2020

An Approach to Twitter Event Detection Using the Newsworthiness Metric

Jonathan Adkins

An Approach to Twitter Event Detection Using the Newsworthiness Metric

by

Jonathan Christian Adkins

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in
Information Assurance

College of Computing and Engineering
Nova Southeastern University

2020

An Abstract of a Dissertation Submitted to Nova Southeastern University
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

An Approach to Twitter Event Detection Using the Newsworthiness Metric

by
Jonathan C. Adkins
August 2020

Currently there exists no clear-cut, commonly understood definition of what an event is in the context of Social Network Analysis (SNA). Events are commonly identified and measured with regards to repeated occurrences of related terms associated with a topic that gradually increase in frequency and then eventually decline. This ebb and flow of keyword frequencies occurs within a continuous stream of user messages in a social media platform such as Twitter. One disadvantage to this approach is that it tends to marginalize the human perspective of communication and event detection in favor of lexical trends. The goal of this study was to develop an alternate event detection technique and apply it to social media discussion venues such as Twitter. What was novel about our approach was that it incorporated the integration of two SNA metrics into a single metric called Newsworthiness. To test our method, we collected two 14-day datasets based on two different trending topics from current events. The first dataset was based on the keyword search "Tulsa+Rally." The second dataset was based on the keywords "Atlanta+Protests." Both datasets were graphed for their corresponding Newsworthiness and keyword frequency trajectories. The results of the two "Tulsa+Rally" graphs demonstrated that the Newsworthiness approach identified events that were undetectable to the keyword frequency approach. Results for the two "Atlanta+Protests" graphs were congruent in that they each identified the same three events. Our contribution to the body of research was threefold. First, we created a single metric called Newsworthiness by integrating Shannon Entropy and Diffusion Centrality. Second, we demonstrated the evaluative benefits of using quartiles to analyze Newsworthiness distributions for outliers and event peaks. Lastly, we demonstrated how to evaluate user activity by analyzing the Shannon Entropy and Diffusion Centrality of a discussion stream over the most efficient time period (p) metric. It has been empirically shown that the proposed metric, along with quartile-based analysis, provides a way to quantitatively identify events on social, political, and cybersecurity Twitter topics, and the performance is superior that of Keyword search. It was evident that the proposed metric has the potential to be applied to other topics and social platforms for event detection.

Acknowledgments

I would like to thank my mother Ann and my sister Alysa for providing me the financial and emotional support I needed to get me through this demanding and challenging endeavor. Aside from my five years of military service, this has been the most difficult, important, and ultimately satisfying accomplishment in my life.

I would like to thank Dr. Li for his guidance and patience with me as I journeyed through the many stages of the research process. I would also like to thank my other two committee members, Dr. Ling Wang and Dr. Ajoy Kumar, for serving on my committee and offering their wisdom toward my dissertation so that it can be the highest quality product that it can be.

I would like to retroactively acknowledge and thank my Army basic training drill sergeants who taught me to be mentally focused, to pay attention to details, and to stop overthinking tasks so that the mission can be accomplished.

Above all else, I would like to dedicate this work to my father who passed away two years ago of Cancer. He will not get to share in the joy of my accomplishment, but he would want nothing more than for me to keep my shoulder to the “wheel of pain” as he called it and complete the mission I started. Dad, I hope that you are proud of me.

Table of Contents

List of Tables

Tables

List of Figures

Figures

Chapter 1

Introduction

Background

Social Network Analysis

An Online Social Network (OSN) is a web-based public platform that allows people to engage in remote social interactions. A discussion stream is a flow of data or content consisting of semi-structured text messages, links, and multimedia (images and videos) that is contributed by users though activities conducted in an OSN (Alkhouli, et al., 2014). Social Network Analysis (SNA) is a discipline which incorporates a set of theories, techniques and tools for studying human behavior and how entities interact with each other. SNA is often used for research in areas such as organizational studies, economics, sociology, psychology, and politics (Serrat, 2017). SNA is also frequently used in research of OSNs such as Twitter to study the dynamics of user influence in social networks (Neves-Silva, et al., 2016; Serrat, 2017). A mathematical graph of a Twitter OSN is depicted as a set of nodes which represents users and a set of connecting edges which represent interactions between the users. Formally, the graph of a Twitter OSN is represented as G = (V,E), where G is an unweighted, undirected graph, where $v_i \in V$ is an individual user in an OSN and $e_i \in E$ represents interactions between users in a social network. In the case of a Twitter OSN, an interaction refers to the reposting of a message that was posted by another user. The graph is represented as subset of connected individual Twitter users. The connected edges between the users represent messaging relationships between the users.

In the context of our research a discussion stream is composed of three objects which are a set of tweets, a set of users, and a period of time. A dataset of tweets is a subset of a discussion stream, represented by the variable *T*. Dataset *T* is composed of a set of individual tweets, defined as $\{t_1, t_2, \dots, t_i, \dots t_n\}$. The second object in discussion stream *T* is a set of Twitter users *U*, defined as $\{u_1, u_2, \dots, u_i, \dots u_k\}$. The third object in discussion stream *T* is a time period *p*, defined as $\{p_1, p_2, \dots, p_i, \dots p_m\}$. In SNA there is no general consensus among researchers as to the official definition of an *event*. In the abstract we define an *event* in the context of SNA as a function of user messaging activity and user diversity that occurs over a period of time *p*. A more detailed definition of *event* will be provided below. Message spreading activity and diversity of participating users are metrics that can be used together to identify events in a Twitter discussion stream. These two metrics are evaluated by using Diffusion Centrality and Information Entropy which are discussed in the following sections.

Entropy

In information theory, *entropy* is officially defined as the measure of the level of disorganization or uncertainty in a system (Laniado, D. & P. Mika, 2010). The mathematical definition of a tweet's entropy is defined as $H(W) = \sum_{i=1}^{n} -\log_2 P(w_i)$, where *P* is a probabilistic model, $W = \{w_1, w_2, \dots, w_n\}$ is a corpus of tweets, and *H* is the entropy of the corpus. A corpus is a term which represents the body or sum of set of textual content that will be analyzed using text mining techniques. A corpus can be the contents of a single document. It can be multiple documents that are aggregated together. It can also represent a collection of several user messages to a social media platform such as Twitter. This formula for entropy will provide a quantitative assessment of the amount

of information within a corpus of tweets (Neubig & Duh, 2013). Entropy is used in several SNA studies in the literature as a metric to evaluate different aspects of social media communications discussion streams. In particular, it is frequently used to evaluate the amount of *surprise* or *diversity* in social media messages that are exchanged by users (Ghosh, et al., 2011; Vajapeyam, 2014). According to many studies, a higher entropy in a discussion stream sample suggests a larger *diversity* of participants contributing to a discussion. In the following sections, events are discussed in terms of how they spread through an OSN. This is done using a combination of entropy and Diffusion Centrality, which will be discussed in the following section.

Diffusion Centrality

Diffusion Centrality ($DC(t_i)$) is a SNA metric which evaluates the level of activity in an OSN with regards to tweeting and retweeting in a discussion stream (Kang, et al., 2016). The metric is a score that is assigned to individual users who are part of a discussion stream. The $DC(t_i)$ score is evaluated based on the connectivity of the users. The more connected a user is to other users who have high connectivity, the higher the $DC(t_i)$ score will be. The score is calculated using the formula $DC(t_i) = \sum_{t=1}^{T} Pr^t$, where t_i is an individual tweet from a discussion stream and Pr^t is the probability that one user can reach an adjacent (neighboring) user in *t* iterations where *T* is the total number of iterations in the time period covered in the discussion stream and *t* is a single iteration (An & Liu, 2016). In the context of SNA there are two ways to view how the $DC(t_i)$ metric evaluates connected users. The first is that $DC(t_i)$ measures how important an individual user is in spreading a message in a discussion stream. The second way to understand how $DC(t_i)$ scores users is that it evaluates how many times a

particular user's message will be seen by the other users in a common discussion stream. If a user has a very high $DC(t_i)$, that user's message will likely be retweeted by a much larger number of other users. $DC(t_i)$ and entropy are used with a time-ordered set of tweets to detect events which are discussed in the following section.

Events and Event Detection

Discussion streams play a role in the dynamics of SA as users interact among themselves through discussions, retweets, and other methods of social media communication (Pinto, et al., 2019). There currently does not exist in the literature a commonly understood definition of what an *event* is (Cui, et al., 2016). To that end we define an *event* as a set of tweets on a related topic within a defined time-period that surpass a threshold defined by statistical measures of *diffusion* and *entropy*. *Event detection* is the identification of *events* that are present in a time-ordered stream of Twitter messages (Cui, et al., 2016; Thapen, et al., 2016).

Problem Statement

Currently in the literature, there is no clear-cut definition of *event detection* in OSN analysis (Zhou & Chen, 2014). One common research thread that is found is the repeated occurrence of related terms associated with a particular topic that gradually increases in frequency and then eventually declines within a continuous stream of user messages (Cui, et al., 2016). Several examples from the literature focus on identifying events as time-ordered clusters of related keywords. (Wang & Goutte, 2017). One drawback to such approaches is that they tend to marginalize the human perspective of communication and event detection and tend to focus on the frequency of topics and keywords (Matei, et al., 2015). OSNs are constantly changing and evolving

communication streams that are fed by human user contributions (Weiler, et al., 2015). There are few if any examples in the research literature which seek to identify events as a time-series smoothed linear trajectory based on the integration of user and message streaming patterns. (Pinto, et al., 2019).

An event is defined as a set of tweets on a related topic within a certain period that surpass a threshold defined by statistical measures of diffusion and entropy. Our approach to identify and measure events was explained using the following abstraction. There is a dataset of tweets and the users who submitted them from a discussion stream *T* which is defined as $\{t_1, t_2, \dots, t_i, \dots t_n\}$. Each element *t* of set *T* represents an individual tweet. There is a set of Twitter users *U* defined as $\{u_1, u_2, \dots, u_i, \dots u_k\}$. In the context of our research there is a one-to-many relationship between the users in *U* and the tweets in *T*. An individual user can tweet a single message, multiple messages, or retweet the messages of other users in the discussion stream within the same time period. When a subset of a discussion stream is created, every tweet has an individual user associated with it. The message associated with the user could be an original tweet or it could be a retweet of another user. (Boyd, et al., 2010).

There are three metrics based on each element *t* that form the basis for event detection in an OSN. These metrics are period, diffusion centrality, and Shannon Entropy. The metrics are each discussed in the following sections. Each member of dataset *T* has a period *p* associated with it defined as $\{p_1, p_2, \dots, p_i, \dots p_m\}$. This paradigm in the context of our research was a discretization of time. The granularity of the discretization for this research is defined in days. The granularity can be further refined to minutes, or it can be expanded to be measured in weeks, months, or years.

Besides the period variable, another aspect of event detection is evaluated by a metric called diffusion centrality which infers the level of activity in an OSN with regards to tweeting and retweeting. Each tweet from dataset *T* has a Diffusion Centrality (DC) score, $DC(t_i)$ associated with it which is a property of tweet *t*. The variable $DC(t_i)$ measures how much influence a tweet has in a discussion stream when a user has tweeted or retweeted it (Kang, et al., 2012). The third metric that is used to identify and measure events is Shannon Entropy, represented in this research as E_s. E_s, also referred to as information entropy (Li, et al., 2015), a measure that was borrowed from Physics which originally measured the level of disorder in a system. Information Entropy was alternatively named after the scientist who converted the metric, Claude Shannon, so it was informally called Shannon Entropy (Li, et al., 2015). In the context of event detection research E_s is used to measure diversity in a discussion stream. The diversity that is measured refers to the number of messages being tweeted and retweeted in a discussion stream, which further suggests the level of diversity in the number of users who are tweeting. (Ghosh, et al., 2011).

In our abstract model for event detection, *p* is a date metric which measures the particular index of time that is being used for the study. In our case, the *p* value was measured in hours with an individual unit index being equal to a single hour on the x-axis. The $DC(t_i)$ value measures the diffusion centrality of an individual tweet in a discussion stream (Kang, et al., 2016). The E_s metric measures how diverse the messages are that are being tweeted and retweeted (Pinto, et al., 2019). This message diversity, in turn, infers the level of contribution made by the users in a discussion stream. (Ghosh, et al., 2011; Matei & Bruno, 2015).

To summarize our abstract model, E_s provides us with inferred information on *who* is spreading the message by telling us how diverse the tweets (and users) are. The $DC(t_i)$ metric tells us *how well* the tweet is spreading through the discussion stream (Kang, et al., 2012), and *p* tells us *when* the message spread occurred. The model for event detection based on Diffusion Centrality and Shannon Entropy could be expressed by the following mathematical formula σ = f($\frac{DC(t_i)}{E_S}$, *p*), where event σ is the result of the function of the ratio of Diffusion Centrality DC and Shannon Entropy per each date index *p*. The use of the ratio $\frac{DC(t_i)}{E_S}$ is a technique consistent with a methodology that was used by Du Jardin, P. (2010) for variable selection used in neural network classification. In our case, it served as a dependent variable for event prediction (Du Jardin, P., 2010). The result was a smoothed linear trajectory which ran longitudinally through the range of dates in the dataset of *T* that depicted peaks and valleys consistent with user activity in the discussion stream. There were examples in the literature which used linear time-series graphs to identify and evaluate events (Guille & Favre, 2015), however to the best of our knowledge there were no existing studies which used a smoothed linear trajectory based on the combination of E_s and $DC(t_i)$. In the context of the formula discussed in the previous section, an *event* could be identified as σ.

1. Calculate $DC(t_i)$ score of each tweet based on the connectivity architecture of the tweet senders and receivers in the discussion stream. The $DC(t_i)$ score calculation is represented by the equation d($DC(t_i)$) where $1 \leq i \leq n$, where DC is a Diffusion Centrality function, *t* represents a single tweet from an OSN stream subset, and t_n represents the nth tweet from the stream (Kang, et al., 2016). The $DC(t_i)$ score is calculated

by leveraging a programming language such as R and providing the appropriate parameters.

2. Next, the Shannon Entropy scores, which make up the denominator portion of the ratio in the σ couplet ($\frac{DC(t_i)}{E_S}$, p), must be calculated. To derive the values of E_s, the text fields for all tweets *t* will be grouped by the corresponding period variable *p*, which for our purposes will be the date. The entropy values will then be calculated by evaluating E_s ($\sum_p t_i$) (Ghosh, et al., 2011; Van der Walt, et al., 2018).
3. Calculate σ by evaluating the ratio of $DC(t_i)$ and E_s and pairing the ratio with a p variable as a couplet, ($\frac{DC(t_i)}{E_S}$, p). The ratio results in a new numeric value we will refer to in this research as *newsworthiness*, represented by *NW*. Higher levels of newsworthiness suggest increased levels of tweet exchange activity and a greater diversity of users in a discussion stream consistent with the occurrence of an event. In the context of this research, the ratio $\frac{DC(t_i)}{E_S}$ was referred to as *NW* for the purpose of discussing the variable in its implementation as a predictor for events in a discussion stream.
4. After researching the literature, we created a baseline based on an existing methodology from previous studies. The baseline consisted of datasets from two different currently trending topics. The accuracy of the baseline methodology was evaluated by comparing its smoothed linear trajectory

graph with the trajectory graph of our approach. The methodologies for the baseline and our approach are discussed below.

- *Baseline Approach:* According to several studies, the basis for many approaches in event detection is called Latent Dirichlet Allocation (LDA) (Figueiredo, & Jorge, 2019; Guo, et al., 2017; Wang, et al., 2012). The model for our approach differed from the methodology used by many existing event detection models in the literature. Unlike LDA, which was dependent on term frequency and word co-occurrence, our approach used an integration of messaging and user activity metrics. Our approach leveraged E_s and $DC(t_i)$ to identify the human influence involved in spreading messages in addition to the diversity of the messages being discussed. LDA is a Bayesian probability-based model which extracts topics from a sample of text by using keywords, term frequency, and probability to group words into parent topics based on the likelihood that certain words will appear together (Figueiredo, & Jorge, 2019). LDA works on the premise that every sample of text can be broken down into a finite number of topics. Under each topic is a group of related terms which are subordinate to a parent topic (Figueiredo, & Jorge, 2019). LDA is used as the foundation for many approaches to identify, track, and classify events from sources such as online discussion streams (Guo, et al., 2017; Wang, et al., 2012). The LDA model is created directly from

a text mining data structure called a Document Term Matrix (DTM), which is a two-dimensional data structure that keeps track of key terms and their frequencies from a dataset of text (Figueiredo, & Jorge, 2019).

- *Our Approach:* The model for our approach used E_s and $DC(t_i)$ as independent variables in order to evaluate the formula σ = ($\frac{DC(t_i)}{E_S}$, p) (Kang, et al., 2012; Li, et al., 2015).

The primary contribution of this research was a novel approach to Twitter event detection that used $DC(t_i)$ and E_s to identify events based on levels of user diversity and tweet exchange activity in a discussion stream (Kang, et al., 2012; Li, et al., 2015). Most current approaches to event detection used methodologies that exploited term frequency and topic extraction aggregated with a time-series (Patil, & Atique, 2013). The novelty of our approach was that events were identified using inferred levels of user contributions to an online discussion (Matei, & Bruno, 2015). Term and topic frequency distributions can be misleading in regard to the conclusions that the numbers suggest. Increased numbers may in fact be the result of smaller groups of users who are contributing larger amounts of messages within a subset of a discussion stream (Pinto, et al., 2019). The novel integration of $DC(t_i)$ and E_s allowed us to infer the amount of diversity in the users and the levels of messaging activity in an OSN subset. As a result, it identified events more effectively than in studies previously demonstrated in the literature.

The existing event detection methodology was driven by an LDA model, which was constructed using keyword frequency data derived directly from a DTM (Figueiredo & Jorge, 2019). Our approach did not use frequency distribution data (Patil & Atique,

2013). Instead it used a ratio of two metrics ($DC(t_i)$ and E_s) as input (Kang, et al., 2012; Pinto, et al., 2019). The most efficient method of evaluating the performances of the different approaches was to plot comparison linear trajectories in a time-series graph to assess which method better identified events (Lee & Sumiya, 2010; Pozdnoukhov & Kaiser, 2011).

Dissertation Goal

Event detection has been applied to several different areas to exploit the real-time format of social media platforms. For example, it has been used to assist in the administration of response planning by filtering Twitter's discussion stream for posts that relate to specific emergencies (Klein, et al., 2013). Event detection is also used to predict results in political elections. For example, events are identified in real-time from a discussion stream to provide trend analysis and public feedback so that news analysts and politicians can make well-informed decisions (Unankard, et al., 2014). In addition to public administration and political science, event detection has been implemented in the areas of cybersecurity and law enforcement. One such proposed application was the modeling of OSN behaviors to train intrusion detection system algorithms to detect malicious user behavior (Amato, et al., 2018). The principal goal of this research was to develop an alternate event detection technique applied to Twitter discussion data. The novelty of our technique was an integration of information entropy and diffusion metrics to evaluate user activity and diversity levels throughout a given time-period.

Research Questions

As we reviewed the research literature on the topics of SNA, Information Entropy, Diffusion Centrality, and event detection, we developed several questions that we wanted to answer at the end of our study. Our research questions are listed below.

1. Is the combined use of Information Entropy and Diffusion Centrality a valid method for the identification of events in a discussion stream?
2. Is the use of quartile analysis a feasible method for isolating average user messaging activity from events?
3. When the two approaches are considered, i.e. word frequency occurrence or user messaging activity, which approach produces more event peaks overall in a smoothed linear trajectory on a graph?
4. In a smoothed linear trajectory that has one or more event peaks, is there a sizable variance in the Information Entropy scores throughout the time period? Does this variance suggest noticeable changes in the diversity of participants in a discussion stream?

Relevance and Significance

Our literature review covered a broad range of topics in the domains of SNA, social network platforms, event detection techniques, machine learning algorithms, and evaluative metrics. Initially we intended to implement four different machine learning classifiers as part of our research. We researched four classifiers, i.e. Artificial Neural Network, Random Forest, Support Vector Machine, and XGBoost (Ren, et al., 2018; Zulfikar, 2019). After a thorough review and a considerable amount of empirical testing, we decided to include machine learning classifiers as part of future research. We found

that many event detection methods measured occurrences of term or phrase frequency over a time period (Patil & Atique, 2013). These methods used techniques which varied from wavelet analysis to measuring clusters (Cordeiro, 2012; Hasan, et al., 2016).

Microblogging, a.k.a. Twitter posting, is a popular source for SNA data according to several studies (Zhou & Chen, 2014). The samples are (usually) subject specific and are limited to 248 characters which makes them ideal for collecting samples (Guille & Favre, 2015). Twitter has its own issues with regards to SNA and data preparation. Some of the problems cited in this area include excessive noise (emojis, profanity, slang terms) and off-topic posts (Boyd, D., et al., 2010; Figueiredo & Jorge, 2019). Information Entropy is a metric that is used in several studies involving Twitter. It is used in SNA studies to measure *surprise* and *diversity* (Ghosh, et al., 2011; Vajapeyam, 2014). When used to evaluate Twitter text samples, entropy can quantify the amount of group participation that individuals contribute toward a common task such as a discussion (Matei, et al., 2015). Entropy is however not sensitive enough as a metric to provide a nuanced evaluation of text (Bentz, et al., 2017). For example, it can't distinguish between two different word orders using the same terms in a string. Diffusion Centrality is a metric that is used in SNA research intended to measure the *semantic importance* of individual users in a network. It takes into account a group of connected Twitter users and a context (Kang, et al., 2012; Kang, et al., 2016). The idea behind the Diffusion Centrality metric is that depending on the topic being discussed, one user may be more influential than another. We found this metric in SNA studies that focused on message virality and user influence (Alp & Öğüdücü, 2018).

In addition to entropy and Diffusion Centrality, *quartiles* are an evaluative technique that are derived from statistical methods (Shih & Liu, 2016). This technique is a form of data exploration which allows researchers to examine distributions so that outliers stand out. The creation of quartiles calculates a series of values which serve as boundaries when viewing data (Domínguez, et al., 2017). Q1 is a lower boundary which separates average data from low value outliers. Q2 is the median of a dataset. Q3 is the upper boundary which separates the average values from high value outliers (Langford, 2006; Shih & Liu, 2016). We found that using Q3 functioned adequately as a boundary between normal data and high-value data points. Outliers often suggest events since they represent tweets that fall outside and above the normal range (Lee & Sumiya, 2010; Pozdnoukhov & Kaiser, 2011). Our method used a smoothed line trajectory to identify the occurrence of events. Peaks in the smoothed line trajectory that formed above Q3 were interpreted as occurrences of events (Weng & Lee, 2011). This method was not quite ideal, but it allowed us to display time series data in a way that isolated average tweets from abnormal tweets. Preliminary empirical testing with the quartile method met with moderate success. Our research was significant for two fundamental reasons. First, the approach would allow entities such as governments, intelligence agencies, and corporations to identify and measure real-world events in an OSN using Twitter data as input (Atefeh & Khreich, 2015). Second, the approach would allow these entities to identify and follow emerging events as opposed to events that have run their course within the media (Cataldi, et al., 2010).

Barriers and Issues

Success in gathering data for our study depended on access to the Twitter platform's API. Enterprise memberships allow users to have privileged access to full archives of tweets along with platform metrics (Puschmann & Burgess, 2013). Rank and file users must abide by the policies put in place by the Twitter administrators. We did not possess privileged access to the platform's API, therefore we were restricted to the amount of data that we could collect for a single request. Our request for tweets was restricted to 10,000 rows for a single instance. If our request in a single instance exceeded 10,000, we received a message stating that our limit had been exceeded. We were forced to wait for a period of 15 minutes until our next available window opened. The issue that we had to consider for this study was that we had to assess the amount of usable tweets that we collected with our 10,000 tweet maximum per 15-minute window.

Assumptions, Limitations and Delimitations

The restrictions placed on rank and file Twitter users was discussed in the previous section with regards to the number of tweets available per 15-minute window. There was an additional limitation which applied to average non-paying users which affected the quality of the data. Whereas enterprise users could traverse the entire available Twitter timeline, non-paying users only had access to Tweets that dated back eight days during a general search. Those with access to the data "firehose" could gather all of the necessary data with one request within minutes. If a study was being conducted and the desired tweets were not available, a user was required to make several requests over a period of hours to collect the required number of tweets covering a desired time-

period. We decided upon a coverage period of fourteen days. We set a goal of 1,000 tweets for each of our two datasets per day over the fourteen-day period.

Definition of Terms

This section provides a list of definitions for specific terms that were used in discussion throughout the document.

Application Programming Interface (API) – A development interface that defines interactions between a user and a social media platform and defines the protocols that are used when requests are made for data.

Diffusion Centrality – A SNA metric that evaluates how frequently a message sent by a particular user is seen by other users participating in a discussion stream.

Discussion Stream – A flow of data or content consisting of semi-structured text messages, links, and multimedia (images and videos) that is contributed by users though activities conducted in an online social network.

Event – A set of tweets on a related topic within a defined time period that surpass a threshold defined by statistical measures of diffusion and entropy.

Information Entropy – The discrete probability distribution of Twitter text to measure the uncertainty or randomness of the data by analyzing its complexity.

Newsworthiness – A SNA metric of user activity that quantifies the distribution of user message spreading actions over the user diversity in a discussion stream.

Quartiles – An evaluative technique that is derived from statistical methods which allows researchers to examine distributions so that outliers stand out.

Social Network Analysis (SNA) – The analytical process of researching social structures through the use of networks and graph theory characterized by networked entities in terms of nodes (users in a network) and the edges, or links that connect them.

Trajectory – The curve articulated in a graph by a line moving through a timeline.

Chapter 2

Review of the Literature

Overview of Topics in Review

Currently in the literature there is no all-inclusive definition of an event in a social media discussion stream. A definition may involve geographic referencing, the occurrence of natural disasters, or possibly documented evidence of a crime. An event could be as simple as a discussion on some topic of popular culture, the Academy Awards for example. It may also be broader and be associated with a window of time. The definition of an event falls in line with the scope and nature of the research being performed. For our research, there were two principal areas of review from the literature. The areas that formed the basis for our methodology are listed below.

1. Currently existing methods of event detection.
2. SNA metrics for evaluating discussion activity.

In the literature, the most common approach to Twitter event detection incorporates several aspects of text mining aggregated with a time-series variable (Zhou & Chen, 2014). The text mining techniques that we reviewed from several studies included the use of unigrams (single words which have meaning in a body of text), bigrams (combinations of two words from a body of text), and trigrams (combinations of three words from a body of text) (Di Eugenio, et al., 2013; Moghaddam & Ester, 2012). Another text mining technique that is frequently used for event detection from the literature is Latent Dirichlet Allocation (LDA). LDA, which was first discussed in Chapter 1, is a Bayesian-based algorithm which breaks down a body of text into its fundamental topics (Lee, et al., 2010). In section one of the literature review, we will

discuss several existing studies which make use of these text mining techniques that are leveraged to identify events in a discussion stream.

Social Network Analysis (SNA) is a wide umbrella of techniques and metrics used by researchers to collect and evaluate information from social networks (Himelboim, et al., 2017). The term *network* can refer to a unit as small as a dozen people in a company E-mail chain. It can also refer to the members of a large Facebook friends list, where the complement may theoretically number in the hundreds or even thousands (Kim & Hastak, 2018). SNA metrics allow researchers to determine who in a network is the most influential and who is the best connected with the group overall (Garcia, 2017). The diffusion of information through a network is another frequently sought metric from social networks (Kang, et al., 2012). Shannon Entropy is another metric that is used in SNA to measure the diversity of users in a discussion stream (Pinto, et al., 2019). Section two of the literature review will discuss a number of the studies from the literature that used the Entropy and Diffusion Centrality metrics in their approach (Kang, et al., 2016; Van der Walt, et al., 2018). The last section of the literature review summarizes the three areas of review and provide insight as to how they led us to our methodology which will be discussed in detail in Chapter 3.

Gaps in the Literature for Event Detection

A thorough perusal of the literature on the topic of Twitter event detection yielded two common threads that existed in the majority of the available research. The first was a clear lack of an all-inclusive underlying definition of *event.* The definition can differ from domain to domain. The term *event* is ubiquitous in disciplines such as criminal justice, psychology, philosophy, computer science, and medicine (Choudhury & Alani, 2014).

The two common denominators of these many definitions are the inclusion of a time-period and an accompanying object that is measured throughout the time period. The second common thread throughout the many studies was the reliance on text-mining techniques to extract and produce features derived from tweet text (Di Eugenio, et al., 2013). The rest of this section will cover the various definitions of events that were found in the literature and the techniques that were chosen to identify them within Twitter discussion streams.

Definitions of Events

As it was mentioned previously, there is no uniform definition of an event (Choudhury & Alani, 2014). Based on the empirical review of several studies on this topic, the definition of event influenced the scope and depth of the study being performed. Depending on the study, an event could be broad and vague such as political issues and matters of public health (Wang & Goutte, 2017). An event could also be more specific and narrow in definition, such as a criminal incident or a personal occasion such as a wedding (Di Eugenio, et al., 2013; Wang, et al., 2012). There were some studies which did not specifically refer to temporal objects as events. The techniques used by the researchers were very similar to other event detection studies, however, alternate nomenclature was used when referring to events. In one such study, a temporal object that had been extracted from a social media discussion stream was referred to as a theme. Event detection in this study was referred to as temporal text mining (Mei & Zhai, 2005).

Two of the more unique identifications of events involved the aggregation of statistical change points and term frequency into their definitions. One study focused on events as consisting of clusters of *subevents* that could be visualized in a discussion

stream. Changes in the discussion stream were identified using change points. The study did not so much identify events as it sought to measure how recurring themes in Twitter changed through time (Wang & Goutte, 2017). The second more unusual definition for an event was one which used the term "bursty topic". According to the study, a topic was defined as *bursty* if it demonstrated a high frequency of mentions in a discussion stream. If the topic was bursty, it was deemed to possess the qualities of an event (Cui, L., et al., 2016; Guille & Favre, 2015).

Event Detection Techniques

The techniques used for event detection in the literature incorporated two fundamental approaches. Both of the approaches involved a form of "dissection" and analysis of Twitter text. The first approach focused on a type of study described in the literature as *n-gram analysis* (Lee, et al., 2010). The expression *n-gram* referred to the isolation of words and word groupings found in a body of text. The most commonly used types of n-grams were unigrams (single words), bigrams (two-word combinations), and trigrams (three-word combinations) (Nayak, et al., 2016). The second approach began with a Bayesian topic model based on the probability of certain words appearing together in Twitter text. The most popular of the topic model algorithms used in the literature was Latent Dirichlet Allocation (LDA) (Moghaddam & Ester, 2012). Both n-grams and topic models incorporated the use of term frequency and time as variables to determine whether a topic was surging or waning (Di Eugenio, et al., 2013; Zhou & L. Chen, 2014). Both of the aforementioned approaches were popular and are still frequently found in the literature in text mining and SNA research.

The first technique which was popular in the literature decomposes a body of text into its fundamental terms. This model of filtering out the most significant contributing words was referred to in the literature in many studies as “bag-of-words” (Moghaddam & Ester, 2012). The term “bag-of-words” was a research colloquialism that was used to describe the finished product of preparing a dataset of Twitter text and filtering out useless words, also known as “stop words” (Nayak, et al., 2016). The goal of creating a bag-of-words was to have a repository consisting only of terms that contributed the most meaning to the summary of an input of text. Once the bag-of-words was created, a frequency matrix was compiled, sorting the most frequently occurring terms in descending order (Moghaddam & Ester, 2012). In several studies, the bag-of-words was organized using all three variations of n-grams (unigrams, bigrams, trigrams) as separate steps (Choudhury & Alani, 2014; Nayak, et al., 2016). In the first technique of event detection, the frequency values and n-grams were used as features for classification tasks (Di Eugenio, et al., 2013; Moghaddam & Ester, 2012). The accuracy scores using the approach were average to above average based on the available studies from the literature. One study employing this method achieved accuracy scores of 86.2% using a unigram model for classification (Di Eugenio, et al., 2013).

The second event detection technique that was popular in the literature used a topic modeling approach as the basis for identifying events in a discussion stream (Zhou & Chen, 2014). The most popular method of topic modeling was an algorithm called LDA, which is discussed above (Weiler, et al., 2015). Instead of using frequently occurring n-grams like the previously discussed method, LDA sought to cluster words that appeared together in a text with greater frequency. The clusters of related words were

named topics. In some studies, *topics* were used interchangeably with *events* (Cui, et al., 2016), asserting that the topics (a.k.a. events) were constructs that were aggregated with time and frequency variables. One study used LDA to cluster topics pertaining to crime using Twitter posts as input. The study used topics generated from existing tweets within a generalized linear model to predict the probability of crimes occurring in the future. The study successfully predicted future hit-and-run incidents, but the study admitted that its confidence interval was rather wide (Wang, et al., 2012). The two previously discussed approaches to event detection are still found in SNA research in the literature. Based on the synthesis gleaned from several studies in this domain, the apparent benefit of techniques such as LDA and bag-of-words is that they provide a bountiful source of features for prediction and classification tasks (Choudhury & Alani, 2014; Cui, et al., 2016). When text mining features are combined with other SNA metrics such as diffusion centrality and Shannon Entropy (Ghosh, et al., 2011), machine learning classification models can be more diverse and nuanced. SNA metrics will be covered in the following section.

Analysis of SNA Metrics Used in Similar Studies

In the context of our research, Social Network Analysis (SNA) is a discipline that articulates relationships between social media users that is based on methods derived from graph theory (Alarcão & Neto, 2016). One of the fundamental goals of SNA is to identify influential and important user nodes in a network (Bonchi, et al., 2016). The task of isolating and documenting these influential nodes eventually led to a construct known as *centrality*, which is a measure of different aspects of network importance. We could find no all-inclusive definition of *centrality* in the literature, so depending on the context

of the study, a centrality metric could evaluate concepts such as influence, authority, and power. In the SNA research literature there were many existing measures of centrality, including *betweenness*, *closeness*, and *eigenvector*. These metrics evaluated user nodes based on efficiency, independence, and how well connected they were when the structural properties of a user network were taken into account (Grando, et al., 2016).

In addition to centrality, another area of research interest in SNA was that of information diffusion. This area of research asked the question: what are the variables that cause information to spread through a network? (Yoo, et al., 2016). A thorough perusal of the literature uncovered studies in this area which focused on a metric with combined aspects of centrality and information diffusion. The metric was called Diffusion Centrality ($DC(t_i)$) and it evaluated a user's influence in the spread of information through an OSN. The $DC(t_i)$ metric differed from other forms of centrality such as *betweenness* and *closeness* in that a node's level of influence could change based on the semantics of a topic being spread (Kang, et al., 2012; Kang, et al., 2016). For example, one user in an OSN might be an authority on politics, but that same user might not be an authority on popular culture.

Another SNA metric that was found in the literature was Shannon Entropy (E_s), also known as Information Entropy. The E_s metric was not a measurement of centrality. It evaluated the amount of information that was present in a dataset of user text (Li, et al., 2015). This in turn could be used to evaluate and infer the amount of diversity that existed in a dataset of user tweets. Diversity in the context of our research could refer to topics or users in a discussion stream (Pinto, et al., 2019). After synthesizing the literature on these topics, two variables stood out as viable candidates for further analysis, which

were user *message spreading influence* and *user diversity*, represented by the two aforementioned SNA metrics (Kang, et al., 2012;Li, et al., 2015). What distinguished these two variables from other SNA metrics was that the end values were not purely dependent on a static network architecture. The outcome could change based on the topic being spread in an OSN. The remainder of this literature review will be divided into two parts. The first part will discuss the $DC(t_i)$ metric and its importance to SNA and information diffusion (Kang, et al., 2016). The second part will discuss the E_s metric and how it relates to measuring user diversity (Pinto, et al., 2019). The section will conclude with a summary discussing the advantages and disadvantages of using the two metrics in the context of our research.

Diffusion Centrality ($DC(t_i)$)

The $DC(t_i)$ metric evaluates how much influence a user node has with regards to the spread of information. The foundation of the metric is that a user's influence can change based on the topic being discussed in a discussion stream (Kang, et al., 2012). Many other measures are static and depend purely on the connectivity of the overall network (Grando, et al., 2016). According to the paradigm of many static centrality measures, an influential user will always be an influential user because he or she is well-connected (Fredericks & Durland, 2005). With the introduction of the $DC(t_i)$ metric, a user who carried a high score in one topic could score much lower with another topic. The agent of change for such a difference in scores was the introduction of a different diffusion model (Kang, et al., 2012). According to the literature, a *diffusion model* was a hypothetical mathematical model which recreated the progressive spread of an object (Yoo, et al., 2016) such as a message from person to person through a discussion stream.

The $DC(t_i)$ metric was not based on any particular diffusion model, but it took the model as input to evaluate the amount of influence that a user node had (Kang, et al., 2012). We surmised that the benefit of this metric on SNA research was that it provided researchers with the mechanism to study the changing and evolving nature of a discussion stream (Java, et al., 2007; Kwak, et al., 2010). After additional research, we found that there were development libraries in the R programming language which supported the implementation of the $DC(t_i)$ metric. The diffusion model aspect of the $DC(t_i)$ formula was built-in to the library as a parameter (An & Liu, 2016).

Shannon Entropy (E_s)

The E_s metric is also referred to as Information Entropy and was introduced in the late 1940's by Claude Shannon (Shannon, 1948). It was adapted from Physics and applied to information theory with the purpose of evaluating the complexity of systems. With regards to communications, the E_s metric was used to measure the structural information content of text (Dehmer & Mowshowitz, 2011). The E_s metric was used in several SNA studies to measure topic diversity in a discussion stream as well as user distribution (Ghosh, et al., 2011; Pinto, et al., 2019). The latter adaptation was of interest to our research. By evaluating the amount of user diversity in a discussion stream, a small E_s could suggest spam activity (fewer users with more activity), while a larger E_s score (more users in the discussion stream) might suggest increased interest in a topic (Ghosh, et al., 2011). Hasan, et al. (2016) published a study in which the E_s metric was used to evaluate both topics as well as user diversity. The approach clustered tweets by topic similarity and then evaluated the clusters using the E_s metric. Clusters with a E_s topic score greater than 2.5 and a user diversity score greater than 0.0 were considered to be

events (Hasan, et al., 2016). This approach used a combination of text mining and user diversity data to identify events. This was an interesting approach, but it did not include a variable that sought to more succinctly quantify the amount of user-generated activity that was taking place in the discussion stream.

Summary of $DC(t_i)$ and E_s

The $DC(t_i)$ and E_s metrics were both used in studies dealing with SNA. While E_s has been used to measure different aspects of diversity in Twitter datasets in multiple studies, we found that the $DC(t_i)$ metric was used mostly to study the identification of opinion leaders and key spreaders of information for specific topics in discussion streams (Gunasekara, et al., 2015; Kang, et al., 2016). Although $DC(t_i)$ was used to study the spread of topics on Twitter (Bingol, et al., 2016), it has not been used specifically for event detection. It is a measure of influence that was intended to be used to reflect changes in key influential user nodes over different or progressive datasets (Kang, et al., 2012). E_s was a metric that was adapted from its original domain in Thermal Physics to measure the amount of information that was inherent in a system (Li, et al., 2015). The metric was further adapted to evaluate the level of diversity that existed in a SNA dataset (Pinto, et al., 2019). We found that a key benefit of using E_s with regards to SNA and Twitter was that it could suggest whether a small number of users were responsible for a larger amount of tweets, or if the Twitter content was the result of several different users (Ghosh, et al., 2011). The implication of this difference was that the former outcome could be the result of possible automated activity such as a bot, while the latter outcome suggested increased interest in a topic (Chu, et al., 2012). The integration of E_s and $DC(t_i)$ was an area of research that was not found in a thorough review of the literature.

In terms of event detection in Twitter the ensemble of $DC(t_i)$ and E_s was an interest to us as an avenue of research because the approach did not depend on word frequencies to identify events.

Summary of Research

The two principal areas that were covered in this review were existing methods of event detection and SNA metrics. There is also a third area, machine learning, which we intend on pursuing in later research. This topic will briefly be discussed at the end of this summary. After a thorough perusal of the literature, it became apparent that there were three items of interest that needed to be highlighted. The first was that there was no existing all-inclusive definition of an *event* (Ghosh, et al., 2011). The definition depended on the scope, time-period, and domain of the research that was involved. The second item was that there was an abundance of SNA metrics available with which to analyze different aspects of social networks (Grando, et al., 2016). *Centrality* is a broad and generic term for a system of metrics that evaluate different aspects of networked users in a discussion stream. Some of the metrics have existed nearly as long as the field of SNA itself. *Closeness*, *betweenness,* and *degree* centralities are foundational measurements that were found in many studies in the literature (Alarcão & Neto, 2016; Peng, et al., 2018). These three metrics quantified different aspects of information transfer efficiency and influence (Grando, et al., 2016; Peng, et al., 2018). One criticism of these metrics was that they were static in nature and did not capture gradual change in a network over time. Another commonly used metric in SNA research was E_s, which was adapted from the field of Physics to information theory (Li, et al., 2015). Scientist Claude Shannon published his paper on this adapted metric in the late 1940s. In its new interpretation, E_s

measured the amount of diversity in topics and users in a discussion stream (Pinto, et al., 2019). The principal difference between E_s and centrality in SNA research was that E_s was evaluated based on the connective architecture of nodes and edges in a social network (Alarcão & Neto, 2016). The E_s metric was used frequently to evaluate user text from sources such as Twitter posts (Ghosh, et al., 2011).

There were examples in the literature of researchers creating new SNA centrality metrics designed to capture aspects of change in a discussion stream. One such metric was $DC(t_i)$, which sought to measure influential users in a network for different topics. The authors of the study emphasized that a user in one discussion stream might not hold the same level of influence for a different topic (Kang, et al., 2012). $DC(t_i)$ was a centrality metric, however it differed from its predecessors in that it was not static like the betweenness, closeness, and degree centrality metrics. It required a diffusion model along with the nodes and edges to explain how information such as tweets was spread from user to user in the discussion stream.

An additional issue that needs to be highlighted is the use of machine learning algorithms to predict the occurrence of events in a discussion stream. Based on a study that was found in the literature, 22% of research conducted into the domain of social media used Support Vector Machine (SVM) models to evaluate data. Approximately 6% of that same pool of research used Artificial Neural Network (ANN) models (Injadat, et al., 2016). The aforementioned study did not have a statistic for ensemble models such as Random Forest (RF) or XGBoost. Ensemble models are rather popular according to the literature, due to the fact that they aggregate the strong points of individual classifiers to produce a more robust score as a result (Dey, 2016).

Based on the studies discussed above, we performed several empirical experiments using all four models with sample data. As our research continued, we compiled enough empirical and documented research to support the use of SNA metrics $DC(t_i)$ and E_s integrated into one metric to detect and identify events (Díez-Pastor, et al., 2015) . We did not, however, have a documented and supported approach to evaluate our metric using machine learning classifiers. We decided to use the quartiles data exploration approach to evaluate our method (Lee & Sumiya, 2010). In future research we intend to implement SVM, ANN, RF, and XGBoost to provide more concrete evaluation data for our approach using confusion matrices.

Chapter 3

Methodology

Overview of Research Methodology

The approach to event detection detailed in this document was the result of assiduous and careful review of the literature on the subject and continuous empirical experimentation with sample data. To help the reader better illustrate the techniques in our approach, an example case study was used in the various sections throughout the chapter. The example study used a sample dataset collected and processed using the same techniques that were discussed in this document.

Our discussion first addressed the research methods that were used in the study. The discussion began with part one of our case study which demonstrated how a dataset was collected by leveraging the Twitter platform's application programming interface and searching for tweets based on a hashtag keyword search. Part two of the case study detailed the attributes that made up a collected dataset that are part of an imported raw dataset file. Part two of the case study then detailed which attributes were used for further calculations and which ones were discarded. Following parts one and two of the case study, the discussion moved to the calculation of the Diffusion Centrality ($DC(t_i)$) attribute. To create the $DC(t_i)$ attribute a dataset of tweets first had to be presented as a graph of users and connecting edges (Otte & Rousseau, 2002). The $DC(t_i)$ scores were then derived by considering the inherent interconnectivity of Twitter users (Proskurnikov & Tempo, 2017). With respect to SNA, the $DC(t_i)$ metric evaluated how many times or how frequently a message spread by a user could be seen by other users in the same discussion stream (Kang, et al., 2012).

Following the $DC(t_i)$ attribute, the calculation of the Shannon Entropy (E_s) attribute was discussed. E_s was a metric that had its origins in Thermal Physics, but it had been adapted to the field of Information Science (Wang, et al., 2018). With regards to our research, E_s was a metric that evaluated the amount of user diversity that existed in an aggregated sample of Twitter text (Ghosh, et al., 2011). The metric lacked the precision for a nuanced analysis of text, however it was useful for evaluating the diversity of a dataset at the macro level (Bentz, et al., 2017). The $DC(t_i)$ and E_s were integrated to form a unique single attribute called Newsworthiness (*NW*) which we discussed at length. *NW* is the ratio of user message spreading over user diversity. When displayed on a graph, increased levels of *NW* suggested the occurrence of events in a discussion stream. The approach to measuring *NW* is discussed in the next section.

The discussion moved next to instrument development and validation. In this section we discussed how we would measure the *NW* attribute. During early research, we experimented with machine learning classifiers and the use of a threshold line to identify events based on increased levels of *NW*. Initial results from our empirical testing were not satisfactory so we decided to pursue the use of machine learning in later research. Ultimately, machine learning classifiers were desirable because they provided concrete evaluative results in the form of confusion matrices (Lokeswari & Rao, 2016). In lieu of classifiers we decided to implement quartiles as an evaluative technique because they were supported in several studies of SNA and Twitter (Pozdnoukhov & Kaiser, 2011). As it was suggested previously, the downside to using quartiles as a metric was that the technique did not consider the dataset as a whole, but as a set of fragmented ranges (Rousseeuw & Hubert, 2011). We opted for quartiles because they allowed for a method

of displaying data which isolated "normal" user spreading activity from outliers (Lee & Sumiya, 2010). This method of data evaluation allowed us to show the full range of the *NW* attribute distribution while using the Q_3 (upper quartile) value as a boundary fence (Rousseeuw & Hubert, 2011). Our suggestion in this case was that events tended to occur in the region above Q_3 as events were associated with elevated levels of *NW*. Based on initial empirical testing, we found that the quartiles method was not ideal, but it provided a sufficient method of measuring a dataset *NW* distribution over a time period that was covered in a study.

The discussion moved next to data analysis. To analyze the efficacy of our approach, we compared our results to the results provided by a popular existing approach to event detection. One of the popular approaches to event detection that was found in the literature was measuring keyword frequency over a time-period. Peaks in keyword occurrence during a particular time index suggested events (Figueiredo & Jorge, 2019; Guo, et al., 2017; Wang, et al., 2012). Initial empirical testing proved that increased occurrence of keywords during a given time index resulted in peaks when the keyword frequency trajectory was shown in a linear graph using two dimensions (time index and frequency). We analyzed the efficacy of our approach by comparing the linear trajectory of *NW* with the trajectory of keyword frequency using the same dataset. Initial experiments with our case study sample data showed that the *NW* trajectory resulted in the occurrence of more defined peaks during the time-period covered.

Following the section covering data analysis, the discussion moved to formats for presenting results. It was mentioned in the previous section that our approach and the existing method of event detection would be evaluated using linear trajectory graphs for

the time-period covered. Based on empirical testing with the case study dataset, the most efficient and effective method of displaying results was a smoothed linear graph that included data points for individual tweets that occurred at their respective time indices. The smoothed line took data from the averages of the points that were plotted at every time index and created a linear representation based on the averages. The points on the graph helped to explain where the weight of the clustering of tweets fell, causing the trajectory to ascend or descend, resulting in peaks and troughs.

The discussion of our method concluded with a section covering required resources and a summary. The summary included our datasets for the study and the time-period that was covered. We made use of two datasets for this research. The time-period covered for the two datasets was fourteen days. The individual time index for the study was a single day. The topic for the first dataset was based on a keyword search using the phrase "Tulsa+Rally" as the search terms. The second dataset was collected using the keywords "Atlanta+Protests." The two datasets had between them fourteen and twelve thousand tweets (based on API availability) and were collected every day over a period of fourteen days. For each day sampled, the API was leveraged multiple times to ensure a completeness of coverage for the 24-hour period. The summary also included a discussion of our future research. In future research we plan to use four machine learning classifiers to evaluate *NW*. Specifically, these are Artificial Neural Network, XGBoost, Random Forest, and Support Vector Machine. We will also integrate sentiment analysis as an additional attribute to the composite *NW* metric. Sentiment analysis will provide an additional layer of evaluation by considering events in terms of user emotion in addition to user diversity and message spreading activity.

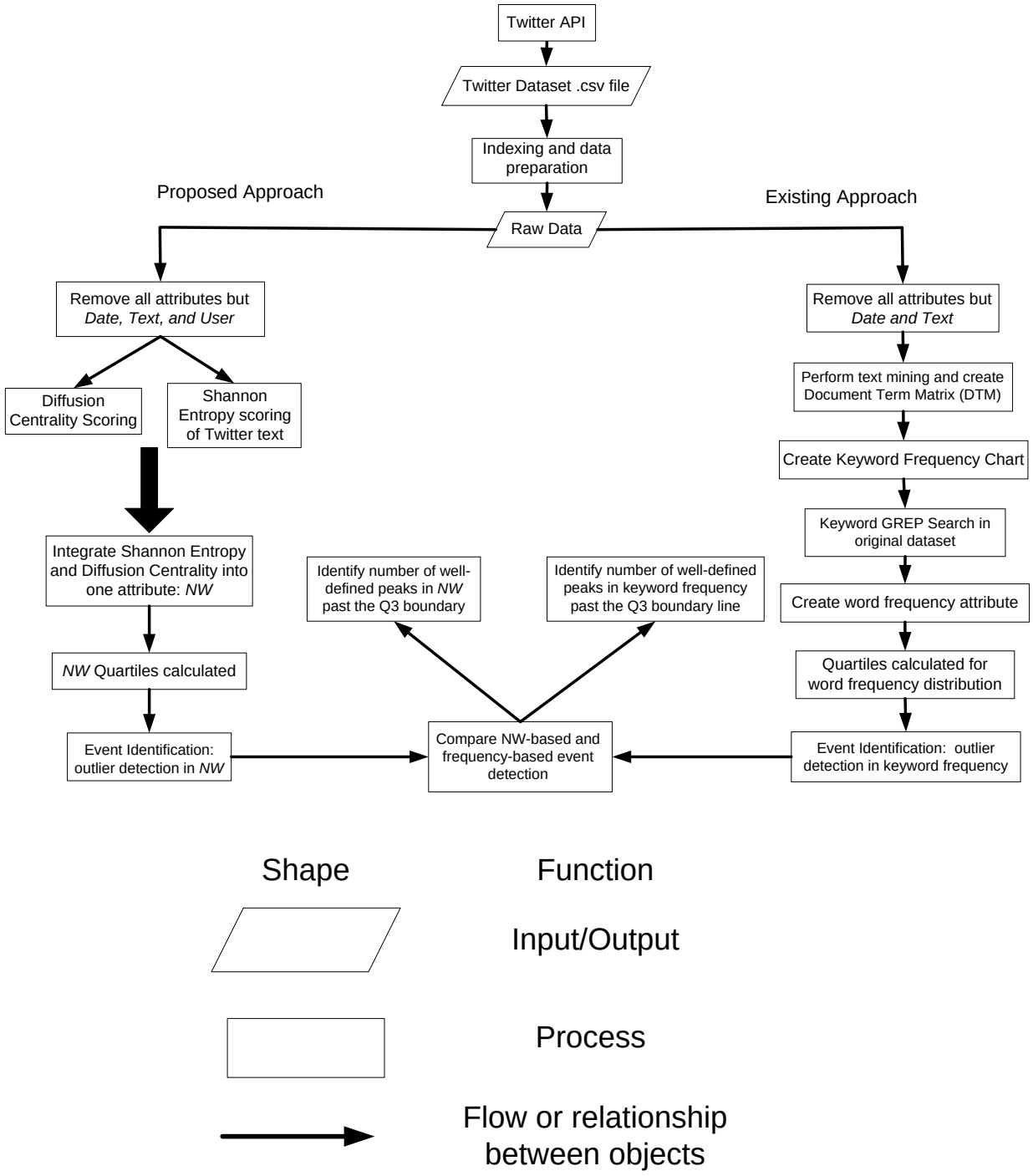


Figure 1. Diagram of the event detection process

Research Methods Employed

In the following section, we discussed the research methods that were implemented in our study. The discussion began with our approach to dataset collection using a hashtag or keyword search within the Twitter API. The discussion continued with the calculations of the $DC(t_i)$ and E_s scores. The discussion of the section concluded with the integration of the $DC(t_i)$ and E_s scores into the single attribute called *NW*. The previously mentioned research methods were explained using our case study which helped to clarify the process by example.

Case Study: Collecting the Dataset Based on a Hashtag Search

To better illustrate the many aspects of our method, for this research we collected a sample dataset as part of a case study which is intended to illustrate the steps involved in the approach. The case study included collecting the dataset, processing it, and graphing the results. In this section we were concerned with acquiring Twitter data, so we began by finding currently trending topics and selecting a sample hashtag as the basis for our search. After a brief perusal of the Twitter interface, we acquired a list of the top 50 topics that were currently trending. We selected the second hashtag from the list, which was *#DowJones*. We used the R programming language to leverage the Twitter API so that it returned a requested sample of 10,000 tweets, which was the maximum number allowed by the API per one-time request. The Twitter API returned the requested number of tweets which spanned a time frame of nineteen hours from their posting time index in the discussion stream.

Case Study: API Table Attributes and the #DowJones Dataset

Prior to creating the four variables that we would need to predict events, some initial preprocessing of the #DowJones dataset needed to take place. The raw collection of tweets that were provided by the API included 16 attributes. Of these 16 raw attributes only four needed to be kept. These attributes were *created*, *text*, *screenName*, and *isRetweet*. Table 1 illustrates the attribute names and their corresponding data types. The remaining 12 raw attributes were not used and were discarded. Table 2 displays the API attributes that were discarded from the raw tweet collection. Of the four attributes that were kept, three were used to create the $DC(t_i)$ attribute. Once this attribute was created, the *screenName* and *isRetweet* attributes were discarded. The *text* attribute was used to calculate E_s. Once the E_s attribute was calculated, the *text* attribute was discarded. The *created* attribute was the only original attribute that was kept throughout the rest of the event detection process.

Table 1

List of Twitter API columns that will be used for event detection

Twitter Data Attribute Name	**Data Type**
created	Time index/Time
text	Character
screenName	Character
isRetweet	Boolean

Table 2

List of Twitter API columns that will be discarded

Twitter Data Attribute Name	
favorited	statusSource
favoriteCount	replyToSN
id	replyToUID
latitude	retweeted
longitude	retweetCount

Calculating $DC(t_i)$ Scores from the Twitter Retweet Graph

Before a dataset could be used to detect events, there would be a total of four attributes. The first attribute, *hour* (converted from the *created* attribute), was carried over from the original tweet samples collected from the Twitter API. The other three attributes needed to be calculated. These attributes were $DC(t_i)$, E_s, and *NW*. To calculate the values for $DC(t_i)$, the tweets collected from the API had to first be visualized as a graph. A Twitter graph is a construct in sociological research that is a *visualized representation of a network* of connected entities, usually people. In the SNA literature, the terms *graph* and *network* are synonymous (Otte & Rousseau, 2002). A graph is formally defined as $G=(V,E)$, where $V=\{v_1, \ldots, v_n\}$ and $E=\{e_1, \ldots, e_k\}$ are finite sets. The individual $v \in V$ elements are *vertices* (individual Twitter users in a discussion stream) and $e \in E$ elements are *edges* (lines that connect user vertices) (Proskurnikov & Tempo, 2017).

The edges of a graph can be weighted or unweighted. A *weighted edge* is a connection from a network that has an associated numerical value that assesses its strength in a category relative to another edge. An example of a weighted edge in the context of our research is a connection between two Twitter users where message sharing occurs several times in a single day as opposed to other connected users who may share only once a day. A connection such as the one just described may be assigned a numerical value to demonstrate the higher rate of message exchange. If a network has no comparisons of relative strength in their connections, the edges are *unweighted* (Malliaros & Vazirgiannis, 2013; Newman, 2004). The edges of a graph can also be *directed* or *undirected*. If the edges between two users in a graph are *directed*, then it is implied that the flow of information is only from one Twitter user to another, not both ways. If the edges of the graph are *undirected*, the flow of messaging is implied to take place back and forth between *both* users (Newman, 2004; Proskurnikov & Tempo, 2017). In the context of our research, the flow of messaging between users in the graph was *undirected.* The edges between users were also *unweighted*, meaning none of the connections between them held any greater emphasis over others. All edges possessed equal weight. The implication of not having any special weights or directed flow between users implied that there were no special considerations to be taken when calculations were evaluated. Calculations would be based on user connectivity, not directional flow or weight.

A Twitter graph was created with each tweet representing a node and a retweet action representing a line connecting the users (Malliaros & Vazirgiannis, 2013). The $DC(t_i)$ scores were determined based on the connectivity of the graph's users (Otte &

Rousseau, 2002; Proskurnikov & Tempo, 2017). The scores were calculated from the two-dimensional retweet graph using the R programming language and the *keyplayer* development package within R Studio. The $DC(t_i)$ scores were evaluated using the formula $DC(t_i) = \sum_{t=1}^{T} Pr^t$, where t_i was an individual tweet from a discussion stream and Pr^t was the probability that one network (user) node could reach an adjacent (neighboring) node in *t* iterations where *T* was the total number of iterations in the time period covered in the discussion stream and *t* was a single iteration (An & Liu, 2016).

The formula mentioned above required a few additional definitions and supplemental discussion to provide clarity. A *sparse matrix* in the context of this study is a two-dimensional mathematical matrix representation of a finite SNA graph where connections and lack of connections between users are represented by zeroes and ones. The matrix is referred to as *sparse* because a large number of its cells contain zeroes (Davis & Hu, 2011). A matrix cell with a *1* value in it represents a connection between users. A sparse matrix is defined as $A = (a_{ij})_{i,j \in V}$, where a_{ij} is a value corresponding to an edge in a graph of a discussion stream and $i, j \in V$ represents two connected users from the finite set of users in a discussion stream. A sparse matrix is the encoding that defines the connections between users in a network graph (Proskurnikov & Tempo, 2017). A *diffusion model* is a small world representation of a Twitter discussion stream that represents all of the potential propagation paths that a message could take through peer-to-peer interactions between users in a network subset (Zhang, et al., 2016).

Based on the above definitions, the following discussion provides further details for the process of calculating $DC(t_i)$ from a SNA graph. $A \in \mathbb{R}^{nxn}$ is a sparse matrix of an undirected network with *n* nodes (users) based on a graph of a Twitter discussion

stream. Sparse matrix *A* contains the structural information for how the users and edges are connected in a network subset. $Pr = A * x$ is a probability matrix created by multiplying an assigned value representing a level of probability *x* by sparse matrix *A* (An & Liu, 2016; Takada, et al., 2010).

Probability matrix *Pr* is a variable which stores values that estimate the likelihood that a user node will spread a message to another user node (An & Liu, 2016). An easier way to understand the formula $Pr = A * x$ is to decompose it in the following manner. *Pr* is a variable which contains the results of $A * x$, which will be used in later calculations. *A* is a sparse matrix that contains the connective information about a Twitter network. Specifically, it contains the mapping data describing which user connects to other users (users and edges). The *x* variable in the formula represents the *probability* that a user will send a message. The *x* variable is multiplied by the sparse matrix containing users and edges, represented by *A*. The multiplication results in a matrix of numbers which represents a probability value that a user will send his message. This value remains the same throughout the number of iterations that a network passes in their time-period. To that end, the value *x* does not change if the same user sends different messages (An & Liu, 2016).

The value of probability variable *x* is an aspect of user behavior prediction that is used in many studies of information diffusion (Han & Tang, 2015). In our research, the value of probability variable *x* was applied to all users in a sparse matrix (An & Liu, 2016). We found examples in the literature which supported this implementation. To this end, we found studies which used a single explanatory variable to simulate a binary user state (Cha, et al., 2010; Diederich & Busemeyer, 2003). For example, a user could *send*

or *not send* a message. A user could also be *informed* or *not be informed*. In those studies, a value such as .4 was assigned to a variable to represent a *probability* that a user would behave in a particular manner (Diederich & Busemeyer, 2003). The model that studies use to implement diffusion probability is defined within a sparse matrix (Heaukulani & Ghahramani, 2013). In the cases we researched, when the probability variable was applied to the sparse matrix the outcome was determined by the topology of connected users (Heaukulani & Ghahramani, 2013). We chose the implementation of probability variable *x* as a single value because the approach helped to simplify the simulation of human decision-making (to send a message or not) in our Twitter discussion stream model (Diederich & Busemeyer, 2003; Han & Tang, 2015).

There was an alternate approach to creating the probability matrix, which was to provide the probability values for each node if the data was available. There was no guidance available in the literature to justify the use of such an approach. We chose the first method described previously, which is a simplified and more generic approach to creating a probability matrix to simulate diffusion. It was not a complex simulation which incorporated a changing probability (which was preferable), but it provided an adequate time series model of diffusion which met our goals of identifying events by analyzing message spread (An & Liu, 2016; Diederich & Busemeyer, 2003). For many studies in the literature, this simplified approach was preferable to individually assigning probabilities to users using complicated diffusion models (Vandekerckhove & Tuerlinckx, 2008).

The $DC(t_i)$ metric is generally intended to work with the connected users and edges of a Twitter network graph (a.k.a the architecture) and a *diffusion model* which

explains how a message spreads through it. Due to the rigidity and difficulty of formal mathematical diffusion models such as *cascade* and *threshold* in SNA studies, researchers have sought to implement alternate methods to simplify the simulation of information diffusion (Diederich & Busemeyer, 2003). As a result, simpler probability matrix methods have been used in SNA research in lieu of the more intractable formal models (Vandekerckhove & Tuerlinckx, 2008). The probability matrix *Pr* in the aforementioned formula simulates a diffusion model using a simpler generic approach (Takada, et al., 2010). This simpler approach uses the product of multiplying sparse matrix *A* with a numerically assigned level of probability *x*, .3 for example (An & Liu, 2016). The researcher is charged with assessing the value for probability *x*. The *T* variable in the $DC(t_i)$ calculation formula is the number of iterations a Twitter network will go through to spread information among users in the network. The *T* variable and *Pr* probability matrix are the two inputs to the equation $DC(t_i) = \sum_{t=1}^{T} Pr^t$.

Case Study: Calculating the $DC(t_i)$ Scores for the #DowJones Dataset

Our research used *hour* for the time index. Samples were taken over a 14-hour period. The first of the attributes that we created as part of our case study using the #DowJones dataset was $DC(t_i)$. As was discussed in the prior section, the tweets had to be visualized as a graph before the $DC(t_i)$ scores could be evaluated. The graph for our case study was visualized using the *igraph* package in R studio. The rows from the *text* and *screenName* attributes were represented as users in the graph. The *screenName* and *text* attribute members from the dataset were each represented as individual users. Edges were displayed in the graph representing connections between the users. These edges represented retweet relationships between users. Individual points in the graph

represented tweets (users) whose value in the *isRetweet* field was TRUE, indicating that all of the tweets in the graph were retweets. After the graph of the "#DowJones" dataset was rendered, the *keyplayer* package was used to calculate the $DC(t_i)$ scores. The scores were saved as the second attribute of the dataset next to *hour.* An example of the $DC(t_i)$ scores can be seen in Table 4. The sparse matrix which is used to calculate $DC(t_i)$ is discussed next.

Case Study: The Sparse Matrix for the #DowJones Dataset

The *sparse matrix,* which was discussed above, was used to calculate $DC(t_i)$ for a dataset. To calculate $DC(t_i)$ a Twitter discussion stream had to first be represented as a graph. The graph was converted from a system of users and edges to a sparse matrix which was used as a parameter to create a *probability matrix*. The probability matrix was used to evaluate Diffusion Centrality measures using the formula $DC(t_i) = \sum_{t=1}^{T} Pr^t$. The probability matrix contained the probability that a user would spread a message to an adjacent (neighboring) user. According to the simulated model that drives the probability matrix, a user has a single chance (probability) to spread its message to an adjacent user (An & Liu, 2016; Saito, et al., 2011). Table 3 below demonstrates a scaled-down sample sparse matrix taken from the *#DowJones* dataset.

Table 3

Example of Sparse Matrix for #DowJones Dataset

	corporatepiggie	kevinsvenson_	matthewryancase	realdonalbtrump	rumanaalvi1
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	1	0	0	0	0
10	0	1	0	0	0
11	0	0	1	0	0
12	0	0	0	1	0
13	0	0	0	0	1
14	0	0	0	0	1
15	0	0	0	0	0
16	0	0	0	0	0
17	0	0	0	0	0

In Table 3 the node and edge configurations are stored in the sparse matrix in a two-dimensional format. The names of the individual users are listed as column names. The sequential numbers on the vertical axis represent the users in the network where a potential connection exists. The cells that lie beneath a user's column in the matrix define whether or not the aforementioned user has a connecting edge with another user. With regards to the sum total of a single user's connections, the data is read vertically from the top down at each intersection point between the user column and the numbered row. All matrix cells are zeroes unless there is a connection between two users in which case a *1* inhabits the cell at the intersecting juncture between the user's column and the numbered row (Davis & Hu, 2011). Sparse matrices differ from other data structures such as adjacency matrices with regards to undirected networks. For example, in an adjacency

matrix an undirected network is identified by the symmetric placement of zeroes in the matrix with respect to the ***1*** values (Wagner & Neshat, 2010; Weisstein, 2007). In a sparse matrix, users and their connecting counterparts are placed along the vertical and horizontal axes. All cells are zeroes unless there is a connection, so there is no symmetric placement of zeroes (Davis & Hu, 2011).

To illustrate the process that was detailed above, consider in Table 3 the username "realdonaldbtrump" that occupies the fourth column of the sample matrix. All of the cells are zero with the exception of the twelfth row. This configuration indicates that user "realdonaldbtrump" is only connected to one other user in this discussion stream. The user that is associated with row 12 is known internally by the function creating the sparse matrix. Alternately, if we consider the user "rumanaalvi1" that occupies the fifth column, we observe that there are two cells with a 1 value in them. This configuration indicates that user "rumanaalvi1" is connected to two other users. By viewing the sparse matrix, it is not explicitly evident who the users are that are connected to "realdonaldbtrump" or "rumanaalvi1" since the connected users are only identified by sequential numbers. The user to user relationships are visually evident when the edge and node connections are restored in graph format. User relationships can be identified by reverse engineering the cells that have a value of 1 in them and matching them to corresponding users. In a very large discussion stream, that would take considerable time.

Table 4

#DowJones dataset with t_i, hour and $DC(t_i)$ columns

t_i	hour	diffusion
t_{1723}	3	2.44
t_{1724}	3	1.8
t_{1725}	3	15.12
t_{1726}	3	41.36
t_{1727}	4	0.28
t_{1728}	4	1.12
t_{1729}	4	63.4
t_{1730}	4	1.52

If we look at the first row in Table 4, the $DC(t_i)$ value for the tweet is 2.44. By contrast, the tweet with the highest $DC(t_i)$ value in the example table is 63.4. In the context of our research $DC(t_i)$ was the expected number of times users in a Twitter discussion stream heard about a message that was spread (Bramoullé & Genicot, 2018). To further extrapolate, $DC(t_i)$ is a measure of the level of *spreading influence* that a Twitter user has with respect to the overall discussion stream (Kang, et al., 2012). A user whose tweet had a $DC(t_i)$ score of 2.44 would not have their message received and further retweeted as often as a user who had a $DC(t_i)$ score of 63.4. In short, the tweet in Table 4 whose $DC(t_i)$ is 63.4 would have a more substantial influence over the discussion stream because his tweet would be seen more times and retweeted more frequently than others.

Calculating E_s Scores from the Tweet Text Attribute

The third attribute of the dataset was Shannon Entropy (E_s). Mathematically speaking, E_s is the discrete probability distribution of Twitter text to measure the *uncertainty* or *randomness* of the data by analyzing its complexity (Wang, et al., 2018).

In several studies, in addition to uncertainty and randomness, E_s has been adapted to measure *surprise* and *diversity* (Ghosh, R., et al., 2011; Vajapeyam, 2014). In the context of this research, E_s was used to measure the average level of user diversity that existed in an aggregation of tweets (Ghosh, et al., 2011; Hasan, et al., 2016).

The Shannon Entropy of a set of aggregated tweets could be calculated using the formula $E_s = -\sum_{i=1}^{N} d_i \log(d_i)$, where *d* data points are sorted into *N* groups based on the time index. The data points *d* in this calculation refer to the individual words that make up the text of a tweet sent by a user (Vajapeyam, 2014). The individual words of a tweet are the fundamental units of communication between sender and receiver according to Shannon's Information Theory (Caballero, et al., 2017). In our research, d_i was a probability variable which was concerned with word frequency in Twitter text. In this context, word frequency in communication related to the minimum amount of words necessary to retain the integrity of information (Vajapeyam, 2014). The d_i variable was calculated after we aggregated all tweets according to the time index of their posting. The aggregated user text that was sorted by time index was then evaluated for E_s, which told us the level of diversity in the users who contributed them.

Case Study: Calculating the E_s scores for the "#DowJones" Dataset

The text attribute from the original dataset of tweets was used to calculate the E_s scores for the *#DowJones* dataset. Our example dataset covered a time-period of 19 hours. All of the tweets for each hour index in the dataset were aggregated and then an E_s score was evaluated for all of the tweets that were posted during that minute index. For example, all tweets which were posted at hour 3 were grouped together and evaluated

to a collective E_s score of 1.604. An example of the *#DowJones* dataset with three of the four required attributes can be seen in Table 5 below.

Table 5

#DowJones dataset with t_i, hour, diffusion, and entropy columns

t_i	hour	diffusion	entropy
t_{1723}	3	2.44	1.604
t_{1724}	3	1.8	1.604
t_{1725}	3	15.12	1.604
t_{1726}	3	41.36	1.604
t_{1727}	4	0.28	1.5244
t_{1728}	4	1.12	1.5244
t_{1729}	4	63.4	1.5244
t_{1730}	4	1.52	1.5244

E_s is essentially the measure of *disorder* in a system (Li, et al., 2015). The E_s metric has been adapted for use in several studies of SNA to measure the amount of *diversity* in a system (Matei, et al., 2015). It is within this context that we used the E_s metric. To this end our goal was to evaluate the level of diversity that existed in Twitter text during a particular hour time index from a dataset (Ghosh, et al., 2011). The literature views the E_s measure in this domain as the level of *collaboration* in a system (Matei, et al., 2015). For our research purposes, diversity was the result of *collaboration* of many users who were *collaborating* in a discussion. Collaboration was evaluated by analyzing the number of participants and the shares of their participation in a Twitter discussion stream. By analyzing the collaboration of users in the discussion stream we derived the *diversity* that existed within the discussion stream (Ghosh, et al., 2011; Matei, et al., 2015). We obtained the collaboration (and consequently the diversity) of the

discussion stream by calculating the average E_s for the aggregated tweets from the discussion stream for a single hour index. In short, the higher the E_s score, the higher the *diversity* which meant that more users were collaborating in a discussion (Matei, et al., 2015).

The E_s score for diversity is a number that ranges from zero to approximately 3.5, which suggests a high level of diversity in the system being measured. A score which ranges from approximately 1 to around 2.5 suggests a diversity that is very low to average. A score of $2.5 < E_s < 2.9$ suggests an above average level of diversity. A score of $E_s >= 3$ suggests a high level of diversity (Ifo, et al., 2016). Based on empirical testing, a score of $E_s >= 4$ has occurred but is rare. The first E_s score in Table 5 is 1.604. This score suggests a low average diversity for the aggregated tweets at the third hour of the *#DowJones* dataset. The smaller average diversity score for the third hour suggests that a smaller group of people posted a larger number of tweets for the time index.

There is one shortcoming to the E_s attribute that should be briefly discussed to provide more clarity as to the capabilities and limitations of the metric. According to several studies in the literature, E_s is used to measure diversity in natural language (Papadimitriou, et al., 2010). The preferred application of this diversity measurement is at the macro level. To this end, E_s is frequently used in closed communities (such as a Twitter discussion stream sample) to measure the amount of participation contributed by users in a larger community. The E_s metric does not have the ability to evaluate language diversity at the micro level (Bentz, et al., 2017; Kalimeri, et al., 2012). Specifically, E_s can measure diversity in textual language, but it is not capable of providing a nuanced evaluation of language based on word order. Two tweets that use the

same words but have different word orders (thus potentially different contexts) will have the same E_s score (Bentz, et al., 2017). This shortcoming excludes E_s from being used in studies which require a detailed evaluation of text at the micro level.

Integrating the $DC(t_i)$ and E_s Variables into a Single Variable

There were many examples in the research literature demonstrating the creation of new variables from existing ones. Díez-Pastor, et al. (2015) used ensembles of variables from machine learning datasets to solve a problem of class imbalance, which is a problem that arises when the proportions of one variable to another are skewed (Díez-Pastor, et al., 2015). Davis and F. Abdurazokzoda (2016) aggregated several different socio-linguistic categories such as population and cultural traits into an individual variable. The intention of the aggregated data was to summarize and bring together many cross-domain elements into a single variable for study (Davis & Abdurazokzoda, 2016). Randall, et al. (2014) aggregated several aspects of a patient's personal information to provide patient cross record linkage across many different distributed medical datasets (Randall, et al., 2014). In all these studies new variables were created using combinations of data aggregation and ratios (Du Jardin, 2010).

The reasons for creating a new variable ranged from solving classification problems, to combining summarized information, to providing multi-variable linkage. In the case of our research, we needed to implement elements of information summary and multi-variable linkage to create a new variable for our study (Davis & Abdurazokzoda, 2016; Randall, et al., 2014). To this end, we combined elements of information spread with a quantitative metric of diversity into a single variable to measure events on Twitter (Ghosh, et al., 2011; Kang, et al., 2012). Next, we further discuss the ensemble of

variables we used. The dataset at this point had three variables, which were time index p, $DC(t_i)$, and E_s**.** Based on a review of the literature (Du Jardin, 2010), the method we decided to integrate the $DC(t_i)$ and E_s variables into a single ratio. The quotient of the ratio of the two variables resulted in a numeric value we called *newsworthiness*, represented by NW. NW is the ratio of $DC(t_i)$: E_s, expressed mathematically as $NW = \frac{DC(t_i)}{E_s}$. It was a metric of user activity that quantified the distribution of user message spreading actions over the user diversity in a discussion stream.

The E_s attribute was a value that frequently fell in the range of $0 < E_s < 3$ (Ghosh, et al., 2011; Hasan, et al., 2016). Based on several test datasets, a sample that contained Twitter text where $E_s > 4$ was anomalous but did occur. Based on our sample datasets, the $DC(t_i)$ attribute was a value that ranged from $0 < DC(t_i) < \infty$. Low E_s scores can result in larger ranges of *NW*. Although, based on the guidance from the literature this appeared to be antithetical to logic, several studies suggested a valid reason for this occurrence. One reason for the disproportion between low E_s and high *NW* was a systematic repetition of a message by a discrete cluster of users in a discussion stream (Gurajala, et al., 2016). If a small number of messages was repeated at a high frequency, the contribution to a discussion resulted in a corpus of tweets void of diversity in content. The dissonance between a small cluster of users and high *NW* could be attributed to spam bots, whose primary objective was message amplification (Gurajala, et al., 2015).

During the integration of E_s with $DC(t_i)$, the E_s attribute was placed in the numerator of the *NW* equation to avoid results that evaluate to very small numbers. To that end, we placed the E_s attribute in the denominator of our equation, resulting in graphs of the *NW* distribution that clearly articulated peaks and valleys that were

consistent with *events*. The *NW* attribute took the information content of Twitter message text and the architecture of connected users as input and provided a metric to evaluate the levels of human activity (Ghosh, et al., 2011; Kang, et al., 2012). It was where the levels of *NW* were highest in a discussion stream those events were identified. The creation of the *NW* metric can be seen in the algorithm below.

Event Detection 1 *NW* Peak Identification

Function *integrateAttributes*
Input: $DC(t_i)$, E_s
Output: *NW*
FOR all rows in $DC(t_i)$
Divide by corresponding rows in E_s
ENDFOR
EndFunction
Function *calculate* Q_3
Input: *NW*
Output: Q_3
Sort *NW* in ascending order
Calculate 0.75 percentile of sorted *NW*
EndFunction
Function *createSmoothedLineGraph*
Input: *NW*, hour
Output: Smoothed linear curve
FOR all rows in *hour*
AND all rows in *NW*
Plot *hour* on x-axis
Plot *NW* on y-axis
ENDFOR
EndFunction
Function *detectEventPeak*
Input: *NW* smoothed line curve
Output: Event_Peak
Event_Peak = Ø
IF *NW* smoothed line curve > Q_3
ANDIF
IN smoothed line curve{Apex_points} THEN

Apex_points ⊆ Event_Peak
ENDIF

Case Study: Calculating the NW attribute for the #DowJones dataset

We calculated the *NW* attribute values in the *#DowJones* dataset using the algorithm presented in the previous section. For every row of the dataset, the $DC(t_i)$ value was divided by the E_s value. An example of the *#DowJones* dataset with the *NW* attribute is seen in Table 6.

Table 6

#DowJones dataset with t_i, $DC(t_i)$, E_s, and *NW* columns

t_i	hour	diffusion	entropy	newsworthiness
t_{2589}	5	24.72	1.5441	16.00932582
t_{2590}	5	33.48	1.5441	21.68253351
t_{2591}	5	24.72	1.5441	16.00932582
t_{2592}	5	24.72	1.5441	16.00932582
t_{2593}	6	0.4	1.5213	0.262933018
t_{2594}	6	0.4	1.5213	0.262933018
t_{2595}	6	24.36	1.5213	16.01262078

In the first row of Table 6 the NW value is 16.00932582. This value was obtained by evaluating the ratio of $DC(t_i)$ for the tweet in question over the average E_s for the hour in which the tweet was posted. The higher *NW* value in this tweet suggests that the combined user spreading activity and average diversity level of the discussion stream is somewhat higher for this tweet than the others. To further extrapolate on this point, the tweet with *NW* 16.00932582 suggests that the discussion stream with which it is associated is higher in messaging activity and has more users. In other words, the discussion stream at the time this tweet was posted was more active overall.

Instrument Development and Validation

Quartiles as a Metric

In our study we looked at a number of metrics to identify one metric to evaluate the *NW* values which would suit our research needs. Our evaluation of metrics shifted from machine learning classifiers to quartile ranges. After further review and initial testing, the latter seemed to be a better fit. We initially searched the research literature to identify an appropriate metric with which to evaluate the *NW* distribution with regards to event detection. We attempted a technique that implemented the use of a threshold line (Aminikhanghahi & Cook, 2017), the location of which was calculated using machine learning classifiers such as Artificial Neural Network, XGBoost, and SVM (Lokeswari & Rao, 2016; Stamp, 2018). Preliminary experimental results were less than satisfactory and not sufficient for us to justify pursuing this approach to identify events. We made the decision to pursue the use of machine learning classifiers to identify an event threshold in future research. In lieu of machine learning, we decided to pursue an approach that instead focused on statistical exploration of the *NW* distribution spread itself (Rosenthal, et al., 2019). This decision to emphasize statistical exploration led us to the use of quartiles which will be discussed next.

Quartile Analysis refers to a statistical method of data exploration in which a data attribute is split into four equal groups after its distribution is placed into ascending order. Each of the four subcomponents is called a *quartile* (Shih & Liu, 2016). The three points that divide the distribution into quartiles are denoted by the variables Q_1, Q_2, and Q_3 (Langford, E., 2006). The values of each variable in reference to the subdivided distribution are Q_1 = 25%, Q_2 = 50%, and Q_3 = 75% (Shih & Liu, 2016). To illustrate the

use of quartiles, consider the following example. There is a dataset, for example {1,2,3,4,5,6,7,8,9,10,11,12}. First, the dataset is placed in ascending order and then it is split into two halves. These are the *upper half* and the *lower half*, for example: lower half {1,2,3,4,5,6} and upper half {7,8,9,10,11,12}. If the dataset does not split evenly into two subsets, the median value is included in both the upper half and the lower half (Langford, E., 2006). The lower half and upper half are then further subdivided into halves resulting in *four subgroups* total, each called a *quartile*, {1,2,3|4,5,6|7,8,9|10,11,12}. The three lines that separate the four quartiles in this dataset are our values of Q_1, Q_2, and Q_3. The three values serve as regional "fences" which partition the dataset into functional regions of equal spread (Domínguez, et al., 2017). Q_1 is the 25th percentile, which for this dataset is Q_1 = 3.25. Q_2 is the 50th percentile, which is Q_2 = 6.5. Q_3 is the 75th percentile, which is 9.75 (Langford, 2006; Shih & Liu, 2016).

In many studies in the research literature, the statistical region of Q_1 to Q_3 is referred to as the interquartile range (IQR). In our example, the IQR includes the set {4,5,6,7,8,9}. The values in the dataset that fall to the left of Q_1 and to the right of Q_3 are where outliers are found (Rousseeuw & Hubert, 2011). There are many studies in the literature which focus on the IQR for data evaluation due to its isolation from outlier influence (El Asri, et al., 2019; Tommasel, et al., 2016). Other studies focus on the areas outside of the IQR because the emphasis of the research is on outlier detection (Lee & Sumiya, 2010; Pozdnoukhov & Kaiser, 2011). Our research fell into this latter category. For our study, we were concerned with tweets that fell within the area above the Q_3 fence in particular. For this reason, we chose the upper quartile outside of the IQR for our

testing emphasis. In the next section, we will further discuss the use of the upper quartile region that was adapted for our research.

Use of Quartiles in Our Approach

Quartiles are an evaluative method that we found in several SNA studies where the research required an established range of values representing a "normal" to "abnormal" range of data points from a dataset (Lee & Sumiya, 2010). Pozdnoukhov and Kaiser (2011) used the ranges delimited by Q_1, Q_3, and IQR to quantify the level of *normality* of crowd behaviors using both long and short-term time-series datasets that were collected from Twitter (Pozdnoukhov & Kaiser, 2011). Lee and Sumiya (2010) used the Q_1 and Q_3 calculated values to define a range of *usuality* in a dataset of geo-tagged tweets to detect the regularity of geographical events in a discussion stream. Outliers from the upper half of the dataset are identified as unusual tweets (Lee & Sumiya, 2010).

Quartiles have limitations when it comes to evaluating distributions of data. They do not consider the data as a whole, since it is examined in fragmented ranges (i.e. *upper half* and *lower half*). The values of upper and lower quartiles are susceptible to the effects of outliers, which makes them often susceptible to undesirable influence from variance. In the case of our research, however, outliers help to identify events since they are sporadic occurrences of data that fall outside of an established "normal" range. We hope to improve upon any limitations imposed by quartiles in future research when we include machine learning classifiers and accuracy metrics in our methodology. We implemented the use of quartiles in our research because it allowed us to split the *NW* distribution into two measurable ranges (upper half and lower half) that were delimited by the fences of

Q_1 and Q_3. The "fenced" areas (for our research, the area above Q_3) allowed us to evaluate the levels of user activity as *normal* and *abnormal* (Lee & Sumiya, 2010; Rousseeuw & Hubert, 2011). The lower half (from the minimum value to the distribution median) represented the *normal* range of user *NW* activity for our study. Q_1 served as the *NW lower fence* in our measurement, however we were not concerned with tweets that fell below this threshold. In our research, events would be found in the *upper half* of the distribution (from the median to the maximum value) above the Q_3 fence. The Q_1 lower fence was kept in our measurement model in order to maintain the integrity of the technique as it was described in the literature. The following section demonstrates the conversion of the *NW* attribute into a fenced-off measurement model using Q_1 and Q_3 as fences which define and delimit our area of interest (Joarder & Firozzaman, 2001; Rousseeuw & Hubert, 2011).

Our decision to use a single value to act as a threshold throughout our time-period to identify events was supported by several studies (Nairac, et al., 1997; Weng & Lee, 2011). Events were identified as patterns that occur within a specified time domain (Weng & Lee, 2011). In order to properly identify irregular patterns that occurred in a temporal trajectory, ranges of *normal* and *abnormal* values had to be established (Nairac, et al., 1997). The use of a *threshold* value was a technique used in many anomaly identification studies known as *novelty detection*. Its purpose was to establish a boundary between normal and abnormal data (Nairac, et al., 1997; Pimentel, et al., 2014). In the context of our research, our Q_3 fence was our *novelty detection* threshold. It was a statistically calculated boundary that separated normal levels of *NW* from outliers which

could form peaks identifying as events. The threshold was uniformly implemented throughout our time-period (Pimentel, et al., 2014).

Case Study: Calculating the Q_1 and Q_3 Fences for the NW Distribution

To calculate the Q_1 and Q_3 boundary lines from the *#DowJones* dataset we ordered all the values from the *NW* attribute distribution from lowest to highest. Next, we split the ordered range at the median of the entire dataset which was 5.39. In the lower half of the *#DowJones* dataset, we found the median of the range which was 0.68 which was our Q_1 fence. Next, we found the median value of the upper half of the *#DowJones* dataset which was 20.8. This value was our Q_3 boundary fence. The Q_1 fence (lower-half median) did not exceed a value of one which suggested that the range of *NW* values in the lower-half of the dataset were all rather small. Q_2, which was the median of our complete *#DowJones* dataset, had a value of 5.39. Since Q_3 was our upper fence boundary, tweets with a score of greater than 20.8 *NW* fell within the *outliers* region of the dataset. As it was mentioned in the section above, even though tweets that fell below the Q_1 fence were considered outliers, for our research they would be considered part of the *normal* range of *NW* activity.

Sample Used

The following section discusses our method for collecting data for this study. Our study used an application programming interface to gather raw tweets that were continuously contributed by users as part of a global discussion stream that was composed of a torrent of dynamically changing and frequently trending topics. The section also discusses our dataset size and unit index of time for the study.

The Twitter API

The Twitter application programming interface (API) is a public-facing layer of the social media platform (https://developer.twitter.com/en/docs). The API allows developers, analysts, and statisticians to collect messages from the platform without having to use unethical methods of data gathering such as scraping, which seek to copy published data from Twitter directly from the browser to an alternate location. The API provides a direct connection to the Twitter microblogging service which can be leveraged programmatically using a development environment such as R-Studio or Python. When considering data collection for our research, the size of the dataset needed to be determined by the scope and duration of the study (Perera, et al., 2010). In the next section, we discuss the duration, size, and subject matter in the datasets used for our study.

Size, Time Period, and Topics for the Two Datasets

In our research, we identified events by studying how topics trended in a discussion stream within a 14-day time window. Some studies used a time-period which lasted several months and used hundreds of thousands of tweets in an individual dataset. Originally, we had committed to using two prior topics: "cybersecurity" and "#DowJones." We queried the Twitter API using both sets of keywords during different sessions and discovered that neither topic produced a sufficient quantity of tweets for a viable dataset. This suggested to us that neither topic had characteristics that were causing them to trend in a discussion stream. To serve as our alternates, the first dataset we collected was done using the keywords "Tulsa+Rally". The second dataset we collected was performed using the keywords "Atlanta+Protests" as search criteria. Each dataset had 14 one-day units per dataset. The topic of each of the two datasets was chosen

by performing a search of current events in the media and performing API sampling to determine if the topic was trending. According to the literature, the life span of an average Twitter topic was quite volatile. A topic could remain relevant in a discussion stream for a little as a day to as long as a month. There was no strict guideline in the literature that dictated a size for a dataset collected from an API. In general, if the duration of a study lasted a significant amount of time (like two months, for example), the size was expected to be larger. Since our datasets covered a period of 14 days, we made use of a dataset size of approximately 14,000 tweets per dataset. The "Atlanta+Protests" dataset contained approximately 12,000 tweets. In the following section, we demonstrated our approach using a sample dataset collected from the API.

Data Analysis

Comparing Our Approach to a Popular Existing Approach

Our methodology for detecting events was compared to an approach that was commonly found in the literature (Figueiredo & Jorge, 2019; Guo, et al., 2017; Wang, et al., 2012) which used keywords and term frequencies measured over a period of time. The approach used an unsupervised classification algorithm called Latent Dirichlet Allocation (LDA). LDA is a method of deconstructing a sample of text into its constituent topics (Lansley & Longley, 2016). To accomplish the deconstruction, the algorithm uses a combination of text mining, probability, and word clustering. Words are clustered into relevant groups based on probability of appearance in the text. Text mining tools are used to prepare the text by removing unnecessary words, characters, and whitespace (Karl, et al., 2015). The result of the text pre-preparation is a data structure called a Document Term Matrix (DTM).

The DTM is a two-dimensional matrix of zeros and ones which documents and enumerates the occurrence of words in a sample of text (Figueiredo & Jorge, 2019). Mathematical algorithms such as LDA use the DTM as the foundation to perform text mining analysis tasks such as classification and clustering (Karl, et al., 2015). For our research, we used the DTM as the basis for mathematical comparison against our method of event detection. The DTM provided us with a frequency distribution which was then translated into word counts for each time index over a time-period.

In the remaining sections of this chapter, we discuss how the DTM of the *#DowJones* dataset was prepared and then converted into quartiles using the same techniques that were used in our approach. First, the method of converting the dataset to a DTM is discussed. Second, leveraging the DTM for a word frequency chart is detailed. Third, the word frequency information from the DTM is converted to a variable called *keyword_frequency*, whose distribution is evaluated mathematically to calculate the fence line boundaries of Q_1 and Q_3 to identify outliers. Finally, the *keyword_frequency* trajectory is plotted in a graph with the hour attribute on the x-axis and the *word count* attribute on the y-axis. The Q_1 and Q_3 fence lines were placed in the graph to delineate normal word count ranges and identify any *event peaks* that occurred above the Q_3 fence (Lee, R. & K. Sumiya, 2010). The chapter ends with a discussion about comparing event detection results of our method and an existing approach. Since the formal process of determining the fence values of Q_1 and Q_3 was explained earlier in this document (Langford, 2006), we used the "#DowJones" case study example dataset to demonstrate the conversion process for the existing approach to event detection.

Word Frequency Chart

In the literature, there are several studies which use single-word (unigram) search approaches to event detection (Choudhury & Alani, 2014; Di Eugenio, et al., 2013). To recreate the unigram single word approach, we created a graph that displayed the frequency of word occurrences in Twitter text based on data retrieved from the DTM (Welbers, et al., 2017). The frequency graph provided the key words that were used most frequently in the Twitter dataset which we used to create a *keyword_frequency* attribute, which subsequently were converted to quartiles for event detection.

Creating the keyword_frequency Attribute from the Word Frequency Chart

There was no available clear guidance in the literature on the issue of choosing an appropriate number of keywords from a frequency graph to predict events. Based on empirical evidence gathered from experimenting with sample data, we compiled a list of approximately 25 keywords for a regular expression search through the original tweet text. The regular expression search results were used to create a word count variable as part of the input for our smoothed line trajectory. The number of times each keyword occurred in a row of tweet text was counted as a numerical value for each tweet.

Case Study: Convert keyword_frequency Attribute Distribution into Q_1 and Q_3

In the previous section, in our approach we used the *NW* attribute of the *#DowJones* dataset to create the calculated values of Q_1 and Q_3 to function as fence boundaries to isolate average activity from event peaks (Lee & Sumiya, 2010; Rousseeuw & Hubert, 2011). To this end, we also used Q_1 and Q_3 to function as fence boundaries in the existing approach to isolate routine data points from event peaks. For the existing approach we used the *keyword_frequency* attribute to calculate our Q_1 and Q_3 fences. The method of event identification for the *existing approach* was the same

method that we used in our approach. We looked for event peaks that formed above the Q_3 boundary fence after Q_1 and Q_3 had been calculated and placed in the distribution graph (Domínguez, et al., 2017; Subramani & Kumarapandiyan, 2012).

To calculate Q_1 and Q_3 for the *keyword_frequency* attribute, the distribution had to first be placed in ascending order (Vega, et al., 1998). The median of the distribution was identified and then the distribution was split in half. In the lower half of the split distribution, the median was identified. The median value of the lower half of the *keyword_frequency* distribution was the value for Q_1. In the upper half of the *keyword_frequency* distribution the median value was identified. The median value of the upper half was the value of our Q_3 fence (Vega, et al., 1998). The calculated values of Q_1 and Q_3 were placed in our *keyword_frequency* distribution graph as our boundary fences (Joarder & Firozzaman, 2001; Lee & Sumiya, 2010). Just as we did in our approach, we used Q_1 and Q_3 to isolate average data points from event peaks (Lee & Sumiya, 2010; Rousseeuw & Hubert, 2011). In the smoothed histogram graph shown in Figure 4, the word count distribution has a minimum value of zero and a maximum value of five. The value of Q_1 for the distribution is one. The median value for the entire distribution is three and the value of Q_3 is four. When the *keyword_frequency* distribution was plotted, the upper boundary fence for the graph was four.

Table 7

Sample of the keyword counts listed with the hour of their occurrence

hour	keyword	frequency
19	dowjones	1103
20	dowjones	912
20	trump	911
20	money	750
19	trump	678
2	dowjones	637
20	stock	445
20	donald	407
20	ethanjsomers	399
3	dowjones	380
20	cdc	370

The *keyword_frequency* attribute in Table 7 displays a sample of the count of the keywords that are found in the original tweets from the *#DowJones* dataset. The tweets are sorted by the hour index in which they were posted in the discussion stream. The keywords evaluated in the count are based on frequencies that are taken from the *#DowJones* DTM (seen in Table 8) frequency table. As an example, the first row of Table 7 shows that at hour 2 the text corresponding to the tweet in this row has a count of five keywords that are mentioned by the user.

Comparing the Two Event Detection Methods

The peak and valley formations that occur beyond the Q_3 fence served as an objective metric by which our method and the existing approach were evaluated for identifying events (Lee & Sumiya, 2010; Pozdnoukhov & Kaiser, 2011). In Chapter 1, we defined an event as *a set of tweets on a related topic within a certain period that surpass a threshold defined by statistical measures of diffusion and entropy.* This definition was expanded in our approach to include the identification of events based on

the occurrence of peaks and plateaus above the Q_3 boundary. The identification of peaks and plateaus applied to both our method and the existing method of event detection. In our discussion, we referred to the occurrence of *both* peaks and plateaus above the Q_3 fence as "event peaks," as event peaks are harbingers of events. The number of well-defined event peaks during the time-period covered determined which of the two methods performed event detection more effectively (Lee & Sumiya, 2010).

As it was mentioned above, *event peaks* were representative of *events* in a timeline, whether the trajectory uses *NW* or word frequency to measure the events. What the event peaks did not concretely quantify was the *magnitude* of the event (Lee & Sumiya, 2010). A peak might form just below the Q_3 boundary fence. Such a peak would be disqualified as an event, but it did not discount the possibility that the peak represented an increase in activity. The quantifying of event magnitude is the subject for later research. Our research focused only on the existence or occurrence of events in a dataset timeline.

Based on the definition of event detection mentioned above, the "statistical measures of diffusion and entropy" that were used in our method were implemented using a metric that we created through the integration of *diffusion* and *entropy* into a single attribute called *Newsworthiness* (*NW*). By performing this integration, a single attribute could identify events by evaluating the ratio of message sharing activities to user diversity over a time-period. The *threshold* from our definition was the fence defined by the value of Q_3, which was located at the median of the *upper half* of the *NW* attribute distribution. The threshold for the existing approach was the Q_3 fence calculated from the distribution for the "keyword_frequency" attribute (Langford, 2006; Rousseeuw &

Hubert, 2011). Tweets which fell above the Q_3 value for both methods were *outliers* (Domínguez, et al., 2017). Event peaks are intrinsically tied to outliers since they refer to groups of data points that occur *outside* of an established "normal" range (i.e. Q_1 and Q_3) (Zubiaga, et al., 2012).

One of the shortcomings of the quartile evaluation metric was that it used ranges and did not provide a single accuracy metric (like a confusion matrix) (Langford, 2006; Rousseeuw & Hubert, 2011). To this end, we discuss two objective methods to identify *events* that have been used in several highly cited studies. These methods were the number of well-defined *event peaks* (Kolchyna, et al., 2016; Yu & Wang, 2015) and temporal bursts (Lappas, et al., 2012). These evaluative techniques could be used to identify events in both our approach and the existing method. These techniques were not as desirable as quantitative methods of evaluation, but they objectively and effectively detected events by visually identifying recurring spatiotemporal patterns throughout the duration of the time index. We briefly discuss these two objective methods and how they related to our contribution to the field of SNA research. The second method we will leave as an open option for later research as it is more complicated, involves more resources, and requires further review of the literature. For now, we simply included it as part of the discussion for our current research.

The first objective event detection method was to evaluate the number of well-defined peaks created by outliers beyond the Q_3 fence in the upper half of the *NW* distribution. (Earle, et al., 2012; Yu & Wang, 2015). Peaks define elevated user activity (*NW*) and increased frequency of keywords (*keyword_frequency*). The method which produces more defined peaks (*NW* vs. *keyword_frequency*) in the upper half (beyond the

Q_3 fence) better identifies the occurrence of events. This evaluative approach was not as concrete as a confusion matrix, however our method mathematically defined a fence between "normal" and "abnormal" activity effectively using the upper half Q_3 value to delineate the separation between normal Twitter activity and abnormal activity (Lee & Sumiya, 2010).

The second objective method for detecting events beyond the Q_3 fence was the occurrence of temporal bursts (Lappas, et al., 2012). A temporal burst (TB) is the occurrence of an unusually high frequency of hashtag usage in a discussion stream during a specific timeframe. A TB has a life cycle of three basic phases which occur at specific points in its timeline. The first phase is the initial *growth* or onset of the TB where the height of the *NW* trajectory line begins a period of increased elevation. This onset is followed by a *peak*, which occurs when the height of the *NW* trajectory line halts its upward motion. This halted upward motion can result in either a well-defined, rounded peak or an extended flat surface called a plateau. The final phase of the TB is a *relaxation* of the *NW* intensity (Kolchyna, et al., 2016). Relaxation occurs when the *NW* trajectory line starts a descent after halting at its highest point. The principal benefit of using a TB as an objective measure is that the duration and periodicity (repeated cycles) of the burst can be evaluated with more scrutiny, since there are three phases which define it (growth, peak, relaxation).

A TB pattern for user activity (*NW*) and keyword frequency (*keyword_frequency*) can demonstrate important *signatures* for research. A *signature,* with respect to graphs, is a unique or distinguishing pattern or frequency of peaks and valleys in a linear time-series trajectory that facilitates the identification and characterization of a phenomenon of

interest (Conte, et al., 2004). A TB can last for several hours, remaining at its highest point for long periods (forming a plateau), before its zenith weakens and declines (Abdelhaq, et al., 2013; Ratkiewicz, et al., 2010). Several burst instances could also occur in tandem, forming a recurring pattern of peaks and valleys (Kolchyna, et al., 2016; Lappas, et al., 2012). TB signatures such as these help to describe what kind of event is occurring in a discussion stream (Abdelhaq, et al., 2013). The approach that better defines the shape and patterns that are inherent in a dataset, better identifies events (Ratkiewicz, et al., 2010). For our research, we focused on the number of well-defined peaks in the time index (Dou, et al., 2012; Zhang, et al., 2015). Our novel event detection approach contributed to the body of SNA research by combining measurable user messaging behavior with spatiotemporal patterns evaluated using a mathematically calculated fence (Q_3) isolating normal from abnormal activity (Langford, 2006; Shih & Liu, 2016). In furtherance of our contribution, we developed a new attribute (*NW*) to evaluate user activity by integrating two existing SNA attributes (Shannon Entropy and Diffusion Centrality) (Ghosh, et al., 2011; Kang, et al., 2012).

Formats for Presenting Results

In this section we discuss our chosen method of graphically displaying the trajectory of the *NW* data throughout its time period. The efficacy of our method was demonstrated using empirical data in our case study. Through many preliminary experiments, we found that the most efficient way of displaying the NW trajectory was using a combination of a smoothed line graph with points. This manner of graphing the data aided in the detection of events by articulating peak formations in the linear

trajectory where events occur. The points in the graph helped to explain why the smoothed line ascended or descended at certain time indices.

Case Study: Plotting the NW Distribution Over the Dataset Time Index

The *#DowJones* hashtag dataset had a total time index of 19 hours. The *NW* attribute distribution that was created in the previous section was plotted in a smoothed line graph with the *hour* attribute on the x-axis and *NW* on the y-axis. The range of *NW* values below the Q_3 fence line represented the *normal* Twitter message spreading habits for the overall time index. The linear trajectory of the graph ranged from 4 *NW* at its first hour and reaches a maximum value of approximately 33 *NW* at hour 9.5. The lowest point of the trajectory was 3 NW at 17.5 hours where the trajectory remained unchanged until the end of the time index. The graph in Figure 2 showed both tweets as data points and a smoothed linear curve. The smoothed curve was created using a moving average of the data points. There was a large number of points at or below the Q_1 fence line at time indices one through seven and again at fifteen through nineteen. At time indices one through fourteen there are also several data points with higher *NW* scores. Due to the wide spread of values in the data points at these time indices, the moving average of the smoothed line curve remained above the Q_1 fence value and does not go below it. Event peaks in the graph are discussed next.

What was of interest to us in our research were data points in the *NW* trajectory that ascended above the Q_3 fence, formed a rounded or plateaued summit, and then descended. Summit formation could occur once or several times during the progression of the *NW* trajectory. The summits which formed above the Q_3 fence (which in this sample graph was 20.8 *NW*) were *event peaks*. An *event peak* is a cresting formation of

NW above the Q_3 boundary fence that results from a crescendo of user-related messaging activity in a diverse social media discussion stream. In the graph in Figure 2 the trajectory showed two *event peaks* which formed above the Q_3 fence line. The first peak developed at half past the 9th hour in the time index. The second event peak formed at half past the 12th hour. At hour 13 the *NW* trajectory began its second descent and crossed below the Q_3 fence line at hour 14.

We could read the *NW* graph in the following manner. From hour zero to the eighth hour, the user messaging behavior was within normal, average ranges. This was evident from the *NW* trajectory during this time period. From halfway past the ninth hour to the eleventh hour there was a surge in *NW* which suggested that the users in the discussion stream had significantly increased their participation in the discussion. This surge and peak in *NW* were interpreted as an *event* on our graph. A second *event* immediately followed the first. The two events were separated by a shallow valley that formed after the eleventh hour.

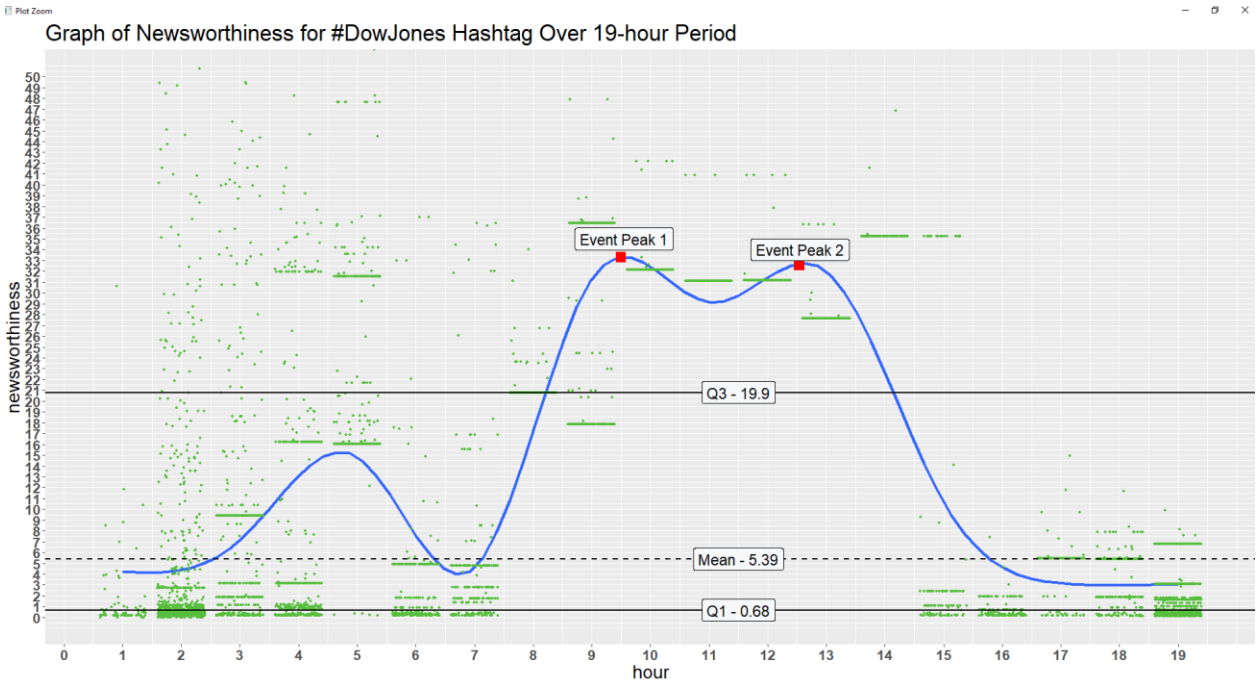


*Figure 2. T*rajectory *for the NW distribution over the 19-hour period*

Case Study: Plotting the the keyword_frequency Distribution.

As seen in Figure 3, the trajectory for the *keyword_frequency* attribute ranged from a minimum of two keywords per tweet to a maximum of just below four keywords per tweet. The smoothed curve averaged the number of tweets per hour index in its linear graph representation. The crest of the trajectory's apex was almost tangent to the Q_3 fence, falling just before the boundary line. The trajectory line descended between the range of two and three keywords per tweet for average frequency. When we used the Q_1 and Q_3 lines as boundary fences with the existing method no events were identified since the apex of the *keyword_frequency* attribute fell just below the Q_3 line. However, the trajectory clearly demonstrated an articulated peak at hour 9 of the time series which could be identified as an *event*. The data from Figure 3 suggested that Q_1 and Q_3 boundary fences were not an efficient objective method for identifying events using a smoothed line trajectory for word frequency distribution. Contrarily, based on empirical evidence, the Q_1 and Q_3 approach did appear to be well-suited to identifying events using the *NW* attribute and a smoothed line trajectory. The statistical evaluative method did not have the same uniform level of efficacy for both attributes.

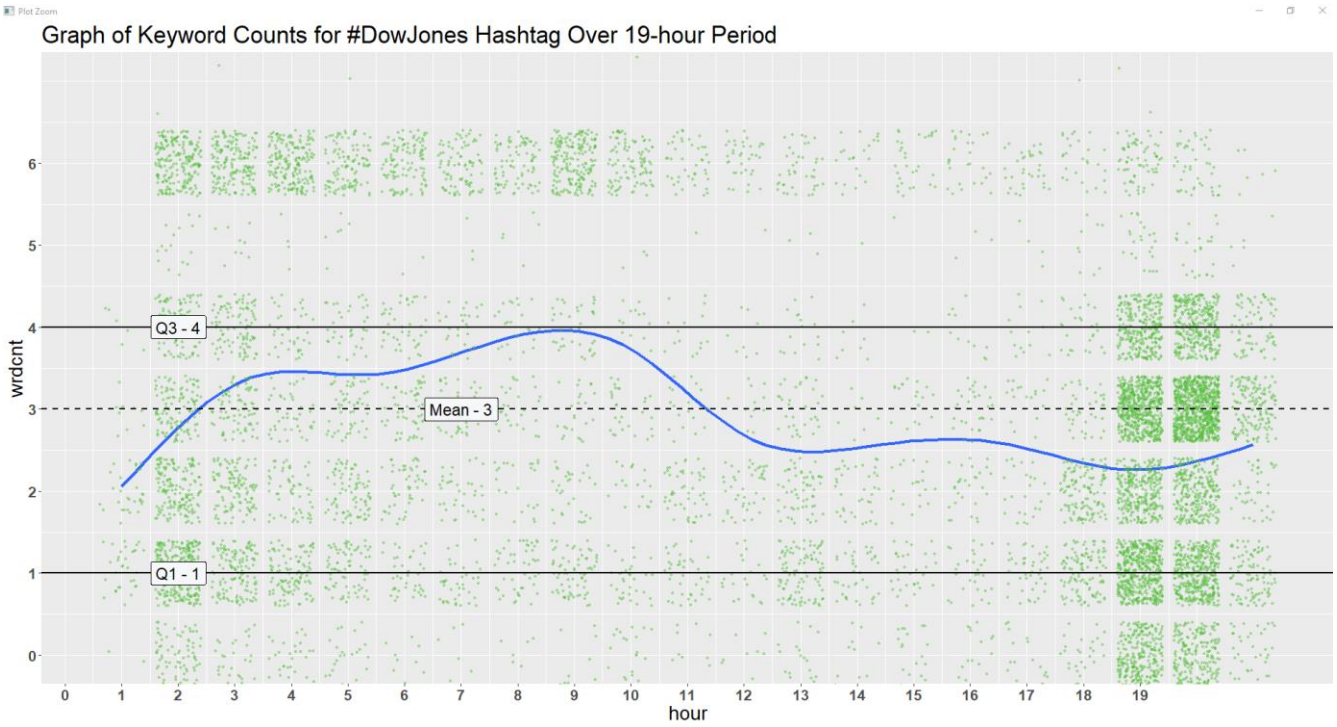


Figure 3. Keyword time-series frequency graph for *wrdcnt*

Resource Requirements

The following resources were used to perform Social Network Analysis computations of $DC(t_i)$ and E_s using two datasets to calculate *NW*. The derived values of *NW* were plotted in two dimensions using a graphing library for analysis.

1. Toshiba Qosmio X70-A laptop computer with the following configurations.
   (a) Processor: Intel Core i7-4700 CPU @ 2.4GHz 2.40
   (b) Hard Disk: 1 TB Disk Drive
   (c) 256 GB Solid State Drive
   (d) Operating System: Windows 10, 64-bit
2. R-Studio Version 1.1.383
3. R i386 3.4.3
4. R libraries used for Natural Language Processing and classification:

(a) library(twitteR)

(b) library(ROAuth)

(c) library(centiserve)

(d) library(igraph)

(e) library(tidytext)

(f) library(ggplot2)

Summary

Research Plan

The research in this document focused on the use of the newly created *NW* attribute to identify event peaks that occur over a specified time index. The approach made use of several techniques and metrics used in common SNA research found in the literature. We plan to pursue four different applications of our approach for future research. These applications include collection of datasets, evaluation of our *NW* metric with machine learning algorithms, sentiment analysis (SA) scores as an additional attribute, and temporal bursts to evaluate event peaks. We wish to collect datasets from many different domains to see if one produces more occurrences of events. For our current research, we collected two datasets. The first was a 14-day sample based on the keyword search "Tulsa+Rally". The second dataset was another 14-day collection based on the keywords "Atlanta+Protests". We will collect future datasets from political, criminal justice, healthcare, and popular culture domains to determine if our approach identifies more events in certain domains over others.

The second application for future research will be the use of machine learning algorithms to evaluate our approach with more concrete precision. Our current method of

evaluating our approach uses statistically created boundary lines and peak occurrence to identify events. This is not an ideal approach to evaluate events, but for this current study it allows us an objective method to identify and quantify them as peak formations in a smoothed line trajectory. We plan to see if Support Vector Machine, Artificial Neural Network, Random Forest, and XGBoost provide concrete evidence to validate our approach. Accuracy scores will provide a metric we will use for evaluation.

SA is the third application that we will pursue in later research. SA scores will provide an additional attribute that can be used to further fortify the *NW* metric. The real benefit of SA is that it will provide an additional dimension to our human user contribution metric that is inherent in the NW attribute. In its current state *NW* tells us *how much users are messaging each other* and *how diverse the composition is* with regards to the number of participants in a discussion stream. SA will add a level of *emotional strength* that is associated with the increased user activity. This addition to the *NW* metric will not only fortify the existing efficacy of the attribute from a statistical standpoint, it will also allow for a sentiment dissection of a trajectory. Specifically, SA data that is incorporated into *NW* will allow us to evaluate not only *where and when event peaks form,* but *what emotions* potentially cause the surge in activity.

The last item we may pursue for later research is the use of temporal bursts (TB) in detecting event peaks. TB are a more detailed and robust way of analyzing all of the phases of an event curve from ascension, to crest formation, to eventual decline. Several studies in the literature use techniques such as wavelet analysis to evaluate TB. We do not plan on using wavelet analysis with our research, but that decision may change. Also, our plan to use machine learning algorithms for *NW* evaluation may render our interest in

TB moot. Machine learning algorithms require different features as input which could potentially render a smoothed linear curve unnecessary. We will keep the study of TB in our future plans as an open option, but our interest in the first three previously discussed applications is more concrete.

Chapter 4

Results

This chapter provides a discussion of the results that were obtained from experiments using five datasets. Three of the datasets were short-term collections that covered the same 13-hour time-period. We collected these shorter datasets with the primary objective of identifying the same events in parallel samples to demonstrate the efficacy and reliability of our approach. The remaining two datasets covered a longer time-period of two weeks. Our goal with the 13-hour datasets was to achieve a minimum amount of bias in the smoothed line trajectories between the three graphs. The experiments compared the smoothed line trajectories of Newsworthiness and keyword frequency during their respective time periods. The graphs of the trajectories were observed for the number of well-formed event peaks which occurred. The first three sections of this chapter will discuss the datasets that were used in the experiments. Specifically, the topics chosen for the datasets and the sizes of the collections will be discussed. Section four will discuss the results of comparing the three trajectories from the 13-hour datasets. Section five will compare the results of the Russian GRU Disinformation Newsworthiness and keyword frequency datasets. Sections six and seven will discuss the results of the Newsworthiness graph trajectories for the two fourteen-day datasets. Sections eight and nine will discuss the keyword frequency graph trajectories for the two fourteen-day datasets. The chapter will end with a summary of the experiment results.

The Russian Disinformation and RollingStones Datasets

We collected four short-term datasets as part of our research effort. The first dataset we collected was a 28-hour sample based on a topic that was trending in news media in the domain of cybersecurity. The remaining three datasets covered a more discrete time-period (13-hours) and came from the domain of popular culture. The 13-hour popular culture dataset will be discussed in more detail later in this section. With regards to the 28-hour dataset, it was reported on July 29, 2020 that the Russian GRU was behind a cyber disinformation campaign designed to spread fake news featuring information pertinent to the coronavirus in order to cause confusion and chaos in the general public (Tucker, 2020). Our sample consisted of 8,466 tweets and 28-hours-worth of discussion. We converted the sample into both Newsworthiness and keyword frequency datasets to compare their trajectories.

With regards to the remaining three short-term datasets, we intended to demonstrate that our *NW* method would not succumb to bias. To accomplish this task, we collected multiple datasets on the same topic that covered the same time-period. By doing this, we would plot the resulting graph trajectories and then compare the graphs. If the smoothed linear trajectories were similar in their paths and shapes, then our results could be successfully repeated and reproduced, thus providing evidence that our approach did not produce random results. When we collected the duplicate datasets, we used an API keyword search from a currently trending topic from popular culture. At the time of collection, one of the topics that was trending in the news media was a story concerning the musical group The Rolling Stones. The band was seeking legal action against the Donald Trump presidential campaign vis-à-vis the campaign's use of the band's

copyrighted music at its political rallies (Kirka, 2020). Due to the limitations of the Twitter API, the maximum number of tweets that could be collected per request was 10,000. We collected three different datasets of 10,000 tweets using the keywords "Rolling+Stones." The three datasets covering the same time periods were collected over a period of approximately six hours on June 29th, 2020 using the Twitter API. We made the decision to collect these parallel samples for two reasons. First, in the interests of expedition and efficiency, we decided to opt for a shorter target period of time for the datasets. Second, we chose the "Rolling Stones" keywords because based on empirical sampling and analysis, topics based in popular culture often demonstrated a more short-lived and intense cycle of public interest.

The three datasets at the time of collection contained a time frame of approximately 10 hours-worth of tweets. To plot a smoothed line graph in R, a minimum of 11 points on the x-axis was required. This minimum number of data points on the x-axis was a limitation imposed by R development environment. If too few data points were used, then an error would result in the R development environment. Based on empirical data using R and several prior datasets, 11 points was a minimum acceptable number for use with the smoothed line function. A second round of dataset collections was performed using the API hours later, resulting in three aggregated datasets that covered a total of 13 hours-worth of tweets. To ensure that the three resulting datasets were unique, each of the three collections were randomized. In each of the three datasets, 8,000 tweets were randomly sampled from their parent dataset of 10,000. The random sampling was performed in such a way as to maintain the integrity of the timeline. Specifically, there were samples taken from each hour of the 13-hour time period. The

three randomly sampled datasets were named respectively, “RollingStones1,” “RollingStones2,” and “RollingStones3.”

The Tulsa Rally Dataset

The selection of a dataset for a longer-term collection needed to be weighed and considered for a number of factors. The two variables that we decided to use for selecting our topics for dataset collection were scope and volatility. For the determination of scope, we needed to choose a topic that was not overly broad, otherwise a longer period of time would be required for a fair analysis. For example, a topic of Coronavirus occurring in 2020 is rather large in scope and would likely require months, if not greater than a year to adequately analyze. Often times, popular culture topics like the previously mentioned “Rolling+Stones” topic are much smaller in scope. Frequently they tend to generate a large amount of short-term interest before they are rendered inconsequential by topics of greater, more lasting consequence.

For our first long-term dataset, we chose a topic that would last at least a week within the mainstream media discussion cycle. After considering several recently trending news topics, we decided on Donald Trump’s Tulsa political rally (Steakin & Pereira, 2020). We chose this topic because at the time of collection, the event was scheduled to take place five days in the future. Eight days followed the rally in our collection timeline. If our approach was successful in identifying events, a spike in the smoothed line trajectory for the dataset would occur in association with day five of the fourteen-day dataset. Essentially, the rally was a highly publicized and advertised event with a fixed date, which was June 20th, 2020. This gave us a point of reference for us to measure peaks in Newsworthiness.

Volatility was the second variable we considered when choosing our topic for the long-term datasets. Many topics that trend on Twitter can last less than a day. We needed to pick a topic that would reliably generate a large number of tweets over a longer period of time. The "Tulsa+Rally" keyword search produced such a large return of tweets that several collections needed to be made during each day of the collection period in order to acquire tweets to represent the entire 24-hour period. For each day of the fourteen-day period, 1,000 tweets were collected, covering the morning, afternoon, and evening periods. At the end of the fourteen days of collection, each of the fourteen datasets of 1,000 tweets were aggregated into a single composite dataset of 14,000 tweets. Day one of the dataset began on June 15th, 2020. The final day of collection was June 28th, 2020.

The Atlanta Protests Dataset

We chose the topic for our second long-term dataset using the same criteria that we used for the first dataset, which were scope and volatility. Three days prior to the start of collection, one of the most significantly trending topics in the news media was that of progressing social unrest in Atlanta, Georgia due to controversial police action in that city. The police action resulted in days of protest and destruction of business and property (McKay, R., 2020). We started collecting tweets from the Twitter API using the search criteria "Atlanta+protests" on June 15th, 2020. The timeline of events for the topic began on June 12th, 2020 with the police shooting of Rayshard Brooks. The timeline ended on June 23rd, 2020 with Brooks's televised funeral, spanning an approximate total inclusive period of twelve days. We used two documented real-life events from the timeline to serve as points of reference as a comparison against the smoothed line trajectory of our Newsworthiness metric. The first of these two points of reference was the public

announcement of charges against the two police officers involved in the police action which were filed on June 17th, 2020. The second point of reference was the televised funeral of the victim of the police action, Rayshard Brooks which took place on June 23 (Cohen, 2020; McKay, R., 2020). Similar to the "Tulsa+Rally" dataset, event peaks would be associated with the points of reference in the timeline.

The "Atlanta+protests" topic trended heavily for the first eight days of collection. During the days of heavy trending, multiple samples were collected throughout the day to ensure that the morning, afternoon, and evening time periods were represented. For the remaining six days in the collection period there were fluctuating volumes of tweets available on the topic. For the final five days, the number of total available tweets had diminished to a degree that we were able to collect all of the available tweets on the topic starting from midnight to 11:59 pm of the 24-hour period. At the end of the 14-day collection period, all of the individual datasets were aggregated into a single composite csv file. For the "Atlanta+protests" composite dataset, we had a final total amount of 12,203 tweets.

Results of the 13-Hour Datasets

This section provides a discussion of the results that were obtained from the three 13-hour datasets that were collected using the search criteria "Rolling+Stones." The three datasets were scored for $DC(t_i)$ and E_s. These two metrics were then converted into the Newsworthiness attribute. An example of the converted metrics is seen below in Table 8. The first column in the table lists the individual time index for the period covered, which for this experiment was the hour index. The Newsworthiness attribute for all three

datasets was graphed separately as smoothed linear trajectories. The resulting graphs of the three 13-hour datasets can be seen in Figures 4, 5, and 6.

Figure 4

*T*rajectory *for the NW distribution of Rolling Stones sample 1*

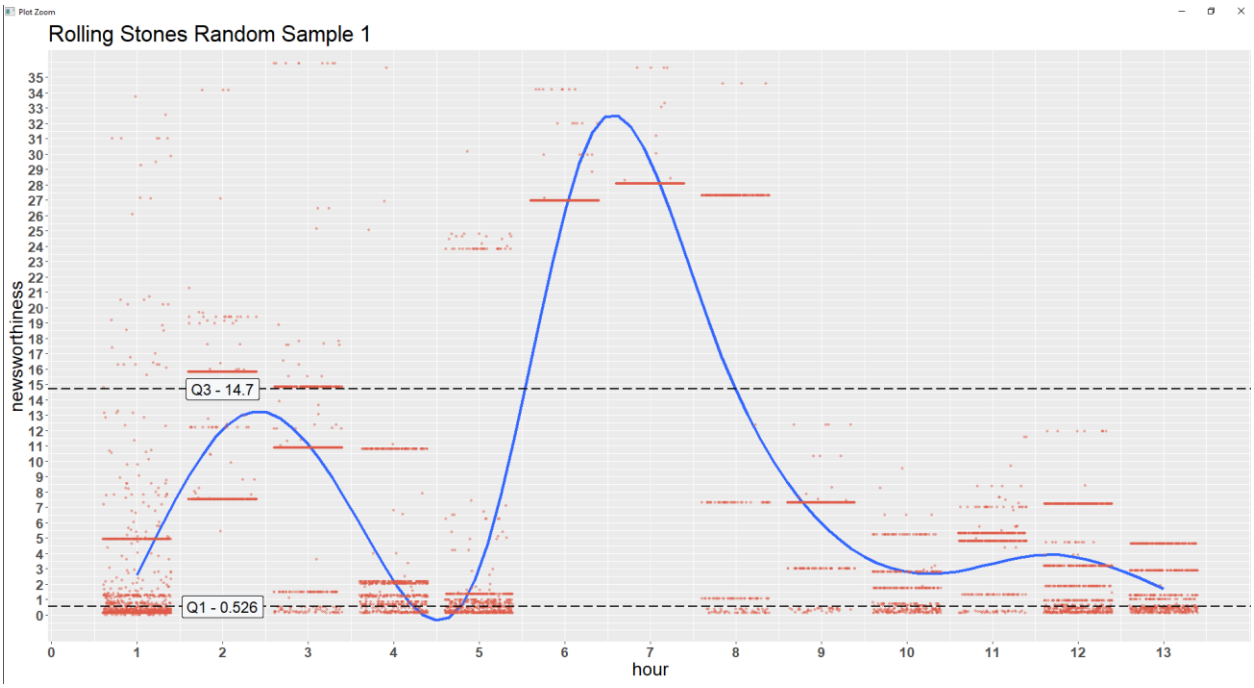


Figure 5

Trajectory *for the NW distribution of Rolling Stones sample 2*

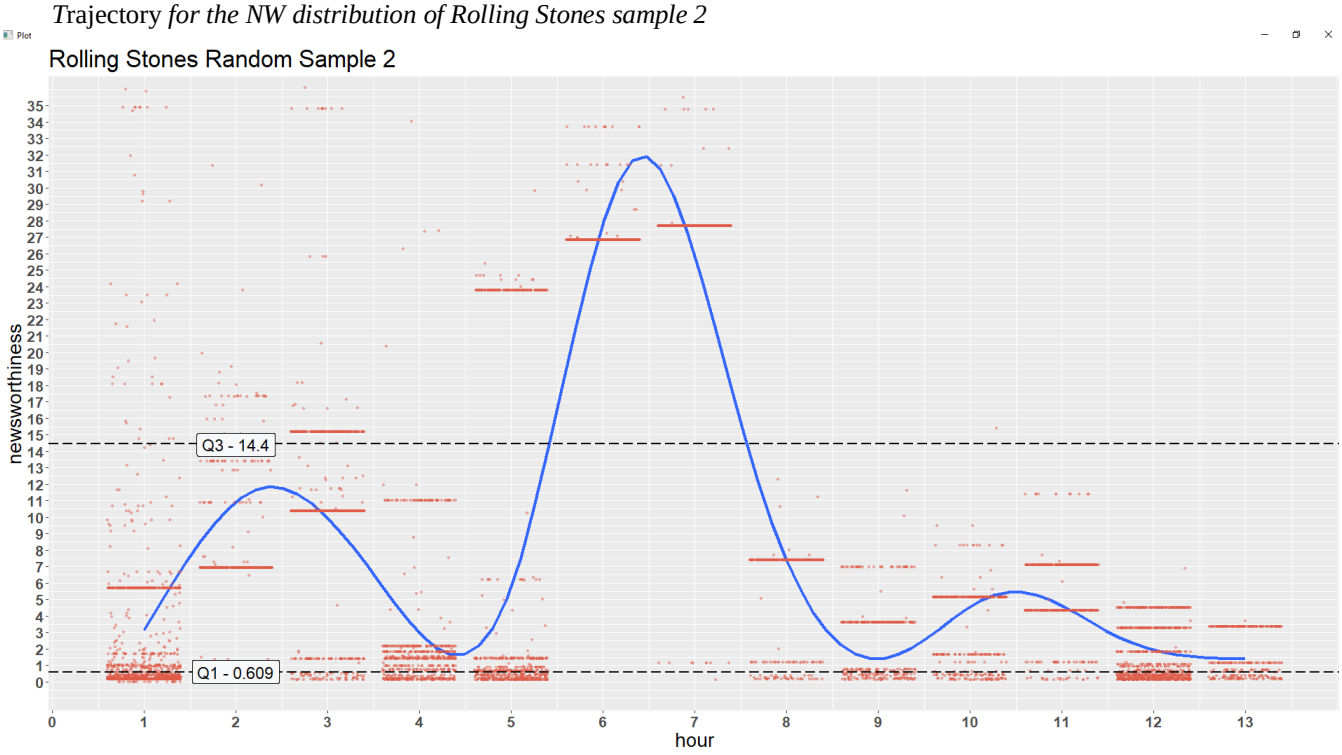
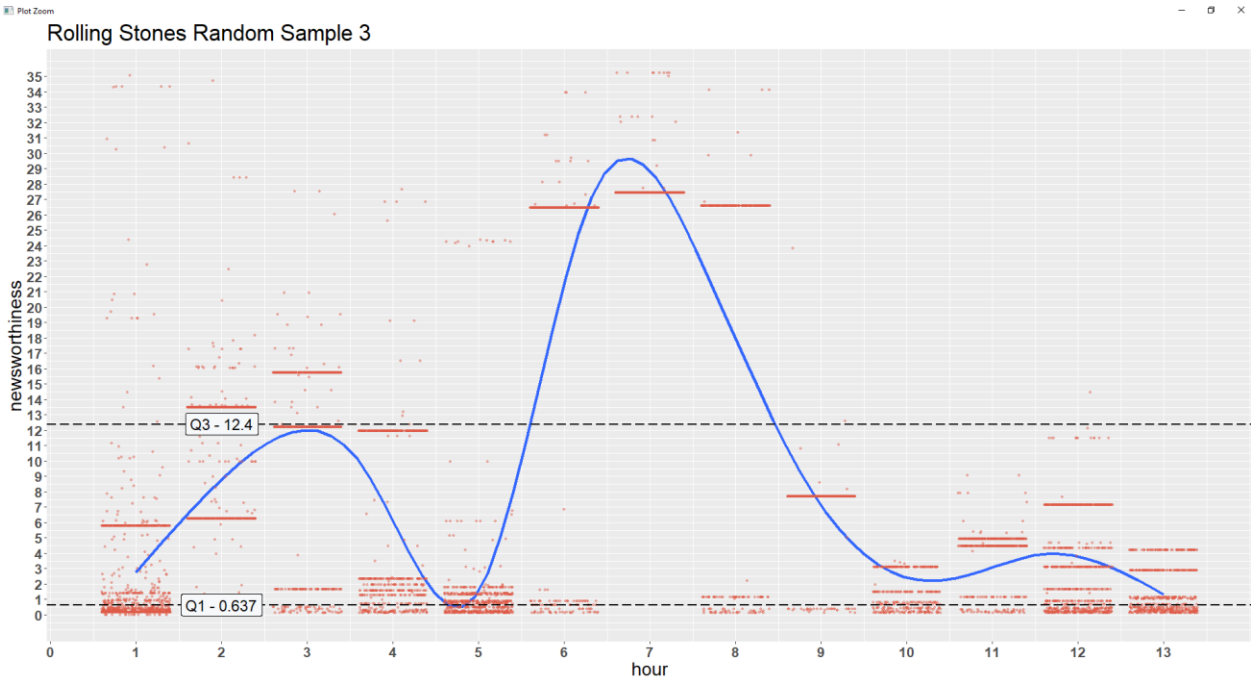


Figure 6

Trajectory *for the NW distribution of Rolling Stones sample 3*



Discussion of the Results

The results of the three graphs seen in Table 9 (below) were very similar in the overall shapes of their trajectories. Each of the three graphs identified two primary events in the 13-hour timeline. The second event was the more significant, as its peak formed well past the Q_3 fence. There were three subtle but noteworthy variances between the three graphs. First were the values of the Q_1 and Q_3 fences. In RollingStones1, the value of the upper fence, Q_3, was 14.7. In RollingStones2 and RollingStones3, the values of Q_3 were respectively 14.4 and 12.4. The second variance between the graphs was found in the location of the peak in the first event. In all three graphs, the first event lasted approximately 4.5 days. In the first graph (Figure 5), the crest of the event formed at approximately 13 units on the Newsworthiness scale. In graphs two and three (Figures 6 & 7) the event peak formed at approximately 12. The third variance could be seen in the peak location of the second event in the 13-hour time period. The crest of the second event for graphs one and two both fell at approximately 32.5 units on the Newsworthiness scale. In graph 3, the crest formed at approximately 30. With all the graphs considered together, the tolerance between the three measured less than 2.5 in Newsworthiness. The results suggested that our approach consistently identified the same events within the 13-hour time-period with a minimum bias between the three smoothed linear trajectories.

Table 8

Sample from RollingStones1 dataset after NW conversion

hour	avgEntropy	diffusion	newsworthiness
4	1.9732	70.24	35.5969
8	1.7917	62	34.6040
8	1.7917	62	34.6040
8	1.7917	62	34.6040
6	1.8138	62	34.1823
6	1.8138	62	34.1823
6	1.8138	62	34.1823

Table 9

Comparing results from all 3 RollingStones datasets

dataset	Metric	Min-Max	Event peaks	Q3	Variance	Standard Deviation
RollingStones1	NW	0-157.90	2	14.7	95.67	9.78
RollingStones2	NW	0-176.95	2	14.4	92.78	9.63
RollingStones3	NW	0-173.09	2	12.4	92.48	9.61

According to the side-by-side comparison of dataset metrics, seen above in Table 9, the maximum Newsworthiness value for the first sample was slightly less than the latter two samples with a value of 157.9 (approximately 18 units smaller than the next highest in Newsworthiness). As mentioned above, the Q_3 upper fence was two units smaller than the others in the third sample. The variance of the first sample was also slightly larger than the other samples by approximately three units. The cumulative effect of the previously mentioned biases between the three samples was a shorter event peak in the first sample for the second event (occurring between hours five and ten). However, the biases just discussed do not alter the number and duration of events in the three

samples. All three samples identified the same two events. The most significant of the two events occurred above the Q_3 upper fence and lasted a duration of five hours.

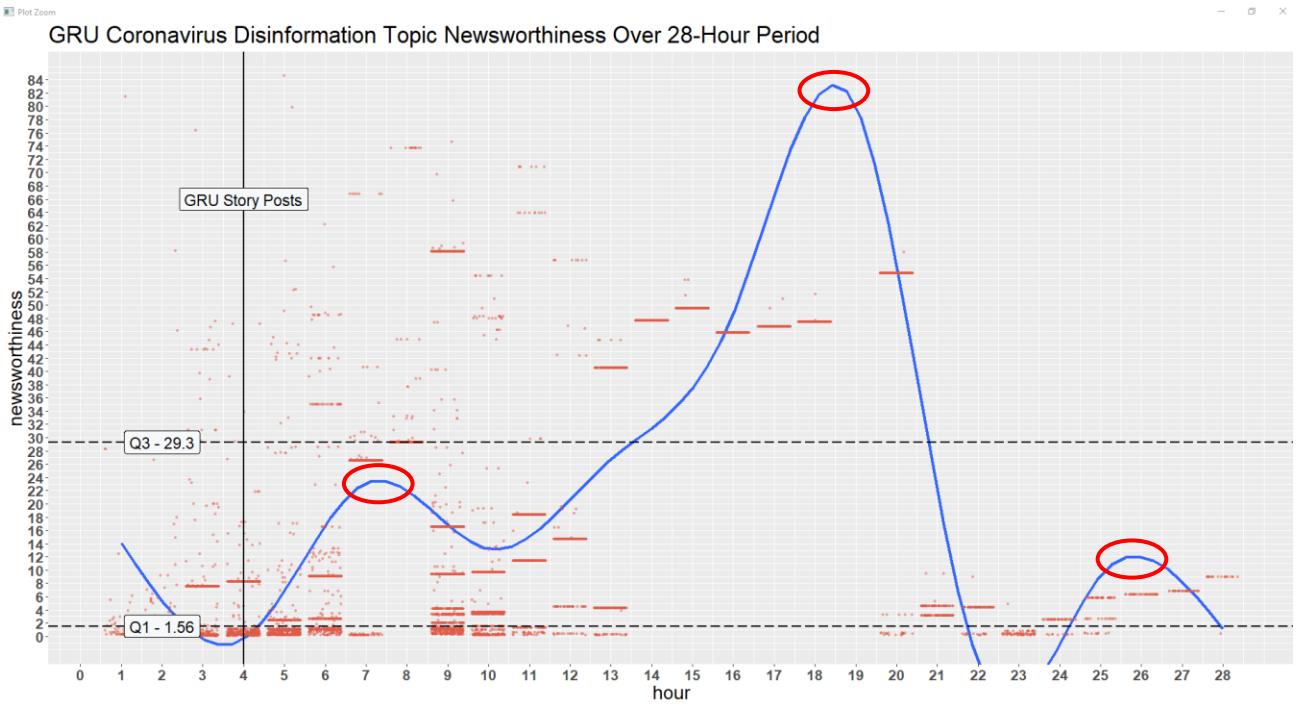
Comparing the Results of the Russian GRU 28-Hour Datasets

"Russian+Disinformation" Newsworthiness Trajectory

The Russian GRU Disinformation dataset consisted of 28-hours-worth of data points. Each hour (p =1 hour) was one index point on the x-axis. For temporal context, we inserted a vertical line in the x-axis at hour 4 which corresponded to the date that the Russian disinformation story broke in the media, which was July 29, 2020. A trajectory for the time-period can be seen below in Figure 7.

Figure 7

Trajectory for the NW distribution of the "Russian+Disinformation" dataset



"Russian+Disinformation" Newsworthiness Dataset Results Discussion

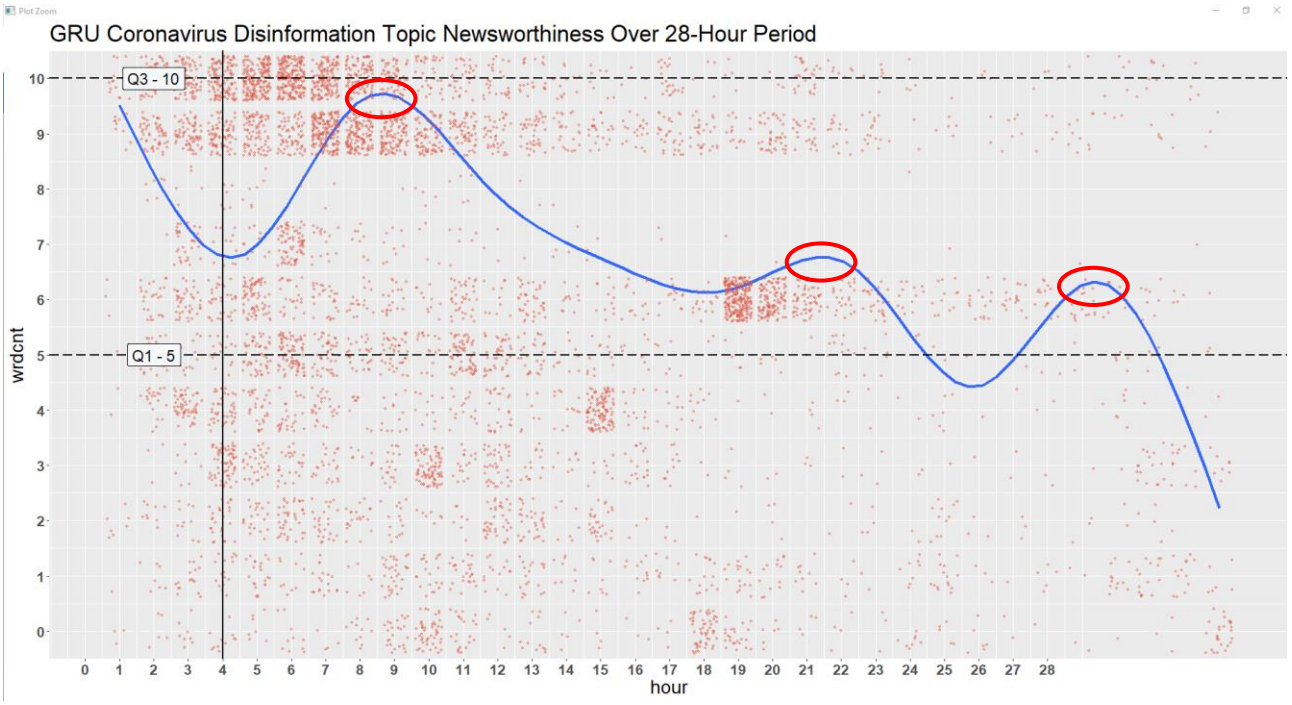
The trajectory for the 28-hour period of the *"Russian+Disinformation"* dataset identified three event peaks. The most significant event of the time-period took place from approximately hour 14 to hour 20. The event peak crested well above the Q_3 boundary at approximately 82 units of Newsworthiness. Two additional minor event peaks occurred prior to and immediately following the most significant peak in the time-period. The first event peak occurred between hour 5 and hour 9. The third peak occurred between hour 24 and hour 28. Event peaks one and three did not break the plane of the Q_3 upper boundary. The vertical line point of reference occurred in a trough prior to the ascent of the trajectory toward the first event peak. This juxtaposition between reference point and peak suggests a possible correlation between the known event and the peak formation.

"Russian+Disinformation" Keyword Frequency Trajectory

There was little agreement between the keyword frequency trajectory and the *NW* trajectory. The keyword trajectory identified three event peaks as did the *NW* graph. None of the three event peaks fell beyond the Q_3 upper boundary. The first event peak was identified as the most significant peak in the time-period. The second and third peaks in the time-period were less significant than the first. The second event peak took place from hour 19 to hour 23. The third event peak took place from hour 26 until the end of the period. The keyword frequency trajectory can be seen in Figure 8 below.

Figure 8

Trajectory *for the keyword frequency distribution of the "Russian+Disinformation" dataset*



"Russian+Disinformation" Keyword Frequency Dataset Results Discussion

As we mentioned above, there is little congruence between the keyword frequency trajectory and the trajectory of the Newsworthiness graph. The Newsworthiness trajectory identified the second event peak as the most significant. The keyword frequency trajectory identified the first event peak as the most significant. Both trajectories shared two fundamental empirical findings. The first was that the known point of reference occurred in both trajectories in a trough before the formation of the first event peak. The second shared finding was that the first event for the two trajectories lasted the same time span, i.e. from hour 4 to hour 10. The second and third event peaks were not in sync between the two trajectories. Event peaks two and three in the Newsworthiness trajectory were more clearly defined. We made two observations when we compared the two trajectories. First, the single mutually shared finding (i.e. the point of reference event occurring just prior to formation of the first event peak) provided more

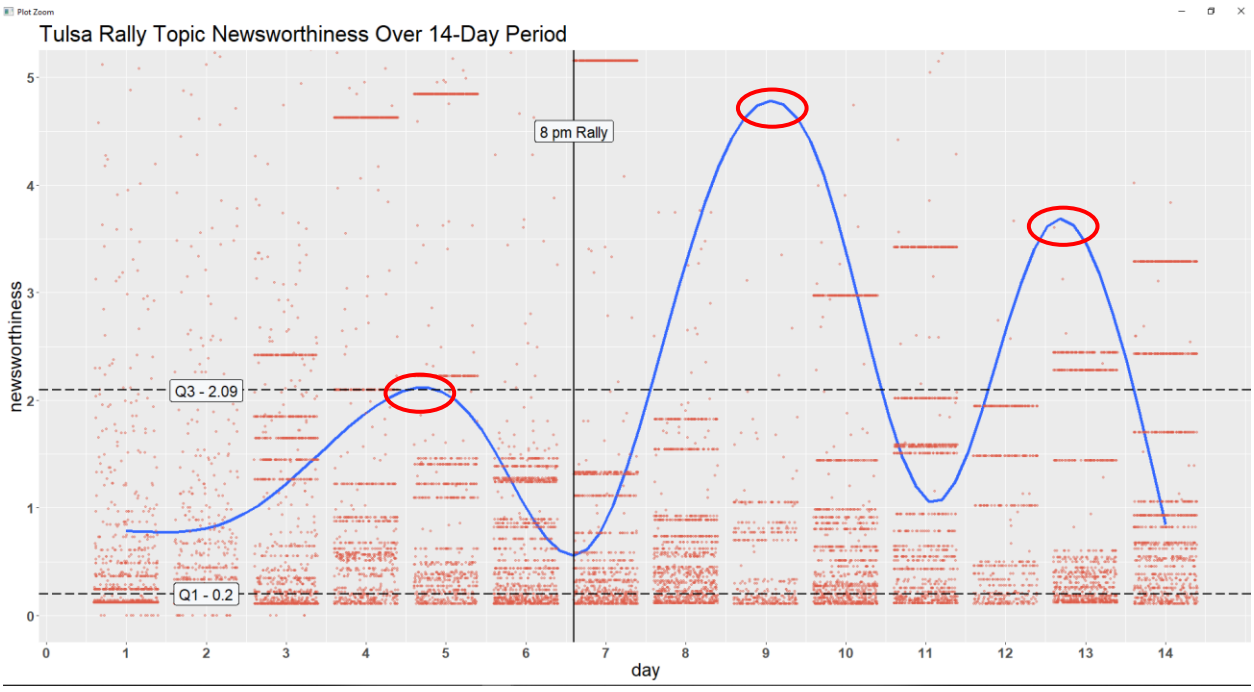
evidence to us suggesting a correlation between the known event on record and the formation of the event peak. Our second observation was that although the two methods captured the same number of events, the keyword frequency method did not appear to be as sensitive to certain discussion stream activity when identifying certain events. Evidence of this could be seen in the two trajectories when comparing the second event peak formations. In the Newsworthiness trajectory, the second peak is very significant and well-formed. In the keyword frequency trajectory, the second event peak is out of sync, less significant, and closer to a ripple than a peak.

Results of the "*Tulsa+Rally*" Dataset: Newsworthiness

The "Tulsa+Rally" dataset contained 14,000 tweets and covered a time-period of 14 days. The rally took place on June 20th, 2020 and we began collection on June 15th, 2020. We completed collection on June 28th. To better place our smoothed line trajectory in temporal perspective, we added a vertical line to the graph with a label denoting the time when the rally took place. The graph displaying the Newsworthiness results of the "Tulsa+Rally" dataset can be seen in Figure 9.

Figure 9

Trajectory for the NW distribution of the "Tulsa+Rally" dataset



"Tulsa+Rally" Dataset Results Discussion

The fourteen-day graph of the "Tulsa+Rally" dataset identified three distinct events. The peaks of all three events occurred either tangent to or above the Q_3 fence line. The rally took place on the sixth day of collection (June 20th, 2020 at 8 pm). A vertical line was placed in this time index to serve as a contextual point of reference, so it was inserted into the graph at approximately three quarters of the distance between day 6 and 7. We interpreted the results of the graph in the following manner with respects to the rally that took place. The first event, which crested during the fourth day of collection, lasted six days. The amount of discussion activity on this topic diminished on the day of the rally. This first event could be attributed to the anticipation and buildup of the upcoming event, which was heavily discussed in news media outlets (Mason, 2020). The second event started its trajectory ascent on the seventh day of collection and crested on the ninth day of collection. The third event began its ascent on day 11 and crested on day

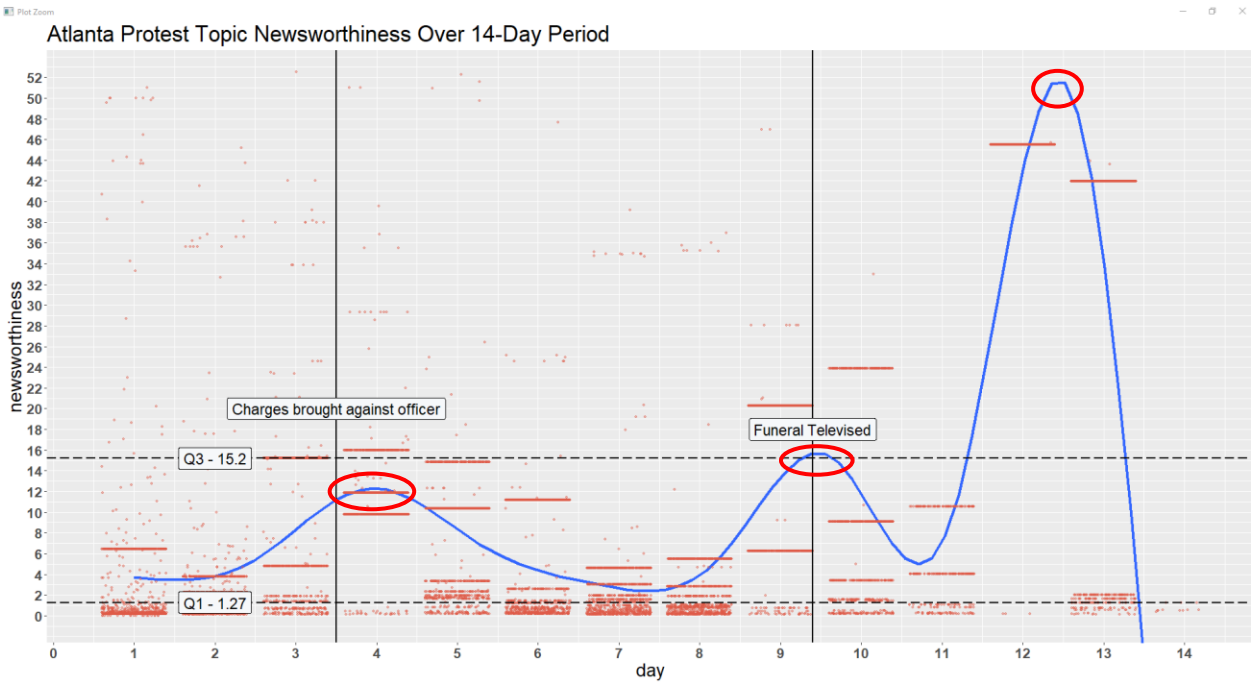
12. We interpreted the occurrence of the second event as the public reaction to the rally, which ended at approximately 10 pm on June 20th. The trough that formed between events one and two we interpreted as a temporary ebb in discussion due to the fact that people were either physically attending the rally in person or viewing it on media. We interpreted the second trough as a temporary reduction of discussion on the rally precipitated by other breaking news.

Results of the "*Atlanta+Protests*" Dataset: Newsworthiness

The "Atlanta+Protests" dataset contained a total of 12,203 tweets, and like the "Tulsa+Rally" dataset it covered a period of 14 days. We inserted two vertical lines into the x-axis of the graph to serve as contextual points of reference. Both points referred to peripheral events which occurred during the 14-day collection period which related to the police action and resulting civil unrest in Atlanta (McKay, 2020). The first of the two reference points was the issuing of charges against two police officers. This event took place on the third day of collection, June 17th. The reference line was placed in the midpoint between days three and four as the charges were announced sometime in midday. The second reference line referred to the televised funeral for police shooting victim Rayshard Brooks, which aired on the 9th day of collection, on June 23. The reference line was placed at the midpoint between days nine and ten since the televised event took place at midday. The graph displaying the Newsworthiness results of the "Atlanta+Protests" dataset can be seen in Figure 10.

Figure 10

Trajectory for the NW distribution of the "Atlanta+Protests" dataset



"Atlanta+Protests" Dataset Results Discussion

The "Atlanta+Protests" graph identified three events over the 14-day time-period. The peak for event one, which formed at day four, did not fall beyond the Q_3 fence line. Event one lasted for a period of six days before it resulted in a trough. The peak for event two occurred just breaking the plane of Q_3. Its peak formed on day nine of the time period. The most significant of the three events was event three which lasted from day 11 through the middle of day 13. Event two lasted two days and event three lasted only two days. Our interpretation of the events as they are depicted in the graph is detailed in the following sections.

Collection of this dataset began three days after the shooting of Rayshard Brooks. Civil unrest was already forming locally in the city of Atlanta. The public discussions of this event had become aggregated in the news media as part of a larger evolving discussion involving civil unrest and the role of law enforcement throughout the country.

Charges against the two officers involved in the shooting were announced on June 17th, day three of collection. Event one occurred during this point of reference. Following the announcement of charges, there was a slight uptick in discussion activity on the collected topic. The activity crested the following day and then gradually leveled off into a trough. The overall profile of event one was a longer time-period and a shorter, wider crest. This evidence to us suggested that the announcement of charges contributed to an existing, growing level of discussion on this topic, but it did not precipitate the increase and peak formation in Newsworthiness.

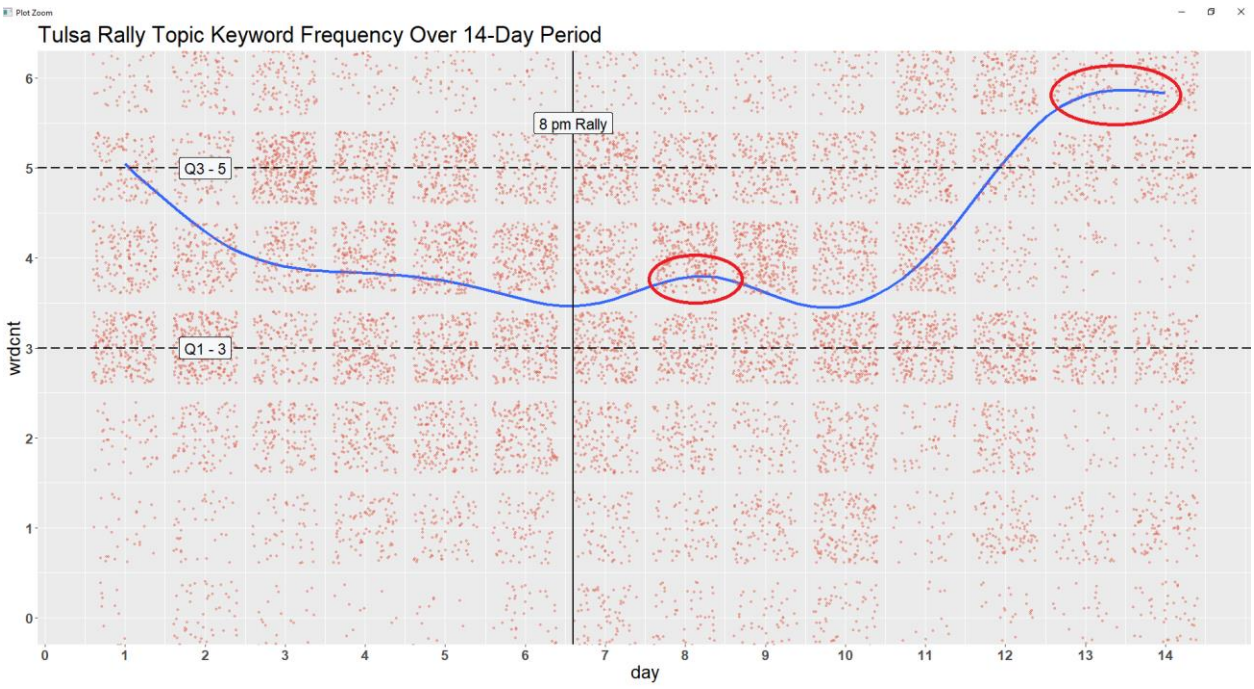
The second event crested right in the middle of day nine of collection. This second event lasted for a period of three days and peaked coincidentally with the second reference point in our timeline, i.e. the televised funeral. The funeral had been discussed on several media outlets and on social media for several days. As with the first event peak in this dataset, we interpreted that the action associated with the reference point did not precipitate the increase in discussion activity but added to an existing discussion. The third event in the "Atlanta+Protests" timeline was the most significant in terms of Newsworthiness score, yet was the shortest in duration, lasting only two days. There was no available data to use as a point of reference to determine the possible precipitation of increased discussion. After analyzing the two long-term datasets we made the observation that there was a possible correlation between the location of the dated reference point and the ascent of the trajectory line to its apex. If the dated reference line was located before the trajectory line begins its ascent, it suggested a possible causal relationship. For example, the reference point for the "Tulsa+Rally" dataset occurred immediately before the ascent of the trajectory line. If the reference line occurred in the middle of the

Trajectory ascent, it was more likely that the occurrence played a contributing role and was not the cause of the increased activity.

Results of the "*Tulsa+Rally*" Dataset: Keyword Frequency

After we completed the two graphs for the "Tulsa+Rally" and the "Atlanta+Protests" datasets using Newsworthiness on the y-axis, we graphed these same datasets using keyword frequency. Keyword frequency was the existing method for event detection that we chose to use to compare against the results of Newsworthiness. We took the tweets from the "Tulsa+Rally" dataset and we used the text mining techniques that were discussed previously in this document to identify the repository of keywords. The keywords were then counted for their occurrence on each date. The graph of the keyword frequency distribution for the "Tulsa+Rally" dataset can be seen in Figure 11.

Figure 11
Trajectory for the keyword distribution of the "Tulsa+Rally" dataset



"Tulsa+Rally" Keyword Frequency Dataset Results Discussion

The keyword frequency graph for the "Tulsa+Rally" dataset only showed two areas where keyword occurrence increased over the 14-day period. The first area of increased frequency occurred on day eight, where there was a subtle ripple in the smoothed line trajectory. When we compared this result with our Newsworthiness graph, there was a sizable discrepancy between the two graphs. Our analysis of the difference between the two graphs is detailed in the chapter summary below. Days 7 through 11 in the Newsworthiness graph showed the most significant event peak in the time-period. The keyword frequency graph for this time period suggested only a mild increase in keyword usage. The second area of increased keyword usage occurred from day 11 to day 14. The graph suggested a gradual increase, ending in a plateau for the smoothed line trajectory. In the Newsworthiness graph, this time-period showed a second event peak which formed on day 12. The smoothed line trajectory for the "Tulsa+Rally" Newsworthiness graph ended on day 14 in a downward slope. In the keyword frequency graph, only days 11 through 14 could be interpreted as an event, since no other well-formed peak could be identified. Ultimately, the results of the keyword frequency graph did not corroborate the results found in the Newsworthiness graph. The Newsworthiness graph identified three distinct event peaks, while the keyword frequency graph identified only one. There was not enough data available to suggest why the same peaks did not form in the keyword graph.

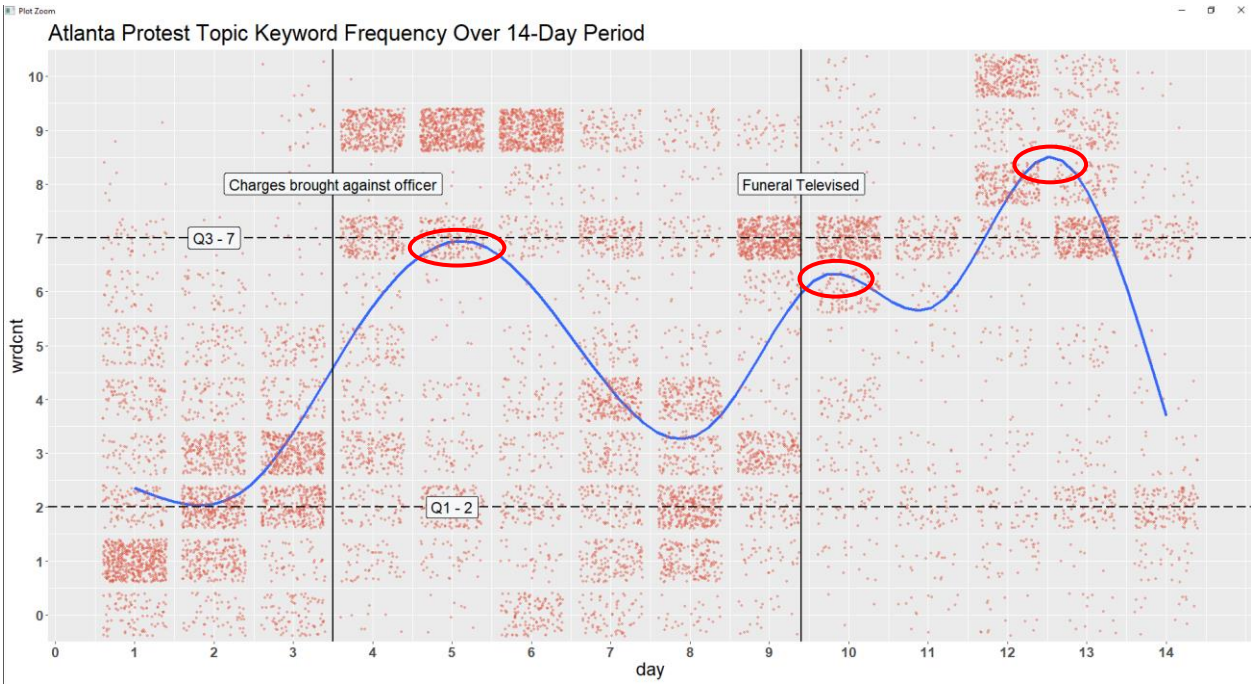
Results of the "*Atlanta+Protests*" Dataset: Keyword Frequency

The results of the keyword frequency graph of the "Atlanta+Protests" dataset validated the results of the Newsworthiness graph. Both graphs showed event peaks

occurring during the same time periods. This level of corroboration was not evident in the two graphs resulting from the "Tulsa+Rally" dataset. The graph of the "Atlanta+Protests" keyword frequency dataset can be seen in Figure 12.

Figure 12

Trajectory for the keyword distribution of the "Atlanta+Protests" dataset



"Atlanta+Protests" Keyword Frequency Dataset Results Discussion

There was considerable agreement in event detection between the "Atlanta+Protests" Newsworthiness and keyword frequency graphs. In both graphs, the most significant event that was identified had its peak on day 12 of collection. The two graphs had trajectories that ended with full descending slopes. They also both identified the first event as lasting approximately seven days before the trajectory descended into a trough. The second event in each of the two graphs lasted from day 8 through day 10. We made an observation regarding the two "Atlanta+Protests" graphs with respects to the two points of reference and their corresponding event peaks. The locations of the

reference points in the developing event peaks suggested that the charges against the officers and the televised funeral contributed to the increased discussion but did not precipitate the increase. In both cases, the trajectories were already moving in an upward direction.

Summary of Results

The following section summarizes the results of the experiments that we performed with a total of nine datasets that we collected using the Twitter API. The first three datasets were short-term collections spanning a period of 13 hours per dataset. The 13-hour datasets were collected in parallel, meaning they were collected over the same time span using the same search term criteria. The purpose of the parallel dataset collection was to validate that our approach to event detection consistently produced the same results in a two-dimensional smoothed line graph with minimal bias. The fourth and fifth datasets were short-term samples that we collected using the keywords "Russian+Disinformation" that covered a time span of 28-hours. The remaining four datasets covered a time-period of 14 days. Two of the datasets were collected using the search criteria "Tulsa+Rally." The remaining two datasets were collected using the search terms "Atlanta+Protests." For each of the collected dataset pairs, two approaches to event detection were evaluated. The first approach was our method, which used $DC(t_i)$, E_s, and Newsworthiness. The second approach was an existing method of event detection which used unigram keyword frequency. The results were promising but demonstrated the need for additional research in the future.

Table 10

Metrics for all 7 NW and keyword count datasets

dataset	Min-Max	Event Peaks	Q3	Variance	Standard Deviation
NW RollingStones1	0-157.90	2	14.7	95.67	9.78
NW RollingStones2	0-176.95	2	14.4	92.78	9.63
NW RollingStones3	0-173.09	2	12.4	92.48	9.61
NW Russian Disinformation	0-217.94	3	29.3	683.43	26.14
WC Russian Disinformation	0-13	3	10	15.99	3.99
NW Tulsa Rally	0 - 76.21	3	2.09	10.38	3.22
WC Tulsa Rally	0 - 13	1	5	5.53	2.35
NW Atlanta Protests	0 - 264.18	3	15.2	224.31	14.97
WC Atlanta Protests	0 - 11	3	7	9.97	3.15

Table 11

Average and Maximum NW and keyword counts for each day when event peak

Dataset	Event	Day	Average NW	Max NW	Average WC	Max WC
TulsaRally NW	1	5	3.31	76.21	N/A	N/A
TulsaRally NW	2	8-9	3.31	76.21	N/A	N/A
TulsaRally NW	3	12	3.31	76.21	N/A	N/A
TulsaRally WC	1	12-14	N/A	N/A	3	3
AtlantaProtests NW	1	3-4	10.79	52.51	N/A	N/A
AtlantaProtests NW	2	9	15.60	60.97	N/A	N/A
AtlantaProtests NW	3	12	45.36	45.66	N/A	N/A
AtlantaProtests WC	1	4-5	N/A	N/A	6.41	10
AtlantaProtests WC	2	9	N/A	N/A	5.36	9
AtlantaProtests WC	3	12	N/A	N/A	8.24	11

Table 12

Newsworthiness frequency table for Alanta+Protests dataset

day	newsworthiness	frequency
3	15.23	497
4	9.81	410
4	11.89	337
3	4.77	198
4	15.98	183
3	1.38	54
3	1.91	48
3	0.76	28
3	0.12	25
3	0.61	22
3	0.70	21

Summary of the three "RollingStones" 13-Hour Datasets

We will refer to Tables 10, 11, and 12 above to explain the results of our experiments. Based on the smoothed line graphs shown earlier in this chapter and the data in the above tables, we have made the following observations. The metrics for the three 13-hour short-term datasets shown in Table 10 ("RollingStones1", "RollingStones2", "RollingStones3") demonstrated that the Newsworthiness methodology consistently identified the same number of events for the three parallel datasets. The most significant difference in min-max values between the samples was 19.05. Specifically, the "RollingStones1" sample had a Newsworthiness that was smaller than the other two. This meant that the more significant event peak (second peak) for "RollingStones1" would be shorter than the other two "RollingStones" samples. Additionally, the Q_3 fence value for the "RollingStones3" sample was less than the other two datasets by two units of Newsworthiness. This meant that "RollingStones3" had a larger range of outliers in the dataset than the other two. The "RollingStones1" sample

also had a variance that was slightly larger than the other two datasets (approximately 3 units greater than the other two). When all these metrics were considered comparatively, the three 13-hour "RollingStones" datasets captured the same number of event peaks with the same durations. The biases between them were relatively minimal.

4.10.2 Summary of the "Tulsa+Rally" Newsworthiness Dataset

The two 14-day datasets provided mixed results with regards to event detection when the Newsworthiness graphs were compared against the keyword frequency graphs. The first of the two longer datasets, the "Tulsa+Rally" dataset, detected three event peaks when the Newsworthiness trajectory was shown on a graph. All three event peaks fell beyond the Q_3 fence line. The first detected event broke the plane of the Q_3 fence, cresting at a Newsworthiness y-axis value of 2.09. The total duration of the first event was six days. A vertical line representing a known event, in this case the 8 pm rally that took place on June 20, was inserted past the midway point between day six and day seven where a trough had formed in the trajectory. The second event crested at day nine and lasted from day 7 to day 11 where a second trough had formed. The third event crested on day 12 and lasted from day 11 to day 14 where the trajectory descended into the Q_1 lower fence.

When we analyzed the metrics for the "Tulsa+Rally" dataset in Table 10, we noticed three numbers that we found interesting. The min-max value (0 - 76.21), Q_3 fence (2.09), and variance (10.38) were all significantly lower than the results found for the "Atlanta+Protests" 14-day dataset. We also analyzed the average and maximum Newsworthiness values for each event that was identified within the time-period. There was no variance in these two metrics for all three of the events in the dataset. In the

"Tulsa+Rally" dataset, as seen in Table 11, events were identified for day 5, days 8-9, and day 12. For all three of these events, the average Newsworthiness was 3.31 and the maximum was 76.21. The smaller values for the dataset observed in Table 10 were the result of smaller $DC(t_i)$ values and relatively larger E_s values. This, in turn, caused smaller overall values of Newsworthiness. We believe that the lack of variance and the smaller values were the result of two influences. First, the smaller $DC(t_i)$ score suggested that there was either a larger number of dispersed individual users with lower messaging influence or a smaller number of user subnetworks discussing the topic. Second, we believed that the average size E_s score was the result of a comparatively smaller user diversity. This average diversity combined with fewer user subnetworks and large volume of messages could result in duplicated messaging. To provide additional clarity to this discussion, we originally defined Newsworthiness as $NW=f(\frac{DC(t_i)}{E_s},p)$. What became clear to us following testing of all seven of the datasets was that clusters of nodes with higher *NW* were more directly correlated with the formation of peaks in a smoothed line trajectory. If $DC(t_i)$ scores were high enough, even after integration with E_s scores, the resulting *NW* values would remain high. What this implied to us was that nodes with higher $DC(t_i)$ scores were connected to larger networks of users. The users who resided at the center of these larger networks held the greatest influence, therefore they were attributed the largest $DC(t_i)$ scores. Based on the evidence, we concluded that the combination of more dispersed user connectivity, average user diversity, and a large volume of messages resulted in a lack of variance for the time-period. We also concluded that the pervasive low $DC(t_i)$ scores in the "Tulsa+Rally" dataset, which contributed to the low *NW*, occurred because the discussion stream for this topic was void of a larger

network of interconnected users. This lower user interconnectivity implied a reduced amount of influence for information spread attributed to all participating users in the discussion stream.

Summary of the "Tulsa+Rally" Keyword Frequency Dataset

The keyword frequency graph for the "Tulsa+Rally" dataset did not corroborate the results of the Newsworthiness graph. The keyword frequency graph showed a gradual decline of keyword frequency over the first three days of the time-period. Days 7 through 9 showed a very mild uptick in keyword frequency, but not enough that we could classify it as an event. Days 10 through 14 of the keyword graph showed a gradual upward slope in frequency. The trajectory for the time-period ended as a flatline plateau. According to the metrics displayed in Table 10, the "Tulsa+Rally" dataset had one event peak lasting from day 12 through day 14. The min-max values for the dataset was a minimum of zero keywords and a maximum of 13. The Q_3 fence value was 5 and the variance was 5.53. According to the data in Table 11, "Tulsa+Rally" had during its one event peak an average keyword count of 3 and a maximum keyword count of 3. We found that the metrics were not very insightful for explaining the differences between the Newsworthiness and keyword graph trajectories. What we found interesting was the overall absence of a repeated pattern in the keyword frequency graph trajectory. The smoothed line graph showed a small amount of variance, which included the event plateau at the end of the time period.

After we analyzed metrics from the Newsworthiness and keyword frequency graphs for the "Tulsa+Rally" datasets, we made an observation. The keyword frequency graph was intended to be sensitive to increases and decreases in the frequency of words

that were inherent in messages circulating in a discussion stream. There may not have been keywords occurring with enough frequency to cause peaks to form in the same locations that formed in the Newsworthiness graph. To further illustrate our observation, the graph displayed in figure 9 showed a trajectory that demonstrated some mild undulation and rippling, terminating with a gradually climbing plateau. The smoothed line graph for the Newsworthiness dataset, which covered the same time-period, showed a trajectory that formed three well-defined peaks. Based on the evidence, we posited that the Newsworthiness graph may have captured event-related activity in its trajectory that word frequency could not effectively capture. There was not enough evidence in the metrics to suggest a possible reason for the dissonance between the two trajectories.

Summary of the "Atlanta+Protests" Newsworthiness Dataset

The "Atlanta+Protests" Newsworthiness graph detected three event peaks in the 14-day period. The first of the three events lasted for a duration of seven days. The peak of the first event did not fall beyond the Q_3 fence. The second event broke the plane of the Q_3 fence with a crest that measured approximately 16 units of Newsworthiness on the y-axis. It lasted from day 8 through day 10. The third event was the most significant in the dataset. It crested with a Newsworthiness magnitude of approximately 51 units and lasted for a total of three days. Two known events were inserted as points of reference into the "Atlanta+Protests" graph. The first reference point was inserted in the middle of day three. It coincided with the time at which formal charges were filed against two police officers involved in the shooting. The second point of reference referred to a televised funeral that had been discussed in the media for several days. This second reference point was inserted in day 9 of the time-period.

With regards to the “Atlanta+Protests” Newsworthiness dataset, there were three observations which we found interesting. The first dealt with the metrics that were produced by the dataset. The second observation which interested us were the specific conditions under which event peaks formed in our experiments. The third observation was the location of event peaks with respect to the known points of reference in the timeline. In the next three sections we will discuss each of these observations in detail. Chapter 4 will conclude with a summary of the results found with the “Atlanta+Protests” keyword frequency dataset.

The first observation we found interesting was the spread of the metrics resulting from our experiment with the “Atlanta+Protests” Newsworthiness dataset. As seen in Table 10 above, the variance for the dataset was 224.31which was significantly larger than what we observed in the “Tulsa+Rally” dataset. We identified three event peaks for the time-period. The events occurred on days 3-4, day 9, and day 12 as shown in Table 11. Between the three event peaks there was a substantial amount of variance. The variance between the three event peaks with regards to Newsworthiness magnitude was significant. The average Newsworthiness score for event peak one was 10.79. Event peak two had an average of 15.60 and the third peak averaged 45.36. What these metrics suggested to us was that there was a significant amount of movement and fluctuation of messaging activity among users in the discussion stream. The decline in the graph’s smooth line trajectory at the end of the time-period suggested a slowing down of activity for the time-period covered.

Previously, we discussed significant differences in Min-Max levels between the “Tulsa+Rally” and “Atlanta+Protests” datasets as shown in Table 10. Even with these

differences in maximum levels of Newsworthiness, both datasets identified three distinct event peaks each in their respective trajectories. This observation brought us to our second point of interest in our results. After analyzing the metrics in Table 10 and Table 11, we eliminated the original notion that higher scores of Newsworthiness alone caused the formation of event peaks. After making this observation, we compiled Table 12, to demonstrate evidence as to what we believed was a contributing factor to the formation of event peaks. We hypothesized that event peaks were formed by a higher frequency of discussion stream nodes that had higher levels of newsworthiness plotted in the same time-period index. In Table 12 the node in the first row had a Newsworthiness of 15.23. What was implied by the table was that 496 additional nodes had the same Newsworthiness score. This cluster of 497 nodes on day 3 contributed to the peak that formed on that day in the time-period. In Figure 8 ("Atlanta+Protests" Newsworthiness smoothed line graph) a row of jittered points can be seen at 15.23 above the event peak for day 3. The relationship between event peaks and the frequency of nodes with high Newsworthiness is a topic we will pursue in much greater detail in future research. The assumption we made concerning node frequency and event peaks will require additional testing to validate.

The final observation that we made was the location of event peaks with regards to known reference points that were inserted into the timeline. As seen in Figure 8, there were two reference points that were discussed previously. The first was the news of charges being filed against two officers involved in a shooting in Atlanta. This known event took place on June 17th, 2020. The second point of reference was the televised funeral for Rayshard Brooks which aired on June 23rd, 2020. In Figure 8, the first known

reference point was marked at the incline slope of the first event, just prior to the peak. The second reference point occurred in sync with the second event peak. What this data suggested to us was that the two points of reference contributed to, but did not directly cause, the two event peaks to form. The two marked and dated reference events contributed to a discussion that was already occurring. What contributed to this observation was that in Figure 7 ("Tulsa+Rally" Newsworthiness dataset graph), the known point of reference was marked in a trajectory low point. The day following the point of reference, the trajectory began an upward ascension toward an event peak. This juxtaposition of reference point and event peak led us to believe that there was a causal relationship between the reference point and the event.

Summary of the "Atlanta+Protests" Keyword Frequency Dataset

The "Atlanta+Protests" keyword frequency graph identified three events on the same days for the same durations as the Newsworthiness graph. The first event in the keyword frequency graph was identified as lasting from day 2 to day 7. The event crested on day 5 and took place in conjunction with the first point of reference (charges brought against officers). The reference point occurred two days before the peak formed. The second event lasted from day 8 to day 11, cresting on day 9. This second event crested one day after the occurrence of the second reference point (televised funeral). The third event peak formed on day 12. This final event ended on day 14 with a trajectory that descended in a downward slope toward Q_1. There was a substantial amount of congruence between the Newsworthiness graph and the keyword frequency graph. This was in direct contrast to the results that were obtained from the two "Tulsa+Rally" graphs.

Summary of the "Russian+Disinformation" Newsworthiness Dataset

The *"Russian+Disinformation"* Newsworthiness dataset lasted for a time span of 28-hours and identified three events in its trajectory. Only one of the event peaks formed a crest above the Q_3 upper boundary, which was the second event. This was the most significant event in the time-period since its peak formed at approximately 84 units of Newsworthiness. Peak one and peak three both formed their crests below the Q_3 boundary. Event one lasted from hour 4 to hour 10. Event three lasted from hour 24 to hour 28 when the time period ended. The most significant event in the period lasted from hour 11 to hour 21. The known event point of reference for the time-period was the release of the "Russian+Disinformation" story in the news on July 29, 2020, which was hour 4 of the collection period. The point of reference occurred in a trajectory trough prior to the ascent and formation of the first event peak.

Summary of the "Russian+Disinformation" Keyword Frequency Dataset

The *"Russian+Disinformation"* keyword frequency trajectory identified three events, which was in agreement with the corresponding Newsworthiness trajectory. The first event peak lasted from hour 4 to hour 10 which was also in agreement with the Newsworthiness approach. The remaining two events did not correspond to the Newsworthiness trajectory in their significance or their locations on the timeline. The keyword frequency trajectory identified the first event as the most significant. The Newsworthiness approach identified the second event as most significant. We observed one additional finding which was in agreement with the Newsworthiness trajectory. The known event point of reference in the "Russian+Disinformation" keyword frequency time period occurred in a trough just prior to the upward movement of the trajectory.

Chapter 5

Summary, Contributions, and Future Work

Research Summary

Events & Event Detection, SNA Metrics, and Newsworthiness

An event is defined as a set of messages on a related topic within a defined time-period that surpass a threshold measured by the statistical values of Diffusion Centrality and Shannon Entropy. Event detection is the identification of events that are present in a time-ordered social media discussion stream such as Twitter. One of the more popular approaches to event detection that we found in the research literature was temporal keyword frequency measurement. This technique involved the measurement of message keyword occurrence throughout the trajectory of a time-period. Increases and decreases in keywords associated with a certain topic have been viewed in many studies as synonymous with the growth and decline of events. One drawback to this approach has been that it tended to marginalize the human behavioral perspective of activity in a discussion stream.

Social Network Analysis (SNA) is an interdisciplinary field that combines elements of sociology and computer science. SNA, as a discipline, is concerned with studying human behavior and how entities interact with each other. Diffusion Centrality ($DC(t_i)$) and Shannon Entropy (E_s) are two metrics that fall under the large umbrella of SNA evaluative tools. $DC(t_i)$ is a SNA value that measures the message spreading influence that individual users in a discussion stream subset have with respects to the network of connected users as a whole. E_s is a metric that was adapted to information science from the field of Physics where it was originally used to measure the level of

disorder in a system. In the field of SNA E_s measures the level of diversity in a system. E_s was used in several studies in the literature with regards to diversity of users and messaging in social media. We found in our research that the E_s metric excelled at the macro level of measurement. Specifically, it was proficient at measuring the levels of diversity of overall participant contributions to a collective project. According to our research, a "project" translated to a discussion stream and a "participant" translated to a user. We proposed the integration of $DC(t_i)$ and E_s into a single metric we called Newsworthiness (*NW*) to identify and measure events in a discussion stream. We defined *NW* as a SNA metric of user activity that quantifies the distribution of user message spreading actions over the user diversity in a discussion stream. We formally defined the *NW* metric as ($\frac{DC(t_i)}{E_S}$, p), where $DC(t_i)$ was a user's Diffusion Centrality score, E_s was the average Shannon Entropy of the message text for the time period index, and p was the individual time period index being studied. After many preliminary trial experiments, we found that using a smoothed linear graph with jittered points was the optimum method to track the trajectory of a discussion stream through its time-period coverage.

Short-Term Dataset Collection

We decided to collect a total of six Twitter datasets to demonstrate the identification of events using our *NW* metric. We performed our collections using the Twitter platform's application programming interface (API). The platform's API included some inherent limitations for average users, which included a 10,000-tweet limit per 15-minute window. There was also a restriction on how far back in time we could collect tweets (8-days at most). The first three datasets were short-term collections that covered a period of 13 hours per dataset. The three datasets were collected in parallel,

meaning all three covered the same topic over the same time span. The reason for collecting the three short-term parallel datasets was to demonstrate that our approach identified the same events in all three smoothed line graphs with minimal bias. The topic that we chose for our 13-hour datasets was based on news that was trending and would likely fade from public interest more quickly. Empirically, we found that topics in popular culture often tended to demonstrate a shorter and more intense public interest based on the news being circulated. We proceeded with the rationale that topics with shorter and more intense cycles of interest would likely produce more well-formed event peaks. With this rationale in mind, we selected the Twitter API keywords "Rolling+Stones ." This keyword search related to a story that was circulating in the news about the musical group The Rolling Stones. The band had issued a cease and desist order to the Donald Trump campaign ordering them to stop using their music at political rallies. Each of the three parallel "Rolling+Stones" samples had 8,000 tweets per dataset.

Long-Term Dataset Collection

In our original research proposal, we proposed two datasets. The original topics were "cybersecurity" and "#DowJones." We had to forego these two topics because neither keyword search produced a sufficient quantity of tweets at the Twitter platform's API. As a result, we monitored the news using outlets that were seen as the most reliable for news and free of bias. According to the Media Bias Chart at https://www.adfontesmedia.com/ , abcnews and Reuters were two of the least biased sources for news, so we used these outlets to find trending topics (Media Bias Chart, 2020). We started collection on June 15, 2020, and on this date two stories were circulating heavily. The first was the Trump rally in Tulsa, Oklahoma, which took place

on June 20, 2020. The second story concerned the shooting of Rayshard Brooks by a police officer in Atlanta, Georgia. On both of these topics, we collected tweets every day for a period of 14 days. For each individual one-day period, we collected throughout the day to ensure that the entire 24-hour period had been represented. At the end of the 14-day collection period, the "Tulsa+Rally" dataset had a total of 14,000 tweets. The second dataset (we used keywords "Atlanta+Protests") had a total of 12,203 tweets.

Known Events as Reference Markers

We decided to include an additional measure to evaluate the occurrence of events in each of the two 14-day datasets. For each of the two datasets, we inserted known events (represented by a vertical line) into the y-axis timeline as points of reference to evaluate the juxtaposition of event peaks with the known events. In the "Tulsa+Rally" dataset, we inserted the vertical line in the middle of the time index that corresponded to June 20, 2020. Collection began on June 15, 2020, so the point of reference was five days from the start of the collection. In the "Atlanta+Protests" dataset, we inserted two points of reference. The first was on the third day of collection, June 17, 2020. At this point of reference, charges were formally brought against the two officers who shot Rayshard Brooks. The second point of reference for the dataset was on the ninth day of collection, June 23, 2020. This second point of reference corresponded to a planned televised funeral for Brooks. Since both event reference points in the "Atlanta+Protests" occurred in the middle of the day, we inserted the vertical lines in the middle of both time period indexes.

Results of RollingStones Parallel Datasets

The results of the three parallel "Rolling+Stones" samples validated our original assumption that the *NW* approach would identify the same events in all instances of the time-period. In all three datasets, two events were identified. We used the Q_3 quartile value for the dataset distribution as a threshold line to delineate the separation of average *NW* values from outliers. Ideally, event peaks formed beyond the Q_3 line, as events were related to the existence of outliers in a dataset. However, event peaks could form tangent to or below the Q_3 line. The location of where the event peak falls is related to the magnitude of the event. Event magnitude evaluation is a topic that we plan to pursue in later research. The first event in the "Rolling+Stones" samples fell below the Q_3 line however the second event developed a well-formed peak above Q_3 at hour 6. We qualified the first event in this dataset series as "less significant" than the second event. Since evaluation of event magnitude is a topic for later research, we will not be able to precisely assess the quantitative differences between the two events at this point. The cumulative bias in *NW* values between all three datasets was minimal.

Results of the "Tulsa+Rally" NW Dataset

When we analyzed the first of our 14-day datasets, the "Tulsa+Rally" *NW* dataset, we identified three events in the time-period trajectory. All three events had their peak formations fall above the Q_3 outlier boundary line. The first event peak barely broke the plane of the Q_3 horizontal line, registering an *NW* score of approximately 2.10 units. The second event was the most significant event in the time-period covered for the dataset. The event peak formed at approximately 4.7 *NW* units. There were two issues with the "Tulsa+Rally" *NW* dataset that were worth mentioning. First, the vertical reference line

(the 8 pm Tulsa, Oklahoma rally) occurred in a trough in the smoothed line trajectory. The day following the known reference there was the beginning of steep incline toward the most significant event peak in the time-period. We interpreted this juxtaposition of reference point to slope formation as a correlation between the two entities.

The second issue of note with the "Tulsa+Rally" *NW* dataset was the significantly lower overall *NW* scores. Lower *NW* scores were caused in part by smaller $DC(t_i)$ scores and perhaps larger E_s scores. Lower $DC(t_i)$ scores suggested an absence (or reduced number) of larger subnetworks of highly connected users. Higher E_s scores suggested a higher diversity among the users in the discussion stream. The p value in the *NW* algorithm served as a qualitative variable by which we could distribute the $DC(t_i)$ and E_s scores over a segmented temporal range for evaluation. For the "Tulsa+Rally" dataset we chose a single day as our value of p. The hundreds of lower scores associated with each 24-hour period of the dataset averaged together to articulate a 14-day period where the *NW* magnitude ranged from approximately zero to five when plotted as a smoothed line graph. Based on this available evidence we concluded that message spread in this discussion stream was conducted through a larger, more diverse group, composed of many smaller, more dispersed subnetworks of users.

Results of the "Tulsa+Rally" Keyword Frequency Dataset

Next, we analyzed the smoothed linear trajectory created by the "Tulsa+Rally" keyword frequency dataset. In the 14-day trajectory we found only one region that we could reasonably classify as an "event." It was not a full formed peak like the three events that were identified in the *NW* dataset. Starting midway through day 10 the trajectory begins an incline. At the end of day 14 the trajectory terminated at a plateau.

This plateau we identified as the graph's one and only event for the time-period. At day 8 there was a very subtle ripple in the trajectory, but it was not enough that we could reasonably call it an event. Overall, the "Tulsa+Rally" keyword frequency dataset had only a small amount of variance in it with regards to keyword frequency values.

We compared the keyword frequency graph with the *NW* graph and we made two observations. First, we saw a mild correlation between the event plateau at the end of the keyword trajectory and the third event peak in the *NW* dataset. The increases in keyword frequency and *NW* between hours 11 and 12 were in sync. However, the *NW* trajectory terminated its path in a trough, while the keyword trajectory remained elevated as a plateau. This was where the lone similarity ended. The second observation that we made concerned the overall efficacy of the keyword frequency technique itself. The keyword technique was designed to be sensitive to increases and decreases in word occurrence to identify events. We hypothesized that the *NW* approach was able to detect events that the keyword frequency method could not. The *NW* trajectory had three well-formed peaks where there was little to no corroboration in the keyword trajectory. More testing will be required during future research to validate this observation.

Results of the "Atlanta+Protests" NW Dataset

We found three event peaks in the 14-day trajectory of the "Atlanta+Protests" *NW* dataset. The most significant of the three events took place on day 12 of the time-period. It had the highest *NW* score at approximately 50 units. Event number two barely broke the plane of the Q_3 boundary with an *NW* score of approximately 15.3. The first event in the time period did not pass beyond the Q_3 boundary line but was the longest lasting event in the time-period with a duration of four days. Event one had an *NW* score of

approximately 12 units. We made two observations concerning the "Atlanta+Protests" *NW* dataset. The first observation concerned the significantly larger *NW* scores in the dataset. The second observation dealt with the juxtaposition of known event reference lines with event peaks.

The *NW* scores for the "Atlanta+Protests" dataset were significantly higher than those of the "Tulsa+Rally" dataset. This increase in scores was attributed to a greater number of users with high $DC(t_i)$ scores and average E_s scores. The reduced range of E_s scores suggested a somewhat smaller diversity of users participating in the discussion stream. While the "Tulsa+Rally" discussion stream had E_s scores that fell within the "average" range, the majority of them were higher (greater than 2.0). The higher average E_s scores indicated greater participation in the discussion. The "Atlanta+Protests" discussion stream had a smaller amount of participation (E_s scores ranged between 1.4 to 2.041). The larger overall $DC(t_i)$ scores suggested there were more interconnected users in larger subnetworks. The largest $DC(t_i)$ score in the "Atlanta+Protests" dataset was a 424.84, which suggested this user was the most influential person in the discussion stream with regards to message spread. He or she was likely the center of the largest user subnetwork. Message circulation in this discussion stream was more efficient and more widespread at a quicker rate than the "Tulsa+Rally" dataset. The higher *NW* scores are indicative of this.

In the "Tulsa+Rally" dataset we made the observation that the vertical line reference marker for the June 20, 2020 rally was rooted in a valley immediately prior to a trajectory ascent toward an event peak. This suggested to us that there was a possible causal relationship between the reference point and the formation of the subsequent

event. In the "Atlanta+Protests" *NW* dataset there were two vertical line reference markers. The first reference point was rooted slightly before the cresting of the first event peak. The second reference point was rooted in sync with the cresting of the second event peak. The juxtaposition of these two reference points in relation to the two event peaks suggested to us that the two known events correlating to the reference points contributed to the formation of the event peaks but did not cause them. In both cases the levels of *NW* were already increasing prior to the occurrence of the known events.

Results of the "Atlanta+Protests" Keyword Frequency Dataset

When we viewed the results of the "Atlanta+Protests" keyword frequency graph, we noticed that there was a lot of congruence between the trajectories of the two graphs. In both graphs, three events were identified. The third event peak in the "Atlanta+Protests" graph was the most significant of the three. This result was in agreement between the keyword frequency and *NW* approaches. Also in agreement was the fact that the first two event peaks in the time period crested beneath the Q_3 boundary line. A third item that was in agreement between the two "Atlanta+Protests" graphs was the fact that the first event in the time period lasted for a duration of five days. The results of the "Atlanta+Protests" keyword frequency dataset validated the events that were identified by the *NW* method.

Results of the "Russian+Disinformation" NW and Keyword Frequency Datasets

The "Russian+Disinformation" samples were part of our short-term dataset collection. The RollingStones datasets consisted of 13-hour time periods. The "Russian+Disinformation" samples consisted of 28-hour periods. By approaching tweet collection using the method we documented in this research, we successfully

implemented event detection from both short and long term perspectives (i.e. 13-hour, 28-hour, 14-day). When we compared the results of the two different approaches for the "Russian+Disinformation" datasets, we noticed first that they both captured the same three events over the time period (p = 1 hour, 28 hour time span). The first captured event peak for both methods lasted approximately the same duration, i.e. 6 hours. Also, in both approaches, the vertical line known event reference marker occurred just prior to the ascent of the first event peak in the time-period. This is where the similarities ended. The *NW* approach for the time-period identified the second peak as the most significant event. The trajectory for the keyword frequency method identified the first event as the most significant in the time-period. The *NW* dataset had a rather large variance in its distribution, i.e. 683.43 as seen previously in Table 10. The maximum value for *NW* in the dataset was 217.94. The large variance size suggested that the conditions were favorable for the formation of more significant peak formations due to the existence of more outliers.

The keyword frequency dataset had a max value of 13 keywords occurring in a tweet during the 28-hour time-period. The variance was 15.99 with a Q_3 value of 10. None of the three event peaks in the keyword frequency trajectory breached the Q_3 boundary. The smaller variance and an unbroken Q_3 boundary suggested an absence of outliers in its distribution, which in turn would lead to an absence of significant peaks. The results of the two approaches suggested to us that the *NW* method was able to capture information about events that were not fully captured by keyword occurrence. The same three events were universally captured, however the relative amplitudes of the three events differed between the two trajectories. The keyword frequency trajectory identified

the latter two events as lesser in significance. The NW trajectory identified the second event peak as very significant. These observations need to be tested further with additional datasets in future research to corroborate our findings.

Contribution

The contributions of this research to the body of knowledge are threefold. Our first contribution was the creation of a new metric called Newsworthiness (*NW*) by integrating two existing SNA metrics, $DC(t_i)$ and E_s. The *NW* metric quantitatively identified events in a social media discussion stream by evaluating the message spreading influence and user diversity of participating users over a defined period of time. Currently, the magnitude of events is determined by evaluating *significance* from least *NW* values to greatest *NW* values. An event with the highest event peak is considered the "most significant" event in the time-period. In future research we will add to the existing event detection algorithm to provide a more concrete method of evaluating an event's magnitude.

Our second contribution was the use of quartiles to evaluate dataset distributions for outliers in the context of analyzing *NW* to identify event peaks. Ideally events in a dataset distribution formed above the Q_3 boundary line, as that region was where outliers in a dataset were found. As it was evidenced in our experiments, event peaks could form beneath the Q_3 boundary. Since peak formation occurred above and below the Q_3 boundary, Q_3 threshold requirement was not deemed a rigid rule to follow. It served more as a guideline for us. The Q_3 boundary allowed us (along with *NW* score) to evaluate the magnitude of dataset distributions by tracking where peaks formed in the trajectory. As a general guideline, using Q_3 as our point of reference, a shorter peak was consistent with

an event of lesser magnitude. A taller peak suggested a greater magnitude. A peak that formed above Q_3 was consistent with the more ideal definition of an event since the peak was formed in the dataset region where outliers existed.

The third contribution of our research was the use of $DC(t_i)$ and E_s to analyze user activity in a Twitter discussion stream. As we previously discussed in our experiment findings, high levels of $DC(t_i)$ in a group of networked users suggested a greater level of interconnectedness between the users and a higher level of messaging activity. The $DC(t_i)$ metric *positively* affects the message spreading influence and levels of activity among the users in a discussion stream. A higher level of diversity (E_s) *negatively* affects message spreading influence and activity in a discussion stream. To expand on this idea further, if there are more people in a discussion, a user's spreading influence would need to be greater to reach the increased number of people in the discussion stream. For this reason, greater diversity adversely impacts the messaging activity among a group of users.

In addition to the user activity data we can derive from E_s and $DC(t_i)$, the p value (individual unit of time) allows us to qualitatively evaluate user activity from a broader, more protracted perspective. A smaller p value (e.g. hour, minute) allows for a more nuanced micro view of a dataset where rapid changes over a more discrete time-period are the focus of study, for example a Twitter hashtag that goes viral. In our case, a larger p value (1 day) allows us to evaluate subtle changes in user activity that develop over longer periods (week, month). The use of E_s and $DC(t_i)$ also provides us with a broader framework which we will further explore in later research. It is the ability to articulate the topology of a discussion stream and sample its composition.

Future Work

When we began this research, the scope was significantly broader than the breadth it currently maintains. We set aside a substantial amount of work we had previously completed in the interests of refining our study and not submitting to research creep. With this in mind, we decided upon the four most important goals we will pursue for future research. Our first goal is the use of machine learning algorithms to identify events. We wish to implement and test four classifiers which are ubiquitous in the research literature. These classifiers are Support Vector Machine (SVM), XGBoost, Neural Network, and Random Forest. When we implement machine learning as part of our research infrastructure, it will provide us with a concrete method of evaluating the accuracy of event identification. We will use the *accuracy* metric from confusion matrices to give us feedback. We have used all four classifiers in empirical testing, and each has its own inherent performance strengths and weaknesses.

Our second future research goal is to use SA as an additional attribute in our event detection algorithm. The NW metric will be the combined integration of $DC(t_i)$, E_s, and SA. The SA attribute can include categorical emotions in its evaluation, such as anger, fear, and joy. It can also include numeric evaluation, such as magnitudes of negativity or positivity. The implementation depends on the SA lexicon that is used for analysis. The NRC lexicon, for example, categorizes message samples by emotion type: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Each of the emotion categories also has a magnitude which can be measured. The AFINN lexicon measures message samples using a numeric scale that ranges from negative five to positive five.

By implementing one or more SA lexicons in our future research, we will be able to evaluate two things with regards to events in a discussion stream. The first is the determination of what emotion (anger, joy) or sentiment valence (positive versus negative) is influencing the formation of an event. We could, for example, collect a sample dataset from a college sports discussion forum after a team wins the national championship. An analysis of the resulting time-period might reveal measurable amounts of *joy* and *surprise*. The second metric we could determine by using SA in a discussion stream is the magnitude of an event. The magnitude could be measured, for example, by using the AFINN lexicon to evaluate *how positive* or *how negative* an event was. If we identified an event peak in a time-period using an ensemble of $DC(t_i)$, E_s, and SA, the SA attribute could be isolated to measure the magnitude of the event peak. A negative event, such as a peak that results after a natural disaster, could measure from negative one to negative 5 on the AFINN scale. Different SA lexicons could also be combined into an ensemble to exploit the benefits of both algorithms.

The third goal of future research is to refine and scale the *NW* metric so that it has a common numeric reading after the $DC(t_i)$ and E_s attributes are integrated. Currently, when *NW* is graphed as a smoothed line trajectory, event peaks can form whether the values of NW are low (the "Tulsa+Rally" dataset) or high (the "Atlanta+Protests" dataset). Our goal is to have a unified metric. However, if we pursue machine learning classifiers as part of our research infrastructure, this goal will not have as much relevance. Currently, the only efficient method of evaluating the inner workings of the *NW* metric is to take each of the two individual attributes and analyze them separately with respects to the *NW* smooth line trajectory. By including SA as an additional attribute

in the algorithm ensemble, it will make the NW metric more versatile as a numeric scoring tool. We touched upon our fourth goal in our *contributions* discussion above. We discussed how the Newsworthiness subcomponents of $DC(t_i)$ and E_s could be used to qualitatively evaluate user activity. For our fourth research goal we will take these two metrics and convert them into an evaluative framework that will use to quantify the topology and composition of a discussion stream as concrete values.

Appendices

Appendix A: List of Acronyms

List of Acronyms Used in Document

Acronyms

1. **ANN** – Artificial Neural Network
2. **API** – Application Programming Interface
3. **DTM** – Document Term Matrix
4. **LDA** – Latent Dirichlet Allocation
5. **OSN** – Online Social Network
6. **RF** – Random Forest
7. **SNA** – Social Network Analysis
8. **SVM** – Support Vector Machine

Appendix B: List of Variable Names

List of Variable Names Used in Document

Variables

1. T – Discussion Stream
2. U – Set of users
3. $DC(t_i)$ – Diffusion Centrality
4. E_s – Shannon Entropy
5. NW - Newsworthiness
6. A – Sparse matrix
7. Pr – Probability matrix
8. Q_3 – Third quartile

References

Abdelhaq, H., Gertz, M., & Sengstock, C. (2013, November). Spatio-temporal characteristics of bursty words in Twitter streams. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 194-203).

Abramowitz, A., & McCoy, J. (2019). United States: Racial Resentment, Negative Partisanship, and Polarization in Trump's America. *The ANNALS of the American Academy of Political and Social Science*, *681*(1), 137-156.

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *algorithms*, *4*(5), C5.

Alarcão, A. L. L., & Neto, M. S. (2016). Actor centrality in network projects and scientific performance: an exploratory study. *RAI Revista de Administração e Inovação*, *13*(2), 78-88.

Alarifi, A., Alsaleh, M., & Al-Salman, A. (2016). Twitter turing test: Identifying social machines. *Information Sciences*, *372*, 332-346.

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433-443.

Alkhouli, A., Vodislav, D., & Borzic, B. (2014, October). Continuous top-k processing of social network information streams: a vision. In *International Workshop on Information Search, Integration, and Personalization* (pp. 35-48). Springer, Cham.

Alp, Z. Z., & Öğüdücü, Ş. G. (2018). Identifying topical influencers on twitter based on user behavior and network topology. *Knowledge-Based Systems*, *141*, 211-221.

Amato, F., Moscato, V., Picariello, A., Piccialli, F., & Sperlí, G. (2018). Centrality in heterogeneous social networks for lurkers detection: An approach based on hypergraphs. *Concurrency and Computation: Practice and Experience*, *30*(3), e4188.

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, *51*(2), 339-367.

An, W., & Liu, Y. H. (2016). keyplayer: An R Package for Locating Key Players in Social Networks. *R Journal*, *8*(1).

An, Y., Ding, S., Shi, S., & Li, J. (2018). Discrete space reinforcement learning algorithm based on support vector machine classification. *Pattern Recognition Letters*, *111*, 30-35.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, *31*(1), 132-164.

Bamakan, S. M. H., Nurgaliev, I., & Qu, Q. (2019). Opinion leader detection: A methodological review. *Expert Systems with Applications*, *115*, 200-222.

Barbieri, F., & Saggion, H. (2014, April). Modelling irony in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 56-64).

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, *19*(6), 275.

Bingol, K., Eravci, B., Etemoglu, C. O., Ferhatosmanoglu, H., & Gedik, B. (2016). Topic-based influence computation in social networks under resource constraints. *IEEE Transactions on Services Computing*.

Bonchi, F., De Francisci Morales, G., & Riondato, M. (2016, April). Centrality measures on big graphs: Exact, approximated, and distributed algorithms. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 1017-1020). International World Wide Web Conferences Steering Committee.

Boyd, D., Golder, S., & Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences* (pp. 1-10). IEEE.

Braga, P. L., Oliveira, A. L., & Meira, S. R. (2008, March). A GA-based feature selection and parameters optimization for support vector regression applied to software effort estimation. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1788-1792). ACM.

Bramoullé, Y., & Genicot, G. (2018). Diffusion centrality: Foundations and extensions.

Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, *11*(1), 13-41.

Caballero, A., Niguidula, J. D., & Caballero, J. M. (2017, November). Analysis and Visualization of University Twitter Feeds Sentiment. In *International Conference on Big Data Technologies and Applications* (pp. 132-145). Springer, Cham.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010, July). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining* (pp. 1-10).

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010, May). Measuring user influence in twitter: The million follower fallacy. In *fourth international AAAI conference on weblogs and social media*.

Choudhury, S., & Alani, H. (2014). Personal life event detection from social media.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811-824.

Cohen, M. (2020, June 17). Fired Atlanta officer who killed Rayshard Brooks is charged with felony murder. Retrieved July 22, 2020, from https://www.politico.com/news/2020/06/17/atlanta-police-rayshard-brooks-326540

Conte, D., Foggia, P., Sansone, C., & Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, *18*(03), 265-298.

Cordeiro, M. (2012, January). Twitter event detection: combining wavelet analysis and topic inference summarization. In *Doctoral symposium on informatics engineering* (pp. 11-16).

Cui, L., Zhang, X., Zhou, X., & Salim, F. (2016, September). Topical event detection on Twitter. In *Australasian Database Conference* (pp. 257-268). Springer, Cham.

Das, K., & Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*, *5*(2), 1301-1309.

Davis, L. S., & Abdurazokzoda, F. (2016). Language, culture and institutions: Evidence from a new linguistic dataset. *Journal of Comparative Economics*, *44*(3), 541-561.

Davis, T. A., & Hu, Y. (2011). The University of Florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, *38*(1), 1-25.

Dehmer, M., & Mowshowitz, A. (2011). A history of graph entropy measures. *Information Sciences*, *181*(1), 57-78.

Devi, D. N., Kumar, C. K., & Prasad, S. (2016, February). A feature-based approach for sentiment analysis by using support vector machine. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)* (pp. 3-8). IEEE.

Dey, A. (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174-1179.

Diederich, A., & Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. *Journal of Mathematical Psychology*, *47*(3), 304-322.

Di Eugenio, B., Green, N., & Subba, R. (2013, September). Detecting life events in feeds from twitter. In *2013 IEEE Seventh International Conference on Semantic Computing* (pp. 274-277). IEEE.

Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C., & Kuncheva, L. I. (2015). Random balance: ensembles of variable priors classifiers for imbalanced data. *Knowledge-Based Systems*, *85*, 96-111.

Docs - Twitter Developers. (n.d.). Retrieved March 1, 2020, from https://developer.twitter.com/en/docs

Domínguez, D. R., Redondo, R. P. D., Vilas, A. F., & Khalifa, M. B. (2017). Sensing the city with Instagram: Clustering geolocated data for outlier detection. *Expert systems with applications*, *78*, 319-333.

Dou, W., Wang, X., Ribarsky, W., & Zhou, M. (2012, October). Event detection in social media data. In *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content* (pp. 971-980).

Du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing*, *73*(10-12), 2047-2060.

Du, N., Song, L., Woo, H., & Zha, H. (2013, April). Uncover topic-sensitive information diffusion networks. In *Artificial Intelligence and Statistics* (pp. 229-237).

Earle, P. S., Bowden, D. C., & Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, *54*(6).

Edla, D. R., Mangalorekar, K., Dhavalikar, G., & Dodia, S. (2018). Classification of EEG data for human mental state analysis using Random Forest Classifier. *Procedia computer science*, *132*, 1523-1532.

Efstathiades, H., Antoniades, D., Pallis, G., Dikaiakos, M. D., Szlávik, Z., & Sips, R. J. (2016, December). Online social network evolution: Revisiting the Twitter graph.

In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 626-635). IEEE.

El Asri, I., Kerzazi, N., Uddin, G., Khomh, F., & Idrissi, M. J. (2019). An empirical study of sentiments in code reviews. *Information and Software Technology*, *114*, 37-54.

Evans, L., Owda, M., Crockett, K., & Vilas, A. F. (2019). A methodology for the resolution of cashtag collisions on Twitter–A natural language processing & data fusion approach. *Expert Systems with Applications*, *127*, 353-369.

Figueiredo, F., & Jorge, A. (2019). Identifying topic relevant hashtags in Twitter streams. *Information Sciences*, *505*, 65-83.

Fredericks, K. A., & Durland, M. M. (2005). The historical evolution and basic concepts of social network analysis. *New directions for evaluation*, *2005*(107), 15-23.

Garcia, D. (2017). Leaking privacy and shadow profiles in online social networks. *Science advances*, *3*(8), e1701172.

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., & Wolff, E. (2018). Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. *IEEE Geoscience and Remote Sensing Letters*, *15*(4), 607-611.

Ghosh, R., Surachawala, T., & Lerman, K. (2011). Entropy-based classification of 'retweeting' activity on twitter. *arXiv preprint arXiv:1106.0346*.

Glasgow, K., & Fink, C. (2013, April). Hashtag lifespan and social networks during the London riots. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 311-320). Springer, Berlin, Heidelberg.

Grando, F., Noble, D., & Lamb, L. C. (2016, December). An analysis of centrality measures for complex and social networks. In *2016 IEEE Global Communications Conference (GLOBECOM)* (pp. 1-6). IEEE.

Guille, A., & Favre, C. (2015). Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach. *Social Network Analysis and Mining*, *5*(1), 18.

Gunasekara, R. C., Mehrotra, K., & Mohan, C. K. (2015). Multi-objective optimization to identify key players in large social networks. *Social Network Analysis and Mining*, *5*(1), 21.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467-483.

Gurajala, S., White, J. S., Hudson, B., & Matthews, J. N. (2015, July). Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. In *Proceedings of the 2015 International Conference on Social Media & Society* (pp. 1-7).

Gurajala, S., White, J. S., Hudson, B., Voter, B. R., & Matthews, J. N. (2016). Profile characteristics of fake Twitter accounts. *Big Data & Society*, *3*(2), 2053951716674236.

Hajibagheri, A., Alvari, H., Hamzeh, A., & Hashemi, S. (2012, August). Community detection in social networks using information diffusion. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 702-703). IEEE.

Han, Y., & Tang, J. (2015, August). Probabilistic community and role model for social networks. In *Proceedings of the 21th ACM SIGKDD international conference on Knowledge Discovery and Data Mining* (pp. 407-416).

Hasan, M., Orgun, M. A., & Schwitter, R. (2016, November). TwitterNews+: a framework for real time event detection from the Twitter data stream. In *International Conference on Social Informatics* (pp. 224-239). Springer, Cham.

Heaukulani, C., & Ghahramani, Z. (2013, February). Dynamic probabilistic models for latent feature propagation in social networks. In International Conference on Machine Learning (pp. 275-283).

Himelboim, I., Smith, M. A., Rainie, L., Shneiderman, B., & Espina, C. (2017). Classifying Twitter topic-networks using social network analysis. *Social Media+ Society*, *3*(1), 2056305117691545.

Ifo, S. A., Moutsambote, J. M., Koubouana, F., Yoka, J., Ndzai, S. F., Bouetou-Kadilamio, L. N. O., ... & Mbemba, M. (2016). Tree species diversity, richness, and similarity in intact and degraded forest in the tropical rainforest of the Congo Basin: case of the forest of Likouala in the Republic of Congo. *International Journal of Forestry Research, 2016*.

Ikeda, Y., Hasegawa, T., & Nemoto, K. (2010). Cascade dynamics on clustered network. In *Journal of Physics: Conference Series* (Vol. 221, No. 1, p. 012005). IOP Publishing.

Injadat, M., Salo, F., & Nassif, A. B. (2016). Data mining techniques in social media: A survey. *Neurocomputing*, *214*, 654-670.

Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and*

1st SNA-KDD 2007 workshop on Web mining and social network analysis (pp. 56-65).

Joarder, A. H., & Firozzaman, M. (2001). Quartiles for discrete data. *Teaching Statistics*, *23*(3), 86-89.

Joyal, C. C., Cossette, A., & Lapierre, V. (2015). What exactly is an unusual sexual fantasy?. *The journal of sexual medicine*, *12*(2), 328-340.

Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F. K., & Papageorgiou, H. (2012). Entropy analysis of word-length series of natural language texts: Effects of text language and genre. *International Journal of Bifurcation and Chaos*, *22*(09), 1250223.

Kanakaraj, M., & Guddeti, R. M. R. (2015, March). NLP based sentiment analysis on Twitter data using ensemble classifiers. In *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)* (pp. 1-5). IEEE.

Kang, C., Kraus, S., Molinaro, C., Spezzano, F., & Subrahmanian, V. S. (2016). Diffusion centrality: A paradigm to maximize spread in social networks. *Artificial Intelligence*, *239*, 70-96.

Kang, C., Molinaro, C., Kraus, S., Shavitt, Y., & Subrahmanian, V. S. (2012, August). Diffusion centrality in social networks. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 558-564). IEEE.

Karl, A., Wisnowski, J., & Rushing, W. H. (2015). A practical guide to text mining with topic extraction. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(5), 326-340.

Khan, M. A., & Salah, K. (2018). IoT security: Review, blockchain solutions, and open challenges. *Future Generation Computer Systems*, *82*, 395-411.

Kim, J., & Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, *38*(1), 86-96.

Kirka, D. (2020, June 28). Rolling Stones threaten to sue Trump over using their songs. Retrieved July 22, 2020, from https://abcnews.go.com/Entertainment/wireStory/rolling-stones-threaten-sue-trump-songs-71497519

Klein, B., Castanedo, F., Elejalde, I., Lopez-de-Ipina, D., & Nespral, A. P. (2013). Emergency event detection in twitter streams based on natural language processing. In *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction* (pp. 239-246). Springer, Cham.

Kolchyna, O., Souza, T. T., Treleaven, P. C., & Aste, T. (2016, December). A framework for Twitter events detection, differentiation and its application for retail brands. In *2016 Future Technologies Conference (FTC)* (pp. 323-331). IEEE.

Krishna, M. H., Rahamathulla, K., & Akbar, A. (2017, March). A feature based approach for sentiment analysis using SVM and coreference resolution. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 397-399). IEEE.

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008, August). A few chirps about twitter. In *Proceedings of the first workshop on Online social networks* (pp. 19-24).

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600).

Kywe, S. M., Hoang, T. A., Lim, E. P., & Zhu, F. (2012, December). On recommending hashtags in twitter networks. In *International conference on social informatics* (pp. 337-350). Springer, Berlin, Heidelberg.

Langford, E. (2006). Quartiles in elementary statistics. *Journal of Statistics Education*, *14*(3).

Laniado, D., & Mika, P. (2010, November). Making sense of twitter. In *International Semantic Web Conference* (pp. 470-485). Springer, Berlin, Heidelberg.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85-96.

Lappas, T., Vieira, M. R., Gunopulos, D., & Tsotras, V. J. (2012). On the spatiotemporal burstiness of terms. *arXiv preprint arXiv:1205.6695*.

Lee, R., & Sumiya, K. (2010, November). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks* (pp. 1-10).

Lee, S., Song, J., & Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computer Information Systems*, *51*(1), 1-10.

Levenson, J. C., Shensa, A., Sidani, J. E., Colditz, J. B., & Primack, B. A. (2016). The association between social media use and sleep disturbance among young adults. *Preventive medicine, 85*, 36-41.

Li, D., Zhang, S., Sun, X., Zhou, H., Li, S., & Li, X. (2017). Modeling Information Diffusion over Social Networks.

Li, L., Situ, R., Gao, J., Yang, Z., & Liu, W. (2017, October). A hybrid model combining convolutional neural network with xgboost for predicting social media popularity. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 1912-1917).

Li, Y., Zhang, G., Feng, Y., & Wu, C. (2015). An entropy-based social network community detecting method and its application to scientometrics. *Scientometrics*, *102*(1), 1003-1017.

Lokeswari, N., & Rao, B. C. (2016). Artificial neural network classifier for intrusion detection system in computer network. In *Proceedings of the Second International Conference on Computer and Communication Technologies* (pp. 581-591). Springer, New Delhi.

Malliaros, F. D., & Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports*, *533*(4), 95-142.

Mason, J. (2020, June 21). Trump slams protests, defends pandemic response as Tulsa crowd underwhelms. Retrieved July 17, 2020, from https://www.reuters.com/article/us-usa-election-trump/trump-slams-protests-defends-pandemic-response-as-tulsa-crowd-underwhelms-idUSKBN23R0G5

Matei, S. A., & Bruno, R. J. (2015). Pareto's 80/20 law and social differentiation: A social entropy perspective. *Public Relations Review*, *41*(2), 178-186.

McKay, R. (2020, June 15). Family demands justice after Atlanta police fatally shoot Rayshard Brooks in the back. Retrieved July 17, 2020, from https://www.reuters.com/article/us-minneapolis-police-atlanta/family-demands-justice-after-atlanta-police-fatally-shoot-rayshard-brooks-in-the-back-idUSKBN23M1VB

McKay, R. (2020, June 18). Atlanta police officer charged with murder in shooting death of Rayshard Brooks. Retrieved July 22, 2020, from https://www.reuters.com/article/us-minneapolis-police-atlanta/atlanta-police-officer-charged-with-murder-in-shooting-death-of-rayshard-brooks-idUSKBN23O1IO

Media Bias Chart. (2020, June 19). Retrieved July 25, 2020, from https://www.adfontesmedia.com/

Mei, Q., & Zhai, C. (2005, August). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 198-207). ACM.

Mertler, C. A., & Reinhart, R. V. (2016). *Advanced and multivariate statistical methods: Practical application and interpretation*. Routledge.

Moghaddam, S., & Ester, M. (2012, October). On the design of LDA models for aspect-based opinion mining. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 803-812).

Nairac, A., Corbett-Clark, T. A., Ripley, R., Townsend, N. W., & Tarassenko, L. (1997). Choosing an appropriate model for novelty detection.

Nayak, A. S., Kanive, A. P., Chandavekar, N., & Balasubramani, R. (2016). Survey on pre-processing techniques for text mining. *International Journal Of Engineering And Computer Science, ISSN*, 2319-7242.

Neubig, G., & Duh, K. (2013, March). How much is said in a tweet? A multilingual, information-theoretic perspective. In *2013 AAAI Spring Symposium Series*.

Neves-Silva, R., Gamito, M., Pina, P., & Campos, A. R. (2016). Modelling influence and reach in sentiment analysis. *Procedia CIRP*, *47*, 48-53.

Newman, M. E. (2004). Analysis of weighted networks. *Physical review E*, *70*(5), 056131.

Nusselder, A. (2013). Twitter and the personalization of politics. *Psychoanalysis, Culture & Society*, *18*(1), 91-100.

Omenzetter, P., Brownjohn, J. M. W., & Moyo, P. (2004). Identification of unusual events in multi-channel bridge monitoring data. *Mechanical Systems and Signal Processing*, *18*(2), 409-430.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, *28*(6), 441-453.

Papadimitriou, C., Karamanos, K., Diakonos, F. K., Constantoudis, V., & Papageorgiou, H. (2010). Entropy analysis of natural language written texts. *Physica A: Statistical Mechanics and its Applications*, *389*(16), 3260-3266.

Patil, L. H., & Atique, M. (2013, February). A novel approach for feature selection method TF-IDF in document clustering. In *2013 3rd IEEE International Advance Computing Conference (IACC)* (pp. 858-862). IEEE.

Peng, S., Zhou, Y., Cao, L., Yu, S., Niu, J., & Jia, W. (2018). Influence analysis in social networks: A survey. *Journal of Network and Computer Applications*, *106*, 17-32.

Perera, R. D., Anand, S., Subbalakshmi, K. P., & Chandramouli, R. (2010, October). Twitter analytics: Architecture, tools and analysis. In *2010-MILCOM 2010 MILITARY COMMUNICATIONS CONFERENCE* (pp. 2186-2191). IEEE.

Peres, A. (1984). What is a state vector?. *American Journal of Physics*, *52*(7), 644-650.

Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215-249.

Pinto, S., Albanese, F., Dorso, C. O., & Balenzuela, P. (2019). Quantifying time-dependent Media Agenda and public opinion by topic modeling. *Physica A: Statistical Mechanics and its Applications*, *524*, 614-624.

Pozdnoukhov, A., & Kaiser, C. (2011, November). Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks* (pp. 1-8).

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221.

Prasad, D. V. V., & Suresh, J. (2019). A quick survey of Artificial Neural Network based face classification algorithms. *Cluster Computing*, *22*(4), 9477-9488.

Proskurnikov, A. V., & Tempo, R. (2017). A tutorial on modeling and analysis of dynamic social networks. Part I. *Annual Reviews in Control*, *43*, 65-79.

Puschmann, C., & Burgess, J. (2013). The politics of Twitter data.

Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). Privacy-preserving record linkage on large real-world datasets. *Journal of biomedical informatics*, *50*, 205-212.

Ratkiewicz, Jacob, Filippo Menczer, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani. "Traffic in social media ii: Modeling bursty popularity." In *2010 IEEE Second International Conference on Social Computing*, pp. 393-400. IEEE, 2010.

Ren, Y., Ji, D., & Ren, H. (2018). Context-augmented convolutional neural networks for twitter sarcasm detection. *Neurocomputing*, *308*, 1-7.

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362-376.

Roopa, V., & Induja, K. (2019, April). Customized Visualization of Email Using Sentimental and Impact Analysis in R. In *International Conference on Advances in Computing and Data Sciences* (pp. 144-154). Springer, Singapore.

Rosenthal, S., Farra, N., & Nakov, P. (2019). SemEval-2017 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.00741*.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 73-79.

Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., & Motoda, H. (2011, June). Learning diffusion probability based on node attributes in social networks. In *International Symposium on Methodologies for Intelligent Systems* (pp. 153-162). Springer, Berlin, Heidelberg.

Sarlan, A., Nadam, C., & Basri, S. (2014, November). Twitter sentiment analysis. In *Proceedings of the 6th International conference on Information Technology and Multimedia* (pp. 212-216). IEEE.

Schubert, E., Zimek, A., & Kriegel, H. P. (2014, April). Generalized outlier detection with flexible kernel density estimates. In *Proceedings of the 2014 SIAM International Conference on Data Mining* (pp. 542-550). Society for Industrial and Applied Mathematics.

Scott, J. (2011). Social network analysis: developments, advances, and prospects. *Social network analysis and mining*, *1*(1), 21-26.

Serrat, O. (2017). The future of social marketing. In *Knowledge solutions* (pp. 119-128). Springer, Singapore.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379-423.

Shanthamallu, U. S., Spanias, A., Tepedelenlioglu, C., & Stanley, M. (2017, August). A brief survey of machine learning methods and their sensor and IoT applications. In *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)* (pp. 1-8). IEEE.

Shih, H. C., & Liu, E. R. (2016). New quartile-based region merging algorithm for unsupervised image segmentation using color-alone feature. *Information Sciences*, *342*, 24-36.

Stamp, M. (2018). A Survey of Machine Learning Algorithms and Their Application in Information Security. In *Guide to Vulnerability Analysis for Computer Networks and Systems* (pp. 33-55). Springer, Cham.

Steakin, W., & Pereira, I. (2020, June 20). Trump refers to 'kung flu,' West Point ramp and 'sleepy Joe Biden' as he returns to campaign at Tulsa rally. Retrieved July 22, 2020, from https://abcnews.go.com/Politics/trump-heads-tulsa-return-rally-amid-pandemic-mounting/story?id=71307799

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting bots on Russian political Twitter. *Big data*, *5*(4), 310-324.

Subramani, J., & Kumarapandiyan, G. (2012). Variance estimation using quartiles and their functions of an auxiliary variable. *Ratio*, *1*, 1.

Takada, T., Miyamoto, A., & Hasegawa, S. F. (2010). Derivation of a yearly transition probability matrix for land-use dynamics and its applications. *Landscape ecology*, *25*(4), 561-572.

Thapen, N., Simmie, D., & Hankin, C. (2016). The early bird catches the term: combining twitter and news data for event detection and situational awareness. *Journal of biomedical semantics*, *7*(1), 61.

Tommasel, A., Corbellini, A., Godoy, D., & Schiaffino, S. (2016). Personality-aware followee recommendation algorithms: An empirical analysis. *Engineering Applications of Artificial Intelligence*, *51*, 24-36.

Tu, B., Yang, X., Li, N., Zhou, C., & He, D. (2020). Hyperspectral anomaly detection via density peak clustering. *Pattern Recognition Letters*, *129*, 144-149.

Tucker, E. (2020, July 29). US jabs Russia over claim of spreading virus disinformation. Retrieved July 31, 2020, from https://abcnews.go.com/Technology/wireStory/russia-rejects-accusations-spreading-virus-disinformation-72049224

Unankard, S., Li, X., Sharaf, M., Zhong, J., & Li, X. (2014, October). Predicting elections from social networks based on sub-event detection and sentiment analysis. In *International Conference on Web Information Systems Engineering* (pp. 1-16). Springer, Cham.

Using Twitter. (n.d.). Retrieved from https://help.twitter.com/en/using-twitter#tweets

Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. (2009, September). A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6214-6217). IEEE.

Vajapeyam, S. (2014). Understanding Shannon's Entropy metric for Information. *arXiv preprint arXiv:1405.2061*.

Van der Walt, E., Eloff, J. H., & Grobler, J. (2018). Cyber-security: Identity deception detection on social media platforms. *Computers & Security*, *78*, 76-89.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior research methods*, *40*(1), 61-72.

Vega, M., Pardo, R., Barrado, E., & Debán, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water research*, *32*(12), 3581-3592.

Wagner, S. M., & Neshat, N. (2010). Assessing the vulnerability of supply chains using graph theory. *International Journal of Production Economics*, *126*(1), 121-129.

Wakamiya, S., Lee, R., & Sumiya, K. (2011, November). Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks* (pp. 77-84).

Wang, S. H., Li, H. T., Chang, E. J., & Wu, A. Y. A. (2018, May). Entropy-assisted emotion recognition of valence and arousal using XGBoost classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 249-260). Springer, Cham.

Wang, X., Gerber, M. S., & Brown, D. E. (2012, April). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction* (pp. 231-238). Springer, Berlin, Heidelberg.

Wang, Y., & Goutte, C. (2017, August). Detecting Changes in Twitter Streams using Temporal Clusters of Hashtags. In *Proceedings of the Events and Stories in the News Workshop* (pp. 10-14).

Wang, Y., & Djurić, P. M. (2015). Social learning with Bayesian agents and random decision making. *IEEE Transactions on Signal Processing*, *63*(12), 3241-3250.

Weiler, A., Grossniklaus, M., & Scholl, M. H. (2015, June). Run-time and task-based performance of event detection techniques for twitter. In *International Conference on Advanced Information Systems Engineering* (pp. 35-49). Springer, Cham.

Weisstein, E. W. (2007). Adjacency matrix.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11*(4), 245-265.

Weng, J., & Lee, B. S. (2011, July). Event detection in twitter. In *Fifth international AAAI conference on weblogs and social media*.

Yoo, E., Rand, W., Eftekhar, M., & Rabinovich, E. (2016). Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. *Journal of Operations Management, 45*, 123-133.

Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets. *Computers in Human Behavior, 48*, 392-400.

Zainuddin, N., & Selamat, A. (2014, September). Sentiment analysis using support vector machine. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)* (pp. 333-337). IEEE.

Zhang, Z. K., Liu, C., Zhan, X. X., Lu, X., Zhang, C. X., & Zhang, Y. C. (2016). Dynamics of information diffusion and its applications on complex networks. *Physics Reports, 651*, 1-34.

Zhang, X., Chen, X., Chen, Y., Wang, S., Li, Z., & Xia, J. (2015). Event detection and popularity prediction in microblogging. *Neurocomputing, 149*, 1469-1480.

Zheng, X., Zeng, Z., Chen, Z., Yu, Y., & Rong, C. (2015). Detecting spammers on social networks. *Neurocomputing, 159*, 27-34.

Zhou, X., & Chen, L. (2014). Event detection over twitter social media streams. The VLDB Journal—The International Journal on Very Large Data Bases, 23(3), 381-400.

Zubiaga, A., Spina, D., Amigó, E., & Gonzalo, J. (2012, June). Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 319-320).

Zulfikar, M. T. (2019). Detection Traffic Congestion Based on Twitter Data using Machine Learning. *Procedia Computer Science, 157*, 118-124.