3-2022

Analysis of Twitter Networks to Aid Open Source Intelligence Capabilities: A Multilayer Network Approach

Austin P. Logan

Analysis of Twitter Networks to Aid Open Source Intelligence Capabilities: A Multilayer Network Approach

THESIS

Austin P. Logan, 2d Lt, USAF

AFIT-ENS-MS-22-M-146

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-22-M-146

ANALYSIS OF TWITTER NETWORKS TO AID OPEN SOURCE INTELLIGENCE CAPABILITIES: A MULTILAYER NETWORK APPROACH

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Operations Research

Austin P. Logan, BS

2d Lt, USAF

March 24, 2022

AFIT-ENS-MS-22-M-146

ANALYSIS OF TWITTER NETWORKS TO AID OPEN SOURCE INTELLIGENCE CAPABILITIES: A MULTILAYER NETWORK APPROACH

THESIS

Austin P. Logan, BS
2d Lt, USAF

Committee Membership:

LTC Phillip M. LaCasse, PhD
Chair

Brian J. Lunday, PhD
Member

AFIT-ENS-MS-22-M-146

Abstract

Open Source Intelligence using social media is a practice which provides military intelligence analysts insight into the thoughts and minds of an online population. Understanding the climate of social media in a region is imperative to any involved party because the power of connected individuals through social media cannot be underestimated. Using Social Network Analysis, user interactions on Twitter will be modeled as a weighted, directed network. Topic modeling through Latent Dirichlet Allocation uncovers the topics of discussion in Tweets and is then integrated into a multi-layer network which allows users to be connected to the conversations with which they have participated. Influential users in this network as well as highly connected groups of individuals are then discovered to provide perspective to intelligence analysts of the online landscape with which they are dealing.

The results of this research demonstrate that the inclusion of topics in the social network allows for more robust findings regarding influential users when analysts collect Tweets from a variety of discussions through the use of more general search queries. PageRank was identified as the best performing influence ranking method for this problem context, and two potential community identification methods were analyzed. Through this research, a framework for a replicable and automated process for high-level analysis of Twitter activity for use by military intelligence analysts has been developed.

To my girlfriend, family, and friends for their unwavering support and encouragement

Acknowledgements

I would like to express my sincere gratitude to my faculty advisor, LTC Phillip LaCasse, for his support and latitude throughout this thesis research process. I would also like to thank my sponsor, LTC Brian Wade, for affording me the opportunity to research these topics. Finally, I would like to thank my reader, Dr. Brian Lunday, for his guidance and oversight in preparing this research document.

Austin P. Logan

Table of Contents

Page

List of Figures

List of Tables

Table Page

ANALYSIS OF TWITTER NETWORKS TO AID OPEN SOURCE INTELLIGENCE CAPABILITIES: A MULTILAYER NETWORK APPROACH

I. Introduction

1.1 Motivation and Background

Military intelligence analysts focus on collecting data and producing insights from a number of different sources including humans, signals, imagery, and open source means. Open Source Intelligence (OSINT) refers to a broad category of intelligence gathering in which the material used is strictly open source and available to the public. OSINT originated as a means of collecting data from publicly accessible sources such as the news, television, and radio; however, given the proliferate use of social media in today's culture, analysts have more recently been provided insight to the thoughts and actions of individuals which was less practical given the sources of OSINT in the past. While OSINT is performed using each of the mentioned sources, the primary focus of this research involves data collected from social media, in particular, Twitter, as this is a growing field of interest within the military, which they are not able to perform as proficiently as others due to its more recent origination. In the current intelligence landscape, the use of social media for OSINT is becoming increasingly popular due to the ease of access to massive amounts of data. The main goals as identified by intelligence analysts when conducting social media OSINT include performing sentiment analysis on Tweets to determine public opinion, identifying the key communities, and ranking the key nodes that make up the network of Twitter users (25th Infantry Division, 2021). This research is part of a larger research effort and

will focus primarily on finding topics in Tweets and network analysis while explaining to a lesser degree the sentiment analysis which will be performed in other research.

With the amount of information that is made public on social media today, OSINT is a particularly powerful tool because it helps to understand the thoughts of a portion of the world's online vocal population. This form of intelligence collection is only limited in that it makes use of the information provided by others. The primary challenge with this task is finding the important insights which are buried among heaps of less useful information. In contrast with other intelligence collection methods, automation is crucial since it is less practical for analysts to manually search for the information they desire. Fortunately, however, due to the public nature of the data, OSINT is not exclusively conducted by military organizations as large companies and individuals make use of this form of analysis as well. The result is that advancements in the field not only occur more rapidly due to the number of interested entities but also appear in frequent research.

The focus of this research is collecting OSINT as it pertains to the region of Indonesia since this is a primary goal of intelligence units within US Indo-Pacific Command (USINDOPACOM). With the fourth largest global population and the sixth largest Twitter population, Indonesia has a significant presence on social media. It is estimated that about 77% of all internet users in Indonesia use Twitter (Carley et al., 2016), meaning that through an understanding of what is being said on the social media site, it is possible to gain insight to the country's population. Indonesian Twitter users are also extremely active, especially in comparison with other countries because they have the highest number of Tweets per user globally as observed in a 2011 study (Poblete et al., 2011).

Indonesia's online presence has come to define a major part of its national identity. Under authoritarian control until as recently as 1998, the internet emerged heavily

controlled by the state. In the decades which have followed, social media has become an integral part of the culture. The evolution of Indonesia from an authoritarian controlled state to a democratic one occurred at the same time as the growth of the internet and social media, leading researchers to associate these two strongly together (Lim, 2003). Through internet cafes, the internet merged online and physical relationships for the lower and middle classes of Indonesia, leading to the frequent political discourse on social media seen in the country today.

A limitation to analyzing the populace through social media is that Twitter users are often not a representative sample of the population. Research has shown that, in the United States, Twitter is often biased toward more affluent and urban populations (Hecht and Stephens, 2014), and this trend is likely mirrored in Indonesia.

Twitter is one of the most used social media sites in the world where users can share their own opinions or the opinions of others through a number of different methods. The most typical method is a Tweet in which one user writes a message and disperses it to all of his or her followers. These Tweets then appear on each follower's home page of the social media site which is known as a timeline. A key aspect of Twitter is the use of mentions using the "@" symbol followed by another user's name within text. By mentioning others, users engage in conversation and share each others' Tweets to their own timelines, effectively spreading the conversation to a wider audience. Similarly, Retweeting is an action which shares another user's post directly to someone's own timeline therefore making it visible to all of his or her followers. These actions are critical to track because they result in the spread of a message beyond only one user's followers and create the potential for highly visible or viral Tweets. While a user with many followers can quickly disperse a message to many people, the ability of users with significantly fewer followers to spread a message is not completely diminished due to these features. Another feature of Twitter is

the hashtag which allows users to categorize their Tweet or identify the topic of it providing them the opportunity to participate in conversations with other users on the same topics. Additionally, by listing the trending topics of conversation within the site, Twitter encourages discourse from previously unengaged users on the most popular topics which are rapidly changing.

Twitter is a particularly interesting source for intelligence gathering because its contents are more text-based than many other common platforms. Other social media sites such as Instagram and Snapchat are image-based platforms, meaning analysis is much more complicated given the challenges of image data in contrast with text. In addition, the unique features of Tweet sharing and hashtags mentioned earlier set Twitter aside from sites such as Facebook where post sharing is a less emphasized feature. The result of these features is that Twitter interactions create an interconnected network of relationships among users.

Data collection from Twitter is typically done through queries to an Application Programming Interface (API) involving some number of filters on the types of Tweets to be collected. While not explicitly covered in this research, queries are an important initial step in social media OSINT because they help to filter out much of the unwanted and uninformative data. While many search methods may be inefficient since they retain much of the unnecessary information which is sought to be overlooked, they provide the data needed to perform network analysis nonetheless. When this model is deployed, intelligence analysts will be responsible for identifying the keywords or filters which will yield results for the desired information.

While OSINT is limited in social media by the content which others choose to make public, recently many extremist political groups in addition to terrorist organizations have used social media to spread information (Alizadeh et al., 2019). Twitter seeks to block and suspend accounts found to be spreading dangerous information.

However, it is inevitable that much of the pervasive information spread by these groups will exist on the site even after the removal of the accounts originally associated with it. These residual relationships on Twitter indicate a possible application of this research: identifying potentially harmful and dangerous associations of accounts linked to discussion about extremist material. The existence of this material online is an unfortunate reality of social media and through post sharing it can become visible to those previously unaware of it.

In recent years, social media has proved to be a catalyst to massive worldwide events because it provides people with the ability to share opinions to large pools of people rapidly. The United States Government recognized recently in the Interim National Security Strategy that much of the emerging technologies in use today are "ungoverned by laws or norms designed to center rights and democratic values" (President of the U.S., 2021). With social media, people possess the ability to spread messages to a wide audience and alter public opinion. By gaining insight to this information, intelligence analysts are able to get a big picture idea of what the most frequently talked about subjects are and what the opinion of the public is on them.

A failure to understand the climate of the public through the social media lens can be extremely detrimental. When negotiating the Transatlantic Trade and Investment Partnership, the United States and European Union failed to consider public opinion and the social media uprise that followed consisted of many parties working in an organized and networked manner which prevented the deal from ever going through (Ashbrook and Zalba, 2021). Through online communities and networks, support for or against a cause can be rallied quickly, which makes understanding the structures of these communities imperative. With the massive number of users on social media sites such as Twitter, the virality of a message may appear difficult to predict. However, through keen insight of the large communities involved it can be understood how and

why certain messages are able to spread so rapidly.

1.2 Problem Statement

Since conducting OSINT using social media is a relatively new practice for lower echelon military formations, many military intelligence analysts lack the tools for automated large scale collection and analysis since it is currently done manually which is extremely time consuming (25th Infantry Division, 2021). The primary goal of this research is to automate a process to conduct network analysis on large amounts of data from Twitter. This network analysis should convey to analysts a detailed picture of the communities in an area of interest, the general sentiment toward the most common topics, and the key nodes in the communities which have the most influence over others.

Network analysis can be broken into four major steps as it concerns this research: the discovery of topics of conversation within a sample of Tweets, the construction of a network given the dataset of Tweets, the identification of influential users and topics, and the detection of communities in the network. This analysis should be conducted in a way which is automated and easily replicable by lower echelon intelligence units to support future OSINT capabilities.

1.3 Research Questions

The main question this research will answer is how a connected network of users can be made from observations of Tweets and data about them including user interactions. To accomplish this outcome, the interactions between users necessary to justify a connection will be explained as well as how the strength of this connection can be quantified. Given that some interactions with a Tweet by users do not constitute a two-way relationship, a directed network must be considered to accurately portray

interactions.

Once a sufficient network has been created, the method by which communities are identified must be decided. Community creation should take into consideration a number of factors including strength of interaction, connectedness of a community, and the topics which unite a community. The methods by which communities are identified will be defined, as well as the metrics used to evaluate the performance of community creation algorithms.

Additionally, the influential users within the network must be identified because their potential to spread a highly infectious message is pivotal to the key problems of this research. Many methods have been used previously to identify key nodes in a network and their effectiveness in this problem context must be evaluated while considering the introduction of new metrics for influence ranking.

1.4 Organization

This research will consist of a review of the related literature in Chapter II of both topic modeling and social network analysis (SNA), including methods previously used in these fields which can be leveraged to answer key research questions. Next, the methodology which will be used to conduct this analysis will be explained in Chapter III to include the creation of a network representative of Twitter users and the metrics used to evaluate this network. The results of this exploratory SNA will then be presented in Chapter IV along with analysis for their implications which pertain to the OSINT that intelligence analysts aim to conduct. Finally, conclusions of this research will be presented in Chapter V and recommendations will be made of how to best employ these insights.

II. Literature Review

The review of the related literature will span a number of fields beginning first in Section 2.1 with the significance of social media in Indonesia as well as characteristics of its usage to ensure that OSINT is practical for this region. In Section 2.2, a brief discussion of previous research in sentiment analysis will be provided since it is one of the analytical methods which will be paired with this research in the future. Next, Section 2.3 will introduce Social Network Analysis and how it should be applied in this problem context. Beginning the review of previous research related to the methodologies in this research, Section 2.4 will present the methods which have been used to identify topics in large documents and how these methods might be translated to Tweets. In addressing the key research questions, Section 2.5 will present methods for detecting anomalies and ranking user influence in social networks while Section 2.6 will introduce methods for finding communities in networks.

2.1 Indonesian Social Media Culture

The relevant literature involving the usage of social media in Indonesia will be reviewed, to include details on the complexities of the language in informal texts and the way in which users of social media in the country interact with Tweets from highly followed individuals.

As alluded to earlier in research by Lim (2003), the transition to democracy in Indonesia had a significant impact on the political discourse seen in online media. Echoing these sentiments, Hermawan (2016) describes how candidates were framed through mainstream media to manipulate votes in the 2014 Indonesian presidential election. A viral Tweet about the candidate Jokowi, which became the second most retweeted Tweet in the world that year, proclaimed him to be the winner of the

Twitter election, and, unsurprisingly he went on to win the election later in the year. Through a high volume of Tweets about the candidate circulating to millions of voters' timelines, he became a salient thought in the minds of the people resulting in his eventual election (Hermawan, 2016).

The analysis of this election framing also included the portrayal of Jokowi's opponent as "assertive" and himself as "humble and down to earth". By viewing the candidates in different frames such as experience and personality, a better understanding of the discourse regarding each candidate was possible. The research also identifies how the anonymity of Twitter pages in contrast with news outlets allowed for users to make baseless claims on candidates with less fear of retribution. This claim is furthered in research by Alatas et al. (2019), which showed that Tweets which included explicit citation of sources for information received less diffusion on the social media site (Alatas et al., 2019). In addition, it appeared that Twitter discourse on the candidates focused much more on candidate personalities than their experience or electability especially when compared with news sources. Overall, this research demonstrates how Twitter is an effective medium for shaping public opinion while also being a news source for many individuals especially in a blossoming online culture such as the one in Indonesia.

To perform many of the natural language processing tasks which will be described in detail later such as topic modeling and sentiment analysis, a firm understanding of the language used by Indonesians on social media is required. Given the fragmented island structure of the country, Indonesia is one of the most ethnically and linguistically diverse countries on Earth having the second most spoken languages amongst all countries. While Bahasa Indonesian is the official language and spoken by 94% of the population, other large dialects are prevalent in many regions of the country. Despite this, the official language is the most seen on social media accounting for

55% of posts while another 41% include Bahasa Indonesian in addition to some other language. Perhaps surprisingly, the most frequent additional language is English, exceeding even the local dialects (Arunarsirakul, 2020).

This intricacy of Indonesian social media is known as code switching in which users switch between two languages, often Indonesian and English, freely within a post. Research by Sutrisno and Ariesta (2019) has shown that this code switching occurs either as a result of word or phrase insertion, idiom insertion, and intra-sentential mixing or the inclusion of words from different languages within a phrase or clause. Through analysis of social media influencers, posters claim they utilize code switching primarily to make their message more accessible to gain further endorsements while also influencing others to adopt the practice and embrace the English language. Generally speaking, code switching with English is trendy and, through observation of influencers using the practice, others will want to emulate it as well. The prevalence of this practice is crucial to the development of natural language processing tools as both English and Indonesian must be understood due to the frequent usage of each.

In addition to the communication seen on social media, an understanding of the way in which users engage with influencers is also critical to understand the culture of Indonesian social media. While not explicitly true, celebrities are often categorized as influencers due to having a large following on social media in addition to some level of rapport with others due to their highly publicized actions outside of the site. To investigate the ways in which celebrities diffuse messages on social media, Alatas et al. (2019) gauged the impact of a variety of messages about an immunization campaign sent by celebrity accounts to determine differences in engagements and attempt to attribute this to a number of factors.

One vital finding of this research was that the origin of a Tweet from a celebrity or non-celebrity account is paramount to diffusion. The example given by Alatas

et al. (2019) compares two situations in which a user not following a celebrity sees a Tweet on his or her timeline. The first of these is when a celebrity writes a Tweet which one of his or her followers Retweet making it visible to each of their followers. The second occurs when celebrities Retweet a particular Tweet making it visible to their followers who then Retweet it similarly. The key difference here is that, in the first scenario, the penmanship of the celebrity is recognizable and provides immediate legitimacy to a message. In the second scenario, although the celebrity did endorse a message to their immediate followers, the users who see the message as a result of a celebrity follower's Retweet are unaware of the celebrity's involvement. This phenomenon attributes a 70% gain in engagements to the first type of interaction when compared with the second and demonstrates how the direct endorsement of a highly influential individual is instrumental in the diffusion of a Tweet (Alatas et al., 2019).

Additionally, as previously referenced, when assessing the impact of an attached source to a Tweet, Alatas et al. (2019) found that the presence of a source significantly decreased the subsequent engagements. By retweeting messages with and without attached sources from celebrity accounts, a decrease in subsequent Retweets by 50% was observed in those messages which included a source. The reasoning for this phenomenon is largely speculative, although some conjectures include that a message with a source may appear to be less authentic when compared with non-sourced Retweets from a celebrity. Also, the existence of a source may serve to negatively impact the notion that a user is able to distinguish between a good and bad Tweet when viewed by his or her followers leading to lower subsequent engagement. Regardless of the rationale, the conclusion is clear, illustrating how Twitter is at times a dangerous medium of information exchange given the prevalence of disinformation and heavily opinionated messages.

2.2 Sentiment Analysis

The process of sentiment analysis involves determining through a number of possible methods what the overall message of a Tweet conveys, often times characterized as positive or negative in tone, as it relates to a topic. One such method discussed by Tsugawa and Ohsaki (2015) uses a dictionary of known positive and negative words and tallies the number of each present in a Tweet. In this instance, rules were made which defined that Tweets with positive and no negative words were labelled positive while those with negative words and no positive were labelled negative. Tweets with neither type of words were labelled neutral and those with both were discarded. Additionally, dictionary-based sentiment analysis can be accompanied by manual analysis for validation to ensure that the rules imposed are reasonable. A noticeable pitfall of this method that is the absence of accounting for slang or other shorthand messages which are prevalent on social media. For instance, Tweets containing a word such as "sick" may be classified as negative while in modern usage such words could certainly be used in a positive meaning. While other much more advanced methods exist for classifying sentiment, this method proposes one of the simplest examples.

Identifying sentiment is not only important for overall analysis of opinion on keywords being observed, but it can also be used to assist in identifying communities within a network. For instance, if one was interested in classifying those members of political parties, it would not be sufficient to collect only keywords since this method would struggle to differentiate between members of different parties. To complement traditional community detection methods, by considering the sentiment of messages involving particular keywords such as a candidate, a more accurate classification can be made (Salehi et al., 2018). Community identification is one of the key components of SNA since it aids in a better understanding of the relationships between users. In addition, by identifying communities, the most influential members of a network as

well as individual communities can be ascertained, which is an important goal of this research.

2.3 Social Network Analysis

The next major topic this research will involve is SNA. As introduced by Scott (2000), there appears a specific logic behind the relationships we choose to form and maintain, which results in a social configuration that can be represented graphically. This method makes use of connections between entities to allow for a graphical representation of a network in addition to mathematical analysis through measuring the strength of connections (Legradi, 2009). The essence of SNA is to distill from a complex structure of relationships the individual connections between entities to assist in both visualization and analysis using graph-theory algorithms. When conducting SNA, Allard (1996) identifies two main goals: developing an understanding of the factors that affect relationships and their correlation as well as ascertaining the affects of these relationships including the possible identification of an informal leader. By utilizing SNA to create connections between users given information about the Tweets that they write as well as those that they can interact with, these end goals will be achieved.

A challenge with identifying communities and clusters as they relate to data from social media is the structure of the network. Network graphs can be divided into two main categories, directed and undirected. The key difference between the graph types is that in an undirected graph nodes are simply connected or unconnected while in a directed graph the direction of the connection between nodes is stipulated. For many practical networks involving the relationships between humans, especially in social media, it is necessary for the network to be directed since many users have a relationship with others through their activities while the other user does not share

this connection with the other. This asymmetry and the challenges it poses for finding communities is discussed at length by Malliaros and Vazirgiannis (2013). While extensive work exists for finding clusters in networks which are undirected, there are fewer common methods used for graphs of the directed kind. The method proposed by Tsopze and Domgue (2021) uses Boolean factorization to create an adjacency matrix of a directed network. Using this matrix, communities can be identified. In addition, a metric called NR-modularity is defined which can be used to measure the performance of a clustering approach as typical measures are inappropriate for directed graphs. This research can be supplemented when creating communities from social media data by accounting not only for a binary indication of a connection between users but also a measure to indicate how strong a connection is.

Upon determining the general structure of the network model, links must be created between nodes in the network. In typical models such as geographical networks, connections are logical since they represent two entities which are directly connected. However, in social media, connections can be a point of debate. The question posed is, "what interactions or similarities between users constitutes a connection in the network between them?" Potential factors to consider include Retweets, mentions, same Tweet Retweets, followership, geographical proximity, topical interests, and profile similarities. The number of factors to consider is limited only by the API through which data is retrieved as only certain aspects of a Tweet or user are available depending on the data source. Specifically, given the API currently used by many intelligence analysts, the list of users which an account is following or followed by is not available which would have presented itself as the most obvious connection.

In exploring other methods of connecting users, Aiello et al. (2010) uses the social media site aNobii to examine link creation amongst users on the site. aNobii is a social media site in which users create lists of books which they enjoy or are interested in

and may follow other users as a result of seeing their reading preferences. Similar to Twitter, links on this website can occur without the explicit consent of the user and are of the directed kind. This research determined that most new connections between users occur due to a geographical proximity and, once a user has established a small following, followers of followers become the next links made because their profiles often exhibit high similarity. It was concluded that through following other users, accounts seem to be influenced by them in modifying their own profile which leads to further similarities with the user who they followed and, therefore, their friends as well. In addition, they determine that preferential attachment occurs toward the beginning of a new user's profile creation in which they begin by following users which already have a large number of connections. This commentary on link creation should be explored on Twitter to determine how profile similarities drive connections within the network.

An alternative approach for creating a network which makes use of multiple dimensions of a user profile is the multi-layer network introduced by Bródka and Kazienko (2012). This method leverages the idea that in the real world people are connected by more than one type of relationship since there are many different aspects to social connections as seen in Figure 1.

In this type of network nodes may exist on a number of different layers each representative of a different dimension of interactions. For each node, connections to other nodes are dependent on the layer and while two nodes may be completely unconnected in some layers they may be directly connected in others. This method attempts to mirror the complexities of real life interactions that a generic network model may not be able to represent adequately.

Applying this multi-layer network model to social media, Ramokhoro et al. (2020) proposes an adaptation which allows for different entities and actions within a network

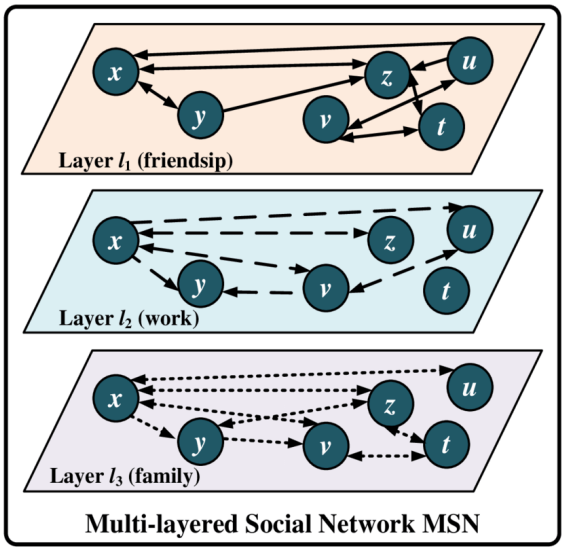


Figure 1. In Multi-layer Networks Users Can Associate With Others Differing Based on the Layer (Bródka and Kazienko, 2012)

to be modeled displayed in Figure 2.

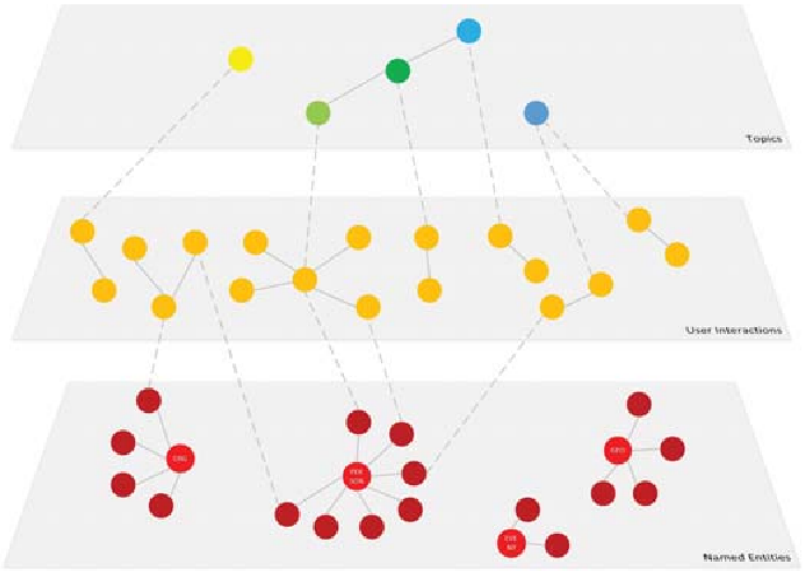


Figure 2. Multi-layer Twitter Networks Allow for Differing Node Types and Interactions Between Layers (Ramokhoro et al., 2020)

This network modifies the original proposal by Bródka and Kazienko (2012) by implementing variations between layers resulting in nodes which represent different elements of the social network at each layer. The layers represent the named entities

or users, the user interactions, and the topics of discussion. For each Tweet, the user is added to the named entities layer, and if they mention or Retweet another user in a Tweet then both of these users are added to the user interactions layer. The uppermost layers consists of the topics of discussion as identified by an LDA model. If Tweets mention any of the topics then connections are made connecting the nodes in the user interactions and topics layers. In addition, connections between topics are made based on similarities of the words contained within them.

This holistic view of the social network allows for a better representation of the intricate relationships as well as the crossover between users discussing multiple topics to be modelled. The research addresses that for large networks a significant challenge is inferring meaning as less measures of multi-layer networks exist, particularly for novel ones. Still, the proposed network model is a promising structure for understanding the diffusion of information across Twitter.

2.4 Topic Modeling

One important consideration of this research is to determine the most frequent topics of discussion on Twitter. While rudimentary methods such as keyword frequencies can accomplish this task, their primary shortcoming is the failure to link similar words which should be viewed as part of one topic together. For instance, keyword extraction might find a number of Tweets containing "Republican" and others containing "Democrat" but has no knowledge to link these two words together unless done manually. Topic modeling seeks to overcome this by grouping words which belong to similar topics such as the example previously given of which short inspection would reveal that this topic must be politics.

Two of the most common methods for performing topic modeling, Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are compared in research by

Kalepalli et al. (2020). In LDA, the most common method for topic modeling, the co-occurrence of words within a document is used to associate words together, and words are assigned to topics based on their probability of belonging to it. By providing a desired number of topics, LDA will find the highest probability words associated with each topic to assign these topics to each of the documents. In LSA, a word occurrence matrix for each document is used along with a process called single value decomposition to determine the similarity of each document and group them into topic categories.

While both methods succeed at distinguishing similar documents, the clear advantage of LDA in the context of this research is the association of the most likely words with each topic. This output should make for easier identification of topics by intelligence analysts rather than needing to search through the grouped documents to determine their common theme.

Applying the method of LDA to a context more similar to this research, Rahmadan et al. (2020) explored topic modeling on Tweets during the Jakarta flood disaster which occurred at the beginning of 2020. The key challenge of topic modeling with Tweets in contrast to other documents is the imposed character limit by Twitter which reduces the number of words in each Tweet allowing for less co-occurrence of words within a document. Despite this, using only 15,000 Tweets, nine clear topics were identified amongst the Tweets with high coherence amongst the words belonging to them allowing for manual labelling of the topics. In addition to aiding OSINT by identifying the topics of discussion, LDA allows for groups of Tweets to be focused on to learn important information about the Tweets and users belonging to each topic.

To further aid in OSINT capabilities, it would be desirable to produce a short explanation of the Tweets belonging to a certain topic rather than requiring the analyst to manually label topics given the list of related words. For instance, if

an LDA model identified a grouping of ten words which belonged to a topic and 1,000 Tweets were found to contain some of these words, the overall summary of these Tweets can be found. While the individual words of a topic on their own are informative to an analyst, a summary of the Tweets containing the words may provide some additional context on why users are talking about the topics. This concept can be extended further to find the summary of the Tweets talking about a topic which have a negative or positive sentiment to provide context on the main reasons users feel a particular way about something.

In research by Jiwanggi and Adriani (2016), a process for performing this summarization on Indonesian Tweets is outlined. Two different types of summaries are presented: extracted and abstracted. Extracted summaries pull word for word sentences from the Tweets to create summaries using the words of others while abstracted methods attempt to create new sentences given the important words in the Tweets. Unsurprisingly, abstracted summaries are much more difficult to produce because they require a higher understanding of the language whereas extracted summaries must simply find what is deemed to be the most important sentences. To bypass this difficulty, a semi-abstracted approach is proposed; it finds a root word of high importance and builds sentences around it using the most common word occurrences in the Tweets. Through human evaluation, these summaries were graded and found to be 64% accurate in that they matched what the actual Tweets were discussing while also being grammatically correct and readable. This method should provide additional insight to OSINT analysts when analyzing topics of discussion on Twitter.

2.5 Key Node Identification

2.5.1 Anomaly Detection

As identified by intelligence analysts, it is desirable to understand those users which stand out in the network for any of a number of reasons. One of these reasons may be evidence the user poses a threat to others in the network. Research conducted on this topic by Sheth et al. (2021) identified multiple dimensions of toxicity that can be used by statistical learning algorithms to identify Tweets which contain toxic messages. This method results in better classification of message sentiment because it takes into account a number of factors and not just a simple lexicon as other methods do. Methods which account for more than word usage are particularly important in analysis of social media data as many users write quite informally which could result in frequent misclassification. The presence of a single toxic word might not always indicate toxicity, which is why both the content and context is necessary for better classification (Sheth et al., 2021). When used alongside sentiment analysis, this method can be utilized to flag Tweets which meet a certain set of criteria or violate policy set by intelligence analysts.

Toxicity identification is closely linked to anomaly detection within a network. An anomaly can have a number of definitions depending on the context of the problem, but for the most part it is defined as a user whose interactions and characteristics are significantly different from what is expected. This definition of anomaly can be adapted to suit a number of contexts such as in the work done by Venkatesan and Prabhavathy (2019). In their research, a number of features which can be used to characterize nodes and detect outliers given a node's subnetwork of immediate neighbors are defined. These features include the number of neighbors, the number of edges, the total weight of edges, and the eigenvalue of the subnetwork matrix. Using these features, different metrics are defined such as the existence of a single strongly

connected pair and if a subnetwork is a clique or a strongly connected community. By comparing the difference between each node's features and their expected values, nodes which are highly anomalous are identified. This method is highly adaptable in the features to consider as well as the cutoff value for identifying outliers allowing for its use to be tuned according to the concerns of intelligence analysts when it comes to the types of users they are interested in as well as the severity of difference from what is expected.

2.5.2 Influence Ranking

To quantify statistics of interest when analyzing networks, it is important to understand how networks can be compared as well as how to identify key nodes and edges. One of the desired outcomes is to determine the most influential users in a network or possibly in communities. Erlandsson et al. (2016) discusses how the definition of influential users may differ depending on the social media being observed and that for Twitter this may be measured by how many likes and Retweets a user received on his or her Tweets. Identifying these users is important because it can inform analysts of the nodes most likely to cause a rapid or broad spread of information through the network. Ranking influence can be done in several ways including using clustering algorithms on influence features, linear regression models, and popularly a PageRank algorithm. Another method proposed by Erlandsson et al. (2016) involves using associations between users to determine influence. For example, if it is observed that User A comments on many of the same posts as User B then it can be concluded that their actions are either due to them having a similar interest or that User B is influencing User A. Relationships such as this can be leveraged in a network model to find those users which exert the highest influence on others.

When determining the influence a user has over others in the network, Bakshy

et al. (2011) presents a method which traces the origin of URLs on Twitter and follows their spread through accounts. Influence is quantified by observing instances where a user posts a URL and one of his or her followers successively posts the same URL. Behavior such as this would indicate the first user likely has influence over the second due to the mimicry in action. This logic can be applied to many different actions taken across Twitter to detect influence such as the topics of conversation in which a user engages. This method unsurprisingly finds that the most influential users often have many followers. However, a user's followers alone is an unreliable method of quantifying influence (Erlandsson et al., 2016). Implementation of a similar method to this allows for a robust platform for analysis on influence since it might relate to many different topics of interest.

In addition to previous influence identification methods, a number of other methods are considered in a comparative study by Bhavnani et al. (2021). They define an influential user as one whose actions have the potential to create significant impact on others through sharing their point of view on a topic. It is also noted that social networks are dynamic and constantly change over time highlighting a concern of the prediction of future influence on new nodes within the network. The similarities between disease spread and idea spread motivates one of the methods defined for dynamic networks which utilizes the idea of Susceptible-Infected-Recovered (SIR) nodes. Nodes which are connected to an infected node are removed from the susceptible state with some probability β while infected nodes recover with probability γ. By seeding an idea at a particular node and observing the number of users which it is able to infect, the influence of a node is quantified.

Another model compared by Bhavnani et al. (2021) extracts several features from a Tweet including the count of favorites, Retweets, and hashtags as well as statistics about the user which posted the Tweet including their number of followers and

following to build a model which quantifies user influence index and Tweet influence index. Findings showed that news agencies were often the top influential users while the most influential Tweets had no mentions. This method is quite rudimentary in that it only considers a limited number of factors about a Tweet which unsurprisingly returned that the most influential Tweets were those with the largest number of Retweets penned by celebrities. This method can be modified however by incorporating network statistics in the ranking such as the ones which will be defined in the research by Pudjajana et al. (2018).

Along with the identification of key nodes, it is of interest to find system metrics which may quantify the network being observed. As defined by Pudjajana et al. (2018) three metrics which can be used to compare nodes are degree centrality, closeness centrality, and betweenness centrality. Degree centrality simply measures the number of connection from one user to all others while closeness centrality measures the minimum distance from all other nodes to a particular nodes. Betweenness centrality measures the number of shortest paths from one node to another which include a node of interest (Sari et al., 2021; Pudjajana et al., 2018). Given that these metrics measure distance between nodes, modifications of the strength of connections between users can be made to function similarly to distance to allow for these to be calculated for each user.

This metric of betweenness centrality is a common tool for finding influential users because it is a direct proxy for the flow of information over a graph. As implemented in the research by Jin (2020), betweenness centrality was used to rank the top influencers on Twitter during the crisis communication surrounding Hurricane Irma. In this study, betweenness centrality is described as "The extent to which a Twitter user works as an information "bridge" along the shortest path between other users." (Jin, 2020). While a slight improvement upon simply using the number of followers to

quantify influence is observed, results show that rankings still reflect that the number of followers is highly significant. Potentially, the number of followers is particularly important in crisis communication due to the rapid dissemination of information and would not be as dominant in the context of political discourse for this research. Despite the detected influence of many highly followed accounts, several accounts with less followers which were not celebrities or news media outlets were identified as influential due to their frequent engagements with others in discussion. While larger accounts drive their influence score through dissemination of information, smaller accounts have significantly more out-degree linkages with other nodes due to their role in conversations as opposed to one-way provision.

A related problem to detecting influence is locating the opinion leader in discourses on social media. This process relates to the discovery of an informal leader as introduced by Allard (1996) earlier since opinion leaders are not always evident to other involved members of discussion since the leader can become obscured over time. Despite the obscurity, it is important to find opinion leaders because they are clearly capable of engaging a large number of other users in discussion and therefore must have a strong level of influence within the network. This problem is approached by Dewi et al. (2017) using many of the node metrics defined by Pudjajana et al. (2018) such as degree centrality, betweenness centrality, and PageRank. By assigning weights to each centrality measure, this method is customizable for use in different domains and problem contexts than just the one originally studied. Results demonstrated that, when combining centrality weighting with various types of edge weighting between nodes, it was possible to identify opinion leaders with some certainty. It was observed that retweeting other user's Tweets had a much stronger impact on idea spread than mentioning a user which should be considered when determining the weighting scheme of edges in this research.

While most methods for quantifying influence leverage centrality measures, other research has explored more sophisticated methods which attempt to better capture the actions of users on social media. In the research by Oro et al. (2018), user interactions are modeled as a directed multi-layer network seen in Figure 3. Here, the three layers of the network model are the users, the items or topics which they have posted about, and the keywords mentioned in the post describing the topic. Inter-layer connections take the form of user u posted about topic i using keyword k. In addition, intra-layer connections represent when users have mentioned or retweeted each other, when topics are similar, and when keywords are similar measured through co-occurrence within Tweets. These connections can all be weighted according to the frequency of times a user has talked about a topic, for example.

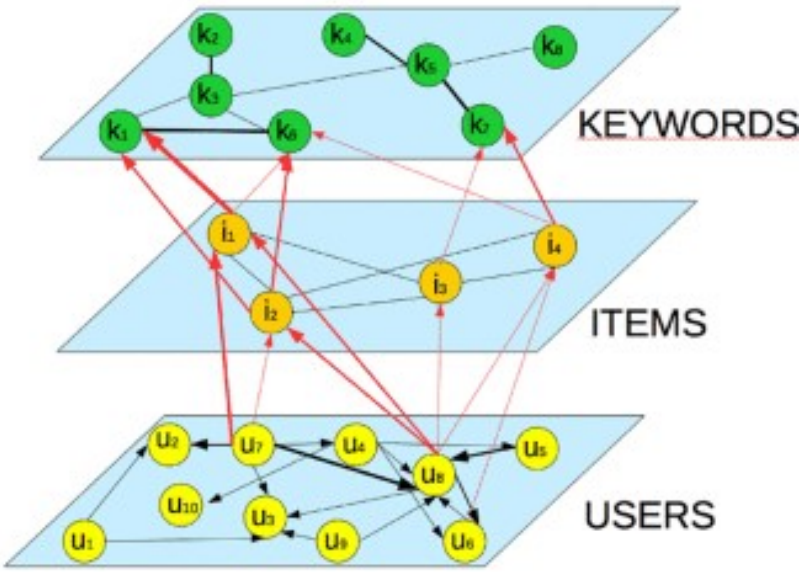


Figure 3. Proposed Multi-Layer Twitter Network Accommodates Users, Items, and Keywords with Connections Between Layers (Oro et al., 2018)

The proposed method for detecting influence called *SocialAU* is compared with another method called *TOPHITS* proposed by Kolda et al. (2005). The primary difference in these methods is that *SocialAU* seeks to consider the importance of each

node within its own layer as the previous method considered only importance within the multi-layer network as a whole. A number of typical metrics for influence are used as a method of comparison for results such as the Retweet and mention ratio of a user's Tweets as well as his or her interaction ratio with other users which can both be normalized to account for the number of Tweets which a user has made to better capture his or her interactions. Results seem to show that *SocialAU* performs better on real data when comparing the users which they found most influential as all of *TOPHITS's* top 20 users showed up in the top 100 for *SocialAU*. Conversely, many of the users which had a large following on Twitter and are most likely very influential, as suggested by other research, were assigned very low influence rankings by *TOPHITS* as these users post much less on the site while *SocialAU* ranked them in its top 20.

2.6 Community Identification

When collecting data for OSINT analysis, a particular area of interest may be the identification of bot Tweets as these users behave much differently from typical users on a site. Bots make up a significant portion of social media users and are typically accounts which are not manually operated by a human. Some bots are programmed to scrape the internet for data and reply to users which have mentioned them while others tweet links to news articles and web pages. As identified by Bessi and Ferrara (2016), bots can play a pivotal role in shaping the opinion of a community through automated Tweets and replies with the targeted goal of supporting or smearing a candidate in an election. They discuss a method which combines factors such as frequency of Tweets from an account and other metadata to flag accounts as bots. This could be a useful preprocessing step as it allows the separation of bots from humans which will aid in the network analysis to be discussed later. When determining the purpose of these

bots, a key factor is the sentiment of the Tweets they are circulating in addition to the target. By identifying sentiment and target, analysts will be able to ascertain the activity of bots in the area of interest as well as the goals of the bots being observed. By identifying communities of bots, it is possible to identify the motives of the anonymous actors deploying them to social media.

A topic of interest for intelligence analysts performing OSINT might be the presence of coordination between users involved in influence campaigns. As discussed previously, much research has been done to study and prevent social media bots from disrupting a network however, often times malicious accounts work in groups in a coordinated effort. The method introduced by Pacheco et al. (2021) involves identifying traces of information from a variety of sources which indicate a lack of independence between users. These sources include the type of content shared, geospatial and time-related activity, identity related features such as handle or profile picture, and a combination of these. By assigning weights to the connections between accounts and the features studied and projecting this bipartite network onto a graph connecting users where edge weight relates to similarity in the user's features, cluster analysis can be performed to identify users displaying high levels of dependence. A visualization of the communities identified when considering users which Retweet many of the same Tweets is seen in Figure 4.

This method of locating communities is useful in this research because it considers features about users other than simply the Tweets they interact with in addition to being designed for a directed graph. This method can be implemented to identify suspicious users in the network and expanded upon by considering pairing this method with a bot identification to better understand the actors at play in the network.

Another method which relates to the classification of nodes is presented in the research by Tang and Liu (2011). The method they propose also attempts to solve

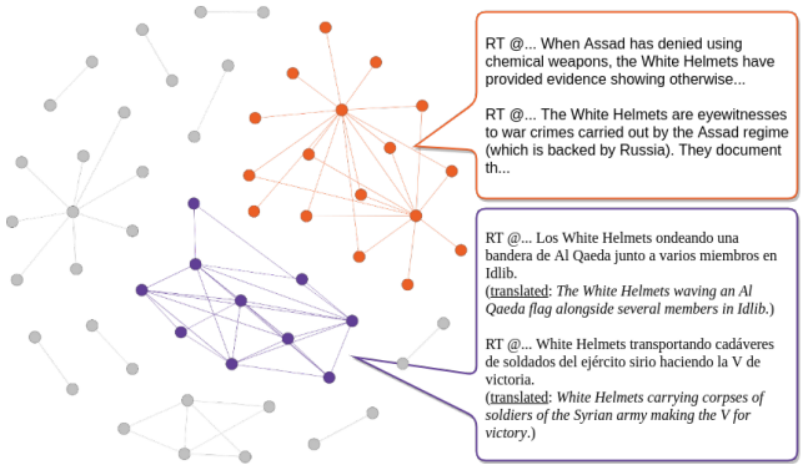


Figure 4. Communities Can Be Found Through Similar Retweets by Different Users (Pacheco et al., 2021)

the problem of typical community identification techniques being inapplicable for networks where the relationships between users is not commutable. For instance, typical graph logic would say that if Node A and B are connected to Node C, they both share the features of C. However, in nonhomogenous networks such as social networks, it is highly possible that A associates with C for some set of their interests while B is associated with C for a separate set. This problem introduces what Tang and Liu (2011) refers to as social dimensions, the utilization of a user's known affiliations to predict some label. In other words, if a user has a set of affiliations S_i, then a probability can be assigned for their likelihood to be labeled y_i. This type of reasoning is similar in logic to the multi-layer network proposed earlier because it considers the multiple different types of interactions each user is engaged in. A combination of this method with other clustering types listed before will allow for a clear picture of the relationships of the network to be discovered.

III. Methodology

The methodology by which the analysis in this research will be conducted will be discussed in this section broken down into a number of steps. First, a description of the datasets which will be analyzed in this research is presented in Section 3.1. Next, the method by which the networks will be created using each of the datasets is described in Section 3.2. Following this, the topic modeling and summarization methods will be described in Section 3.3 along with the integration of these topics into the created networks. Finally, the methods by which many of the essential research questions will then be shown in both Section 3.4 where the influential user discovery methods will be described and in Section 3.5 where the community discovery methods will be discussed.

3.1 Description of Data

The data which this research analyzed consists of thousands of Tweets including information about the authors of the Tweets originating from several different open-source API scrapes. Some of these user features include username, number of followers, date of account creation, and the verification status of the user. To gain a better understanding of the interactions on Twitter and how they impact SNA, four almost exclusively English datasets were used which were collected at different times and with unique search features. The Tweets from this data span a wide range of topics of discussion on Twitter including sports, entertainment, COVID-19, and politics. The ultimate goal of analyzing such different spheres of discussion was to learn how they might lead to differences in both influential user findings as well as general network metrics. These datasets were used to verify the network creation and topic modeling methods before transitioning to analysis on the Indonesian datasets.

A description of the datasets used in this research as well as information concerning the types of Tweets which they contain can be seen in Table 1.

Table 1. Summary of Datasets and Key Features

Dataset Name	Number of Tweets	Contains Retweets?	Contains Replies?
2018 World Cup	530,000	Yes	No
Game of Thrones	760,660	No	Yes
2016 US Election	42,013	Yes	Yes
COVID-19	179,108	No	Yes
Indonesian Data	6,087	Yes	Yes

A key difference between the datasets which plays a crucial role in determining the resulting networks is the search criteria used to find the Tweets. Most API's provide dozens of filters for querying Tweets including but not limited to the text found within the Tweet, the hashtags used, the accounts which the user mentions, the language and country of the user, and the media contained in the Tweet. Queries can be very broad and collect all Tweets which simply use a certain word or phrase, as seen in many of the Indonesian datasets which are queried using only the word "Indonesia". Conversely, queries can be as specific as only Tweets which mention a specific Twitter user resulting in a densely connected network of Tweets which are all centered around one user.

Accordingly, analysis of the Tweets in a dataset is not sufficient on its own to understand the dynamics of a network. Certain Tweet query practices are highly influential on the type of data which will be returned and are a crucial first step in SNA. A key focus of this research is identifying what effects some of these practices might have on the resulting network allowing for future work in SNA to place more of an emphasis on querying when conducting analysis.

In addition to the datasets already described, significant analysis was performed on data created from an amalgamation of the four English datasets. By sampling an equal number of Tweets from each source, the resulting dataset can be viewed as

a more accurate representation of Twitter with many different users having conversations about different topics. In addition, the dataset helped to verify community creation methods because it is expected that Tweets originating from the same discussion would be found in the same communities. Artificial data creation also assisted in verifying topic modeling as the differentiation in topics should lead to a more clear division between the words belonging to each topic.

Finally, the Indonesian datasets provided by USINDOPACOM intelligence units were analyzed to ensure the methodology is robust to differentiation in language, querying practice, and online cultures. The techniques used on the other datasets were specifically limited only to those which would be possible given the user and Tweet data available in the Indonesian data. For instance, information about a user's location was often unreliable or missing in the Indonesian data and was therefore not leveraged. Additionally, if the data included any pre-extracted information about users which a Tweet was in response to or users who were mentioned in a Tweet the information was not used. Ultimately, the methods utilized only the username of the author of a Tweet and the text of the Tweet because these were the only consistent and useful features present in all datasets.

3.2 Network Creation

After obtaining the data and cleaning it to ensure a username and Tweet were available for each entry, the network was created. In this network, the nodes can be viewed as users on Twitter while the edges represent some sort of relationship between the two users which they connect. As alluded to in Chapter II, a directed network was chosen to model the interactions because it captures the essence of relationships on Twitter better than an undirected network. A directed network allows for the distinction between users which have a high number of in-degree relationships such

as celebrities who do not tweet a lot but receive much attention and users with a high number of out-degree relationships such as bots and spammers. An undirected network would view these two users as identical and would therefore struggle to quantify influence in users. In using a directed network, actions taken by a user to show his or her relationship with others was modeled as an outbound link while users which are being tweeted about or in response to by others was modeled as in inbound link.

The next phase of modeling the network involved determining a weighting scheme for the connections between users. The reasoning behind a weighting scheme for each connection was to reflect the strength of connection that an interaction constitutes. In the datasets analyzed there are three primary interactions that can be seen: a user mentioning another user, a user retweeting another user's Tweet, and a user replying to another user's Tweet. Unfortunately, no formal data exists to define the effect that each of these interactions might have on the true connection between users. As a result of this, a proxy statistic must be created which captures this effect to appropriately weight interactions.

Ideally, a universal distribution of the rate at which each of these three interactions occurs on Twitter would serve as a proxy. To demonstrate the validity of this approach, consider this example. A Twitter user has just seen a Tweet from another user that they find intriguing and they want to interact with this Tweet. Excluding liking the Tweet, which is an interaction not captured in the data used for this research, Twitter users can either directly respond, Retweet, or pen a Tweet of their own mentioning the user whose Tweet they saw. All of these interactions are equally accessible by all Twitter users and one need not be following the other user to engage in these interactions with them. While it is true that some users may prefer to reply to posts rather than writing their own with a mention, on average the distribution with

which each occurs describes the general preference of the entire Twitter community.

Despite a barrier to entry not existing for each of the interactions that would warrant stronger connections such as needing to be following a user to send them a direct message, it is fair to say that in general users understand that some interactions imply a closer relationship. This fact is evident by observing the ratio of likes to Retweets for nearly every Tweet on the platform. Tweets consistently have far more likes than Retweets implying that in all circumstances users understand that a Retweet is more significant than a like. In a similar way, Tweets almost always have fewer replies than both Retweets and likes indicating a similar understanding exists for this interaction. Seeing as this trend is prevalent throughout Twitter, the frequency with which each interaction occurs should serve as a sufficient proxy for the strength of connection implied through an interaction.

An additional benefit of using data emerging from several different conversations and cultures as well as a variety of querying practices is that the combination of these datasets should present a general idea of the distribution of the interactions. Through inspection of all datasets which contain each of the three interactions, the distribution of interactions consisted of 6.97% replies, 39.49% mentions, and 53.54% Retweets. While more data pulled from broad queries would help bolster these findings, for this research the inverse of the percentage of all interactions that each constitutes will be used to weight edges between users in the network. In addition, if a user interacts with another user several times in the data then this weighting will be summed to reflect all of the interactions between the users.

While a number of other factors can be used to connect users including but not limited to Tweet liking, following a user, and mutual followers, this information is not available in the Indonesian datasets as well as many other datasets. By using only information retrievable from two essential aspects of a Twitter scrape, this method

of network creation is robust to many data collection methods. In addition, all of the interactions used require active engagement whereas information about who a user follows does not.

The methodology described is used to create the basic user network on which analysis was performed for each dataset. In addition to these user networks, a multi-layer network adaptation was explored to determine if significant insights could be gleaned through the addition of a layer to the existing user network. In Section 3.3, the inclusion of topic nodes will complete the multi-layer weighted and directed network.

3.3 Topic Modeling and Summarization

The goal of implementing topic modeling alongside social network analysis is two-fold. Firstly, topic modeling provides an analyst with a general overview of the discussions contained within the dataset of Tweets. Secondly, topic modeling can be used in conjunction with the existing network to connect users through the conversations which they are having, a feature which cannot be extracted directly from Twitter data. Through the inclusion of topics in the network, a multi-layer network can be created which more accurately portrays the connections on the diverse social media platform.

Topic modeling will be performed in this research according to the approach first described by Blei et al. (2003) known as LDA. This process is described as "a generative probabilistic model of a corpus" (Blei et al., 2003) where a number of topics are synthesized and populated with the words that have the highest probabilities of belonging to them. The three key elements of an LDA model are the topics, documents, and corpus. In this research, the topics are the eventual bins into which words from the Tweets will be placed, the number of which k will be decided by the analyst. The

documents will be each of the individual Tweets appearing in the data. Finally, the corpus will be the complete collection of documents. Across all documents, V unique words will exist with the ultimate goal of LDA being to find the $k \times V$ probability matrix β, where β_{ij} represents the probability of the word j belonging to the topic i.

A focal point of preparing the documents for LDA is the removal of stop words, lemmatization of words, and tokenizing. This process involves stripping the original Tweet of words found in a dictionary of common, typically non-meaningful words, reducing words to their root, and creating a list of the resulting words. An example of this process on a Tweet can be seen in Figure 5. Through preprocessing each Tweet in this manner, only the most relevant words are left to be assigned topic membership probabilities.

Figure 5. Preprocessing for LDA Reduces Tweets to Only the Essential Words

The only required input from the analyst needed to produce results from LDA is the number of topics. While the optimal number of topics is highly dependent on the data being analyzed, a coherence measure called C_v exists to grade the effectiveness of an LDA model and can be used to determine the optimal number of topics. C_v is a measure first proposed by Röder et al. (2015) which aims to assign a score to the set of highest probability words in an LDA topic based on their similarity to each other and interpretability by a human. This measure combines both the joint

probabilities of any pair of top probability words in all documents with a confirmation measure showing how well any word set supports another as described by Syed and Spruit (2017). C_v also incorporates a sliding window to create virtual documents when calculating joint probabilities in an effort to place more importance on close proximity of word pairs in documents rather than simple co-occurrence anywhere in a document.

Through inspection of the highest probability words associated with each topic, analysts can gain an understanding of the conversations within the Tweets. While manual labeling of topic names is necessary, ideally topic modeling should make this task quite simple given familiarity with the language of origin for the Tweets. Next, LDA will be utilized to create connections between users and topics in the multi-layer network through a novel method which maps Tweets to similarity scores with each topic to determine which conversation the user is most involved in based on his or her Tweet.

After performing LDA and deriving β, the similarity of words within Tweets to each of the k topics can be calculated using a dot product. First, a vector of length V where V_d represents the count of token d which appeared in a Tweet is created, called s. Next, by calculating $\beta_i \cdot s$ for each topic i, a singular value can be obtained per topic representing the Tweet-topic similarity strength. This strength can then be used to weight the edge connecting a user node to a topic node for which their Tweet is related. It should be noted that, since the edges in this network are directed, two edges will be added for each discovered user and topic connection. This method aims to model scenarios where users scrolling through a topic of conversation on Twitter using the search function are directed to a user through his or her Tweets which include keywords consistent with the topic.

In this research, only the strongest connection with a topic was established in the

network for a user based on each of his or her Tweets. While a single Tweet could be linked to a number of different topics which were discovered by LDA, to simplify computation time, especially in instances with a large number of topics, only the most relevant topic for a Tweet was considered. In addition, only Tweets which had topic strengths in the top 25% were kept to avoid establishing weak connections between users and topics. While no formal research exists to determine what the minimum level of similarity between sentence and topic word probability vector is to establish a relationship, future work could utilize labeled training data with a machine learning approach to make better decisions in this space. This method has its shortcomings, particularly that the order of the words is of no importance. However, for topics with a significantly different distribution of words, word order should be of much less importance than the count of words due to minimal overlap in high probability words across topics. Additionally, if using this method for longer documents than Tweets, there may be a bias when classifying the longest documents because they contain more words and therefore have a higher likelihood of having a large similarity score. In this research, this bias is unlikely to have a significant effect as Tweets can be no longer than 280 characters.

After establishing connections between the user layer and the topic layer of the multi-layer network, the final step for completing the network model was finding the connections between the topics discovered by the LDA model. Connections were created by finding cosine similarities between each of the topic word probability vectors with each other. Cosine similarity measures the similarity between two vectors by taking the cosine of the angle between them and is calculated using Equation (1).

$$similarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} \tag{1}$$

Cosine similarities are in the range of [-1, 1]. However, given all probabilities for

words belonging to topics must be positive, the similarities between topics will be in the range of [0,1]. The interpretation of these values is that a similarity of 0 means the angles between the vectors are orthogonal so no relationship exists, while a value close to 1 indicates similar angles and therefore similar word distributions for the two topics. These similarities were used to quantify the connection weight between two topic nodes in the network. Differing from the Tweet-topic connections, topics were allowed to have connections with multiple other topics so long as the cosine similarity between the two was greater than 0.5. Given that these edges must be directed, if Topic A connects to Topic B, the inverse connection will also exist with the same strength. The result is a network with both a user and a topic layer with weighted and directed edges connecting user to user based on pre-defined interactions, users and topics between the layers based on the vocabulary of a user's Tweets, and topics with topics based on word probability distribution cosine similarities.

Leveraging the Tweet-to-topic classification already performed, topic modeling can be improved through summarization of the Tweets linked to each topic. Of two primary types of text summarization, extracted and abstractive, extractive summarization will be used since it is less computationally expensive. While extractive methods are generally viewed as inferior to the best abstractive methods due to their inability to produce unique thoughts, they excel in their simplicity. In addition, when compared with the methods which Twitter uses to create their trending topics, they often present the most relevant Tweets involved in the conversation which produces a very similar result to extractive summaries. The method used is a variation of the TextRank algorithm as proposed by Barrios et al. (2016). The original TextRank algorithm operates by creating a graph where sentences represent nodes and the edges are weighted by sentence similarity which can be calculated by Equation (2) as provided by Barrios et al. (2016):

$$similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \tag{2}$$

This rudimentary method only counts the number of shared words between the two sentences, weighted by sentence length, to compute similarity. Barrios et al. (2016) improve upon this method by incorporating inverse document frequency of words to boost the similarity between sentences which contain words that are rare in the full document. Using the created network, PageRank was then used to calculate the most important sentences which can then be returned to produce an extractive summary containing the most important sentences within the body of Tweets related to a topic. This step will ensure analysts are provided with not only a list of the highest probability words for each topic, but a short summary of the Tweets to provide additional context and reduce the human effort required to analyze Twitter data scrapes.

3.4 Influential User Discovery

Finding influential nodes in a network is a common task which can be accomplished using many methods; however, the performance of different methods on a network is highly dependent on the network itself. For smaller connected networks, many different methods might produce reasonable results. Conversely, on larger disconnected networks, like the ones created in this research, it is likely that some methods will produce far better results than others. For this reason, a number of influential node algorithms were implemented to determine which method sees the best performance in general.

First, eigenvector centrality was used to measure node importance in the network. This method leverages the idea that connections to nodes which are of high importance are more valuable than connections to nodes of lesser importance. Eigenvector

centrality is calculated by creating an $N \times N$ adjacency matrix $\mathbf{A}$ where N is the number of nodes in the network and $\mathbf{A}_{ij}$ is equal to the weight of the connection between nodes i and j. Next, the equation $\mathbf{A}x = \lambda x$ is solved where λ is the largest eigenvalue of the adjacency matrix and x is the vector of influence scores for each of the nodes. Aside from measures such as in-degree or out-degree of a node, eigenvector centrality is one of the most simplistic measures of influence. Several of the other influence methods which will be discussed later are a modification of eigenvector centrality.

The next influence measure is the same PageRank algorithm used to calculate the most important sentences when finding extractive summaries. PageRank is an algorithm first proposed by Page et al. (1999) which modifies a random walk on a graph to determine long term stochastic probabilities of residing at any one node. In a random walk, at each time step a simulated entity travels from one node to a neighboring node based on its neighbors and their associated connection weights where higher weighted edges have a greater likelihood of being travelled. Through repetition of this process, the likelihood of being at any node in the graph can be found. This method is modified by Page et al. (1999) to account for situations where two nodes have outbound edges pointing exclusively to each other creating an inescapable loop. At each time step, a probability is included that the entity will randomly jump to any other node in the network regardless of whether it is connected to the current node so that these loops do not occur. This random jump is the essential modification from eigenvector centrality. Just as in eigenvector centrality, PageRank assigns more influence to nodes who are neighbors with other more influential nodes.

Next, betweenness centrality is used as an influence metric. This metric is calculated by determining the number of shortest paths between any two nodes s, t in the network which involve going through the node of interest v as seen in Equation (3).

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{3}$$

In this method, the inverse of weights are used as edge distances such that edges with larger weights indicating strong connections are modeled by a shorter connecting distance. A notable downside to this method is that it involves calculating shortest paths between each possible combination of nodes which can be very computationally expensive for larger networks.

The final node importance metric, *HITS*, is an influence algorithm first proposed by Kleinberg (1999) which extends PageRank using the idea that some nodes are hubs directing to other nodes and might have very few inbound edges. In contrast, an authority is a node to which many other hubs direct. These two views of nodes allow for scores to be calculated in both a node's ability to direct to other nodes and the number of nodes which link to it. This algorithm operates by querying some small subset of nodes and adding to this subgraph all nodes which are linked from the initial subset as well as some other nodes which link to the subset. For each node, hub scores are updated for the number of other nodes which are pointed to while authority scores are updated for the number of other nodes pointing at the node. Since this method produces two scores, this research summed the two to consider a node's merit as either a hub or an authority when assigning influence.

3.5 Community Discovery

Identifying groups of highly connected individuals in social networks is a task which can leverage a number of user features, as discussed in Section 2.6. In this research, communities will be found using solely the information contained within the network's nodes and edges. Given the nature of social media data, especially in the case of broad queries of Tweets with a large number of unique users, generally net-

work structures can be very fragmented and consisting of several components. When querying by common keywords, the chances of capturing a back-and-forth conversation through Tweets between two or many users is exceedingly small considering the millions of Tweets which are made each day. Accordingly, attempting to find clique networks, or small subgraphs of users complete with every possible edge, is an unreliable method for detecting communities. Rather, social network analysis in practice is better suited to find the implicit or underlying networks of users who are not in direct conversation with each other yet share the same topics of conversation or connections with mutual users.

One method used to create communities in this research was a greedy method first proposed by Clauset et al. (2004), which seeks to maximize the modularity of the network which will be referred to as the Greedy Modularity Algorithm (GMA). Modularity is a metric which is designed to measure the strength of community classification in a network and can be calculated by Equation (4).

$$Q = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \gamma(\frac{k_c^{in} k_c^{out}}{2m})^2 \right] \tag{4}$$

In this equation, L_c is the number of links within each community c, m is the number of total edges in the graph, k_c is the total sum of in degree edge weight or out degree edge weights in a community, and γ is a parameter which balances the importance of edges within a community and edges connecting communities. By initializing each node into its own community and joining the two which present the largest increase in modularity, the overall community structure of the network is discovered.

The challenge of many other community discovery methods in the context of this research is that the networks created are disconnected and directed. Both of these features are crucial to network structure and provide valuable information; however,

they present significant challenges to traditional community and cluster identification methods. Transforming the network to one with undirected edges would allow for more community discovery at the cost of a loss of significant information about users. In a similar way, disconnected fragments of users make up a significant portion of the network and are crucial to network structure.

To employ a community detection method intended for undirected and connected networks, two transformations were applied, as discussed by Malliaros and Vazirgiannis (2013). The first of these transformations involves the addition of a minimal number of low weight edges between disconnected fragments of the graph to maintain the integrity of the original network while connecting its components. The resulting network has almost entirely the same structure as the original disconnected network with an additional number of edges equal to one fewer than the number of disconnected components. In addition, by assigning weights to these edges lower than any other edge weights in the network, they serve to artificially connect components while minimally advocating that the newly connected users have actually interacted with each other. Next, directed edges are transformed to undirected edges while maintaining the total edge weight between the nodes. This transformation is potentially more damaging to network structure because it implies two-way connections between each user, a reason why the network was made to be directed initially. Despite the removal of directed edges, an undirected graph allows for a larger number of community detection methods to be explored and their results to be compared against those found using the original network structure.

After these transformations were applied, the Leiden Algorithm (Traag et al., 2019) was applied. This method improves upon the Louvain Algorithm, a modularity-based approach. Louvain assigns each node to its own cluster and joins the node with the neighboring cluster that would yield the largest increase in modularity. Next, it

consolidates the network by mapping communities to nodes in an aggregate network and proceeds with joining these community nodes together. After repeating this process until no improvements can be made, the final community structure is discovered. The issue with this method identified by Traag et al. (2019) is that communities are often created which by inspection are inherently disconnected. Disconnected communities occur when the node which bridges two disconnected components of a community is moved to another community as seen in Figure 6. In Figure 6a, Node 0 has been placed in the red community and connects all other nodes to each other. However, in Figure 6b, the Louvain Algorithm found improvement when moving Node 0 to the blue community. Despite this overall improvement, the result is that Nodes 1-3 and Nodes 4-6 now exist in the same community despite not being connected through any other nodes in the community with them, a problem which the Louvain Algorithm is unable to correct.

(a) Connected Community **(b) Disconnected Community**

Figure 6. Louvain Algorithm Creates Inherently Disconnected Communities Through Movement of Joining Nodes (Traag et al., 2019)

The Leiden Algorithm implements an additional step before aggregation where each community is refined to ensure these disconnected communities are not created when moving joining nodes later in the algorithm. This process consists of splitting each of the communities into subcommunities which would improve overall modularity, shown in Figure 7.

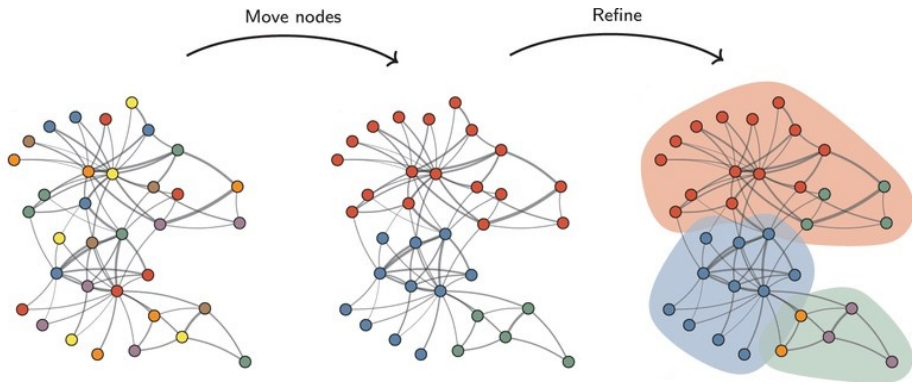


Figure 7. Leiden Algorithm Implements Refinement to Eliminate Disconnected Communities (Traag et al., 2019)

Refinement is not a strictly greedy process meaning some non-optimal changes may be made to promote a broader search of the community structure. This improvement eliminates many of the inherently disconnected communities discovered by Louvain and results in better overall clustering.

Performance of community discovery algorithms can be quantified with a number of metrics including the aforementioned modularity, partition coverage, and partition performance. Partition coverage is equal to the ratio of the total number of intra-community edges to total edges in the graph across all communities therefore favoring partitions which have minimal outgoing edges from nodes in one community to another. Partition performance is equal to the ratio of the number of intra-community edges and possible inter-community non-edges to the total possible edges between any two nodes in the graph. For the networks in this research, it is expected that both coverage and performance scores should be quite high as networks tend to be highly disconnected therefore meaning that most graph partitions would find communities which isolate many of the fragmented components. In addition, networks often have thousands of nodes with minimal linkages between them meaning that they are often not very dense resulting in a much larger number of potential edges than actual edges.

IV. Results and Analysis

In describing the results of this research, findings will be presented for each of the major research questions. First, high level insights discovered by network creation will be discussed in Section 4.1 to include visualizations of networks from different datasets. Next, the topics discovered when creating LDA models of datasets in addition to their implications on network structure when including a topic layer will be presented in Section 4.2. Accordingly, topic summarizations will be shown for many of the LDA model's topics and the success of this method will be discussed. After, a comparison of the results for each of the influential user discovery methods will be shown in Section 4.3 including a discussion of the impact of user verification on these influential user findings. Next, community discovery using each of the previously described methods will be presented in Section 4.4. Finally, a discussion of the impact of querying practices on each of the aforementioned findings will be shown in Section 4.5 along with results showing the Tweet sample sizes required for consistent results from differing datasets.

4.1 Network Creation

When creating networks, 15,000 Tweets were sampled from the larger English datasets and graphs were first created using only the defined interactions to connect user nodes. In Section 4.2, topic modeling will introduce the additional topic layer to these base networks to determine its effect on both influential user and community discovery. Due to the varying frequency of interactions in different datasets, the same number of Tweets can produce networks of different sizes as seen in Table 2.

On its own, network size is not enough to make conclusive statements about Twitter interactions as a number of other factors, namely query practice, can have

Table 2. Summary of English Tweet Networks

Dataset Name	Number of Nodes	Number of Edges
2018 World Cup	15,073	16,522
Game of Thrones	6,416	4,432
2016 US Election	12,565	13,603
COVID-19	8,457	6,565

a substantial effect. For instance, the 2016 US Election data was gathered primarily from users which were engaged in a number of discussions with political correspondents tracking the projected winner of the presidential election. For this reason, a larger number of nodes and edges than many of the other networks is a natural result, given it is far more likely that these users were interacting with the correspondents rather than writing Tweets with no mention of other users. This phenomenon can be seen in the visualization of the election network in Figure 8.

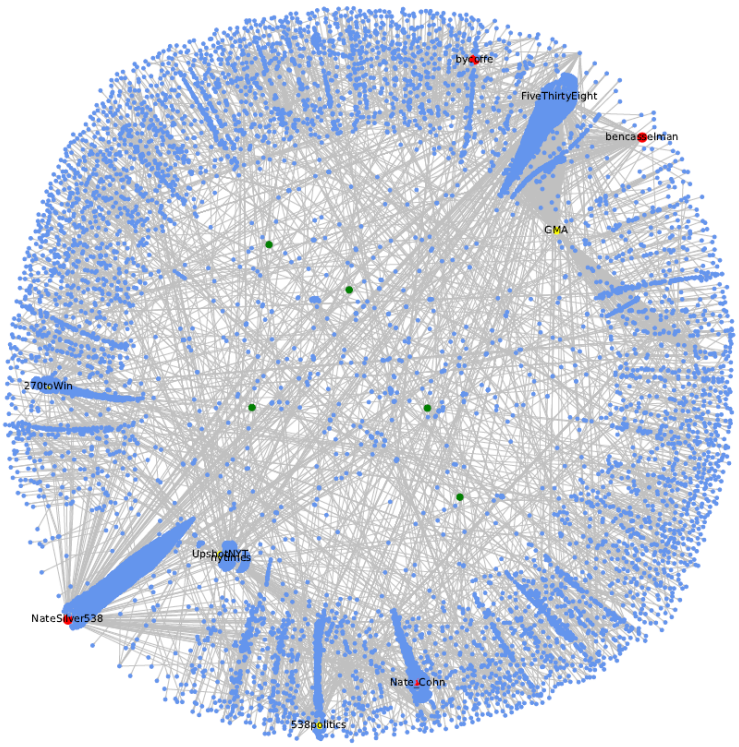


Figure 8. Force-Directed Algorithm on 2016 US Election Network Shows Large Clusters of Nodes Around Influential Users

In this visualization, red nodes signify the top five most influential users, yellow represent the five users with the highest number of incoming edges, and green represent the users with the highest number of outgoing edges. This graph was created using the Fruchterman-Reingold force-directed algorithm where nodes which are connected by an edge try to stay close together while other nodes are repelled from each other ideally minimizing crossover of edges and placing nodes in hubs based on who they interact with. While visualizations on their own can be misleading since node placement is still somewhat random and arbitrary, graphs such as these show a number of users clearly have a strong relationship with a large community of others.

Intelligence analysts can derive quick insights purely from inspection of these graphs given that a query practice has been established resulting in an expected network view. Large clusters of users in these graphs reliably demonstrates influence of central hub nodes and community membership which will be conferred in Sections 4.3 and 4.4. While visualizations cannot provide a complete understanding of a network, there are clear differences between them which can be seen when contrasting the election network in Figure 8 with the COVID-19 network in Figure 9.

In this network, communities are much more difficult to identify by inspection and influential users lack the obvious cluster of nodes surrounding them which they had before in the election network. Although of traces of communities can be seen in the lines of blue nodes near an influential user, the densely packed outer shell of nodes suggests a far more fractured network of individual users communicating with a small number of other users. To illustrate the necessity for both visualizations and network metrics, consider the summary of network statistics in Table 3.

Despite similar network densities, the COVID-19 network and election network display very different user behaviors when visualized. More consequential to the visualization is the average node degree or the average number of edges connected

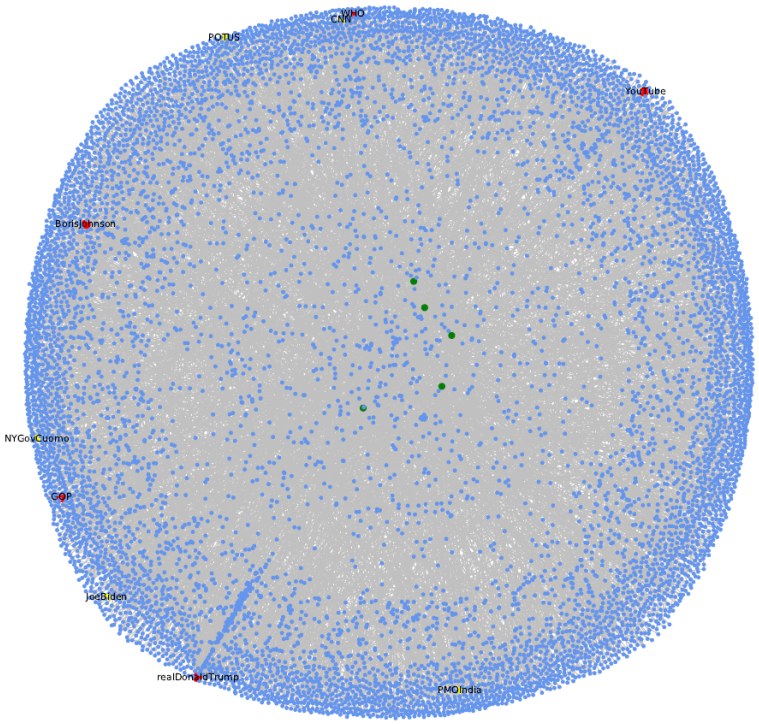


Figure 9. Force-Directed Algorithm on COVID-19 Network Lacks Distinct Clusters

Table 3. English Tweet Network Statistics

Dataset Name	Density	Average Node Degree	Average Weighted Node Degree
2018 World Cup	$7.36 * 10^{-5}$	2.190	0.295
Game of Thrones	$1.08 * 10^{-4}$	1.363	0.628
2016 US Election	$8.63 * 10^{-5}$	2.169	0.375
COVID-19	$8.97 * 10^{-5}$	1.533	0.338

to a node. The higher degree in the election network shows that on average each user is interacting with multiple other users, in contrast with the lower degree in the COVID-19 network which confirms the fractured network visualization as previously referenced. Average weighted degree is calculated as the total weight of all connected edges to a node and larger values represent stronger connections on average between users such as in the Game of Thrones network. Paired with the lower average node degree, this network can be characterized by each user having a strong but limited number of connections with others.

When considering the impact of these relationships on future work by intelligence

analysts, it is important to note that networks often are not fit to be compared to each other if they are found using vastly different queries. As explained earlier, it is a natural expectation that the election network would have larger groups of communities surrounding influential users meaning this finding is less consequential. If a similar pattern were observed in a network queried from a non-specific keyword, these large clusters would be of higher interest. Analysts must first establish a baseline expectation through repeated similar queries such that the status-quo is understood. Through repeated analysis over time, clear deviations from the status-quo should then be more apparent. Using just these basic metrics and visualizations, analysts can quickly identify patterns which may not have been observed in previous networks.

When creating the network from the dataset of Indonesian Tweets which were queried using the keyword "Indonesia", many of the observable features in the more specific query networks were still visible, as seen in Figure 10.

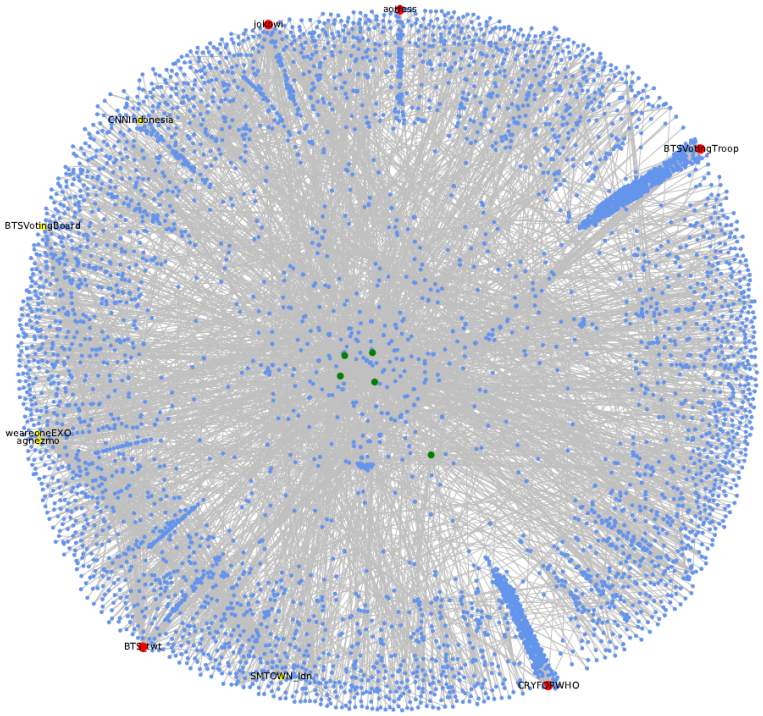


Figure 10. Force-Directed Algorithm on Indonesian Network Shows Distinct Clusters Despite Generic Keyword Query

This network validates the claim that discernible interactions between users can still be found even when the Tweets which are used to create a network are not geared toward a specific topic. Using only 6087 Tweets, this network contained 5335 nodes and 6164 edges. In addition, the network density was $2.17 * 10^{-4}$, average node degree was 2.311, and the average weighted degree was 0.448. When assessing both network metrics and the visualization, users appear to interact with several other users on average as indicated by the high node degree and density. When compared with other Indonesian networks, one focus should be determining if the large clusters of users around influential users are persistent throughout time or perhaps related to a rise in popularity of a specific topic or current event.

4.2 Topic Modeling and Network Integration

After creating the base networks, LDA was then used to find the topics of discussion in each dataset and connect Tweets to their associated topics. The primary burden on the analyst when performing LDA is determining the optimal number of topics to extract. The optimal number of topics is highly dependent on the data as a collection of Tweets from a very specified discussion will contain fewer topics while a broad collection of Tweets might contain many more topics. LDA performance can be quantified using coherence where a model with larger coherence should yield the highest interpretability when a human inspects topics. To illustrate the differentiation in optimal topic number, consider Figures 11 and 12 depicting the plots of coherence against topic number for both the World Cup and Game of Thrones datasets.

While the coherence score continues to climb as the number of topics has increased past ten for the Game of Thrones data, minimal improvement is seen after only six topics for the World Cup data. The reasons for this differentiation can be varying but intuitively it may stem from the World Cup data containing Tweets which make

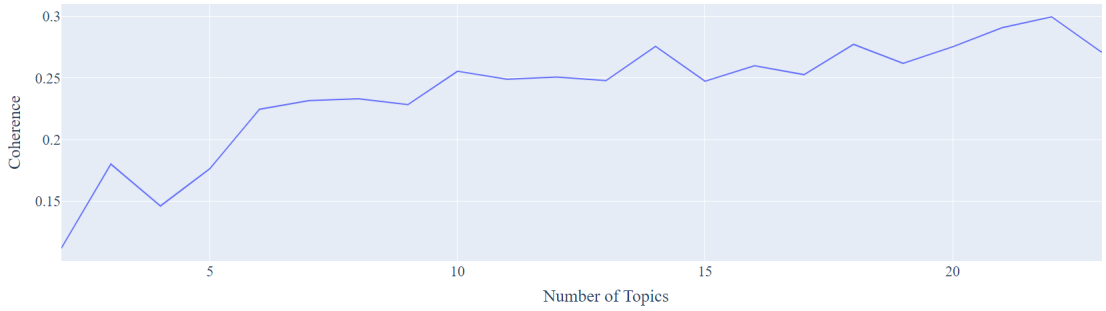


Figure 11. LDA Coherence Increases Up to Ten Topics in Game of Thrones Data

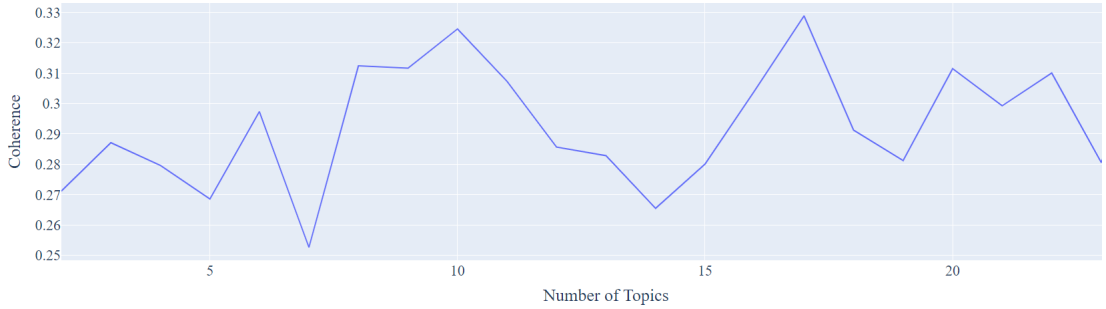


Figure 12. LDA Coherence Increases Stagnate After Only Six Topics in World Cup Data

reference only to a specific number of games from the tournament and containing more factual information about what occurred in games, resulting in less variation in language. Alternatively, the Game of Thrones data contained mainly opinionated reactions to the series finale of the show resulting in more differentiated language across the Tweets. Additionally, while both of these datasets have roughly the same optimal coherence score of about 0.3, this score is particularly low for LDA models potentially given the highly similar nature of the Tweets and their short length.

As a result of these poor coherence scores, topics often appear very similar and contain many of the same words as each other as seen in Table 4.

While the words contained in each topic display some level of cohesion with each other, the reuse of some words in several of the topics suggests that many of the Tweets, regardless of their underlying message, contain the same verbiage. As a result, the LDA model struggles to create clear separations between topics. For

Table 4. Selection of Topics from LDA Model of World Cup Data

Topic Number	Top 10 Highest Probability Words
Topic 0	world cup good sorry russia champion fifa congratulation fra win
Topic 1	eng ronaldo messi world final england complete premierleague cup paulpogba
Topic 3	penalty frabel win team save proud france eng time threelion
Topic 4	mbappe fra player kylian young award golden fifa score ball
Topic 9	team france congratulation win african dear khaledbeydoun cut racism xenophobi

these reasons, when using LDA on a collection of Tweets queried from a specified conversation, intelligence analysts should expect topics to be more cyptic and difficult to manually label.

Additionally, if coherence is low for the LDA model, Tweets are less likely to share a strong connection with a topic since the words with the highest probability of topic membership will have less semantic connection with each other. As a direct result, Tweets will be categorized into topics with which they do not necessarily fit more often. For example, in the World Cup dataset using the LDA model presented in Table 4, the Tweet "Kylian Mbappé will donate everything he earns playing for France at the World Cup to charity" was connected with Topic 0. At a glance it would appear this Tweet is more suited for Topic 4 since it makes reference to the player which is the focus of the topic, however, the term "World Cup" which showed up in many topics results in a stronger connection with Topic 0. Due to such erroneous connections when implementing topic nodes in a network, the effectiveness of both influential user and community discovery can be diminished.

To determine the effectiveness of LDA on a dataset more representative of a generic query of Tweets, samples were taken from each dataset and joined to form a broad collection of Tweets. In this dataset, the increase in diversity of word usage and topic discussion allowed for LDA models to attain higher coherence values as seen in Figure

13.

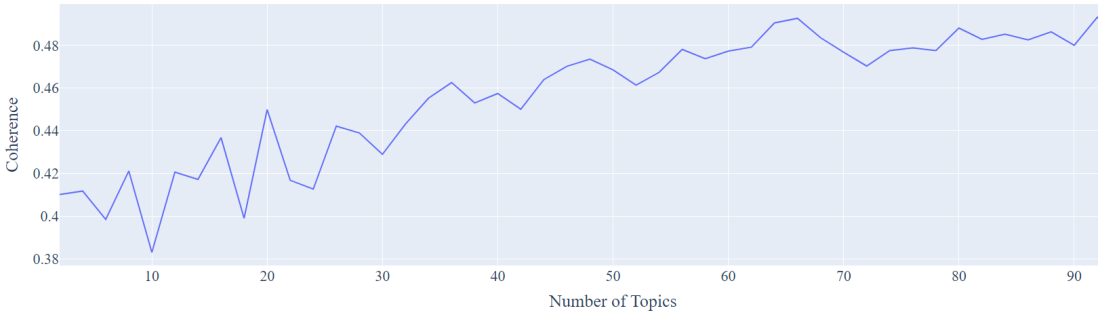


Figure 13. LDA Coherence Increases at Larger Number of Topics in Joint Data Than Other Networks

The increase in coherence past 50 topics indicates that, as expected, a larger amount of topic separation is possible as the diversity of Tweets in a dataset increases. This finding is critical for the creation of a network which integrates topic nodes as larger coherence scores are necessary to justify connections from Tweets to topics. This topic separation is apparent when inspecting word membership in topics in Table 5. In this model, most words strongly associated with the different topics appear to come from each of the different datasets and much more topic separation is apparent.

Table 5. Selection of Topics from LDA Model of Joint Data

Topic Number	Top 10 Highest Probability Words
Topic 2	good gt clinton tonight time today year vote forecast popular
Topic 4	win france team chance congratulation cut african senate dear racism
Topic 6	poll forecast clinton chance fivethirtyeight give final presidency late favorite
Topic 7	watch like not season episode final go eng night good
Topic 8	election live forecast result fifaworldcup new presidential york times people

When topics are well separated and higher coherence scores are observed, topic summarization is able to extract the most relevant sentences from Tweets related to a topic to provide an analyst with a concise summary of the Tweets. This capability overcomes the main shortcoming of LDA as further manual work would otherwise be

required to extract meaning from the related words to a topic. For example, when inspecting the topics in Table 5, Topics 6 and 8 both appear to be discussing the forecast of the 2016 presidential election and an analyst may not be able to gain any more insight than this using just the related words. Topic 6 was summarized as "FiveThirtyEight's polls-only forecast gives Clinton a 71% chance of winning the presidency #ElectionDay. RT @NateSilver538: Our final forecast of the year just published", while Topic 8 was summarized as "Trump favored to 77%—-Live Presidential Forecast - Election Results 2016 - The New York Times." When considering these summarizations, the added context that Topic 6 is mainly concerned with the conversation surrounding a forecast projecting Clinton to win, while Topic 8 is discussing a different poll projecting Trump as the winner allows for an analyst to gain a deeper understanding without the need for manual inspection of any Tweets. This context might provide further usage when considering the general sentiment of Tweets related to topics when they would have otherwise appeared to be indistinguishable. Another example of an informative and concise summary from this LDA model is Topic 4's summary, "80% of your team is African, cut out the racism and xenophobia. Africa did not win the #Worldcup France did. Africa did not even win it for France", revealing the controversy the topic was centered around while providing both stances of the users engaged in the conversation.

By implementing the topic layer of the multilayer network, users are connected to each other through the topic layer without the need for them to have been in direct communication with each other. This intricacy enables networks to more accurately portray the dynamics of a social network since users who are engaged in similar conversations have a higher likelihood of having seen each other's Tweets either in the past or future yet have not directly interacted, justifying the weaker and more distant connectivity between them. A visualization of the topic layer for the LDA

model referenced in Table 5 can be seen in Figure 14.

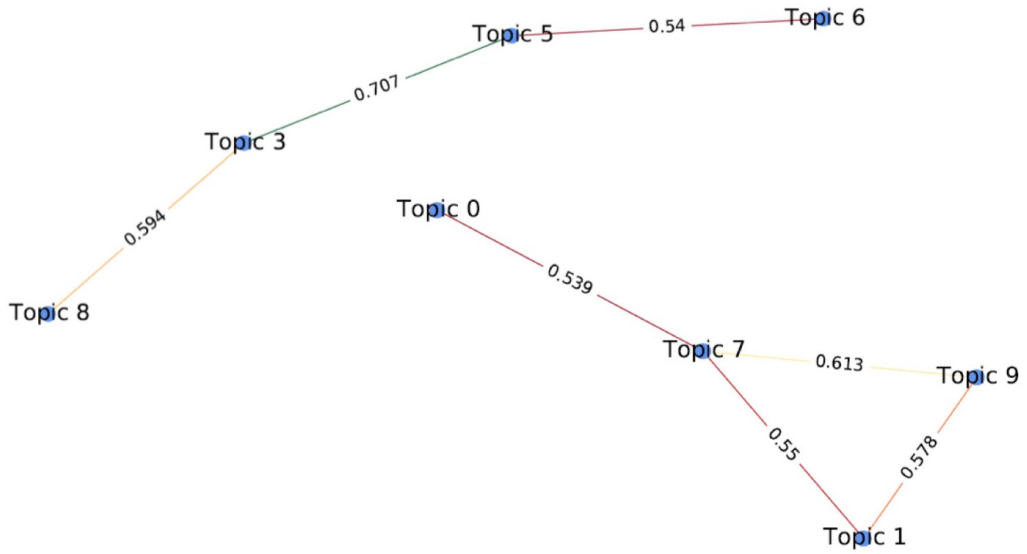


Figure 14. Topic Layer Connections from Joint Data Show Affiliations of Topics and the Strength of Their Connections

The weights on the edges in Figure 14 represent the cosine similarities between topics while edges are color mapped to show the relative strength of connection to other topics. Through inspection of the network, many of the election topics such as 6 and 8 are not directly connected; rather, they are connected through other topics with which they directly share stronger similarity reflecting this method's ability to capture intricacies in the conversations.

4.3 Influential Users

To assess each of the influential user discovery methods, results must be compared across datasets as well as within datasets after implementing the topic layer of nodes which should aid in revealing the potential benefit of its inclusion. As discussed in Section 3.4, four methods were identified for finding the most influential nodes in a weighted, directed network which will be compared in Section 4.3.1. Following a discussion of the description of a typical influential user in different networks, the

effect that a Twitter user's verification status has on his or her influence will be explored in Section 4.3.2.

4.3.1 Comparison of Methods

When considering the findings of influential user discovery on networks from the same dataset both with and without topic inclusion, the quality of an LDA model is a driving force in determining impact. As referenced in Section 4.2, low coherence scores and poor topic separation can result in weak and incorrect connections between users and topics which have a visible impact on the resulting influential users which are identified. Consider the results of influential user ranking on the COVID-19 dataset for the PageRank and eigenvector centrality methods in Table 6.

Table 6. Topic Layer Inclusion Results in More Bots and Smaller Accounts in Top Influential Users for COVID-19 Networks

Rank	PageRank	PageRank w/ Topics	Eigen	Eigen w/ Topics
1	realDonaldTrump	realDonaldTrump	Unverified user	News Blog
2	YouTube	ANI	realDonaldTrump	ANI
3	BorisJohnson	Bot	KamalaHarris	Bot
4	WHO	News Blog	JoeBiden	Bot
5	Change	GlobalPandemic .NET	TamaraMcCleary	Journalist
6	thehill	Journalist	Business Writer	Data Bot
7	CDCgov	Medical Journal	Journalist	Data Bot
8	GOP	Data Bot	CPHO_Canada	Medical Journal
9	narendramodi	Journalist	BorisJohnson	GlobalPandemic .NET
10	JoeBiden	Data Bot	DrRPNishank	Journalist

Findings show significant differentiation in rankings across methods and networks, and while ground truth data is not available to strictly assess the quality of methods, some assumptions can be made about the characteristics of influential users which should aid in determining performance. For instance, both methods identified several highly influential bots when applied to the topic networks which would appear to be

unlikely given that these bots either scrape data or share news articles but do not actively engage in discussion or hold views which would stimulate conversation from other users. In addition, many of these bots operate by replying to a user which mentioned them with a requested statistic or piece of information creating an artificially high amount of connections with other users. While this can be useful information to an analyst, bots such as these often cannot hold opinions and are therefore of less interest for this research. When considering the results of PageRank and eigenvector centrality on networks without topics, a number of high profile politicians and government organizations are present which is logical given the nature of data concerning COVID-19. While the results of these methods both appear similar and logically possible, the appearance of an unverified user with less than 1000 followers as the most influential user when using eigenvector centrality seems unlikely given the other influential users between the methods. In addition, PageRank does not recognize this user as highly influential calling into question the likelihood of this result.

Table 7. Top Influential Users for Joint Data Network Without Topics Are Varying Across Each Method With PageRank Displaying the Most Logical Results

Rank	PageRank	*HITS*	Betweenness	Eigen
1	NateSilver538	realDonaldTrump	GMA	538politics
2	538Politics	Unverified user	FiveThirtyEight	matthewjdowd
3	FiveThirtyEight	GOP	Ginger_Zee	RobMarciano
4	FIFAWorldCup	Unverified user	ringer	rickklein
5	Nate_Cohn	Unverified user	imarleneking	FiveThirtyEight
6	270toWin	Unverified user	SkyNews	Ginger_Zee
7	KhaledBeydoun	Unverified user	Professor	Peggynoonannyc
8	BBCMOTD	Unverified user	Lawrence	GMA
9	GMA	Unverified user	Author	Author
10	YouTube	Unverified user	ABC San Diego	rachel_handler

While topic inclusion appeared to have a negative impact on findings when the LDA model was poor, results are more promising when considering the impact on the joint dataset which has better topic separation and coherence. When comparing the

rankings in Table 7 for the base network to the rankings in Table 8 for the network with topics, the effect of both the topic layer and the method of influence ranking is evident.

In the base network ranking in Table 7, each ranking has very little overlap with few users appearing in two different ranking methodologies. Each ranking method appears to be heavily dictated by different node properties and, with the exception of $HITS$, each finds a set of users which consists mainly of verified users.

When including topics, the rankings in Table 8 show that, with the exception of PageRank, each method identifies far more low profile unverified users as highly influential.

Table 8. Top Influential Users for Joint Data Network with Topics Finds Good Results Using PageRank and Poor Results With Each Other Method

Rank	PageRank	$HITS$	Betweenness	Eigen
1	FIFAWorldCup	Unverified user	FiveThirtyEight	Unverified user
2	NateSilver538	Unverified user	NateSilver538	Unverified user
3	FiveThirtyEight	Unverified user	Unverified user	FiveThirtyEight
4	KhaledBeydoun	Unverified user	Unverified user	Unverified user
5	BBCMOTD	FiveThirtyEight	Unverified user	Unverified user
6	ManUtd	Unverified user	katz	Unverified user
7	realDonaldTrump	Unverified user	Unverified user	Unverified user
8	YouTube	Unverified user	Unverified user	Unverified user
9	brfootball	Unverified user	Unverified user	Unverified user
10	HNS_CFF	Unverified user	Unverified user	Unverified user

In both of these networks, PageRank appears to clearly outperform all other methods given its consensus between both networks on many of the top influential users. Many of these users identified by PageRank can be characterized by a verification on Twitter in addition to having hundreds of thousands if not millions of followers. These users are frequently mentioned and retweeted by other users often making up a significant portion of the nodes in the network with the highest in-degree count. The primary difference between the two rankings is that the exclusion of topics from the

network appears to result in more influential users in the political sphere while the inclusion of topics results in more influential users from the sports sphere. Given these networks are disconnected and PageRank calculates the long-term probability of being at any node in the graph, the dominance of the election and World Cup datasets is expected as these datasets have a higher frequency of interactions as referenced earlier in Table 2.

In applying the results of this analysis to the Indonesian Tweets dataset, PageRank was used to evaluate the performance of a network with and without topics because it produced the best results in the English datasets. The rankings for both networks can be seen in Table 9.

Table 9. Top Influential Users by PageRank for Indonesian Networks Are In Line With Previous Results and Affirm Inclusion of Topic Layer

Rank	PageRank	PageRank w/ Topics
1	Music Voting Bot	jokowi
2	Music Fan Page	BTS_twt
3	BTS_twt	Music Voting Bot
4	jokowi	CNNIndonesia
5	Show Fan Page	Music Fan Page
6	weareoneEXO	Music Charts Page
7	CNNIndonesia	Fahrihamzah
8	SMTOWN_Idn	agnezmo
9	Business Executive	detikcom
10	Music Fan Page	Music Fan Page

When using ten topics for LDA, coherence scores around 0.5 were observed which were consistent with the best scores obtained by models made with the joint dataset. Noting that this score is significantly higher than many of the scores obtained by models on the specific English datasets, it can be said that using the generic keyword of "Indonesia" produced a selection of Tweets from a broad category of conversations which allowed for better performance when separating topics. Due to the strong LDA model, the rankings on both networks found significant overlap and a number of users

who were either verified or had a large following. These accounts spanned many categories but were primarily involved in either music or politics. With a limited knowledge of Indonesian culture, it is difficult to determine which rankings might better represent the interests of social media users. Despite this, when referencing the characteristics of influential users in the English datasets, both networks produce rankings which are in line with previously obtained results supporting the inclusion of topics in networks to produce a more robust list of influential users.

4.3.2 Impact of Twitter Verification on Influence

Given the predominance of verified users in the influential user rankings from Section 4.3.1, particularly PageRank, the factors which might lead to higher numbers of influential verified users is desirable. While users on Twitter are free to interact with any other users which they please, Twitter verification might create a sense of authority which makes them disproportionately more influential. In addition, Twitter utilizes algorithms to present the most relevant Tweets to users on their timeline and within trending conversations which may also provide verified users a stronger platform for sharing their Tweets.

Although the impact of verification cannot be directly quantified, a strong prevalence of verified influential users across a number of networks would suggest that, at the least, these users have a higher probability of having a high influence rank. By bootstrap sampling two networks which have a large difference in density and node degree, the Game of Thrones and 2016 US Election networks, the appearance of verified users in different networks was calculated. Accordingly, 5,000 Tweets were sampled with replacement from each dataset for fifty trials and networks were created. Following this, PageRank was used to find the ten most influential users in each network and the number of verified users among them were counted.

Results showed that in the Game of Thrones networks which were characterized by fewer total nodes and higher density, 90.6% of influential users were verified with only eight unique unverified accounts appearing across the samples. When inspecting the US Election networks which were characterized by more nodes, lower density, and queries which specifically targeted Tweets interacting with political correspondents, 98.2% of influential users were verified with five unique unverified accounts appearing.

These results suggest a significant majority of influential users in these datasets are verified. Although users with the most followers are able to share their Tweets with the largest audiences, verification may provide a boost to users with fewer followers in dispersing ideas resulting in these findings. Similarly to follower count, a user's verification alone cannot be concluded to be the driving force behind his or her influence. Despite the lack of direct correlation, these observations highlight the need to pay particular attention to these users when conducting SNA given their higher likelihood of ranking high in influence despite making up a much smaller portion of the Twitter users.

4.4 Communities

Discovering communities in social networks uncovers relationships between large groups of users which can aid in conversation and interest group identification. A key aspect of SNA is identifying the largest groups of related users which, similar to topic modeling, uncovers the structures of users in a network and identifies what unifies them. While the addition of topic modeling to a network can help in finding which users are discussing particular topics, community identification finds the groups of users which are related to each other through their interactions. In using community identification algorithms on networks with topics, the goal is to identify larger groups of related users than what is possible on a network without topics and to identify

relationships which might exist between users in different conversations.

Communities were found using two methods, as discussed in Section 3.5, on networks with and without topics to determine the impact of assigning users to a topic of discussion on the ability to cluster nodes. The critical difference between these two clustering algorithms is the need to convert the network to a fully connected and undirected network to use the Leiden Algorithm. In contrast, the Greedy Modularity Algorithm (GMA) is able to cluster given the disconnected and directed networks which were created. The effect of this transformation and the addition of topics on community algorithm statistics can be seen in Table 10.

Table 10. Topic Layer Inclusion Decreases Modularity Using Both Methods in Election Networks

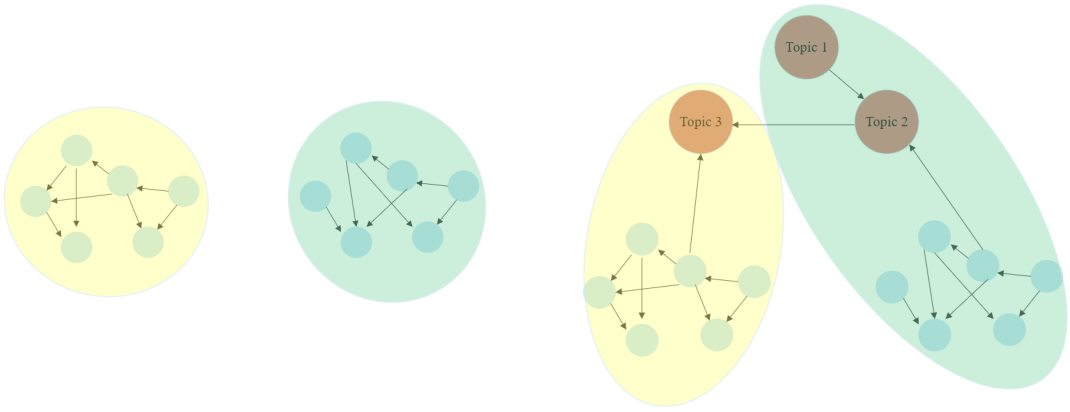
	GMA	GMA w/ Topics	Leiden	Leiden w/ Topics
Communities	1125	1059	822	172
Modularity	0.881	0.589	0.880	0.619
Coverage	0.969	0.861	0.921	0.853
Performance	0.901	0.922	0.901	0.924

Through the addition of topics, components of the network become connected through the topic layer resulting in smaller numbers of communities using both algorithms. Furthermore, by converting the network to an undirected one and connecting each of the disconnected components, the Leiden Algorithm is able to find the smallest number of communities as there are no longer several fragmented sections of users in the graph. In the base network with no topics, 1045 disconnected components exist while topic modeling reduces this number to 991. As expected, GMA finds only a slightly larger number of communities, 1125, than the number of components by only making partitions in the largest components and leaving most of the disconnected components to be in their own community. The results of community discovery are consistent with the influential user discovery for datasets with low coherence in that a decrease in performance is seen.

Table 11. Modularity is Lowest in the Large Communities in Election Networks

	GMA	GMA w/ Topics	Leiden	Leiden w/ Topics
Communities - Large	10	10	11	13
Modularity - Large	0.736	0.574	0.740	0.593
Communities - Mid	64	43	64	120
Modularity - Mid	0.974	0.966	0.974	0.990
Communities - Small	1051	1006	747	39
Modularity - Small	0.998	0.997	0.998	0.939

Modularity is also shown for a number of categories of subgraphs in the network in Table 11. Large communities are defined as those communities with more than 100 nodes, medium communities span the size of 10 to 100 nodes, and small communities are those with less than 10 nodes. Communities of these sizes were isolated into subgraphs and modularity scores were calculated to isolate performance across community sizes. By comparing modularities on this level, it is clear that the largest communities have the most significant negative impact on overall performance across all methods and networks. The reasoning for the reduction in performance, particularly in networks which include topics can be visualized in Figure 15.



(a) Communities Without Topics **(b) Communities With Topics**

Figure 15. Modularity Decreases Artificially Due to Inclusion of Topics in Networks

The inclusion of topic nodes in the networks forces many of the largest communities, which likely contain some users who are linked directly to a topic, to contain

topic nodes, reducing overall modularity. The decrease in modularity as a result of larger community subgraphs is evident in that the modularity scores for the medium and smaller sized community subgraphs in the topic network are comparable to those in the networks without topics. In summary, although modularity scores, which are typically used as a measure of a graph partition's performance, are lower in the topic networks, the cause is likely not due to a substantial decrease in performance. This claim is evidenced by the similar partition coverage and performance scores. Instead, modularity scores see a somewhat artificial reduction due to the classification of topic nodes into communities with user nodes with which few connections exist. This same flaw can be seen in Table 12 depicting the community statistics in the joint data, although the overall modularity scores for topic networks are noticeably higher likely due to the better performance of these LDA models.

Table 12. Modularity is More Comparable In Joint Networks With Topics Due to Higher Coherence Scores

	GMA	GMA w/ Topics	Leiden	Leiden w/ Topics
Communities	3192	2608	1990	585
Modularity	0.969	0.799	0.973	0.824
Coverage	0.983	0.945	0.870	0.920
Performance	0.978	0.959	0.983	0.961

When inspecting the modularity scores for the subgraphs of differing community sizes in Table 13, large communities still see the largest reduction in modularity while medium and small community subgraphs in topic networks maintain scores similar to those without topics. It is also clear in both the 2016 US Election and joint networks that the category responsible for the large decrease in total communities when using the Leiden Algorithm is the small communities.

In addition, despite the addition of low weight edges to connect the components of the network, the modularity scores of partitions created by the Leiden Algorithm score higher than those of the GMA communities. The significance of this improvement

Table 13. Medium and Small Community Modularities are Similar With and Without Topics in Joint Networks

	GMA	GMA w/ Topics	Leiden	Leiden w/ Topics
Communities - Large	15	11	19	12
Modularity - Large	0.823	0.770	0.826	0.775
Communities - Mid	74	38	84	338
Modularity - Mid	0.976	0.954	0.967	0.996
Communities - Small	3103	2559	1887	235
Modularity - Small	0.999	0.999	0.999	0.995

as it relates to the required graph transformations would require further research to ascertain.

Similarly, when performing the same analysis on the Indonesian dataset, overall modularity scores are decreased in topic networks as seen in Table 14. The larger decrease when compared to the joint networks, for which coherence scores are similar, could be attributed to the overall smaller size of the networks as this would mean that topic connections have a more significant effect since fewer total edges and nodes exist.

Table 14. Overall Modularity is Lower in Indonesian Networks Despite High Coherence Due To Smaller Number of Nodes and Edges

	GMA	GMA w/ Topics	Leiden	Leiden w/ Topics
Communities	669	637	538	204
Modularity	0.944	0.693	0.948	0.752
Coverage	0.967	0.969	0.901	0.939
Performance	0.971	0.957	0.972	0.965

Just as before, similar modularities in the medium and small sized community subgraphs in both the base and topic network can be observed in Table 15.

The benefit of community detection as it relates to SNA is the discovery of connected groups of users which in many ways is similar to the design of connecting users through their Tweets to the topic layer. In community detection, the Tweets sent by the groups of connected users can be analyzed by content as well as emotion

Table 15. Community Subgraph Statistics in Indonesian Networks

	GMA	GMA w/ Topics	Leiden	Leiden w/ Topics
Communities - Large	9	8	10	13
Modularity - Large	0.810	0.638	0.831	0.686
Communities - Mid	52	39	50	109
Modularity - Mid	0.953	0.923	0.953	0.984
Communities - Small	608	590	478	82
Modularity - Small	0.996	0.992	0.996	0.979

classification to learn more about the interests and sentiments of the group. When including topics in a network, the users who are connected to topics and those they interact with have already been identified as members of a community by design allowing for the same future analyses. Through either approach or a combination of the two, intelligence analysts are afforded a number of options for analyzing a subsection of Tweets from what are identified to be strongly connected users.

4.5 Impact of Data and Queries on Networks

Vital to all of the aforementioned social network analysis is the querying method by which data is collected. As shown through comparison of the topic networks to the base networks, LDA model topic separation and coherence are crucial to finding useful results when considering both influential user and community discovery using a network which includes both topics and users. One of the biggest driving forces of coherence scores appears to be the scope of the Tweets being collected because it directly impacts the number of conversations found within a dataset. While LDA can be useful tool for finding distinct conversations as illustrated using the joint data, it is significantly less useful when trying to divide already specific conversations into smaller ones. For instance, even though Tweets discussing the two major political parties in the United States are logically part of different conversations, LDA will often struggle to differentiate them since both types of Tweets typically contain much of

the same language. While influential users and communities can still be found in datasets which contain specific conversations, results show that these networks are best when not including topics. From the limited number of datasets analyzed, this cutoff point for finding meaningful results with topic networks appears to be when LDA coherence is roughly 0.5.

In a similar way, if Tweets are collected from queries which mandate references to a small number of users such as in the election data, it should be expected that these users will appear as the most influential. In addition, if these users have a broad and active following on Twitter, significant redundancies in data should be expected in the form of Retweets from many users on the same Tweet. For instance, in the election data, despite over 42,000 Tweets, only about 15,000 unique Tweets exist in the data due to Retweets while four Tweets appear over 1,000 times. While redundancies in data are useful and expected due to viral Tweets which elicit many user interactions, they can have a substantial impact on many of the findings. When considering the number of Tweets in a sample required to find consistent results, these redundancies significantly reduce the sample size required.

To illustrate the necessary sample size, the top ten influential users were found in the election data using all observations and, in the joint data, 50,000 observations were sampled to establish a ground truth expectation provided large amounts of data. Next, bootstrap samples of various numbers of Tweets were taken from each dataset 50 times, and the average number of identical influential users between the sample and baseline was calculated. The results of this analysis can be seen in Table 16.

These results show that when specific queries are used to collect data, fewer Tweets are often required in a sample to produce similar results to those in larger datasets. However, if more general queries are used which produce data from a number of different conversations, more Tweets are needed to find results consistent with larger

Table 16. Less Samples Are Required to Find Higher Percentage of Similar Influential Users in Election Data Due to Specific Query

Tweets per Sample	2016 US Election Data	Joint Data
250	61.6%	38.2%
500	65%	40.6%
1000	65.6%	44.2%
2500	67.6%	46%
5000	71%	46.8%

datasets as more variability is present.

V. Conclusions and Recommendations

In concluding the work of this research, the key questions asked in Chapter I will be revisited and the findings associated with each will be summarized in Section 5.1. Additionally, the methodologies presented and their effectiveness in this problem context will be assessed and recommendations for use of this research by intelligence analysts will be provided in Section 5.2.1. Finally, potential future work will be laid out in Section 5.2.2 to include a discussion of the integration of this research and another which concerns sentiment analysis of Indonesian Tweets.

5.1 Conclusions

This research sought to provide insights to the the military intelligence community regarding open source intelligence, specifically social network analysis of Twitter data from the region of Indonesia. Twitter is a social media platform which gives users the capabilities to share their opinions as well as share the opinions of others to large audiences. Through this medium, groups of users share their support or disdain for current events and other topics and have had success in the past of effecting real-world policy changes as a result of their social media activity. As a result, it is necessary intelligence analysts have a strong understanding of the climate of social media in regions of interests toward certain topics as all decisions will be judged by online communities.

Through previous research in the field of SNA, this research broke down the process of gauging the online climate of social media users into a number of steps. First, LDA topic modeling was used to identify the main conversations present on social media and text summarization techniques were applied to topic to gain a better understanding of the context of Tweets concerning a topic. Next, networks of users

were created with edges representative of the interactions between users including mentioning, replying, and retweeting. Amending this network, an additional layer of topic nodes with intra-layer connections representing topic similarities and inter-layer connections representing Tweet-topic similarities was implemented. Both networks were then analyzed to determine if the topic addition allowed for better understanding of user activity. When analyzing these networks, a number of methods for both influential node and community discovery were presented to determine the most important users in the network as well as the groups of users who were strongly connected. Finally, connections were identified between the methods by which Tweets were collected and the resulting network structure to guide future data collection.

In conducting this research, it was determined that LDA topic modeling was only appropriate for datasets of Tweets which were collected using a broad query such as a generic keyword, date, or location because it yielded Tweets which discussed a number of different conversations or users and allowed for high coherence scores and topic separation. Accordingly, when constructing networks, topic inclusion provided a more diverse list of plausible influential users when coherence scores were high. A weighting scheme was presented for quantifying user interactivity which leveraged both the frequency of interaction as well as the quality of interaction as defined by the frequency of each potential user interaction on Twitter. Additionally, both user-to-topic connections and topic-to-topic connections were quantified through text similarity metrics. When comparing influential node discovery methods in the context of this research, it was determined that PageRank found the most promising results in each of the datasets and when including topics in the network. Two community discovery methods were compared including one which transformed the network to a connected and undirected graph and modularity scores were assessed to show that each method presented plausible results pending future research using more well de-

fined data with pre-determined community membership. Data queries from an API were also investigated revealing that the more broad queries allowed for versatility through topic layer inclusion in the network while specific queries such as interactions with particular users yielded networks with large, predictable clusters and influential users. All methods were found to be consistent across English and Indonesian data.

These findings represent a baseline approach for performing SNA on Tweets scraped using an API as well as a number of guidelines for interpreting results and gathering future data. Using the insights of this research, intelligence analysts can conduct OSINT and gather a number of high level findings which can be tracked over time to better understand the network of users in an area of interest. Topics of conversations and the users who are involved with them can be identified as well as communities of users to perform further analysis such as emotion classification to determine climate toward contentious issues. In addition, users central to the diffusion of ideas in the network can be identified to determine the likely beacons of information and their stance on similar issues to the ones of interest.

5.2 Recommendations and Future Work

5.2.1 Recommendations

When deploying this research for use, intelligence analysts should ensure that sufficiently generic queries are used to collect data such that the resulting sample of Tweets is representative of a number of conversations concerning a topic. If the desired area of interest is quite specific, Tweets should be gathered broadly either across time or region as this will allow for greater diversity of opinion and word usage in the Tweets. If neither of these collection methods are possible, intelligence analysts should treat results from topic modeling and accordingly networks including topics with skepticism and rely more heavily on networks without topics.

Since the results of these methodologies are entirely dependent on a random sample of Tweets, the most meaningful results will most likely be found when collecting Tweets from a similar query across multiple days. Focal points on social media shift daily and, while the appearance of an influential user in a network does suggest that they might be a relevant actor of interest, a period of influentiality across multiple days and samples of Tweets is more suggestive of influence. The appearance of users in a community across many days provides more evidence of their belonging than a single sample observation. When inspecting visualizations as they relate to influential users, changes over time will be more relevant. For instance, large groupings of users around an influential user not previously identified in other samples should be investigated to determine if a particular Tweet or current event led to his or her rise in influentiality.

As previously mentioned, topic modeling should not be used unless LDA models are able to achieve above a 0.5 coherence score when inspecting the coherence plot by number of topics. It should be worth noting that coherence scores generally increase monotonically as topics increase however, analysts should seek to find the point in the graph where substantial improvement is no longer seen as topic number increases. This elbow point should be used as the basis for whether or not to include topics in the network as well as the optimal number of topics to use in the LDA model.

When finding influential users, PageRank has been identified through the findings of this research as the most reliable method for producing plausible results. As such, it should be used in future applications by analysts in both networks with and without topics. Additionally, both community identification methods appeared to show success and currently there is not a clear distinguishable better method for use on social networks. Results using both methods should be found and the Tweets of the users in them should be analyzed as desired to determine potential differences in

the methods before one method can be adopted exclusively.

5.2.2 Future Work

Gaining access to the Twitter API would allow for customized query searching of Tweets and would aid substantially in providing recommendations to intelligence analysts of the best practices. This capability would also permit repeated queries over time to study the changes in resulting networks allowing for more informed recommendations of how to analyze similar networks over time. Additionally, API access would present the opportunity for gathering data from different spheres of conversation from Indonesia as much of the data in this research was quite limited in scope. When conducting analysis, API access would allow for advanced search of user features such as how often they Tweet, their number of followers, and their verification status providing the data needed to build predictive models necessary to understand which user features have the greatest impact on influentiality. In addition, the impact of the presence of an influential user on the size of a community could be investigated to provide insights to analysts on other factors which are instrumental in ranking influence.

Additionally, with access to a larger number of datasets, the quality of the Leiden Algorithm in this context could be better quantified through determining the proportion of users found in each community which originated from the same dataset when creating a network from Tweets sampled from each dataset. While crossover between datasets should be expected to some degree, a good community discovery method on a network which was artificially connected should be able to distinguish which dataset most Tweets came from.

While this research treated topic modeling and community discovery as separate endeavors, future work can utilize the community creation to attempt to find

topics amongst the communities. This method may help to overcome one of the most significant barriers of topic modeling with Tweets which is the small number of words contained within each document. If community separation finds groups of users Tweeting about similar topics, these Tweets can be combined to form documents consisting of many Tweets, and LDA can be used with these larger documents to attempt to find better topic separation and coherence scores.

When considering connections made between nodes in the network, specifically concerning topic layer connections, further research should explore the cutoff values which justify a connection between topics or a topic and a user. While this research limited topic connections to those with a cosine similarity greater than 0.5 and topic-to-user connections as those with strengths in the top 25% of these connections, more detailed and possibly manual analysis could identify more fitting values. In addition, more Tweet features could be included when quantifying connection strengths between a topic and user. For instance, a user's number of followers or verification status could be a more representative way of modeling this connection. Since this situation attempts to capture instances where a user is scrolling through Tweets related to a topic and finds one with which he or she are interested, popularity on Twitter could be a significant factor since Twitter often promotes Tweets which already have a high number of likes.

A broad category of future research in this topic would be practical applications of the results which were presented. For instance, if intelligence analysts want to understand the general emotion of Tweets which were interacting with an influential user, topic, or community, this research should be paired with the ongoing research on this topic concerning emotion classification of Indonesian Tweets. Through this application, it is possible to learn more substantial differences between topics which appear similar in word usage as well as influential users whose profiles have similar

characteristics. Similarly, by tracking emotion toward a topic over time, analysts could learn how public opinion has changed and Tweet summarization can be used at inflection points to learn more about why this might have happened.

Bibliography

25th Infantry Division (2021), 'OSINT analytics'.

Aiello, L. M., Barrat, A., Cattuto, C., Ruffo, G. and Schifanella, R. (2010), 'Link creation and profile alignment in the aNobii social network', *2010 IEEE Second International Conference on Social Computing* .

Alatas, V., Chandrasekhar, A. G., Mobius, M., Olken, B. A. and Paladines, C. (2019), 'When celebrities speak: A nationwide Twitter experiment promoting vaccination in Indonesia'.

Alizadeh, M., Weber, I., Cioffi-Revilla, C., Fortunato, S. and Macy, M. (2019), 'Psychology and morality of political extremists: evidence from Twitter language analysis of Alt-right and Antifa'.

Allard, K. (1996), *Command, control, and the common defense*, rev. edn, Washington, D.C. : National Defense University.

Arunarsirakul, A. (2020), 'Language preference on social media - survey report'.

Ashbrook, C. C. and Zalba, A. R. (2021), 'Social media influence on diplomatic negotiation: Shifting the shape of the table', *Negotiation Journal* pp. 83–96.

Bakshy, E., Hofman, J., Mason, W. and Watts, D. (2011), Everyone's an influencer: Quantifying influence on Twitter, pp. 65–74.

Barrios, F., López, F., Argerich, L. and Wachenchauzer, R. (2016), 'Variations of the similarity function of TextRank for automated summarization'.

Bessi, A. and Ferrara, E. (2016), 'Social bots distort the 2016 U.S. Presidential election online discussion', *First Monday* **21**.

Bhavnani, V., Galphat, Y., Bhawsinghka, G. and Golani, J. (2021), 'A survey on detecting influential user in social networking', *2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)* .

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent Dirichlet Allocation', **3**(null), 993–1022.

Bródka, P. and Kazienko, P. (2012), 'Multi-layered social networks'.

Carley, K. M., Malik, M., Landwehr, P. M., Pfeffer, J. and Kowalchuck, M. (2016), 'Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia', *Safety Science* **90**.

Clauset, A., Newman, M. E. J. and Moore, C. (2004), 'Finding community structure in very large networks', *Physical Review E* **70**(6).

Dewi, F. K., Yudhoatmojo, S. B. and Budi, I. (2017), 'Identification of opinion leader on rumor spreading in online social network Twitter using edge weighting and centrality measure weighting', *2017 Twelfth International Conference on Digital Information Management (ICDIM)* .

Erlandsson, F., Bródka, P., Borg, A. and Johnson, H. (2016), 'Finding influential users in social media using association rule learning', *Entropy* **18**, 164.

Hecht, B. and Stephens, M. (2014), A tale of cities: Urban biases in volunteered geographic information, *in* 'Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014', Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, AAAI press, pp. 197–205.

Hermawan, A. (2016), 'Framing the 2014 Indonesian presidential candidates in newspapers and on Twitter', *repository.arizona.edu* .

Jin, X. (2020), 'Exploring crisis communication and information dissemination on social media: Social network analysis of Hurricane Irma Tweets', *Journal of International Crisis and Risk Communication Research* **3**(2), 179–210.

Jiwanggi, M. and Adriani, M. (2016), 'Topic summarization of microblog document in Bahasa Indonesia using the phrase reinforcement algorithm', *Procedia Computer Science* **81**, 229–236.

Kalepalli, Y., Tasneem, S., Teja, P. D. P. and Manne, S. (2020), 'Effective comparison of LDA with LSA for topic modelling', *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* .

Kleinberg, J. M. (1999), 'Authoritative sources in a hyperlinked environment', *J. ACM* **46**(5), 604–632.

Kolda, T., Bader, B. and Kenny, J. (2005), Higher-order web link analysis using multilinear algebra, *in* 'Fifth IEEE International Conference on Data Mining (ICDM'05)', pp. 8 pp.–.

Legradi, J. (2009), An exploratory social network analysis of military and civilian emergency operation centers focusing on organization structure, Master's thesis, Air Force Institute of Technology.

Lim, M. (2003), 'The internet, social networks and reform in Indonesia', *Contesting Media Power. Alternative Media in a Networked World* p. 273–288.

Malliaros, F. and Vazirgiannis, M. (2013), 'Clustering and community detection in directed networks: A survey', *Physics Reports* **533**.

Oro, E., Pizzuti, C., Procopio, N. and Ruffolo, M. (2018), 'Detecting topic authoritative social media users: A multilayer network approach', *IEEE Transactions on Multimedia* **20**(5), 1195–1208.

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A. and Menczer, F. (2021), 'Uncovering coordinated networks on social media: Methods and case studies', *Proc. AAAI Intl. Conference on Web and Social Media (ICWSM) 2021* .

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999), 'The PageRank citation ranking: Bringing order to the web.', (1999-66). Previous number = SIDL-WP-1999-0120.

Poblete, B., Garcia, R., Mendoza, M. and Jaimes, A. (2011), Do all birds tweet the same?: characterizing Twitter around the world, pp. 1025–1030.

President of the U.S. (2021), *Interim National Security Strategic Guidance*, Washington, DC.

Pudjajana, A. M., Manongga, D., Iriani, A. and Purnomo, H. D. (2018), Identification of influencers in social media using social network analysis (SNA), *in* '2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)', pp. 400–404.

Rahmadan, M. C., Hidayanto, A. N., Ekasari, D. S., Purwandari, B. and Theresiawati (2020), 'Sentiment analysis and topic modelling using the LDA method related to the flood disaster in Jakarta on Twitter', *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)* pp. 126–130.

Ramokhoro, M., Maboea, P., Holtzhausen, T. and Khoza, P. N. (2020), 'Towards an analytical probe for Twitter information flow micro-structure', *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* .

Röder, M., Both, A. and Hinneburg, A. (2015), 'Exploring the space of topic coherence measures', *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining* pp. 399–408.

Salehi, A., Ozer, M. and Davulcu, H. (2018), Sentiment-driven community profiling and detection on social media, *in* 'HT 2018 - Proceedings of the 29th ACM Conference on Hypertext and Social Media', Association for Computing Machinery, Inc, pp. 229–237.

Sari, D., Ahmad, J., Hergianasari, P., Pratiwi, C. and Nur, A. (2021), 'Quantitative study of the cyber-nationalism spreading on Twitter with hashtag Indonesia and Malaysia using social network analysis', *Media Watch* **12**, 161–171.

Scott, J. (2000), *Social Network Analysis; A Handbook*, SAGE Research Methods, SAGE Publications Ltd.

Sheth, A., Shalin, V. and Kursuncu, U. (2021), 'Defining and detecting toxicity on social media: Context and knowledge are key'.

Sutrisno, B. and Ariesta, Y. (2019), 'Beyond the use of code mixing by social media influencers in Instagram', *Advances in Language and Literary Studies* **10**, 143.

Syed, S. and Spruit, M. (2017), Full-text or abstract? examining topic coherence scores using Latent Dirichlet Allocation, *in* '2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)', pp. 165–174.

Tang, L. and Liu, H. (2011), 'Leveraging social media networks for classification', *Data Min. Knowl. Discov.* **23**, 447–478.

Traag, V. A., Waltman, L. and van Eck, N. J. (2019), 'From Louvain to Leiden: guaranteeing well-connected communities', *Scientific Reports* **9**(1).

Tsopze, N. and Domgue, F. G. (2021), 'Boolean factor based community extraction from directed networks with the non reciprocal link relationship', *Information Sciences* **569**, 544–556.

Tsugawa, S. and Ohsaki, H. (2015), 'Negative messages spread rapidly and widely on social media', *Proceedings of the 2015 ACM on Conference on Online Social Networks* pp. 151–160.

Venkatesan, M. and Prabhavathy, P. (2019), 'Graph based unsupervised learning methods for edge and node anomaly detection in social network', *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)* .

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
24–03–2022	Master's Thesis	August 2020 — March 2022

4. TITLE AND SUBTITLE

Analysis of Twitter Networks to Aid Open Source Intelligence Capabilities: A Multilayer Network Approach

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Logan, Austin P., 2d Lt

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
Wright-Patterson OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-22-M-146

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

The Research and Analysis Center-Monterey
Wade, Brian, LTC, PhD
1 University Circle
Monterey CA 93943
brian.wade@nps.edu

10. SPONSOR/MONITOR'S ACRONYM(S)

TRAC-MTRY

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

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION A:
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Open Source Intelligence using social media is a practice which gives military intelligence analysts a window into the thoughts and minds of an online population. Using Social Network Analysis, user interactions on Twitter will be modeled as a weighted and directed network. Topic modeling through Latent Dirichlet Allocation uncovers the topics of discussion in Tweets and is then integrated into a multi-layer network which allows users to be connected to the conversations with which they have participated. Influential users in this network as well as highly connected groups of individuals are then discovered to paint a picture for intelligence analysts of the online landscape with which they are dealing.
The results of this research demonstrate that the inclusion of topics in the social network allows for more robust findings in influential users when analysts collect Tweets from a variety of discussions through the use of more general search queries. PageRank was identified as the best performing influence ranking method for this problem context and two potential community identification methods were analyzed.

15. SUBJECT TERMS

Social network analysis, networks, multi-layer networks, natural language processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			LTC Phillip M. LaCasse, AFIT/ENS
U	U	U	UU	94	19b. TELEPHONE NUMBER *(include area code)* (262) 470-7549; phillip.lacasse@afit.edu