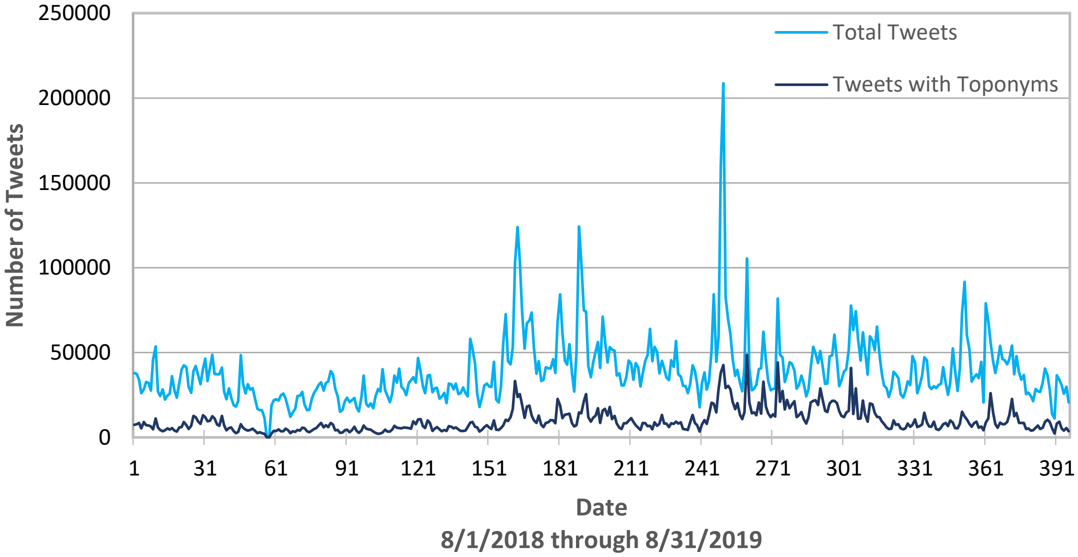


Figure 28 Total Tweets vs Tweets with Toponyms Per Day

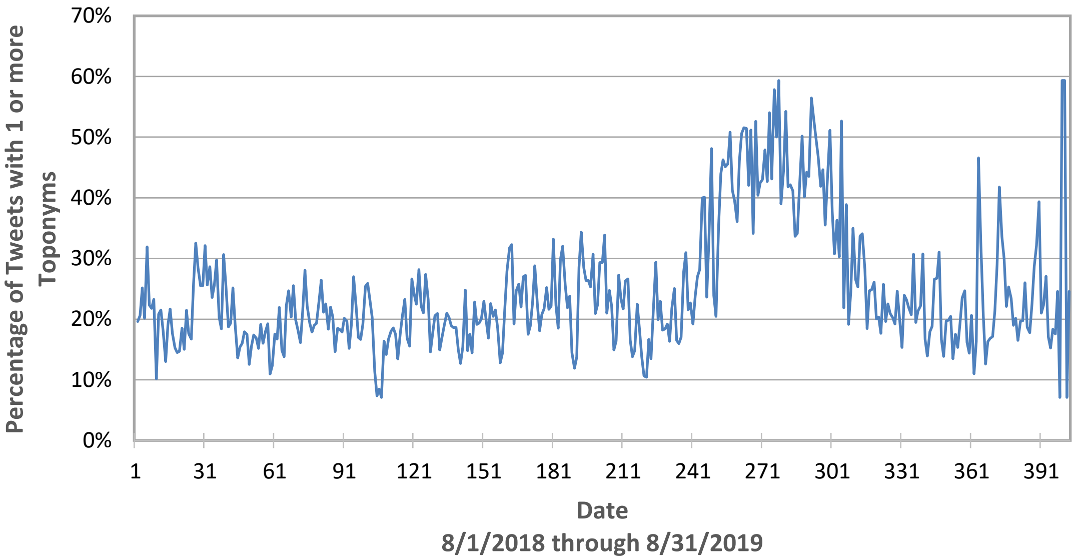


Figure 29 Daily Percent of Tweets with 1 or More Toponyms

There was an average of 24.58% of tweets with toponyms per day in the dataset. May 4th, 2019 had the highest with 59.34% of tweets containing toponyms and November 14th, 2018 had the lowest with only 7.10% of tweets with toponyms. May 4th, 2019 likely had such a high percentage of tweets with toponyms since that day was the 145th annual Kentucky Derby, and by far the most controversial one to date. This is because the winning horse was disqualified and the new winner, Country House having 65-1 odds for him to win. This is was a major moment in horse racing history and had Twitter buzzing over the event. It is also interesting to note that most of the jockeys that ride the horses in the Derby come from Latin American countries. This could be another explanation for why this event was so prevalent in the dataset. As a whole, the percentage of tweets with a toponym present in them maintains an average of 24.58%, or almost 1 in 4 tweets, throughout all 396 days analyzed.

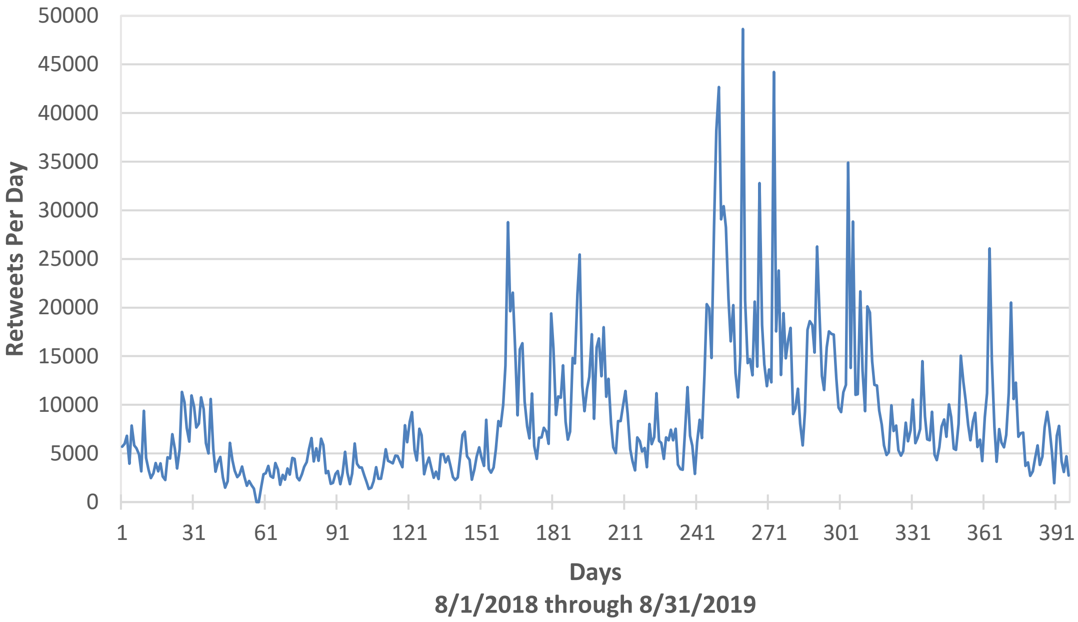


Figure 30 Retweets with Toponyms Per Day

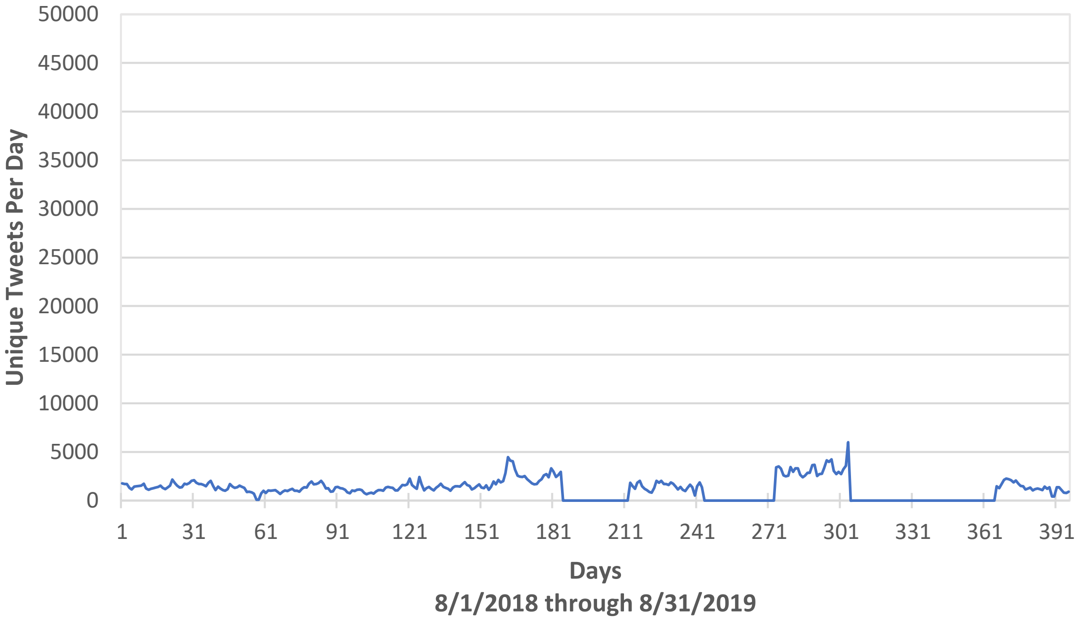


Figure 31 Unique Tweets with Toponyms Per Day

Figure 30 and Figure 31 show that toponyms were more prevalent in retweets than in unique tweets. This matches expectations because 80.34% of our dataset was retweets while only 19.66% of the dataset was unique tweets.

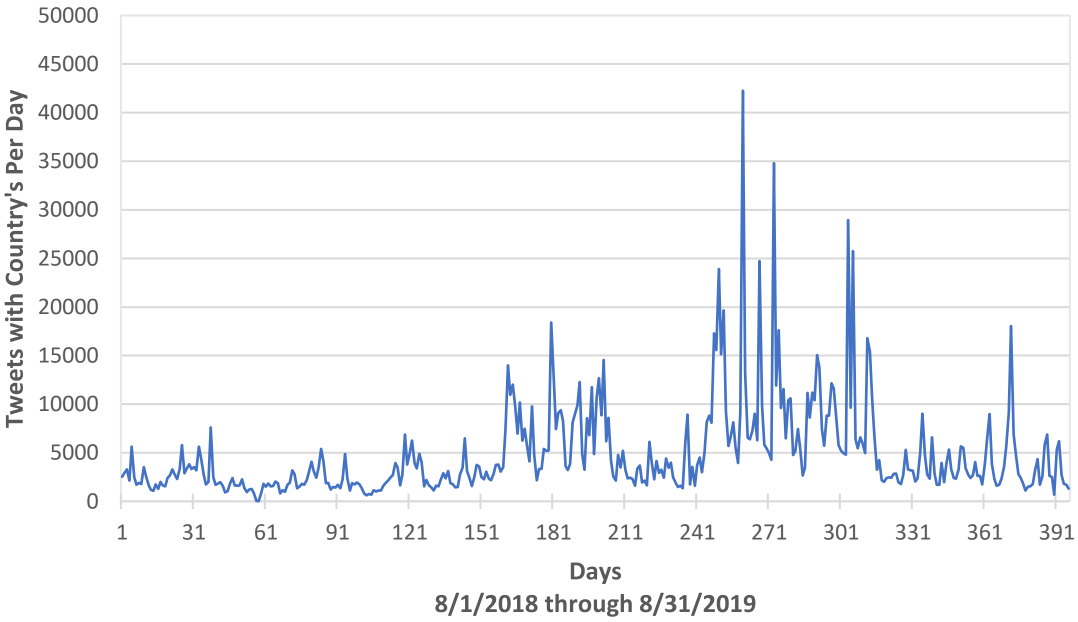


Figure 32 Total Number of Tweets with a Country Mentioned Per Day

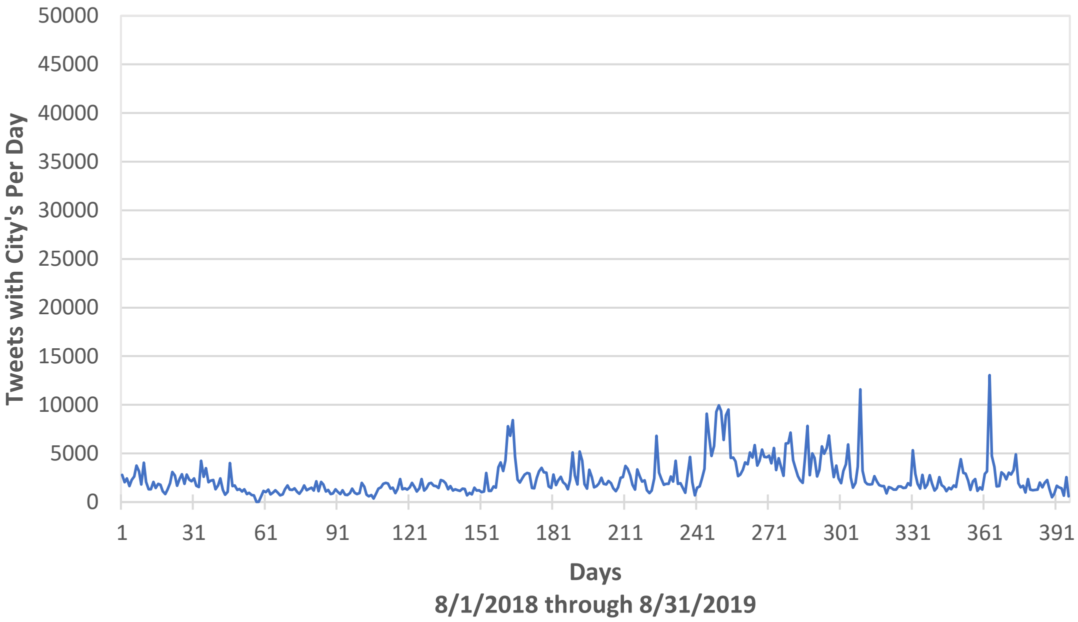


Figure 33 Total Number of Tweets with a City Mentioned Per Day

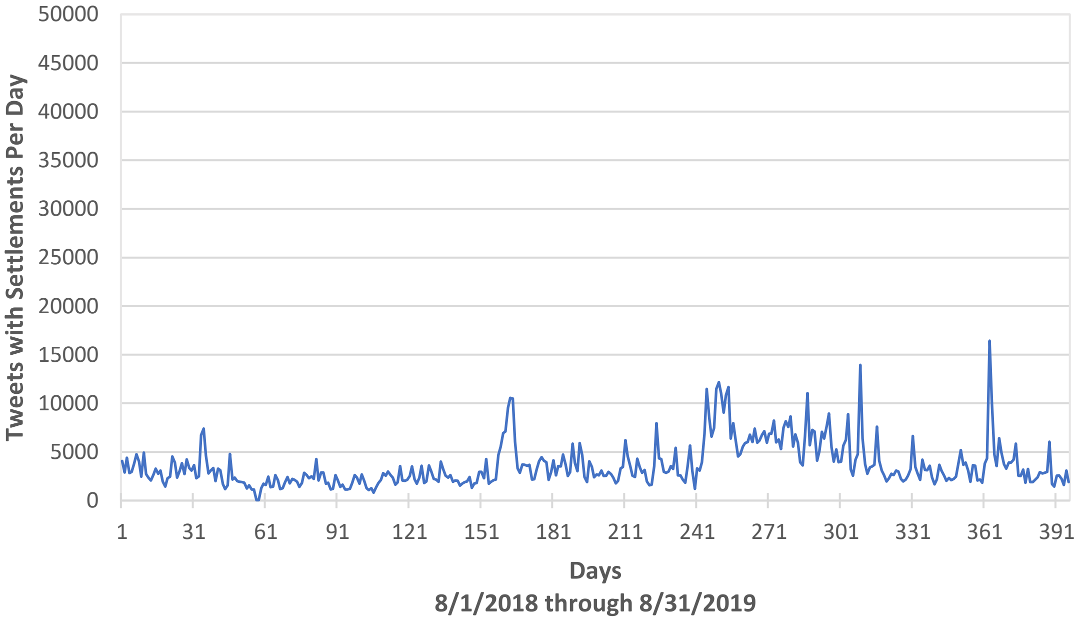


Figure 34 Total Number of Tweets with a Settlement Mentioned Per Day

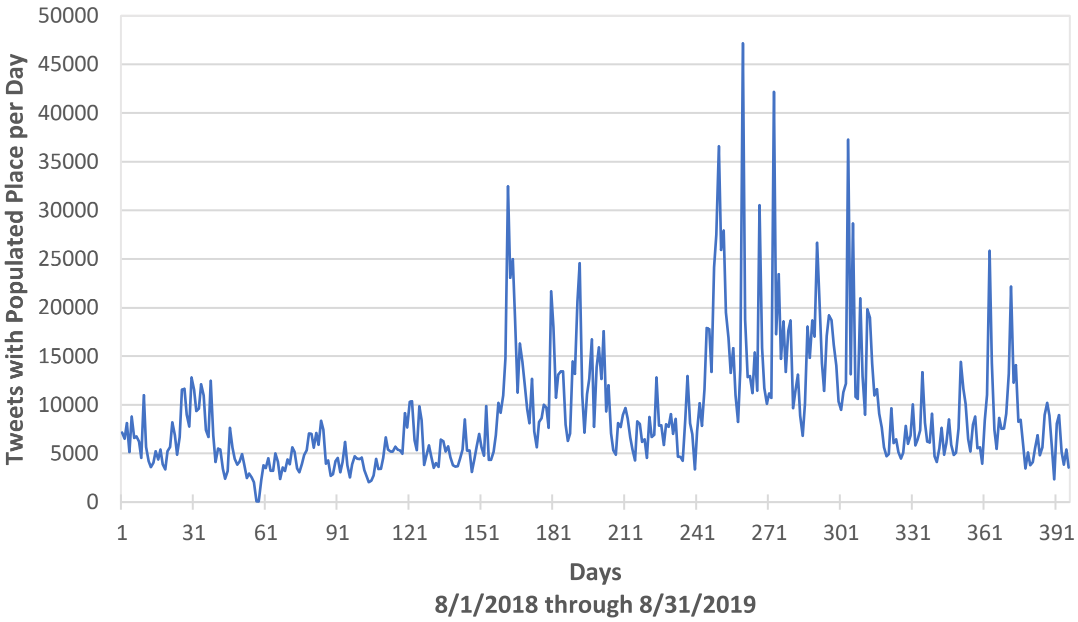


Figure 35 Total Number of Tweets with a Populated Place Mentioned Per Day

Figure 32, Figure 33, Figure 34, and Figure 35 show that country and populated place were the most common types of toponyms found in the dataset. City and settlement only accounted for about 10% of the toponyms. This again suggests that people tweet about more broad locations than specific locations.

This analysis shows that ultimately, toponym usage remained relatively constant with the exception of April and May 2019. On average, 1 in 4 tweets had a toponym present in them. Upon closer examination, certain days with higher than average percentage of tweets with toponyms typically had some sort of event that happened on that day. Accordingly, the most common type of event to cause a rise in the percentage of toponyms in tweets was sporting events. The Kentucky Derby on May 4th, 2019 highlights this phenomenon. A probable explanation for this is that most professional sporting teams, regardless of the sport, have a toponym in their team name and sporting events often have toponyms in their titles.

6. CONCLUSION AND OUTLOOK

After reviewing the literature, it was found that there is a gap on the analysis of the extent of toponym usage in tweets. This study found that Twitter data is a rich source of toponym content. By using a NER tool like TextRazor, toponyms can be extracted and geolocated to attach coordinates to tweets. These coordinates provide a great deal of information about the context of the tweet. They give location context to not only where the author is tweeting from, but also to the places mentioned in the tweets themselves. The following discussion covers all 3 research topics analyzed in this thesis: the extent in which people use toponyms in tweets and the resolution of these toponyms, the types of users most likely to use toponyms, and how toponym usage changes over time.

6.1 Discussion

After analyzing 13 months of the Northern Triangle Twitter dataset, I found that almost 1 in 4 tweets, 24.58%, contains at least one toponym. This reveals that performing a NER analysis on tweets in order to extract toponyms provides a much larger sample size than just using precisely geolocated tweets which accounted for 0.21% of the dataset. While most tweets, 97.11%, have only 1 or 2 toponyms, there were tweets with up to 39 toponyms present in a single tweet. After manually inspecting these tweets, it was found that most tweets that had 12 or more toponyms present were using country flag emojis.

TextRazor turned these country flag emojis into the associated 2-letter country code which was then tagged as a country toponym.

This analysis also showed that people are more likely to tweet at a coarser resolution than a fine resolution. The frequency of country names was almost 10 times more than the frequency of cities in our dataset. This could be in part to the structure of TextRazor and the DBpedia ontology hierarchy, shown in Figure 6, used for the tags on types of places.

Figure 11 and Figure 12 show the frequency of country toponyms in the unique tweets and retweets respectively. Figure 12 makes the frequency of country names present in retweets appear more dispersed than in the unique tweets. This is due to the order of the countries on the x-axis staying the same for both figures. The same applies for the frequency of cities in Figure 13 and Figure 14. These figures also show that the retweets are amplified by the public. This means that countries and cities that appear higher up in the retweet figures than the unique tweet figures were more popular among the public and this was amplified through retweets.

Figure 13 and Figure 14 show the top cities mentioned in unique tweets and in retweets. Como, Italy, a small town in the Lombardy region, was tagged as the most mentioned city in both unique tweets and retweets. This demonstrates one of the downfalls of using an NER tool on microblog posts. Due to the lack of context in the short length of the tweets, Como was tagged as a location in Italy. “Como” is also Spanish for “How.” When “Como” was found in tweets, it was likely the beginning of a question in Spanish, but due to the capital “C,” TextRazor tagged it as a location and provided the link to the Wikipedia page for Como, Italy.

Figure 22 and Table 4 show the Northern Tringle countries, cities, and their total count of mentions from the dataset. The table shows that while all three countries were mentioned frequently throughout the dataset, the none of the cities from Northern Triangle countries had a significant number of mentions. This shows that while this dataset includes information at the country level, it is lacking a more refined insight on what is happening throughout the cities in Northern Triangle countries.

Figure 23 and Figure 24 are histograms showing the distance from the precisely geolocated user to the location of the toponym present in their tweet. This analysis was only performed on tweets that had 1 toponym present within the tweet. There were 1,468 users, 35.04%, who included a toponym in their tweet talking about a place that was 1,000 kilometers or less away from where they were tweeting and 2,721 users, 64.96%, who mentioned a location that was 2,000 kilometers or further away from them. This shows us that most users tweet about places that are far away from them, meaning that users are more concerned with global issues than local issues.

Once the top 100 prolific users of toponyms were identified and manually evaluated, Table 5 shows that 80% of the most prolific users were individuals. This spread of user types was a surprising find. The original expectation was that the most common user types would be news and government agencies. After manual inspection, the results reveal that most of the top authors in our dataset are simply individuals who retweet an abundance of news tweets that were not present in our dataset. This is then counted as a unique toponym for the author that retweeted it. However, it was not surprising to find that 97% of tweets with one or more toponyms in the dataset is either Spanish or English. This was expected based on the keywords that were used to collect the data.

Figure 29 shows the percent of tweets with one or more toponyms present per day throughout the entire 13 months of data. There is a noticeable peak for April and May of 2019. In these 2 months, the lowest day has 30.3% of tweets with a toponym and these months contained the highest rates per day across the entire dataset, May 4, 2019, with 59.34%. In April 2019 Notre Dame caught fire, over 200 were killed in an attack in Sri Lanka, and New Zealand changed their gun laws [39]. In May 2019 North Korea launched missiles, same-sex marriage was legalized in Taiwan, and the US Embassy in Honduras caught fire [40]. These are just a few of the notable world news events that could be associated with the increased use of toponyms on Twitter during that timeframe.

Figure 31 shows the unique tweets with toponyms per day. This figure shows that there are three gaps in the data where there are zero unique tweets with toponyms present. After manual inspection, these days contain a significant number of retweets in which the original tweet is not present in our dataset. For this reason there are large spikes on these days in Figure 30, but a lack of unique tweets on the same days in Figure 31.

6.2 Conclusion

The analysis of this data showed that the most common resolution of toponyms present in a tweet is a country as shown in Section 5.2. It also demonstrated that the most common user to tweet the most unique toponyms in tweets is an individual as shown in

Section 5.3. Lastly, it showed that the use of toponyms does not display significant variance over time with an average appearance of one out of every four tweets with a peak in April and May as shown in Section 5.4. There are days with major events that caused spikes in the percentage of tweets with toponyms. There was also a general increase in the use of toponyms for April and May 2019, but it was pretty consistent across the remaining 11 months analyzed.

Overall, this study provides extensive insight into the prevalence, resolution, and user classification in tweets. While the results are conclusive, they do not necessarily prove that this dataset would be able to be used to locate and track drug cartels and gangs from the Northern Triangle into the United States. This is due to the lack of prevalence of refined locations within Northern Triangle countries.

6.3 Limitations

Although TextRazor was a helpful off-the-shelf NER tool, there are some limitations in using this kind of tool to analyze a dataset of multilingual tweets collected globally. First, there was the aforementioned issue in which "Como" appeared to be the most mentioned city in the dataset. The lack of context present in microblog posts from Twitter causes similar issues to be more prevalent than on other social media platforms. Second, places like "Hell" were tagged as a place by TextRazor. While this is place people refer to, it is not a toponym in the sense of the word for this study. Places like this caused inflation in the total count of toponyms mentioned in tweets. The resolutions of places were not as detailed as was originally expected. In spite of DBpedia's extensive ontology and long list of tags for types of places, TextRazor results only included five basic resolutions used in this study for majority of the toponyms tagged. Lastly, since the tweets were not cleaned before being run through TextRazor, emojis remained in the tweets. This caused the country flags to be turned into the two letter country codes as shown in Section 5.2.

Another limitation of this study was working with a multilingual dataset as this led to an overreliance on Google Translate when manually evaluating the top 100 prolific users. The spread of languages also affected the NER tools that were able to be used on the dataset. This caused issues with identifying places, see “Como” example above. In addition, it affected the ability to determine the user types. In addition to this, the NER tool was also unable to identify colloquial place names.

Finally, the Wikipedia API used to attach coordinates to the toponyms identified in the tweets served as a limitation. The only coordinates used were those in the top right corner of the Wikipedia page shown in Figure 5. While this served the purpose of maintaining consistency and accuracy throughout the analysis [18], there were places that did not have coordinates listed in that location, but instead were listed elsewhere in the article. This, in conjunction with a multilingual dataset made toponym disambiguation, and the use of colloquial, rather than formal names, harder for TextRazor to distinguish. Consequently, this led to false positives in the results.

6.4 Future Work

The impact from this research enables future studies to utilize the NER and geoparsing system as a method for determining toponym prevalence and resolution in a dataset. Once the toponym prevalence is determined, this method should increase a researcher’s confidence in using toponyms located in his or her Twitter data for further research and evaluation of the locations found.

Future work could break the dataset up by language and run a single language through the NER tool at a time. This could allow for better toponym disambiguation. It would also be beneficial to strip the emojis from the data before running them through the NER tool to see how the total toponym count varies. Future work could use more specific regional gazetteers in order to avoid the issue with the Wikipedia coordinates and limited resolutions of places in DBpedia. In addition, Scott McDermott’s work explains how a heuristic approach can assist in the identification and disambiguation of colloquial place names [44]. *Crowdsourcing Urban Form and Function* discusses the typology of implicit and explicit form and function content that builds upon traditional and crowdsourced data [45]. This paper discusses that while “explicit form and function data are purpose-driven and target-specific usages and users, implicit form and function content results through the repurposing of open-source data to address a variety of usages and user types” [45].

Finally, future work could also include using this dataset to map the unique toponyms used by prolific users over time. This would show the hot spot areas that are of interest to the prolific users in the dataset. The current dataset provides many additional opportunities for study of the relations between multiple toponyms mentioned in a single tweet and the varying resolutions of them.

With this being written, this thesis demonstrates that Twitter is a prolific source of toponym content. These toponyms can be used to attach specific coordinates that provide a great deal of information to tweets.

APPENDIX A- COUNTRY INDEX

1	United States	31	Nigeria	61	Guyana	91	Serbia
2	Mexico	32	Panama	62	Costa Rica	92	Liberia
3	Venezuela	33	Syria	63	Holy See	93	Jordan
4	Colombia	34	Korea	64	Netherlands	94	Belgium
5	China	35	Qatar	65	South Korea	95	Gibraltar
6	Russia	36	North Korea	66	Dominican Republic	96	Barbados
7	Spain	37	Afghanistan	67	Egypt	97	Norway
8	Canada	38	Puerto Rico	68	Denmark	98	Rwanda
9	Kenya	39	Bolivia	69	Democratic Republic of the Congo	99	Monaco
10	United Kingdom	40	Japan	70	Yemen	100	Bulgaria
11	El Salvador	41	Republic of Ireland	71	Algeria	101	Albania
12	Iran	42	Turkey	72	Hong Kong	102	Vatican City
13	Argentina	43	Zimbabwe	73	Taiwan	103	Austria
14	Honduras	44	Iraq	74	Uganda	104	Malta
15	Brazil	45	Jersey	75	Poland	105	Korea
16	Cuba	46	Haiti	76	Ethiopia	106	United Arab Emirates
17	Israel	47	Paraguay	77	Sweden	107	Romania
18	Chile	48	Pakistan	78	Ghana	108	Francoist Spain
19	Guatemala	49	Switzerland	79	New Zealand	109	Ecuador
20	Peru	50	Francia	80	Malaysia	110	Morocco
21	Germany	51	Portugal	81	Crimea	111	Francia
22	Nazi Germany	52	South Africa	82	Jamaica	112	Mauritius
23	Australia	53	Greece	83	Indonesia	113	Tanzania
24	India	54	Somalia	84	Hungary	114	Malawi
25	Saudi Arabia	55	England	85	Sudan	115	Finland
26	France	56	Libya	86	Thailand	116	Russian Empire
27	Ecuador	57	Philippines	87	Lebanon	117	Latvia
28	Nicaragua	58	Ukraine	88	Bangladesh	118	Rusia
29	Italy	59	Vietnam	89	Sri Lanka	119	Belize
30	Middle East	60	Armenia	90	Singapore	120	Kazakhstan

121	Nepal	181	Sierra Leone	241	Trinidad and Tobago
122	Angola	182	Cambodia	242	Isle of Man
123	Palau	183	Yemen	243	Marshall Islands
124	Slovenia	184	Slovakia	244	Czechoslovakia
125	Iceland	185	Falkland Islands	245	Turkmenistan
126	Andorra	186	Maldives	246	Uzbekistan
127	Mercia	187	West Bank	247	Montenegro
128	Nigeria	188	Mozambique	248	Lesotho
129	State of Palestine	189	Yugoslavia	249	Texcoco (altepetl)
130	Bolivia	190	Lithuania	250	Djibouti
131	Chad	191	Palmares (quilombo)	251	Central African Republic
132	Cameroon	192	Tunisia	252	Ukrainian People's Republic
133	Niger	193	ChÅ«zan	253	Phoenicia
134	Myanmar	194	Vichy France	254	Organisation internationale de la Francophonie
135	Czech Republic	195	Tlatelolco (altepetl)	255	Autonomous Republic of Crimea
136	Cayman Islands	196	Namibia	256	Miranda (state)
137	The Bahamas	197	Suriname	257	Tanzania
138	Wales	198	Macau	258	Somaliland
139	Luxembourg	199	Benin	259	Zambia
140	Bosnia and Herzegovina	200	Bahrain	260	North Vietnam
141	Estonia	201	Azores	261	Nauru
142	International Criminal Court	202	Liechtenstein	262	Ghana
143	South Sudan	203	Dominica	263	Polish People's Republic
144	Kosovo	204	Georgia (country)	264	Kyrgyzstan
145	Gaza Strip	205	Biafra	265	Zulu Kingdom
146	Croatia	206	Madagascar	266	Mozambique
147	Queensland	207	Burkina Faso	267	United States Virgin Islands
148	Zambia	208	Saba	268	Niue
149	Madeira	209	Portugal	269	Vanuatu
150	Montserrat	210	Eritrea	270	Guam
151	Guinea	211	Gran Colombia	271	Burundi
152	Togo	212	Tonga	272	Bonaire
153	Saint Lucia	213	Singapore	273	Guinea-Bissau
154	Basque Country (greater region)	214	Serbia	274	Eswatini
155	Bermuda	215	Grenada	275	Martinique
156	Fiji	216	The Gambia	276	Mauritania
157	Uganda	217	Equatorial Guinea	277	Andorra
158	Mongolia	218	Moldova	278	East Timor
159	San Marino	219	Comoros	279	Belice
160	Antigua and Barbuda	220	Kingdom of LeÃ³n	280	Qing dynasty
161	Paraguay	221	Cape Verde	281	Madagascar
162	Ivory Coast	222	Papua New Guinea	282	Weimar Republic
163	Mali	223	Bhutan	283	Dejima
164	North Macedonia	224	Republic of the Congo	284	Sri Lanka
165	Montenegro	225	Hong Kong	285	Solomon Islands
166	Cyprus	226	Herm	286	Austria-Hungary
167	Botswana	227	Empire of Japan	287	Mongolia
168	Kuwait	228	Albania	288	Grand Duchy of Baden
169	Sparta	229	Moab	289	Esparta
170	CuraÃ§ao	230	Tajikistan	290	Nordic countries
171	Seychelles	231	Crown of Castile	291	Saint Helena
172	Sardinia	232	Ashanti Empire	292	Mongolia
173	Oman	233	Iraqi Kurdistan	293	Ascension Island
174	Greenland	234	Kosovo	294	Nepal
175	Burundi	235	Zanzibar	295	Viceroyalty of New Granada
176	Laos	236	Samoa	296	Sark
177	Azerbaijan	237	Joseon	297	Kurdistan
178	Palestinian territories	238	Austria	298	Antarctic
179	Belarus	239	Hispania	299	San Marino
180	Gabon	240	Brunei	300	Autonomous Province of Kosovo and Metohija

301	Kingdom of Galicia
302	Nauru
303	Southern Nigeria Protectorate
304	Katanga Province
305	Cambodia
306	Namibia
307	South Vietnam
308	Zaire
309	West Berlin
310	Seychelles
311	ChimÃº culture
312	Saint Vincent and the Grenadines
313	Christmas Island
314	Kingdom of Asturias
315	Transnistria
316	Transkei
317	Kuwait
318	American Samoa
319	Samoa
320	Brandenburg
321	Government of National Unity (Hungary)
322	California Republic
323	Collectivity of Saint Martin
324	French Guiana
325	Republic of Maryland
326	Republic of Canada
327	Samma dynasty
328	Tlaxcala (Nahua state)
329	Brunei
330	Near East
331	Caliphate of CÃ³rdoba
332	QwaQwa
333	Anguilla
334	Madeira
335	Velay
336	Kingdom of Navarre
337	Eritrea
338	Kiribati
339	French First Republic
340	Coalition Provisional Authority
341	Turks and Caicos Islands
342	German Empire
343	Tonga
344	Assyria
345	Galmudug
346	Spanish protectorate in Morocco
347	Sovereign Military Order of Malta
348	State of Katanga
349	Burkina Faso
350	Akkadian Empire
351	Han dynasty
352	British Virgin Islands
353	Socialist Federal Republic of Yugoslavia
354	Cook Islands
355	Republika Srpska
356	Republic of Genoa
357	Dutch East Indies
358	Duchy of Schleswig
359	Duchy of Lorraine
360	Pitcairn Islands
361	First French Empire
362	Freetown Christiania
363	Faroe Islands
364	ChÅ«zan
365	Kingdom of Hungary
366	SÃ£o TomÃ© and PrÃ-ncipe
367	Kingdom of Granada (Crown of Castile)
368	Tuvalu
369	Kingdom of Aragon
370	Hispania Tarraconensis
371	Bailiwick of Guernsey
372	Alodia
373	Tondo (historical polity)
374	Yuan dynasty
375	Sam'al
376	Svalbard
377	Syrian Republic (1946â€“1963)
378	Sint Maarten
379	Sayn
380	Senegambia Confederation
381	Second French Empire
382	Serbia and Montenegro
383	Urartu
384	Islamic Emirate of Afghanistan
385	Irish Free State
386	Inini
387	Korea under Japanese rule
388	Khanate of Kalat
389	Empire of Japan
390	Empire of China (1915â€“1916)
391	Empire of Brazil
392	Duchy of Neopatras
393	Dos Pilas
394	Demerara
395	Democratic Kampuchea
396	Guernsey
397	Conch Republic
398	Cospaia
399	French Third Republic
400	French Polynesia
401	Free Territory of Trieste
402	Glamorgan
403	General Government
404	Ghana Empire
405	Bhutan
406	Panama Canal Zone
407	Bonaire
408	Republic of Entre RÃ-os
409	Republic of the Rio Grande
410	Saint Kitts and Nevis
411	Saint BarthÃ©lemy
412	Qing dynasty
413	Macedonia (region)
414	Neutral Moresnet
415	New Caledonia
416	Northern Region, Nigeria
417	Buganda

REFERENCES

[1] Renwick, Danielle, and Rocio Cara Labrador. “Central America's Violent Northern Triangle.” Council on Foreign Relations, Council on Foreign Relations, 1 October 2019, www.cfr.org/backgrounder/central-americas-violent-northern-triangle.

[2] Return and Reintegration in the Northern Triangle Program. USAID, 2019, https://www.usaid.gov/sites/default/files/documents/1862/Fact%20Sheet%20-%20Return%20and%20Reintegration%20in%20the%20Northern%20Triangle%20Program.pdf, Accessed 19 Apr. 2019.

[3] Bruns, A. (2012). How long is a tweet? Mapping dynamic conversation networks on Twitterusing Gawk and Gephi. Information, Communication & Society, 15(9), 1323-1351.

[4] Aguirre Tobon, Katherine and Muggah Robert. “Citizen Security in Latin America: Facts and Figures”, Igarapé Institute, https://igarape.org.br/wp-content/uploads/2018/04/Citizen- Security-in-Latin-America-Facts-and-Figures.pdf, Accessed 20 Apr. 2019.

[5] Western Hemisphere. U.S. Department of State. Access 20 Apr. 2019.

[6] Cara Labrador, Rocio and Renwick, Danielle. Central Americas Northern Triangle, 26 June 2018, https://www.cfr.org/backgrounder/central-americas-violent-northern-triangle, Accessed 1 Apr. 2019.

[7] Northern Triangle Migration Information Initiative. USAID, 2019, https://www.usaid.gov/sites/default/files/documents/1862/Fact_Sheet_-_Northern_Triangle_Information_Management_Initiative.pdf, Accessed 19 Apr. 2019.

[8] Plan of the Alliance for Prosperity in the Northern Triangle: A Road Map. Inter-American Development Bank, 2014, http://idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=39224238, Accessed 21 Apr. 2019.

[9] Greenfield, V., Mitch, I., et al. “Human Smuggling and Associated Revenues”. Rand Corporation, https://www.rand.org/pubs/research_reports/RR2852.html, 2019, Accessed 24 Apr. 2019.

[10] Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. GeoJournal, 78(2), 319-338.

[11] Molina, B. (2017, October 26). Twitter overcounted active users since 2014, shares surge on profit hopes. Retrieved from https://www.usatoday.com

[12] Aramaki, E., Maskawa, S., & Morita, M. (2011, July). Twitter catches the flu: detecting influenza epidemics using Twitter. In Proceedings of the conference on empirical methods in natural language processing (pp. 1568-1576). Association for Computational Linguistics.

[13] Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake: Twitter as a distributed sensor system. Transactions in GIS, 17(1), 124-147.

[14] Buntain, C., & Golbeck, J. (2015, May). This is your Twitter on drugs: Any questions?. In Proceedings of the 24th international conference on World Wide Web (pp. 777-782). ACM.

[15] Galindo, Y., (2017, October 25). Machine Learning Detects Marketing and Sale of Opioids on Twitter. Retrieved from https://health.ucsd.edu

[16] Leidner, J.L. and Lieberman, M.D., 2011. Detecting geographical references in the form of place names and associated spatial natural language. SIGSPATIAL Special, 3 (2), 5–11.

[17] Abdallah, Zahraa Said et al. “Multi-domain evaluation framework for named entity recognition tools.” Computer Speech & Language 43 (2017): 34-55.

[18] “Wikipedia API.” Wikipedia, Wikimedia Foundation, en.wikipedia.org/w/api.php?action=help&amp;modules=query%2Bcoordinates.

[19] “The Natural Language Processing API.” TextRazor, 1 Feb. 2019, www.textrazor.com/technology.

[20] DeLozier, Grant et al. “Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles.” Association for the Advancement of Artificial Intelligence (2015).

[21] Leidner, Jochen L.. (2007). Toponym Resolution in Text. ACM SIGIR Forum. 41. 124.10.1145/1328964.1328989.

[22] Gritta, Milan & Pilevar, Mohammad Taher & Limsopatham, Nut & Collier, Nigel. (2017). What's missing in geographical parsing?. Language Resources and Evaluation. 52.10.1007/s10579-017-9385-8.

[23] Fize, Jacques & Shrivastava, Gaurav. (2017). GeoDict: an integrated gazetteer.

[24] "About Geonames." Geonames, 2019, https://www.geonames.org/about.html

[25] "Directory of Cities, Towns, and Regions in El Salvador." Fallingrain, 2019, http://fallingrain.com/world/ES/

[26] A. Toral and R. Munoz. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In EACL.

[27] Hill L.L. (2000) Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: Borbinha J., Baker T. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2000. Lecture Notes in Computer Science, vol 1923. Springer, Berlin, Heidelberg

[28] "About." DBpedia, 2019, wiki.dbpedia.org/about.

[29] "Ontology Classes." Dbpedia, mappings.dbpedia.org/server/ontology/classes/.

[30] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," Inf. Process. Manag., vol. 51, no. 2, pp. 32–49, Mar. 2015.

[31] Rettinger, Achim, et al. "Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO." Semantic Web Journal, 2018, www.semantic-web-journal.net/.

[32] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., et al. Media meets semantic web - how the BBC uses dbpedia and linked data to make connections. In Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E., editors, The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings, volume 5554 of Lecture Notes in Computer Science, pages 723–737. Springer, 2009. DOI https: //doi.org/10.1007/978-3-642-02121-3_53.

[33] Sandhaus, E., Abstract: Semantic technology at The New York Times: Lessons learned and future directions. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part II, volume 6497 of Lecture Notes in Computer Science, page 355. Springer, 2010. DOI https://doi.org/10.1007/978-3-642-17749-1_27.

[34] Castillo, L., “Northern Triangle of Central America.” *Dame 1 Minuto De*, 7 Oct. 2016, dame1minutode.org/northern-triangle-of-central-america/.

[35] “Uprooted in Central America and Mexico.” *UNICEF*, Aug. 2018, www.unicef.org/child-alert/central-america-mexico-migration.

[36] Croitoru, A., Crooks, A.T., Radzikowski, J. and Stefanidis, A. (2013), GeoSocial Gauge: A System Prototype for Knowledge Discovery from Social Media, International Journal of Geographical Information Science, 27 (12): 2483-2508.

[37] “United States.” Wikipedia, Wikimedia Foundation, 17 Nov. 2019, en.wikipedia.org/wiki/United_States.

[38] Καλά, Ολά. “22. Citez Les Différents Drapeaux Des Pays Que Vous Avez Visité.” Twitter, Twitter, 14 Jan. 2018, twitter.com/akamevil/status/952677187727056896.

[39] “April 2019 Current Events: World News.” Infoplease, Infoplease, Apr. 2019, www.infoplease.com/april-2019-current-events-world-news.

[40] “May 2019 Current Events: World News.” Infoplease, Infoplease, May. 2019, www.infoplease.com/may-2019-current-events-world-news.

[41] “Mission.” Department of Homeland Security, 3 July 2019, www.dhs.gov/mission.

[42] Klassen KM, Borleis ES, Brennan L, Reid M, McCaffrey TA, Lim MS (2018), What People “Like”: Analysis of Social Media Strategies Used by Food Industry Brands, Lifestyle Brands, and Health Promotion Organizations on Facebook and Instagram. J Med Internet Res 2018;20(6):e10227.

[43] Aucott, P., Southall, H. (2019), International Journal of Humanities and Arts Computing, Volume 13 Issue 1-2, Page 69-94, ISSN 1753-8548 Available Online Oct 2019

[44] McDermott, S. (2017), Frequency and Proximity Clustering Analyses for Georeferencing Toponyms and Points-of-Interest Names from a Travel Journal, Spring 2017

[45] Crooks, A.T., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, Smith, D., A., Karagiorgou, S., Efentakis, A. and Lamprianidis, G. (2015), Crowdsourcing Urban Form and Function, International Journal of Geographical Information Science. 29(5): 720-741(pdf)

BIOGRAPHY

Molly Phillips graduated from Rockwall High School, Rockwall, Texas, in 2014. She received her Bachelor of Science from the United States Air Force Academy in 2018. She is employed as a 92-T0 Pilot Trainee by the United States Air Force and plans to receive her Master of Science in Geoinformatics and Geospatial Intelligence from George Mason University in December 2019.