PREPROCESSING TECHNIQUES TO SUPPORT EVENT DETECTION DATA FUSION ON SOCIAL MEDIA DATA

THESIS

Brandon T Davis, Captain, USAF

AFIT-ENG-MS-16-J-001

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-MS-16-J-001

PREPROCESSING TECHNIQUES TO SUPPORT EVENT DETECTION DATA FUSION ON SOCIAL MEDIA DATA

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Computer Engineering

Brandon T Davis, BS

Captain, USAF

June 2016

AFIT-ENG-MS-16-J-001

PREPROCESSING TECHNIQUES TO SUPPORT EVENT DETECTION DATA FUSION ON SOCIAL MEDIA DATA

THESIS

Brandon T Davis, BS
Captain, USAF

Committee Membership:

Dr. K. M. Hopkinson
Chair

Dr. B. J. Borghetti
Member

Dr. M. E. Oxley
Member

AFIT-ENG-MS-16-J-001

Abstract

This thesis focuses on collection and preprocessing of streaming social media feeds for metadata as well as the visual and textual information. Today, news media has been the main source of immediate news events, large and small. However, the information conveyed on these news sources is delayed due to the lack of proximity and general knowledge of the event. Such news have started relying on social media sources for initial knowledge of these events. Previous works focused on captured textual data from social media as a data source to detect events. This preprocessing framework postures to facilitate the data fusion of images and text for event detection. Results from the preprocessing techniques explained in this work show the textual and visual data collected are able to be proceeded into a workable format for further processing. Moreover, the textual and visual data collected are transformed into bag-of-words vectors for future data fusion and event detection.

Table of Contents

List of Figures

List of Tables

PREPROCESSING TECHNIQUES TO SUPPORT
EVENT DETECTION DATA FUSION ON SOCIAL MEDIA DATA

I. Introduction

For decades, television news media has been the main source of immediate news worthy events. However, the information conveyed on these news sources is sometimes delayed and vague on details due to lack of proximity to the event. In the last decade, social media users have started to bridge this delay of information by transmitting first hand information from the scene of the event. Many news outlets initially rely on information transmitted on social media sources like Twitter and Facebook to acquire this information early in the event's progress.

Social media has become a mainstay of mass communication in many peoples' lives. Over 2.5 quintillion (10^{18}) bytes of data is generated by society online each day [50]. A large portion of this data is generated on such social media services like Facebook, Twitter, Instagram, and Google+. Most social media services allow users to perform actions like upload a message, upload images, upload videos, tag other users in messagse, hashtags, make use of current location and many others. These actions allow users to upload information by a touch of their mobile devices or personal computers. This ease of access to social media information has spurred the use of data mining algorithms and techniques to utilize the information in these Big Data entities.

Big Data, generated by these large scale social media networks, has been used in a wide range of domains to model and predict real world phenomenon. However, the majority of research on mining large scale data streams from social media networks has

focused primarily on textual data streams that suffer from the following challenges: limited resource mining such as specific website forums [3], relying solely on extracting data from one aspect of a piece of data like geo-location [20], and poor assumptions about unbiased social media users [33].

Twitter offers a robust venue to detect and discover knowledge relating to any on-going threats. Twitter users generate more than 500 million tweets per day [1]. Any given tweet may contain useful information such as textual content, geo-location of the tweet, images, videos, etc. Other authors have researched the accuracy and truthfulness of social media networks ranging from predicting product demand [47] to online medical diagnoses [9].

When referring to threat detection on social media outlets, it is not as simple as "Search all messages that contain thread word of interest X.". There are several issues with social media from this aspect. Not all messages communicated through social media are related to news worthy event. An example is "Person X was shot at location Y", while another social media user may be more indirect such as: "Why would someone shoot such a great person, such a loss for the world." Another issue with social media users is an individual's ability to assess if the information being relayed is real or fake. The speed in which information disseminates through a social media network may influence how a social media user responds. An example is how fast the 2013 hoax "White House Bombing" spread across social media to cause a 140 point drop in the Dow Jones (DOW) in five minutes. This was approximately] 1% of the DOW at the time [8]. Another challenge specifically for image data mining is that not all users tag or properly address an image in a given message. An example of a vague message such as "How Cute!" gives little context outside of the two words.Lack of accurate context for a given image discredits the message or can cause the image

[1]https://about.twitter.com/company

to be filtered out from any meaningful analysis processes.

This thesis presents a preprocessing method for utilization on collected live social media data, specifically Twitter. This preprocessing method is developed to support an effort to combine textual and visual data for event detection over social media. This method is tested on three selected, widely covered news events. This method is validated by the creation of parsed data files along with acquired images. The data is also analyzed from the time of the events occurrence to the proceeding hour.

II. Literature Review

2.1 Social Media Networks

Social media networks have become globally-recognized domains to gather large amounts of information for various usage. A study conducted by Parker et al. involved the use of Facebook and Twitter to track and model health epidemics in a given area or worldwide [37]. Social media users leveraged their connections across the many social mediums in January 2010 to raise eight million dollars to aid those affected by the Haiti earthquake [16]. Abbasi et al. focused on the analysis of the specific social media source of web forums[1]. Abbasi wanted to correlate the usage of a certain lung cancer drug to side effects and moods through the collection and analysis of these forum posts. Data gathered through social media networks has also been utilized by Sakaki et al. to detect anomalous activities in social media usage (such as Twitter) during natural disasters such as earthquakes in a specific area [44]. Liang et al. also completed a study involving pinpointing an earthquake by utilizing geo-tagged tweets in two earthquake prone areas [30].

Political-driven research has prompted the use of social media to determine what politicians and political topics are trending amongst social media users [46]. Bukhari et al. monitored Facebook, Twitter and blogs for any visual information pertaining to Super Bowl XLVI and analyzed for future marketing studies [10]. There have even been studies to determine if there is a supply chain risk for a given market utilizing data from social media [15]. Work conducted by Li et al. proposed the use of a batch mode called HybridSeg to split tweets into meaningful tweets [29]. This splitting of the tweets helped identify the local and global context of a batch of tweets. X.Liu et al. proposed the use of a textual data cube to analyze and represent many overlapping features found in a given piece of textual information [32]. He accomplished this work

through the use of linguistic extraction algorithms along with several machine learning algorithms. Yanai et al. utilized the Twitter API to mine all food related images based on the textual data and classify the images based on color with a Histogram of oriented Gradients (HoG) and Support Vector Machines (SVM) [52]. This data was trained on a food dataset that classified each image into one of several food data classes. Liu et al. conducted a study involving the mining of Flickr images for use in cross-referencing major disasters such as Hurricane Katrina [31]. Liu concluded images uploaded by what he called "citizen journalists" provide a significant, first-hand account of these events of interest. Though many of these works utilize the numerous aspects of social media, this research focuses on gathering images and text for preprocessing for event detection. This analysis requirement makes the previous methods mentioned insufficient for the preprocessing and mining research.

2.2 Event and Anomaly Detection

The detection of anomalies (events) helps aid in discovering patterns that do not occur in normal data streams. The purpose of these anomalies is they provide a high amount of information about the potential peaking event from what is usually expected from a given event. Early needs for cyber event and anomaly detection detected fraud [36] along with detection of network intrusion [53]. Guille et al. proposed the use of Mention-Anomaly-Based Event Detection (MABED) to discover peaks of anomalous Twitter tweet patterns based on tweets alone [17]. A method proposed by Anantharam et al. aimed to detect and correlate tweets from Twitter to an anomalous event by evaluating URLs (Universal Resource Locator) provided in a given tweet for related information [6]. Becker et al. proposed several methods and algorithms to identify and classify occurring events based off of mined information from multiple social media sources like Facebook, Youtube, etc [7]. Event and anomaly detection

has also been considered for manufacturing processes and control systems. Allen et al. designed a new anomaly detection solution for event-based systems [4]. Allen's solution generates models of the system, detect faults, and utilizes the models to detect anomalies in the new event streams. Chae et al. created an interactive visual analytics system based on automated message evaluation to detect abnormal events [11]. Chae utilized Latent Dirichlet Allocation (LDA) to extract and rank major topics contained in the textual parts of his social media data. Reuter et al. designed a system to classify incoming social media streaming into already known events or, if needed, create a new event class [40]. Reuter showed this method worked effectively with large amounts of data and scaled accordingly. Li et al. utilized known historic events from Wikipedia to detect events in their collection of 4.3 million tweets. Li created a segment-based event detection system that employed the use of Term Frequency - Inverse Document Frequency (TF-IDF) to transform the tweets into a more meaningful form [28]. Watanabe et al. proposed the use of their local event detection system called Jasmine, to better geo-locate Twitter users from the context of their tweets. Watanabe's study also found only 0.7% of tweets are geo-tagged [48]. Inclusion of such a system could prove valuable for most research in this area as most social media users disable geo tracking on their account. While the research on anomaly and event detection is extensive, the different methodologies proposed are mostly centered around textual analysis. The addition of image analysis adds a new level of complexity which requires either the creation of classes of images or the utilization of already existing classes from elsewhere.

2.3 Textual Analysis Algorithms

In 2012, IBM claimed around 2.5 exabytes of data are created each day [21]. Because of this large amount of information, researchers are trying to create and op-

timize algorithms to efficiently process and collect this data for various usage. Zhao et al. developed an improved depression detection model based on sentiment analysis algorithm due to the performance and accuracy of their currently developed algorithm being somewhat inefficient [54]. Chen et al. proposed an index-based ranking estimation algorithm to improve query processing performance as well extending the several involved query processing algorithms to include and support the missing objects of interest [12]. Wu et al. created a hybrid algorithm called Text Segmentation algorithm based on Hierarchical Aggolomerative Clustering - Discrete Particle Swarm Optimization (TSHAC-DPSO) to improve linear text segmentation's accuracy and lower the computational complexity [49]. Jiang et al. aimed to develop a fuzzy self-constructing feature clustering algorithm that would categorize text with a higher accuracy, precision, and recall compared to other feature reduction methods by reducing the dimensionality of the features in the text classification [22]. Previous work completed by Bodnar et al. utilized the textual information of Twitter data to increase the veracity of event detection on social media networks [8]. This data was first mined based on if users had geo-location enabled. Then a keyword search was utilized on the collected data for approximate time frames and locations to acquire all tweets about a known threat event. This work utilizes several of the concepts introduced in work completed by Bodnar et al. by including the collection of the image data along with the textual data.

2.4 Image Analysis Algorithms

In the past several years, the volume of images uploaded to social media outlets has increased drastically [34]. Several other areas of research have utilized this influx of image data for analysis. Xu et al. presented the Multi-entry Coupled Object Similarity (MeCOS) algorithm to analyze the relationships been social media images

on their coupling attributes [51]. Xu measured the similarity between non-IID (Independent and Identically Distributed) data objects which he describes in several inter-related attributes. Work completed by Gupta et al. designed an algorithm to detect reported known fake Hurricane Sandy images that had propagated through Twitter. This work is based on the known existence of the fake images from media sources and other people. Gupta also observed in the case of the fake images, Twitter users tended to retweet trending topics whether they follow the person tweeting the information or not [18]. For image comparison and processing, Hare et al. created a tool called PicSlurper that processes images over a given time to detect trending images. Hare accomplishes this through a combination of Locality Sensitive Hashing (LSH), Hamming distance calculations, and Euclidean distance between each image feature [19].

However, up until the last few years, much of the research that analyzed these mass amounts of images utilized many of the same methods such as creating HoG (Histogram of Gradients) for classifiers and trained SVMs (State Vector Machines) [7] [42] [52] . More recent work approaches the use of Convolutional Neural Networks (CNNs) and Deep Learning to conduct mass image and video recognition [26] [43]. The utilization of these CNNs have proven to be more robust and accurate compared to previous image recognition methods. CNNs must be trained on vast amounts of data in order to avoid overfitting. CNNs are used in large scale challenges such as the Pascal Visual Object Classification Challenge (PVOCC) and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [43]. These two challenges provide venues to test and train the participants' image recognition networks. The PVOCC requires the recognition of 20 classes while the ILSVRC contains up to 1000 different classes. Though other methods were utilized in these competitions, CNNs have been consistent winners.

Though these competitions provide a novel venue to test a CNN, these competitions provide a select group of classes that represent a wide range of images for testing. This is not the case for social media network images. The quality and professionalism of the photos in the existing challenge classes are not likely to exist on social media networks.

2.5 Online Trust Metrics and Frameworks

One challenge when studying social media data is the trustworthiness of a given user and their provided information. Online information seekers must validate this social media data themselves. They accomplish this through personal judgment of the reliability and quality of the content that is provided. Several previous works designed frameworks around distinguishing these reliability issues. Kim et al. designed a trust prediction framework that distinguishes online experiences as being trustful or distrustful [25]. Kietzmann et al. created a honeycomb framework describing how seven functional blocks of social media are the foundations for building trust between users [24].

Another factor to consider when utilizing social media data for trustworthiness is the physical location of the source from a given event. Croitoru et al. argued that if an individual is closer to a given event then they are more likely to be accurate on reporting the event [13]. This make sense because if individuals are further way from the event, they likely did not get a first-hand view or directly interact with the event. This same idea can be applied to time metrics. Sakaki et al. designed earthquake detection model which required a threshold number of messages before the model considered the reports to be true [44]. This poses the issue with non-earthquake events. A relevant event may not have many messages due to a low number of participants. This poses the loss of such messages could be detrimental

for discovering the existence of events whether they are large or small. Ruan et al. conducted a trust measurement study on the effects of Twitter followers for the group FinancialTimes. By using the tweets from these followers, the study focused on giving each user a weight based on their connectivity to other users as well as weighting each tweet as positive, neutral or negative. This approach is biased to only the followers of a specific group and seems inadequate to directly relate tweets in this group to the fluctuation in the stock market [41].

One study completed by Gupta et al. discovered during Hurricane Sandy three malicious fake images were being disseminated and retweeted on Twitter to supposedly increase government response to the incident. The study discovered 86 percent of the fake images were retweets from bot accounts attempting to propagate the fake images [18]. Though this research does not directly determine the trustworthiness of a user or information, scoring trustworthiness offers another metric for evaluating collected messages.

2.6 Data Fusion

In the sections, several of the proposed text and image systems work efficiently. However, these works ignore other associated media content within the data like audio, video and image data. This is where the concept of data fusion comes into play. Most of the current research in regards to data fusion is focused on fusing data, like images and other graphics, in the medical field [2] [23] [39]. Most of these studies, however, focus more on combining information from other medical instruments into existing data plots and graphs in an attempt to reduce the number of sources needed to read critical health information.

Other studies, specifically geared towards available information on the Internet, pose more viable sources of data fusion techniques. A study conducted by Alqhtani et

al. focused on event detection through social media, specifically, Twitter. Alqhtani utilized Histogram of Gradient (HoG) descriptors for tweeted images and bag-of-words models with Term Frequency-Inverse Document Frequency (TF-IDF) methods for the text. Alqhtani aimed to fuse these two different data types to show higher accuracy for their event detection model when image data is fused with the text data. Alqhtani fails to explain how the data becomes fused as well as lack of explanation on many steps in the entire method of the research [5].

Work completed by Moulin et al. corresponds closely to the aim of this work [35]. Moulin utilized the bag-of-words model to represent both textual and image data. Moulin focused on determining the proper weight for a modality whether the modality is the text, audio, video, or image data. Lee et al. aimed to correlate available online images to geographical locations by fusing images and text [27]. Lee's experimental results showed the proposed method enhanced the task of location-based knowledge discovery given multiple images with associated text information to a known geographic location. Poslad et al. designed an ontology model that fused visual word vectors from images along with textual properties extracted from the images [38]. This research was an effort to create a more accurate Image Retrieval System (IMR) which returns more relevant results given a certain image. The overarching goal of this work is to fuse the textual and visual data collected from social media. This research designs a basic methodology step similar to Moulin et al. for proposed completion of the data fusion in future work.

III. Methodology

3.1 Methodology Overview

This section describes the methodology for preprocessing data from social media for event detection. This preprocessing feeds into work amenable to data fusion once complete. This work defines an event to be an occurrence during a particular span of time t and location l. Some events during this time span may be uninteresting, but some may be anomalous to the known steady state. This research gathers the anomalous information through the use of local nodes. These local nodes are social media users who communicate their perception and understanding of the event based on their approximation to the location l at time t. These nodes communicate whether the nodes are close to the epicenter or further away.

First, the interaction between social media users is described in Section 3.2. Section 3.3 describes the framework for data collection, processing, feature extraction and event detection. Section 3.4 explains the choice of Twitter as the platform for data collection for this work. In Section 3.5, the tools utilized to collect Twitter data as well as the parameters set to collect the target data are discussed. Section 3.6 describes how the data is preprocessed and parsed for further analysis. Section 3.7 describes the methods for extracting features from both the textual and image data collected from Twitter. Finally, Section 3.8 briefly discusses the overarching concept for this work which involves the data fusion of the textual and visual data. This concept will be taken into consideration for future work to determine whether or not event detection is improved through the inclusion of visual data fused with the textual data versus just the textual data alone.

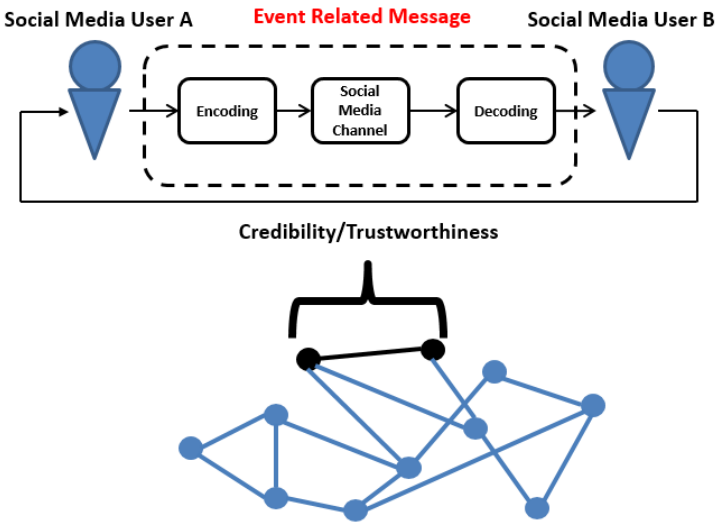


Figure 1. Shannon Model in Social Media. [45]

3.2 Social Media Communication

In order to better understand social media network communication, this research considers a social media model of Shannon's information theory in Figure 1 [45]. Figure 1 depicts a typical message sent from Social Media User A to Social Media User B. For the purposes of this research, encoding and decoding is the transformation of the intended message from typed text or other information into digitized information. This information is "decoded" or recompiled into the original format intended for viewing from the source so the recipient may view the information. Social Media User B gives feedback to Social Media User A by either accepting the transmitted information as real or a hoax based on their believed/perceived legitimacy of the information. User B has the option to pass the information to others in the social network or in real life. Figure 1 depicts a general case of a social media network. A given user (or node in this case) can have one or many intended recipients of a given message. This same message has the potential to be propagated through the same social network given the recipient (User B) believes the contents of the message. These

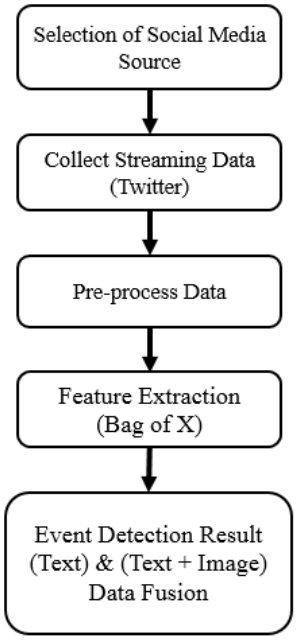


Figure 2. Proposed Framework.

messages potentially contain data such as text and images. This research focuses on capturing these messages for preprocessing and analysis. The overarching hypothesis is to determine whether or not the fusion of the text and image data further indicate the occurrence of an event versus the just textual analysis. The hypothesis for this work consists of designing a method for preprocessing captured social media data for event detection with fused data.

3.3 Methodology Framework

Figure 2 depicts the overview of the methodology for this research. This framework is based on a similar design for the Knowledge Discovery in Database (KDD) process [14]. The first step of the research involves choosing the source from which the data will be collected for analysis. In reference to this framework, social media networks are the main consideration for this research. For the second step, a combination of code and scripts must be established to actively collect streaming data for analysis. In

the third step, the data must be processed in order to remove extraneous information or to make the data compatible with the methods described in the fourth step. This is because some data comes in a JavaScript Object Notation (JSON) or other formats that have the potential to make further processing more difficult than necessary. This difficulty is caused by large file sizes accumulated by the streaming collection program. The fourth step aims to extract features from the processed data that was collected over a period of time. This textual and visual data will be extracted and vectorized using the bag-of-words and bag-of-visual words methods. The final step describes a brief data fusion process which will be explored in future work. The goal of this step is to calculate (detect) the chance of an event occurring during over a given time period in a set of collected data stream by combining textual and visual data. This method will be compared to the analysis of just the text alone.

3.4 Selection of Social Media Source

There are many popular social media networks like Facebook, Flickr, Foursquare, Google+, Instagram, LinkedIn, MySpace, Weibo, and Tumblr. These networks have millions of users around the world. However, in order to gather data from these networks, many connections with the other network members (to be "friends" in most cases) must already exist. Gathering data solely on existing connections would limit the scope of collectible data to only those members within an established "friend" group. Also, several of the larger social media networks limit or completely restrict the amount of data that can be collected from their network. Because of these limitations, social media networks like Facebook and MySpace become unrealistic choices.

One social media network, in particular, caters to application developers to make use of the data that comes across their social media platform. Twitter offers devel-

Figure 3. Example of a Tweet.

opers the tools and access to their data streams to build new applications[1]. Several APIs (Application Program Interface) already exist that allow developers to access Twitter streams to develop new software and applications. The developers only need to contact Twitter to acquire developer unique keys. These unique keys allow the developer to access Twitter streams and historical data. When a message is posted on Twitter it is called a tweet. A tweet can contain up to 140 characters. Figure 3 depicts an example tweet seen on Twitter[2]. The user can upload images, embed links to other websites, and even retweet another user's tweet onto their Twitter feed. Users can mention other users in a tweet which links that user to that tweet. Based on Twitter's encouragement for developers to work with their platform along with the various piece of metadata collectible from a single given tweet (discussed in Section 3.5), Twitter was chosen as the social media platform for data collection.

[1]https://dev.twitter.com
[2]http://mcphee.com/shop/horse-head-squirrel-feeder.html

Figure 4. Example Raw Captured Tweet.

3.5 Data Collection

The reason this research utilizes tweets is due to the availability of already developed Application Program Interfaces (APIs) (with the proper developer keys from Twitter) that interact with Twitter streams. Each message uploaded onto Twitter is called a tweet. Each tweet is stored in the form of a JSON (JavaScript Object Notation) objects on Twitter servers. These JSON objects contain many types of metadata about the each individual tweet such as date of creation, if the user has geolocation enabled , where the user is located, etc. Figure 4 depicts a tweet with notable pieces of metadata. Below lists some example fields of metadata/data that exist in a raw JSON object tweet.

1. The unique identification number for the tweet. Actual tweet number since first tweet.

2. The textual information entered by the user. This field can contain links and hashtags as well.

3. The unique URL (Uniform Resource Locater) for the tweet. This link allows access to the original posted tweet.

4. The screen name of the user who posted the tweet.

5. The geo_enabled field indicated if the user had geo-location enabled during the time of posting. If enabled, this field is populated with latitude and longitude numbers from where the tweet was made/uploaded.

6. Indicates what time zone the user is in when the tweet is made.

In order to collect these raw JSON objects (tweets), the Twitter4J API was utilized. The application utilized by this research was originally utilized by Bodnar et al. to specifically grab any tweets that contained geo-location information [8]. Since this application also enabled the collection of tweets by keywords in the textual communications, this research utilized the same application to pull streamed tweets based on keyword stream collection.

3.6 Preprocessing the Data

The Boston Bombing event tweets were provided in a CSV (Comma Separated Value) file with all of the tweet IDs from previous work [8]. Following the concept of the CSV file structure, future data stream collections are taken; then the tweet ID numbers are collected into a CSV file. These event CSV files are provided to an API called Twython. With Twython, a user can perform various tasks with Twitter data. In the case of this research, Twython provides the function to query older tweets if the tweet ID is provided. With the provided ID, a user can query various

fields associated with the tweet ID. For this research, six fields of data are identified as important metadata to collect. These fields are:

1. "id_str" This provides the tweet ID number.

2. "created_at" This field provides the day and time the user created the tweet. This time is different from a retweet.

3. "followers_count" This field provides the number of followers the user of the tweet currently has. This data is collected for future use in weighing the validity and/or trustworthiness of a tweet based on the follower count.

4. "user" Retrieves the user name of the tweet. This metadata is collect for future efforts to remove user tweets that originate from "spam accounts". An example of these spam accounts are gun control lobbyist accounts that constantly tweet anti-gun messages which can be collected from keyword searches.

5. "text" This field provides the actual content uploaded uniquely for a user's tweet. This field contains a variety of data from text, to links to websites, to hashtags.

6. "media_url" If present, this field provides the URL to a media image.

With the selected fields of interest, a Twython script is ran to collect this metadata into a text file. Due to special characters, like the hyphen, comma, and quotes being commonly used in text messages, the plus sign "+" was utilized to delineate between each field. Figure 5 shows a small snippet of a few tweets that were collected through this script.

Another part to this script also collects the image URLs into a separate text file. Python contains libraries that allow the parsing and downloading of images from URLs. These libraries were utilized to download the images into multiple folders.

```
665371424618483713 + Sat Nov 14 03:31:24 +0000 2015 + 853      +      hellamariz       +   RT @BobOngQuotes: What a sad news to wake
665371426195574788 + Sat Nov 14 03:31:24 +0000 2015 + 451      +      RheenalynP       +   RT @aldubdomination: Our prayers and thou
665371426371702785 + Sat Nov 14 03:31:24 +0000 2015 + 45       +     manashpratim83    +   RT @ArvindKejriwal: Paris attack is an at
665371424849182720 + Sat Nov 14 03:31:24 +0000 2015 + 83       +     felyolea55        +   RT @renalyn0915: #ShowtimeHarana https://
665371425109340160 + Sat Nov 14 03:31:24 +0000 2015 + 48       +     luigistar64       +   RT @DFlawZ: I designed a twitter header f
665371426342354944 + Sat Nov 14 03:31:24 +0000 2015 + 1912     +      nblhabuhssn      +   RT @tomzsz: people who are using this to
665371426447208452 + Sat Nov 14 03:31:24 +0000 2015 + 2174     +      naznasa93        +   RT @MovieGirl1: Praying for the safety of
```

Figure 5. Example snippet of queried tweet fields.

There was one caveat to downloading images in Python. Due to an internal limitation of the Python software, Python crashed initially when attempts were made to download all images from one event, which was around 50,000 image URLs at once. In order to get around this issue, the images were broken up into ten minute chunks according to their timestamps and downloaded into separate folders. Python is also utilized to get a binary return if each keyword is contained within each tweet. These binary results for each tweet are stored in a Microsoft Excel spreadsheet for further analysis.

3.7 Feature Extraction - The Science

Now that the text and images from the tweets have been collected, the data must be broken down further to understand the corpus as a whole. From background studies on data fusion along with text and image analysis, the concept of bag-of-X model for feature extraction appeared in several previous works [35] [5] [38].

For the textual information on each event, the bag-of-words representation is utilized to further parse the text into meaningful features. Since algorithms cannot understand the textual symbols directly, they must be transformed into numerical feature vectors. For this, the scikit-learn module in Python provides several built-in functions to perform these tasks. This step utilizes the following bag-of-word functions to transform the textual data:

1. Tokenize - This converts each string from the event into a list of tokens. This

Figure 6. Two Example Tweets Utilizing Proposed Text Feature Extraction (no N-grams).

tokenization treats white-spaces and punctuation as token separators. Token are stored as a contiguous set of characters/numbers

2. Stop word filtering - Identify common use words which add no value if identified such as "the", "and", "for" etc.

3. Counting - Counts the occurrences of each token (word in most cases).

4. Stem Filtering - This reduces each word to its stem, removing suffixes and prefixes.

5. N-grams - Bag-of-words cannot capture phrases and dependent words (like San Bernardino). Bag-of-words also does not catch misspellings or word derivations. N-grams provides to ability to build a collection of n-grams where the occurrences of pair of words are counted. This work utilized the bi-gram collection (n=2). An example use of bi-gram collection is the word "Golden Gate Bridge". This results in the tokens "Golden Gate" and "Gate Bridge".

Figure 6 depicts an example bag-of-words feature extraction on two differently for-

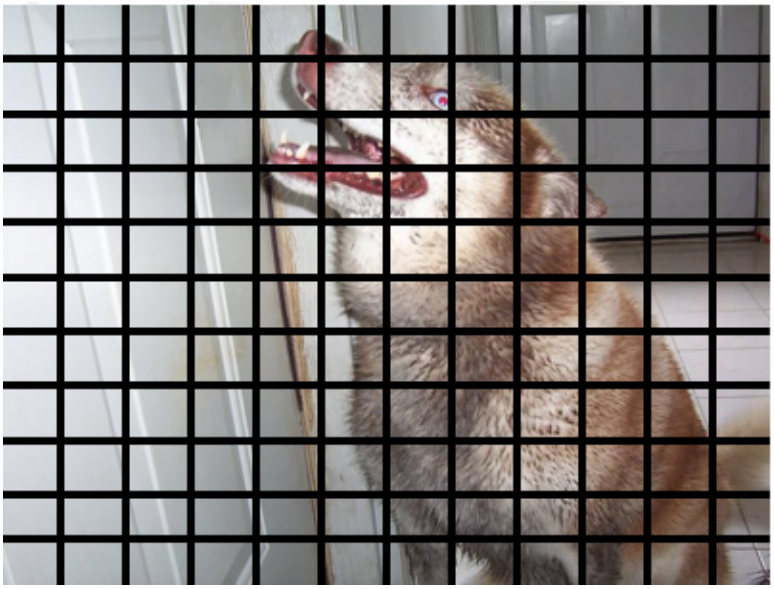


Figure 7. Image Grid Feature Extraction Example.

matted tweets without n-gram vectorizing. The first tweet depicts a readable sentence which is reduced to stemmed words and stop words removed. The second tweet depicts a retweet of a retweet containing two trending hashtags, a target person to include in the tweet, and the link to the retweeted tweet. After running the proposed bag-of-words model, the remaining tokens consist of several non-english words. Though this tweet's content is related to the event, scoring these types of tweets proves a challenge.

Also the associated text, this work utilizes the method of TF-IDF (Term Frequency - Inverse Term Frequency) to statistically determine the importance of a word in a given text corpus. In this research, all tweets from a given event are considered the corpus of words for the document. The scikit-learn package in Python offers a function to perform the TF-IDF transformation of the tweets.

In order to process the images, the images must also be vectorized. Due to the complex nature of image processing and analysis, this research utilizes a simple NumPy function to turn each image into a 1-D array. Figure 7 depicts and example grid

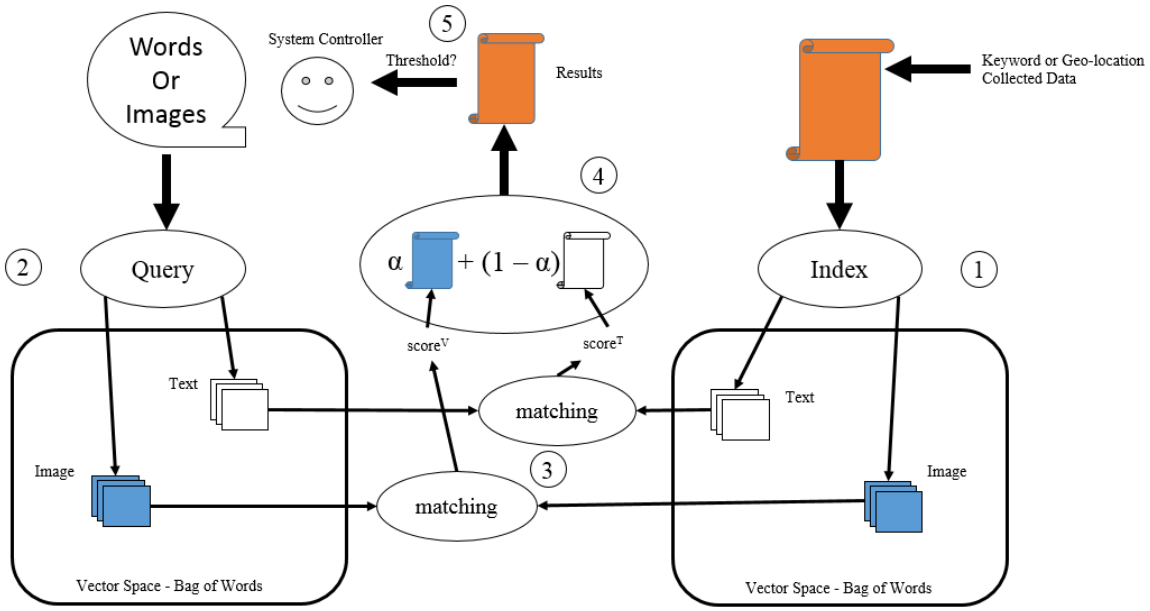


Figure 8. Proposed Data Fusion Model.

dissection of the image. Each grid section is transformed using the bag-of-visual words method. Each section is numerically interpreted and stored into a vector. This provides the data fusion step initial vectorized images.

3.8 Data Fusion

The overall goal of this work is to show the fusion of both textual and visual media in social media messages lead to a higher likelihood of event detection. Even though this research does not focus on data fusion, this concept is explored to show proper preprocessing is accomplished to facilitate meeting this goal. This research considers a similar data fusion model proposed by Moulin et al. [35]. Moulin proposes the transformation of both the image and textual data into a bag-of-X model where both the textual and visual data are broken down into vector space. Moulin proposed a linear combination of both the textual and visual vectors through scoring. Figure 8 illustrates the different steps the proposed data fusion model would utilize:

1. This step indexes the collected social media data, both textual and visual, with the bag-of-words modeling proposed in Section 3.7. The text and images are processed independently in this step.

2. The user provides a set query of text, and possibly in future design, images, to the system. This step also utilized the same bag-of-words modeling.

3. This step calculates a ranking score (V for visual, T for textual) between the user's query and every tweet provided from the collection.

4. This step combines the scored data in order to collect a group of tweets that are relevant to the user's query. Both the textual and the visual are weighed differently by a given value α.

5. This step determines, based on the results from the last step, if an event occurring based on a calculated threshold. The exact threshold parameter will be determined in future work.

IV. Results and Analysis

4.1 Event Selection

For this work, three security-related events were chosen. Two events occurred in the United States: the Boston Bombing in 2013 and the San Bernardino shooting in 2015 [8]. The other event considered was a terrorist attack that occurred in Paris in 2015. These three events were extensively covered by the news media. In the Boston Bombing, 6 people were killed and 280 people were injured due to a terrorist attack. In Paris, 130 people were killed in a terrorist attack in multiple nearby locations and 368 people were injured. In San Bernardino, 14 people were killed in a terrorist related event and 22 people were seriously injured.

These events were collected using the live tweet mining program (API) discussed in Section 3.6. Since Twitter requires specific parameters in order to collect streaming data, certain collection parameters were selected during mining. This work utilized keywords, along with specific times and days, to limit and mine specific event-related tweets. The Boston Bombing tweets were previously gathered for work completed by Bodnar et al. These tweets were provided in a CSV tweet ID file. [8]. These tweets were first gathered from geo-location enabled tweets during the event time period. Then the tweets were further filtered to keep tweets that contained the word "Boston". For the Paris terrorist attack, the keywords "terrorist", "bomb", "explosions", "Hollande", "#PrayforParis", "Paris", and "Bataclan" were used to mine the related tweets. For the San Bernardino shooting the keywords "San Bernardino", "shooting", "shooters", "SanBernardino" were used.

Due to the period of time that had passed since each of the events, a certain percentage of the tweet IDs could not be queried. This is due to a number of factors such as: the user deleted the tweet, Twitter removed the tweet, and erroneous return

Table 1. Initial Collected Event Data.

Event	Tweet Collected	Actual Tweets	Retained Percentage	Images	% Images	Original Images	% Original
Boston	10,948	8,068	73.69%	263	3.26%	139	52.85%
Paris	180,000	152,167	84.53%	57,705	37.92%	19,182	33.24%
San Bernardino	184,275	161,271	87.51%	30,545	18.94%	6,098	19.96%

due to API connection issues. Table 1 shows the number of original tweets IDs collected for each event. This table depicts that each event returned less tweets as event occurrences proceeded back in time, with the Boston Bombing being almost three years ago at 74% recall of the information. The most notable statistic here showed that the Boston Bombing contained 46% "original" images links. This number was not a good indication of actual original images. Users can retweet images from different platforms and programs other than Twitter These alternate platforms caused the collected image URLs to change to a unique name even though the image already exists in the tweet corpus under a different name. Early examinations of the unique image links also showed many users used enhancement programs or cropped trending event images just to re-upload them to Twitter. Certain applications also changed resolutions of the image as well which resulted in a new URL link for the same image.

The next three sections will discuss the results from the data collection for each of the previously mentioned events. For each event, approximately an hour of data was collected with given keywords to the Twitter API. The Boston Bombing differed in collection through collection by geo-location enabled tweets first, then a keyword search for the word "Boston". A data analysis of ten minute periods was chosen for a few reasons. First, several previous works generally analyzed a few hours of work at a time in similar intervals [8] [10] [44]. During collection, it was observed that during the studied events, conservatively, 15,000 to 20,000 tweets were collected every 10 minutes. This provided plenty of data for processing as downloading the images from the URLs in a tweet corpus this large takes several minutes. This work is intended to

Table 2. Boston Bombing Keyword Search Breakdown per 10 Minute Interval.

Boston Keywords	Minutes Past Event						
	10	20	30	40	50	60	Total
Explosion	13	131	425	500	504	459	2032
Boston	81	295	1060	1729	2312	2591	8068

assist in the detection of an event on social media in a certain period of time. Each event is analyzed in 10 minute intervals. This is based on the likelihood the event will be known to the main new sources within minutes, assuming a national level event.

Each event was analyzed for text and image density over a one hour interval in ten minute slices. The top three most tweeted images were included to show example Twitter images that had high retweetability. The image density was also broken down into unique URL tweets per ten minutes, removing duplicates of the same image link. Keyword counts were collected as a binary value for each tweet that contained the keyword in Python and stored in Microsoft Excel.

4.2 Boston Bombing

For the Boston Bombing event, the CSV containing the tweet IDs from Todd Bodnar's work were collected (recalled) for the information fields indicated in Section 3.6 with the Twitter API [8]. Tweets collected between the times 18:49:00 and 19:48:59 (Twitter stamped times) were preprocessed and organized in Microsoft Excel. Table 2 shows the frequency of the keyword "Boston" and "explosion". Though "explosion" was not an initial keyword search term, the word was included due to the noted frequency of use in the tweet corpus as another data point. This table shows the tweet count rate for each keyword over an one hour time period. Each ten minutes a keyword count was taken from each collected tweet to determine how often the words "Boston" and "explosion" occurred in the tweet corpus.

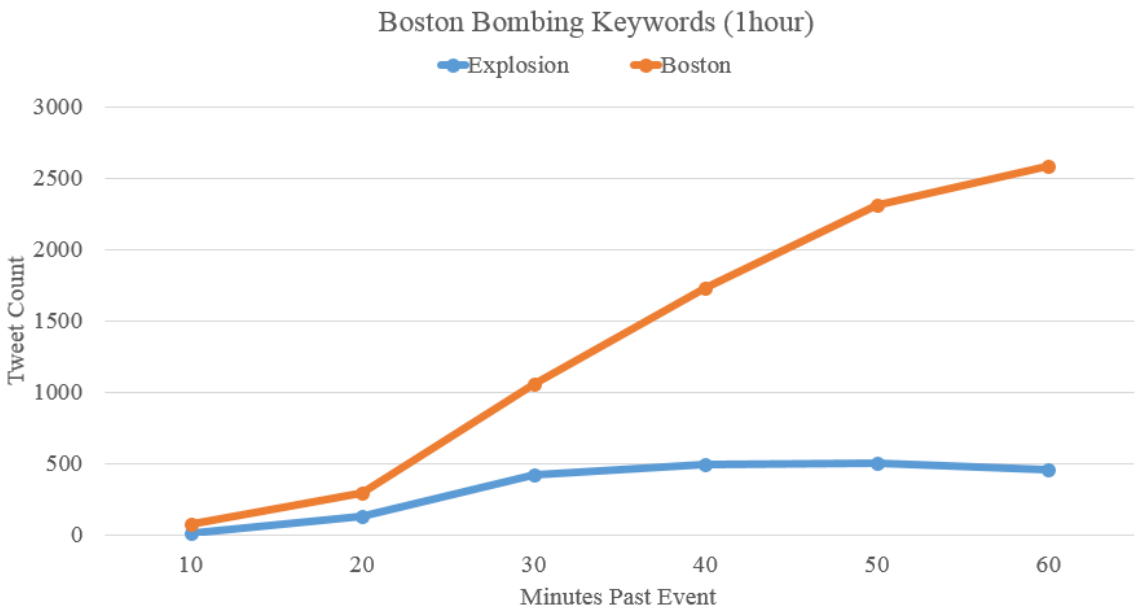


Figure 9. Boston Bombing Keyword Frequencies Over 1 Hour Period.

Another major difference between this event and the other two collected events was the number of tweet IDs collected during stream collection. The initial motivation for searching by keyword over geo-location was due to the lack of Twitter users who enable geo-location[18]. The previous work completed by Bodnar et al. included 3 other geo-located events; but these events had only a couple of hundred tweets each [8]. Luckily, this event was already being televised to some extent on major media outlets and had a high number of people either participating or viewing the marathon. The propagation of this event knowledge was fairly quick from the initial explosion over social media. Another caveat to this collection was Bodnar mentioned he performed "filtering by hand to remove irrelevant tweets". This was not a realistic action to consider for the other two events since any ten-minute-collection interval for both of those events was double the size of the Boston Marathon tweet collection. With this stated, the approximate 10,000 tweets for this event were enough to consider for study.

Table 3. Boston Bombing Unique Images and Image count per 10 Minute Interval.

Boston Images	Minutes Past Event						
	10	20	30	40	50	60	Total
Images	4	23	67	60	67	38	259
Unique Images	4	8	29	33	34	31	139

Figure 9 depicts the occurrence of the keywords "Boston" and "explosion" from the start of the Boston Bombing over a one hour period. In the first twenty minutes, only a few hundred people tweeted about the explosion in the area. Almost every ten minutes preceding this point the tweets relating to Boston doubled in reference to the attack that had occurred. By the thirty minute mark the word "explosion" was mentioned in the Boston geographical area almost 500 times every ten minutes.

Table 3 shows a breakdown of the image collection over the one hour event in ten minute intervals. Though not depicted here, most of the first twenty minutes of images from this event were of Boston Marathon runners participating or finishing the event. The first images of the attack did not appear until near the end of the twenty minute collection. Most of the images occurred between the thirty to fifty minute marks. Figure 10 shows the same increase in tweeted images occurring around after the twenty minute mark. The significant rise in images compared to unique image count conveyed many Twitter users were retweeting the same messages involving the same image.

[t]

Figure 11 depicts the top three images retweeted across Twitter. The numbers below the images are the counts of how many times the image was retweeted in the Boston area. The first two images originated from a gentleman who lived in a second story window right above the area of attack. The first image was right after the explosion which depicts people helping the injured. The second picture depicts the

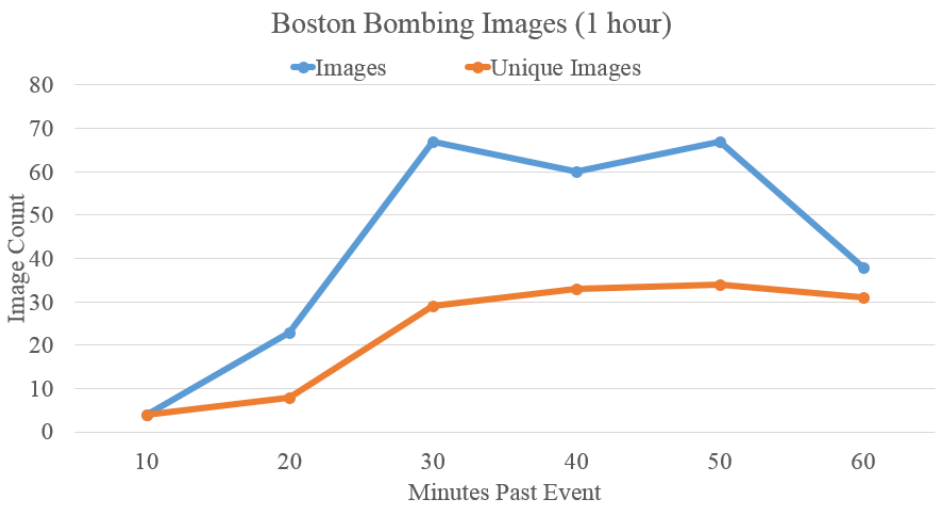
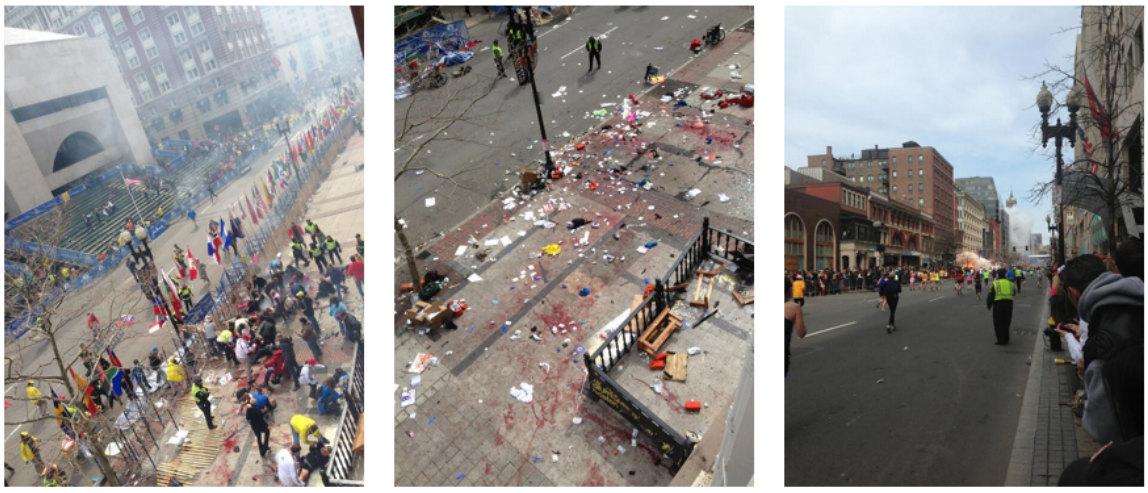


Figure 10. Boston Bombing Unique Images and Image Frequencies per 10 Minute Interval.

bloodied sidewalk minutes later. The third picture was taken by a marathon supporter on the sidewalk a few blocks down from the explosion. This picture shows the cloud of fire and smoke right as the explosive triggers. These three pictures accounted for 41.8% of the total pictures collected in one hour. This 41.8% does not include cropped or different resolution versions that existed under different URL media links. The depiction of blood and explosions appeared to convince social media users an event was likely occurring, which lead to the retweeting of the same images over the time span of one hour.

4.3 San Bernardino Attack

The San Bernardino attack tweets were collected with the keywords mentioned in Section 4.1 in the Twitter API parameters. From this collection of tweets, tweet IDs between the times of 21:00:00 and 21:59:55 are collected into a CSV. From there, the CSV is processed as explained in Section 3.6. The resulting file is loaded into Microsoft Excel with the delimiter symbol removed.



(65) (27) (18)

Figure 11. Top 3 Tweeted Images During the Boston Bombing.

Table 4. San Bernardino Attack Keyword Search Breakdown per 10 Minute Interval.

San Bernardino Keywords	Minutes Past Event						
	10	20	30	40	50	60	Total
San Bernardino	7173	6925	7143	6774	7263	7443	14640
Shooting	15501	15325	15195	16093	15681	15846	42721
Shooter	2723	3160	2793	2452	2354	2323	15805
SanBernardino	2489	2119	2185	2605	2596	2646	93641

Figure 12 and Table 4 depict the results of the keyword collection performed on streaming Twitter feeds. Upon collection, the keywords “San Bernardino” and “SanBernardino” were expected to result in the most tweets collected. However, ”shooting” resulted in almost double of both “SanBernardino” and “San Bernardino” combined. After investigating a small sample of the text and images, there appeared to be many spam, or noisy tweets involving gun control. An initial consideration after collection was the use of the location “San Bernardino” as a keyword may have biased the collection. However, the data collection showed the action word “shooting” surpassed the unique location by a good margin. This indicated the word “shooting’ was trending far more frequently over the one hour time span then the attack location.

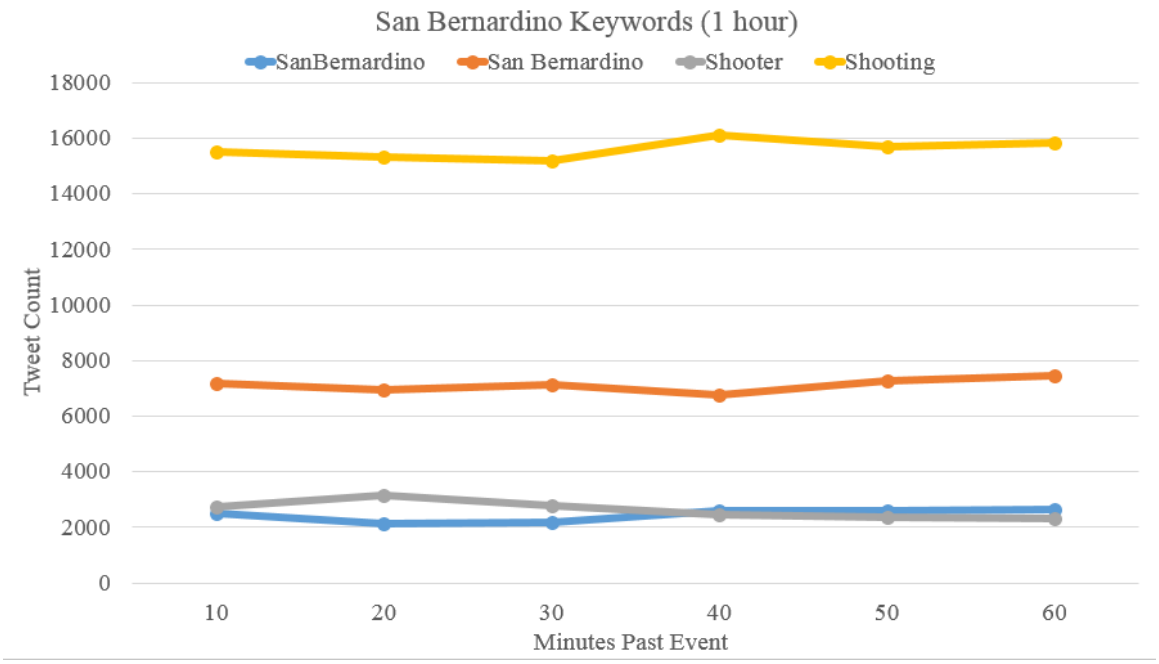


Figure 12. San Bernardino Keyword Frequencies over 1 hour period.

As for the interpretation of the tweets over the ten-minute intervals, the results were not as expected compared to the Boston Bombing keyword collection. Almost all four keywords maintain an even number of tweets that contain the keywords. The collection for this event data started only minutes after the confirmed start of the attack. The only noticeable spike in tweets was during the forty minute mark. This was likely due to the police department officially increasing the announced number of killed people from 4 to 12. For the first half of the hour the only known information was many were wounded and a few people were killed. Upon close inspection of the first few minutes of tweets, most people were tweeting or retweeting the shooting took place in San Bernardino.

Figure 13 depicts the top three images tweeted during the first hour of the San Bernardino shooting. Notice the two most tweeted images are of the geographic area of the United States. The first two images accounted for approximately a sixth of the total images collected for the San Bernardino attack. These tweets generally

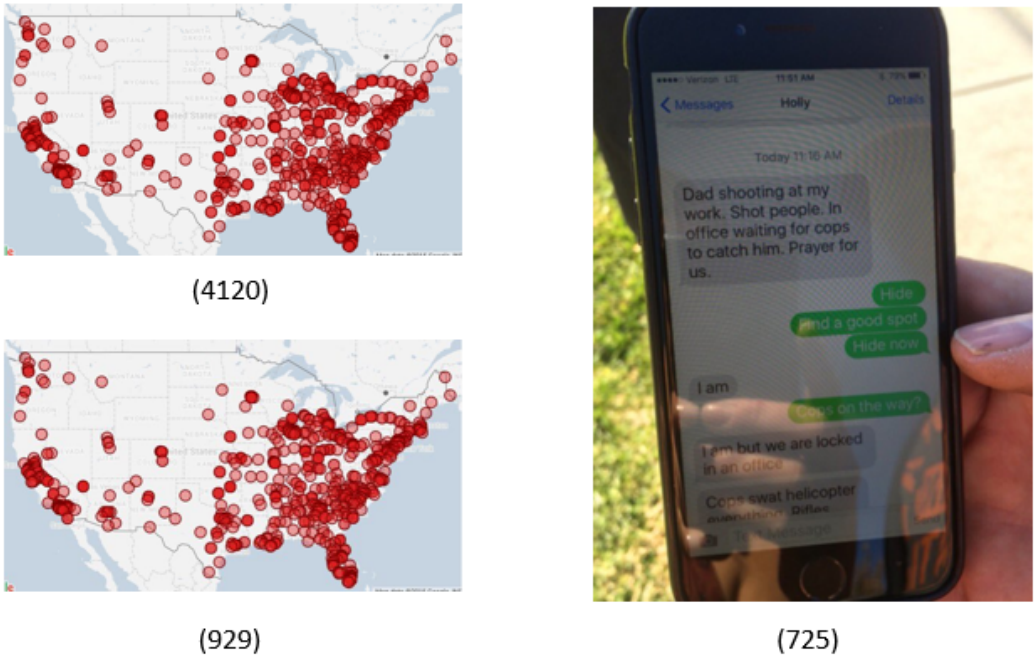


Figure 13. Top 3 Tweeted Images during the San Bernardino Attack.

involved a very brief mention of the, at the time, San Bernardino shooting that was occurring at that moment followed by advocating for more gun control. The red and pink dots indicated the location of a "mass shooting" in the last 20 years. The third picture depicts a father communicating with his daughter located in the social service building as the gunman continued their assault. The frequency of the gun control images and tweets implied the need to develop a user filter function. This function would be utilized to remove tweets from the collected tweets of Twitter users that are known bot or spam accounts.

For the images, Figure 14 and Table 5 depict the spread of images over the one hour collection of tweets. Similarly to the keyword search, the images and unique images held similar levels across the entire hour of collection. Unlike the Boston Bombing images, there were more than three times as many retweeted images as

Table 5. San Bernardino Attack Unique Images and Image count per 10 Minute Interval.

San Bernardino Images	Minutes Past Event						
	10	20	30	40	50	60	Total
Images	4543	4624	4752	4793	4635	4233	27580
Unique Images	948	900	978	944	915	967	5652

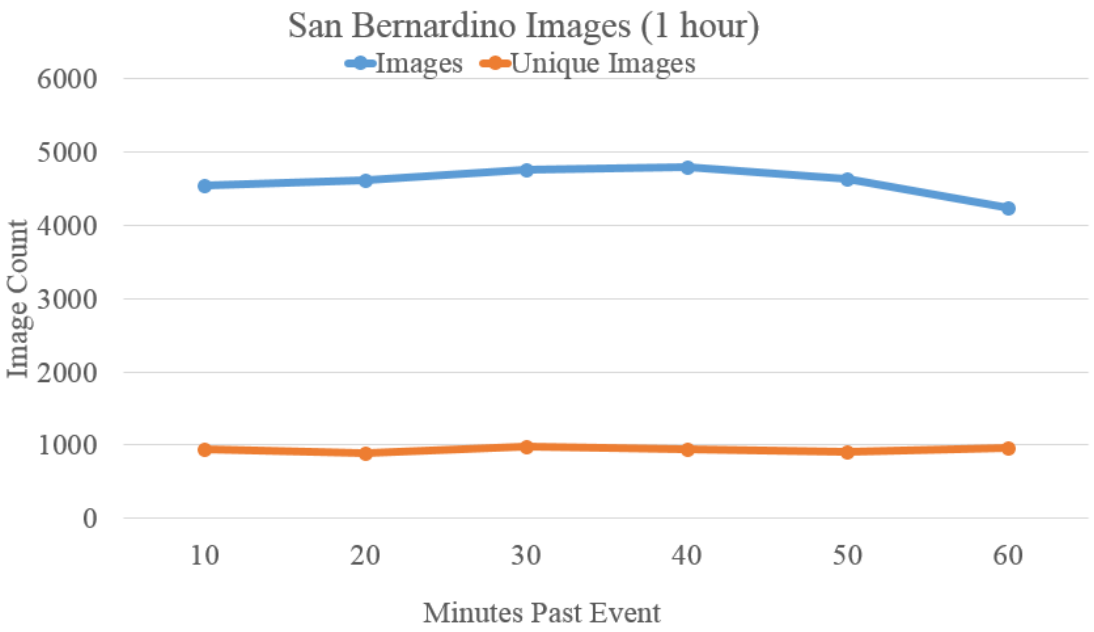


Figure 14. San Bernardino Attack Unique Images and Image Frequencies per 10 Minute Interval.

there are unique images throughout the hour. At no point in the Boston Bombing collection is this ratio this high. This is likely due to ease of retweeting a message over conducting a search to ensure the event was actually occurring.

4.4 Paris Terrorist Attack

One important caveat for the collection of the Paris Terrorist Attack was the data were collected almost four hours after the event occurred. Efforts to utilize other APIs to retrieve older tweets only resulted in the ability to grab a handful of random

tweets from the event day. These random collections included the other twenty-three hours of the event day which were irrelevant data. The "*" indicated on the figures and tables indicate that the data were taken well after the actual event start time. This data consisted of a one hour slice several hours later.

The collection for the Paris Terrorist Attack started roughly four hours after the event with the seven keywords indicated in Section 4.1. One hour of tweets were collected which resulted in approximately 180,000 tweets. Figure 15 and Table 6 show the collection of tweets based on keyword collection over the hour period. Based on the number of people killed, and the multiple locations in Paris attacked, both Paris-based keyword were the top two collected tweets. Since these tweets were collected almost four hours after the collection, the almost steady count of keyword tweets was an expected result. One noticeable data point about this graph was, although "terrorist" was the third highest collected keyword, the tweet collector only found around 1,200 tweets every ten minutes with this word. Since this was claimed as a terrorist attack fairly early into the event, the expected value of tweets to contain this word were much higher. Bataclan and Hollande were target venues where these attacked occurred. These two keywords resulted in significantly lower counts of tweets over the one hour collection in comparison to other keywords like "terrorist" and "prayforparis" This unwavering data collection was due the event attack ending several hours before collection. The event was globally televised by the time collection started. Most Twitter users focused their tweet content on grieving and praying over conveying the occurrence of the event by the time collection occurred.

The image analysis is depicted in Table 7 and Figure 16. Similarly to the keyword analysis, the image collection was constant throughout. However, in comparison to the San Bernardino image analysis, the Paris collection is almost double the number of images. In comparison of the unique image links, the Paris image collection contained

Table 6. Paris Terrorist Attack Keyword Search Breakdown per 10 Minute Interval.

Paris Keywords	Minutes Past Event						
	10	20	30	40	50	60	Total
Explosion	132	119	132	123	112	128	746
Bataclan	554	499	593	568	569	440	3223
Bomb	389	400	361	412	383	506	2451
Prayforparis	9447	8971	9182	8924	8800	9139	54463
Hollande	232	234	294	261	334	374	1729
Terrorist	1213	1296	1193	1239	1330	1258	7529
Paris	22246	21471	21517	21521	21464	21545	129764

Table 7. Paris Terrorist Attack Unique Images and Image count per 10 Minute Interval.

Paris Images	Minutes Past Event*						
	10	20	30	40	50	60	Total
Images	9381	9402	9232	8979	8895	9449	55338
Unique Images	2944	3103	3121	3092	3142	3284	18686

almost triple the number of unique image links. This proves interesting due to the corpus of the Paris tweets per hour are comparable to the San Bernardino tweet corpus: both contain approximately 180,000 tweets. Section 4.1 Table 1 shows the Paris event contained almost double the number of collect image URLs.

Figure 17 depicts the top three most tweeted images during the hour collection of the Paris Terrorist Attack. After browsing the collected images for the Paris Terrorist Attack, a majority of the images contained a significant portion of grieving or praying for the victims of the attack. The three images in Figure 17 depict various groups of people showing resilience to the attack that occurred. Most of the other collected images from the Paris attack included images of various buildings with the France flag colors and images of the Eiffel Tower.

This event collection presented a harder task of data filtering and manipulating as the event occurred so long after the collection. The filtering of sympathy/griev-

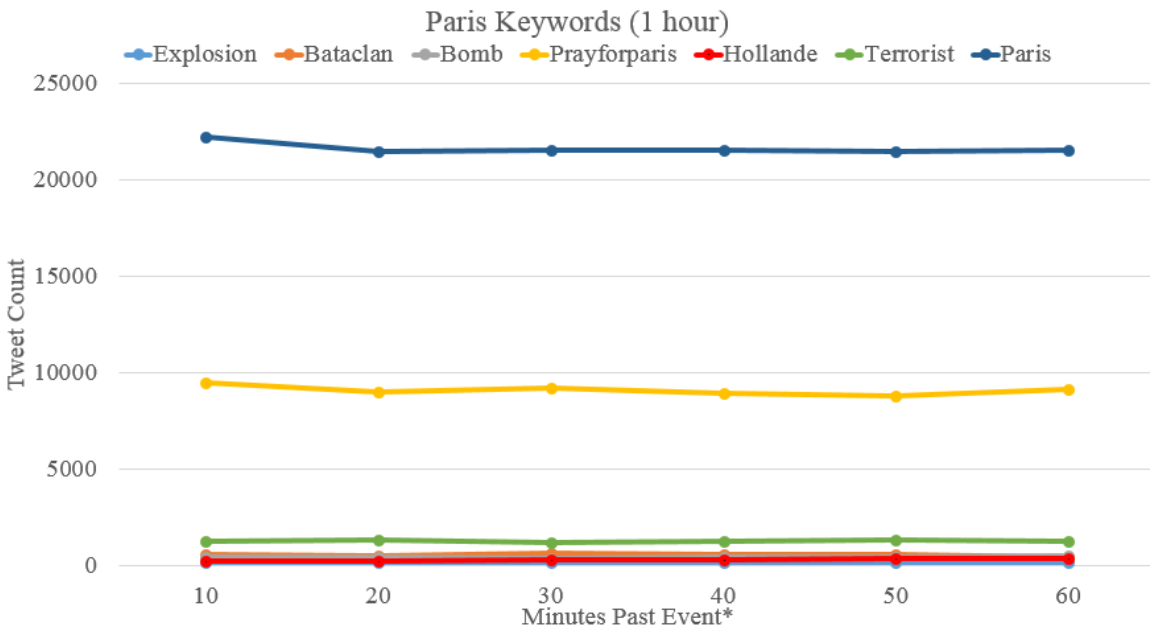


Figure 15. Paris Terrorist Attack Keyword Frequencies over 1 hour period.

ing/resiliency tweets costs extra time and effort. This prove problematic as the desired system for this work requires the ability to detect the event in a small amount of time. Another advocation for early collection was several users uploaded nice, high resolution images to show their support for what happened in Paris. This caused the image downloading for each ten minute collection to take longer than ten minutes since other Twitter users also retweeted these high resolution images.

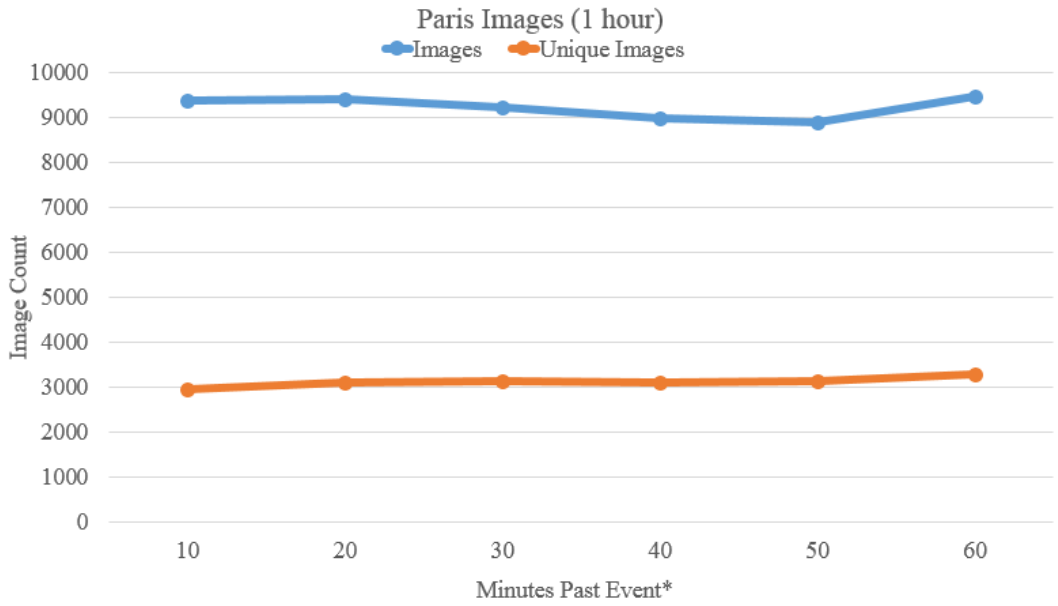


Figure 16. Paris Terrorist Attack Unique Images and Image Frequencies per 10 Minute Interval.

(706) (685) (622)

Figure 17. Top 3 Tweeted Images during the Paris Terrorist Attack.

V. Conclusion

5.1 Parameter-based Selection

This research employed a preprocessing methodology to collect, parse, preprocess, and extract features from the collected text and image data. Three historic events were selected for collection. The data for these events were collected through an API utilizing a keyword collection function. The JSON objects collected from these events were parsed for desired metadata and data such as text, images URLs and time of submission. The textual and image data from these events were vectorized using the bag-of-words model as well as the TF-IDF method for the text to support scoring of the collected tweet corpus. These features were extracted to facilitate future work for event detection through data fusion.

Analysis of the three events revealed the keyword-based collection of Twitter data gathered a large amount of data with both textual and visual information. The Boston Bombing showed a significant increase in tweet count over the one hour collection from the initial attack time. A similar trend was observed with the associated images from the collected tweets. 41% of the images for this event consisted of the site of explosion. Total image count versus unique image count was explored to enumerate how many tweets contained retweeted images. This preprocessing method offers a backbone for future metadata collection given a set of parameters such as keywords or geo-location.

The San Bernardino Terrorist Attack collection yielded approximately 180,000 tweets and approximately 30,000 images. Though collected within minutes of the known event start time, the trend of the keyword collection remained level across the collection hour. A similar trend was observed for the images. Approximately a ratio of 3:1 images were retweeted during the event collection hour. One-sixth of the images pertained to a particular gun control image though a majority of the tweets

still referenced the event.

The Paris Terrorist Attack yielded approximately 180,000 tweets. These tweets were collected four hours after the event occurrence. The keyword count for this event produced a steady number of tweets for all collected keywords. This collection yielded almost double the number of images collected from the San Bernardino event. A majority of the images collected from this event related to grieving or supporting the attack victims.

5.2 Implications and Recommendations

This preprocessing method, along with available API also open the possibilities to fuse or analyze other metadata. An example would be weighing the importance or trustworthiness of a tweet based on the user's follower count. The same concept could be considered if the user has a website linked to their Twitter account (like a business website link).

One flaw in the method of data collection was the reactive approach versus proactive. Future collections should run two versions of the collection API. One constantly collects general keywords like "shoot" and "bomb" utilizing the same methodology. There should be a script that stops the API and saves a ten minute file named after the time and day. A second similar script should be employed to collect tweets with geo-location enabled. This script could run for an hour before stopping, saving, and running again due to the lack of users who enable geo-location.

5.3 Future Work

One piece of metadata that would greatly improve this work is the geo-location of the user where the actual tweet is created. Though Twitter has a geo-location option, most Twitter users by default have this option disabled. Work completed by

Watanabe et al. could be incorporated into each tweet to better locate the users by unique keywords like “Paris” and “Bataclan Hall” [48].

Since the addition of images for tweet analysis is one of the main additions to tweet analysis beyond the text data, the work completed by Gupta et al. should be considered as a future edition of functionality for this system [18]. The propagation of false information and images on Twitter came up during the background research for this work. This work discovered several of the same images were retweeted from other users or the same image was uploaded as a new instance of the image.

Initially the parsing work focused on utilizing built in functions of parsing libraries to pull the desired fields from the raw JSON objects. However, depending on the program or method a Twitter user tweeted their message, initial attempts at parsing out certain information, such as the image URL, proved difficult. If a user uses a platform other than Twitter to tweet, this caused some fields to be repeated multiple times in a tweet. The use of special characters in textual information like “+” and multiple quotes in the same line would stop any further parsing for the text field. Though the use of the Twython API in Python successfully pulled the correct information, this incurred unnecessary re-querying from Twitter and cost a good bit of time. The use of the original JSON files is proposed for future preprocessing versus direct re-querying of the Twitter servers.

Another related issue to fake image identification is the recognition of bot, or fake accounts. During the parsing of the Paris tweet data, several images and accounts constantly tweeted scenic pictures of Paris or were attempting to sell a product. Identification of these accounts or some type of initial filter for these kinds of tweets could keep the quality of the data analysis much higher. In order to really judge whether or not a tweet is related to an event, each tweet would need to be read by a person to catch undertone like sarcastic posts and ironic posts. Many of the

reviewed works contained the caveat that their tweets were hand rated and tagged for relevance.

A future implementation to help detect or ignore these kinds of troublesome tweets would be to try to automate classification of simple tweets and prompt input from the system administrator. The administrator could train the system on the more questionable tweets so less person interaction would be required. Originally this work hinged on the utilization of geodesic object proposals on images that would be then be fed into a Convolutional Neural Network (CNN) utilized in work completed by Matt Dering. However, after feeding the Boston Bombing images through this system, it was determined the CNN needed more classes of images like blood, bombs, guns, etc. The system proved fairly accurate at determining if a person or dog was present in an image but these are basic objects in most tweeted images so there was little to infer from these types of object detections. Setting up the Twitter4J API to constantly collect both keyword based tweets and geo-location enabled tweets should be the future goal for this work. This work demonstrated the feasibility of the setup but not the full setup of this detection system. Creating a batch file or another module to work with the Twitter4J API would be the first step. Having the collection terminate every hour and saving the file name to be the time and day would make the event detection on the data more true to the research as specific location names cannot realistically be used to collect the data.

Bibliography

1. A. Abbasi, D. Adjeroh, M. Dredze, M. J. Paul, F. M. Zahedi, H. Zhao, N. Walia, H. Jain, P. Sanvanson, R. Shaker, M. D. Huesch, R. Beal, Wanhong Zheng, M. Abate, and A. Ross. Social media analytics for smart health. *Intelligent Systems, IEEE*, 29(2):60–80, 2014.

2. T. Adali, Y. Levin-Schwartz, and V.D. Calhoun. Multimodal data fusion using source separation: Application to medical imaging. *Proceedings of the IEEE*, 103(9):1494–1506, Sept 2015.

3. A. Akay, A. Dragomir, and B. E Erlandsson. Network-based modeling and intelligent data mining of social media for improving care. *Biomedical and Health Informatics, IEEE Journal of*, 19(1):210–218, 2015.

4. L.V. Allen and D.M. Tilbury. Anomaly detection using model generation for event-based systems without a preexisting formal model. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 42(3):654–668, May 2012.

5. Samar M. Alqhtani, Suhuai Luo, and Brian Regan. Fusing text and image for event detection in twitter. *CoRR*, abs/1503.03920, 2015.

6. Pramod Anantharam, Krishnaprasad Thirunarayan, and Amit Sheth. Topical anomaly detection from twitter stream. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 11–14, New York, NY, USA, 2012. ACM.

7. Hila Becker, Mor Naaman, and Luis Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 291–300, New York, NY, USA, 2010. ACM.

8. T. Bodnar, C. Tucker, K. Hopkinson, and S.G. Bilen. Increasing the veracity of event detection on social media networks through user trust modeling. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 636–643, Oct 2014.

9. Todd Bodnar, Victoria C. Barclay, Nilam Ram, Conrad S. Tucker, and Marcel Salathé. On the ground validation of online diagnosis with twitter and medical records. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 651–656, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

10. I. Bukhari, C. Wojtalewicz, M. Vorvoreanu, and J.E. Dietz. Social media use for large event management: The application of social media analytic tools for the super bowl xlvi. In *Homeland Security (HST), 2012 IEEE Conference on Technologies for*, pages 24–29, Nov 2012.

11. Junghoon Chae, D. Thom, H. Bosch, Yun Jang, R. Maciejewski, D.S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 143–152, Oct 2012.

12. Lei Chen, Xin Lin, Haibo Hu, C.S. Jensen, and Jianliang Xu. Answering why-not questions on spatial keyword top-k queries. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 279–290, April 2015.

13. Arie Croitoru, Andrew Crooks, Jacek Radzikowski, and Anthony Stefanidis. Geosocial gauge: a system prototype for knowledge discovery from social media. *International Journal of Geographical Information Science*, 27(12):2483–2508, 2013.

14. Piatetsky-Shapiro Fayyad. From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining*, pages 1–34, 1996.

15. X.J. Fu, R.S.M. Goh, J.C. Tong, L. Ponnambalam, X.F. Yin, Z.X. Wang, H.Y. Xu, and S.F. Lu. Social media for supply chain risk management. In *Industrial Engineering and Engineering Management (IEEM), 2013 IEEE International Conference on*, pages 206–210, Dec 2013.

16. Huiji Gao, G. Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *Intelligent Systems, IEEE*, 26(3):10–14, May 2011.

17. A. Guille and C. Favre. Mention-anomaly-based event detection and tracking in twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 375–382, Aug 2014.

18. Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 729–736, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

19. Jonathon S. Hare, Sina Samangooei, David P. Dupplaw, and Paul H. Lewis. Twitter's visual pulse. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ICMR '13, pages 297–298, New York, NY, USA, 2013. ACM.

20. Yan Huang, Zhi Liu, and Phuc Nguyen. Location-based event search in social texts. In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 668–672, 2015.

21. IBM. "ibm what is big data? - bringing big data to the enterprise". 2012.

22. Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee. A fuzzy self-constructing feature clustering algorithm for text classification. *Knowledge and Data Engineering, IEEE Transactions on*, 23(3):335–349, March 2011.

23. M L Kessler. Image registration and data fusion in radiation therapy. *The British Journal of Radiology*, 79(special_issue_1):S99–S108, 2006. PMID: 16980689.

24. Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 54(3):241 – 251, 2011.

25. Young Ae Kim and Muhammad A. Ahmad. Trust, distrust and lack of confidence of users in online social media-sharing communities. *Knowledge-Based Systems*, 37:438 – 450, 2013.

26. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

27. Chung-Hong Lee and Shih-Hao Wang. An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery. *Expert Systems with Applications*, 39(10):8954 – 8967, 2012.

28. Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 155–164, New York, NY, USA, 2012. ACM.

29. Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He. Tweet segmentation and its application to named entity recognition. *Knowledge and Data Engineering, IEEE Transactions on*, 27(2):558–570, February 2015.

30. Yuan Liang, James Caverlee, and John Mander. Text vs. images: On the viability of social media to assess earthquake damage. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 1003–1006, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

31. Sophia B Liu, Leysia Palen, Jeannette Sutton, Amanda L Hughes, and Sarah Vieweg. In search of the bigger picture: The emergent role of on-line photo

sharing in times of disaster. In *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*, 2008.

32. Xiong Liu, Kaizhi Tang, J. Hancock, Jiawei Han, M. Song, R. Xu, V. Manikonda, and B. Pokorny. Socialcube: A text cube framework for analyzing social media data. In *Social Informatics (SocialInformatics), 2012 International Conference on*, pages 252–259, Dec 2012.

33. D. Mahata and N. Agarwal. What does everybody know? identifying event-specific sources from social media. In *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, pages 63–68, Nov 2012.

34. M. Meeker. Internet trends 2014, code conference. 2014.

35. Christophe Moulin, Christine Largeron, Christophe Ducottet, Mathias Gry, and Ccile Barat. Fisher linear discriminant analysis for text-image combination in multimedia information retrieval. *Pattern Recognition*, 47(1):260 – 269, 2014.

36. EWT Ngai, Yong Hu, YH Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011.

37. J. Parker, Yifang Wei, A. Yates, O. Frieder, and N. Goharian. A framework for detecting public health trends with twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 556–563, 2013.

38. Stefan Poslad and Kraisak Kesorn. A multi-modal incompleteness ontology model (mmio) to enhance information fusion for image retrieval. *Information Fusion*, 20:225 – 241, 2014.

39. Hongliang Ren, D. Rank, M. Merdes, J. Stallkamp, and P. Kazanzides. Multi-sensor data fusion in an integrated tracking system for endoscopic surgery. *Information Technology in Biomedicine, IEEE Transactions on*, 16(1):106–111, Jan 2012.

40. Timo Reuter and Philipp Cimiano. Event-based classification of social media streams. In *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 22:1–22:8, New York, NY, USA, 2012. ACM.

41. Yefeng Ruan, L. Alfantoukh, and A. Durresi. Exploring stock market using twitter trust network. In *Advanced Information Networking and Applications (AINA), 2015 IEEE 29th International Conference on*, pages 428–433, March 2015.

42. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564–2571, Nov 2011.

43. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, AlexanderC. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2015.

44. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. pages 851–860, 2010.

45. C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

46. A. Tsakalidis, S. Papadopoulos, A. I. Cristea, and Y. Kompatsiaris. Predicting elections for multiple countries using twitter and polls. *Intelligent Systems, IEEE*, 30(2):10–17, 2015.

47. Suppawong Tuarob and Conrad S Tucker. Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data. In *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 2013.

48. Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2541–2544, New York, NY, USA, 2011. ACM.

49. Ji-Wei Wu, Judy C. R. Tseng, and Wen-Nung Tsai. A hybrid linear text segmentation algorithm using hierarchical agglomerative clustering and discrete particle swarm optimization. *Integr. Comput.-Aided Eng.*, 21(1):35–46, January 2014.

50. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on*, 26(1):97–107, Jan 2014.

51. Zhe Xu, Ya Zhang, and Longbing Cao. Social image analysis from a non-iid perspective. *Multimedia, IEEE Transactions on*, 16(7):1986–1998, Nov 2014.

52. K. Yanai, T. Kaneko, and Y. Kawano. Real-time photo mining from the twitter stream: Event photo discovery and food photo detection. In *Multimedia (ISM), 2014 IEEE International Symposium on*, pages 295–302, 2014.

53. Hao Zhang, Maoyuan Sun, Danfeng Daphne Yao, and Chris North. Visualizing traffic causality for analyzing network anomalies. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics*, pages 37–42. ACM, 2015.

54. Yiming Zhao, Kai Niu, Zhiqiang He, Jiaru Lin, and Xinyu Wang. Text sentiment analysis algorithm optimization and platform development in social network. In *Computational Intelligence and Design (ISCID), 2013 Sixth International Symposium on*, volume 1, pages 410–413, Oct 2013.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
16-06-2016	Master's Thesis	August 2014 — June 2016

4. TITLE AND SUBTITLE

Preprocessing Techniques to Support Event Detection Data Fusion on Social Media Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Davis, Brandon T. Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENG-MS-16-J-001

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Intentionally left blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This thesis focuses on collection and preprocessing of streaming social media feeds for metadata as well as the visual and textual information. Today, news media has been the main source of immediate news events, large and small. However, the information conveyed on these news sources is delayed due to the lack of proximity and general knowledge of the event. Such news have started relying on social media sources for initial knowledge of these events. Previous works focused on captured textual data from social media as a data source to detect events. This preprocessing framework postures to facilitate the data fusion of images and text for event detection. Results from the preprocessing techniques explained in this work show the textual and visual data collected are able to be proceeded into a workable format for further processing. Moreover, the textual and visual data collected are transformed into bag-of-words vectors for future data fusion and event detection.

15. SUBJECT TERMS

Event Detection, Social Media, Data Fusion

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. K. M. Hopkinson (ENG)
U	U	U	U	57	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x4579 kenneth.hopkinson@afit.edu