Analyzing COVID-19 Twitter Data Using A New Graph-based Modeling Technique

Ola Karajeh
Computer Science
Virginia Tech
Blacksburg, VA, USA
okarajeh@vt.edu

Arwa Alebrahim
Computer Science
Virginia Tech
Blacksburg, VA, USA
arwaa@vt.edu

Chris Luersen
Computer Science
Virginia Tech
Alexandria, VA, USA
cluersen@vt.edu

Lauren Smith
Population Health Sciences
Virginia Tech
Blacksburg, VA, USA
laur12@vt.edu

ABSTRACT

With the rise of social media, massive amounts of information have become available to researchers, giving special insight into the thoughts and behaviors of society not accessible before. This can be of particular importance to government and health officials for predicting trends in populations, in this case, urban settings. Using various text-based modeling implementations in tandem with machine learning algorithms, we tested different approaches for predicting informational and non-informational tweets about Covid-19 [1]. These approaches included 2 methods from relevant literature: A Term Document Matrix (TDM) and a graph-based, PageRank, HITS, and graph density (GB-PHD). While we propose a new method, using graph-based text representation with a centrality weight model (GB-CW) [2, 3]. Then we applied three algorithms, a J48 Decision Tree, a Support Vector Machine, and a Naive-Bayes Classifier, with different k-cross values to the models to determine the F-measure value. We also discuss the downsides of using multi-classifiers versus J48 for determining accuracy, as well as comparing build time on the different model tests. The implications of this study are to provide insight into the type of public response by locality using the geodata associated with the Twitter ID, to help health officials plan and act out an appropriate pandemic response.

INTRODUCTION

The current issue we are looking to address is providing an accurate prediction on informational tweets in the sphere of social media data, in this case, Twitter.[1] Measuring public opinion can prove difficult, factors like fact-checking, context and even the syntax of text-based representation can significantly affect the accuracy of the data extracted. To solve this, we first look at the source of the data which comes from re-hydrated Twitter files(.tsv). Rehydrating is the process of extracting the text data with a corresponding ID, and then, to improve accuracy, the data is cleaned, which includes the removal of URLs, punctuation, hashtags, and converting all letters to lowercase. Now, classifying the data, comes in two parts, first, using an appropriate model to represent the text input, and second comparing machine-learning algorithms, to determine after building the model which produces the most accurate output. The text-based representation models we tested include a Term Document Matrix (TDM), graph-based PageRank, HITS, and graph density (GB-PHD), and our newly proposed model, graph-based centrality-weight (GB-CW). [2][3] The first model listed, TDM, is a matrix that describes the frequency of all terms occurring in the text by building a numerical matrix in which each row represents the vector form of tweets, and each column represents the set of words in the corpus. This method is proven to be effective with high accuracy, however, scalability becomes an issue with the increase of data because if any change to the matrix occurs the whole M matrix must be rebuilt again[4]. Also due to the presence of sparse matrices like grammatical errors, slang, and informality, TDM fails to regard the relationship between words and the order of terms within a tweet. To avoid these issues, we also tested two models using the advantages of graph theory to approach text representation. The next model is GB-PHD, which is mainly used with topic classification. While GB-PHD does solve the issue of scalability that we had with TDM through combining and averaging the outputs of the PageRank, HITS, and graph density algorithms, it now could present a trade-off of low accuracy. Since the importance of analyzing large data with high accuracy can depend on an increased emphasis on centrality measures and weight between edges, we developed a new model, GB-CW[5][6][7]. The centrality is determined by comparing the set of elements (vertices or nodes), and the relations between these vertices (edges). In this case the central nodes of importance that we chose surround the context of Covid-19. Furthermore, the weight between edges, used in a weighted directed graph, was calculated using these nine metrics: Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC), Harmonic Closeness Centrality (HCC), and Eccentricity Centrality (ECC), In degree (In), Out degree (Out), the weight of an edge (W), the number of words in each tweet (n). This model is purported to analyze large datasets with high accuracy while retaining vertices relationships between elements.

After determining the models, we implemented three machine learning (ML) algorithms, with 2 k-cross values, one with 5 folds, and the other with 10. These three classifiers were a J48 Decision tree algorithm (J48), a support vector machine (SVM) algorithm, and a Naive-Bayes classifier(N-Bayes) algorithm. J48 is a popular algorithm used in classifying medical data, and while it requires large memory usage and produces low accuracy when used in small instances with small amounts of training data, it is one of the best for large data sets that need to be examined categorically and continuously. An SVM is another popular model largely used with

classification, particularly between two classes along a hyper-plane, which fits our interests in differentiating between informational versus non-informational tweets. N-Bayes is the last algorithm we tested and works well with text classification because of its simplicity and conditional independence assumption. However, since it converges quicker than other algorithms, requiring less model training time, it sacrifices accuracy when compared to the other two when using large amounts of data.

This paper looks to answer the following questions: are the graph-based models able to reduce the dimensionality of the dataset? Does the dataset extracted from the graph model take less time in building the model compared to the TDM based model? Do the graph models achieve higher accepted accuracy than the TDM model? Between the two graph-based models which are more efficient in terms of training time and accuracy? This paper focuses on the proposed graph-based text representation model, GB-CW, for tweet analysis, with special consideration on node centrality and the weight of the edges, determined by the nine metrics listed above, to represent key features with regards to Covid-19 information. What differentiates this paper from similar studies is the approach. Relevant literature uses a variety of different models to test correlations between Twitter data and Covid-19 information. The most similar using TDM and GB-PHD to feed ML algorithms, hence why we test those two as well. However, these two models put more emphasis on word embedding and keyword extraction, with TDM requiring long training times and GB-PHD utilizing different methods for its modeling. Our contribution depends on just a few attributes for modeling, which helps cut down on training time[8] [9] [10] [11]. While utilizing the centrality-weight method goes a step beyond word embedding and keyword extraction by retaining the relationships between words in the text, which will ultimately help produce more accurate results in a faster time frame.

RELATED RESEARCH

Since December 2019, COVID-19 has been spreading rapidly across the globe, giving rise to related conversations all over social media platforms as a result. These social media platforms have become the new method for information sharing and consuming among the public and are specifically important during times of a wide-scale disaster or emergency such as this pandemic. People tend to rely on these platforms for their main source of science news, with 59% of Twitter users describing Twitter as good or extremely good for sharing preventative health information [12]. Social media content is not only news sharing for these individuals but also a tool for various stakeholders such as researchers, first responders, and government officials during times of an emergency disaster. Available content can provide insight into the economic and social patterns throughout the situation, identify certain needs of targeted areas and assess the level of damage that has occurred. For example, public health professionals can use data retrieved from social media posts along with their geotag to determine the scale of the event or the number of people impacted in certain areas [15]. There can also be direct application such as an emergency responder identifying information about an individual emergency case [15]. In the best interest of both the general public understanding and stakeholder applications, information accuracy and efficiency is crucial; however, with social media platforms, it is common that posts relating to emergency disasters contain false information, conspiracy theories, and irrelevant content.

Due to the extent of health information and news regarding COVID-19, Twitter has released datasets of related tweets for open research purposes. This data has many implications in COVID-19 research as Twitter has been the predominant social media platform in providing medical information [18]. Several studies have utilized COVID-19 Twitter data thus far to assess various research questions pertaining to COVID-19 communication and spread. The Twitter messages contain certain keywords, IDs, and tags that allow researchers to gather desired information by training machine learning models or perform any analysis [12]. This can span from identifying misinformation or public views to identifying locations and the flow of information. One study used a machine learning approach to identify common unigrams and bigrams, salient topics and themes, and sentiment analysis [19]. This study used Unsupervised Machine Learning Latent Dirichlet Allocation (LDA) to analyze the patterns and themes of the tweets and a sentiment analysis, or a natural language processing approach (NLP), to classify the emotions expressed in the tweets [19].

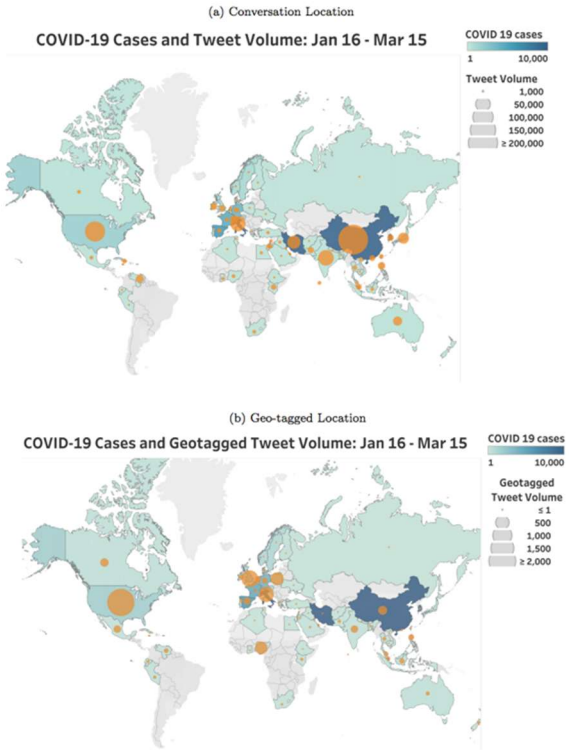

Figure 1: Tweet Volume and COVID-19 Cases [12]

Spatial and/or temporal analysis of the COVID-19 Twitter data provides additional analysis which has applications for studying topics such as geographical distribution of COVID-19 information and cases, flow of COVID-19 information on Twitter, and epidemiological comparison of related tweet patterns with actual virus patterns. With the user's permission, Twitter is able to access the user's location with a Global Positioning System (GPS) to provide data on the geo-coordinates of tweet location; this creates the geo-tagged tweets for spatial and temporal analysis [13]. Figure 1 was produced by a study that was able to use these geo-tags along with country keywords to map confirmed COVID-19 cases with conversation location tweets and with geo-tagged tweet mentions of COVID-19 [12]. Another study used geo-tags with COVID-19 sentiment scores to map the world view of COVID-19, seen in figure 2 [13].

Figure 2: World view of COVID-19 Sentiment [13]

These useful mapping techniques can be applied by public health responders and government officials to understand areas of necessity. To understand the flow of information, a study on risk and crisis communications of government agencies and stakeholders used temporal data from the COVID-19 Twitter data [14]. This study used a dynamic network analysis with 67 nodes to represent federal stakeholders, health departments, state health departments, and the WHO [14]. Several aggregated communication networks were formed including figure 3 showing the agencies and figure 4 showing the flow of

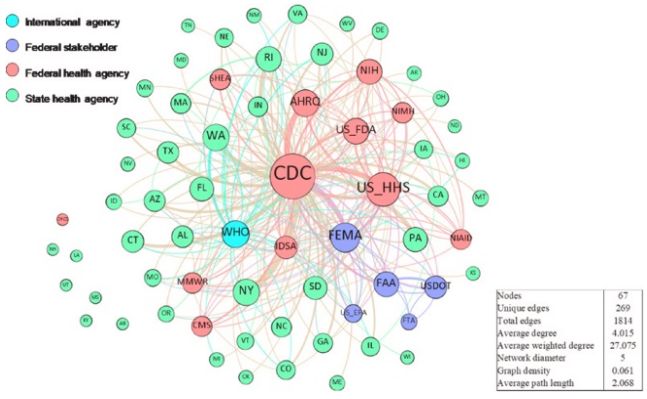

Figure 3:Aggregated communication network [14]

information over 16 weeks [14]. For another spatial and temporal application, a separate study looked at comparing the Twitter data, a COVID-19 simulation, and COVID-19 confirmed

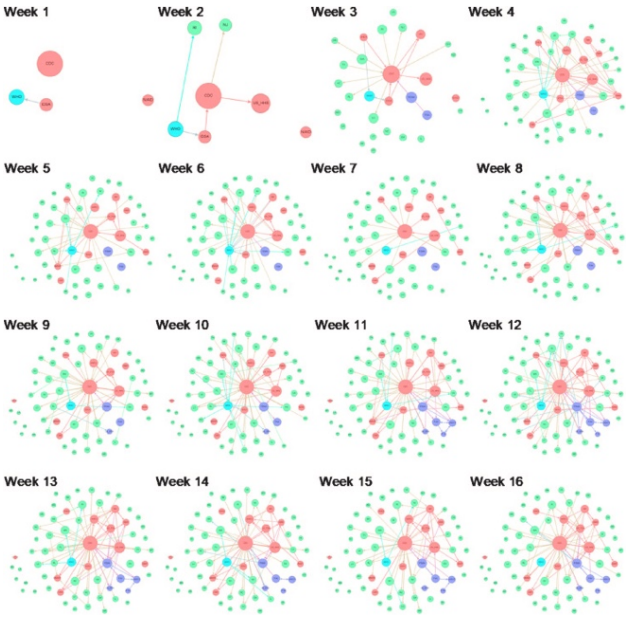

Figure 4: Flow of Information [14]

cases [16]. This study applied an epidemiological method, SIR (Susceptible, Infected, or Removed), to model the diffusion of the three forms of information [16]. Correlations were analyzed between simulation SIR, empirical SIR, and three cascades of tweets (retweet, quote tweet, reply tweet) in terms of both infected and removed nodes [16].

As a result of the COVID-19 pandemic, there has been a diffusion of conspiracy theories and false information throughout social media; also known as an infodemic. Several studies utilizing the COVID-19 Twitter data set have looked into identifying and assessing the misinformation contained in tweets. A study on the "infodemic" twitter network took a clustering approach to look at the network typology, information sources, and messages portrayed [20]. This study conducted a social network analysis (SNA) using NodeXL, clustered the infodemic network using the Clauset-Newman-Moore algorithm, and visualized the network using the Harel-Koren Fast Multiscale layout algorithm [20]. A separate study addressed the specific infodemic topic of the 5G COVID-19 conspiracy theory, a popular theory that has linked 5G to the spread of COVID-19 and resulted in public destruction of 5G towers [17]. This study was able to graph the social network of "5Gcoronavirus" (figure 5) and identify the most influential Twitter users ranked by their betweenness centrality score [17]. Similar to the other study, NodeXL was used to develop a SNA and the clustered network was done using the Clauset-Newman-Moore algorithm; there was additionally use of the Harel-Koren Fast Multiscale layout algorithm in graph development [17].

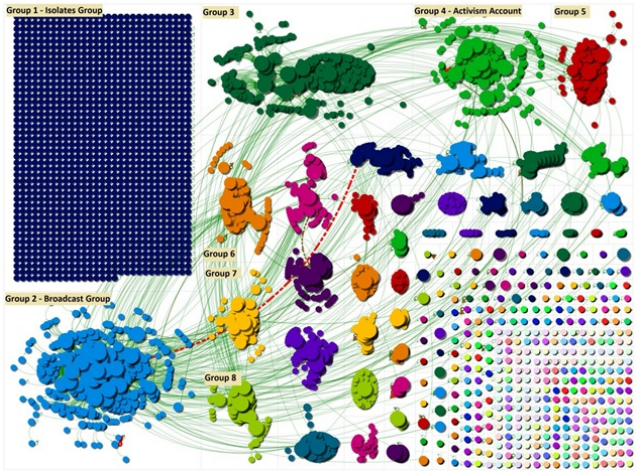

Figure 5: Social network graph of "5Gcoronavirus" [17]

To utilize COVID-19 Twitter data for categorizing and modeling health information, text mining methods for large-scale datasets are necessary. In previous studies on health information extraction from media, the most popular method for building the text matrix include the Vector Space Model (VSM), or the Term Document Matrix (TDM) [21]. TDM is able to extract important terms to form a term-by-sentence matrix which is commonly fed into machine learning algorithms to analyze and cluster the information [23]. A newer method for extraction is graph-based text representation, which uses PageRank, Hyperlink-induced Topic Search HITS (Hub and Authority) and Graph Density (GD) for topic classification [22]. Both methods include appropriate processes for Twitter data classification, with different extraction methods. Further research is necessary to develop the optimal text representation method for classifying COVID-19 Twitter data.

METHODOLOGY

1 Preprocessing

1.1 Dataset

Twitter datasets related to COVID-19 from Panacea Lab [27]on March 22-23,2020 have been used for analysis. 1000 tweets have been chosen randomly to ensure having a statistically representative sample. These tweets are chosen to be in the English language since most of the reliable health organizations and education provide their information in English.

1.2 Data Annotation

The tweets have undergone a rehydration process. For confidential purposes, people provide the Tweets IDs instead of the tweets themselves. To retrieve the text of the Tweets, we must follow a well-known process, which is called the re-dehydration process. After that, we are going to label this dataset into two classes: informational and non-informational. We labeled the tweets manually, each member in the group labeled 250 tweets to information and non-information. There were five main heuristics utilized to classify a tweet as informational: first heuristics the tweet content about symptoms; Second, a tweet has related to causes; third, the tweet referencing medication, vaccines, or remedies; fourth, the tweet about the effects, risks, or consequences of COVID-19; last heuristics general scientific information related to COVID-19. Otherwise, the tweet is considered non-informational. Here is an example of an informational tweet:

"Stay Home and Stay Safe. * Clean your hands often * Avoid close contact * Stay home if you are sick * Avoid crowded places. DEFEAT COVID -19. https://t.co/V8jttpHK3o "

example of a non-informational tweet:

" I just drank a coronavirus without a lime this afternoon and it still tasted better than 99% of the beverages I've ever had #Covid_19

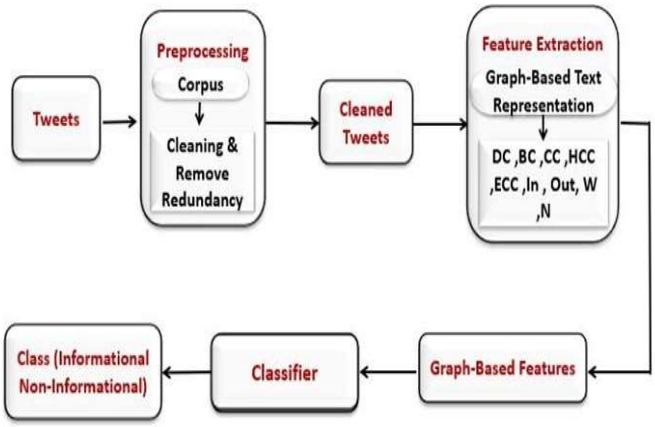

Figure 1: The Preprocessing Flow Chart

1.3 Data Cleaning

The objective of the cleaning step is to remove irrelevant information to facilitate the classification process. In this project, the data have been cleaned by removing URLs and punctuations such as period, comma, question mark, colon, and exclamation mark. Also, removed retweet (RT), that Twitter feature means to help easily share other tweets with your followers. Removed # denotes a hashtag used to index keywords or topics on Twitter. Furthermore, removing the stop-words are a set of commonly used words in any language, some examples of stop words are: "a," "and", "but", "how", "or", and "what.". Use Porter stemmer [26], which means removing the commoner morphological and inflectional endings from words. In addition to that, part of the cleaning process is to convert all letters to lowercase.

2. Features Extraction

Our work utilizes a new graph-based text representation model that can effectively and efficiently detect informational tweets, which can provide accurate knowledge of a person with COVID-19. A graph can be defined as a structure that specifies relationships

between a collection's elements, where a node is the number of words in each tweet, and the relations among nodes it is called edge.

2.1 Matrices

We propose a graph-based tweets analysis approach, which represents the text of the tweet as a weighted directed graph. Our contribution depends on nine metrics, which are the following:

- Degree Centrality (DC)

means calculates the number of the direct edge between nodes by the equation [25] :

$$\text{CD}(v) = \frac{\sum_{j=1}^{n} Xij}{(n-1)(n-2)} \quad (1)$$

- Betweenness Centrality (BC)

is a means to measure the centrality of a node to another node [25], uses the following equation:

$$\text{Betweenness}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma\, st(v)}{\sigma\, st} \quad (2)$$

- Closeness Centrality (CC)

is based on the length of the average shortest path between two nodes [25].

$$\text{Closeness}(v) = \frac{n}{\sum_i d(v,xi)} \forall i \in \{1, \dots, n\} \quad (3)$$

- Harmonic Closeness Centrality (HCC) the sum and reciprocal operations in the closeness centrality:[25]

$$\text{harmonic}(v) = \sum_{v \neq x} \frac{1}{d(v,x)} \quad (4)$$

- Eccentricity Centrality (ECC) [24]

$$\text{ecc}(v) = \frac{1}{\max i\, d(v,xi)} \forall i \in$$

Error! Bookmark not defined. (5)

- In degree (In) is the number of edges coming into a node.
- Out degree (Out) is the number of edges going out to the node.
- the weight of an edge (W) between two nodes can be defined as the co-occurrence of pair of two nodes can be called as v node and w (W)
- The number of words in each tweet (N)

EXPERIMENTS

Data

The experimental results for the proposed approach are conducted with one dataset which is related to the COVID-19 and crawled from a Twitter social networking site. This dataset contains 1000 tweets, and it is unbalanced; in which 27% are informational and the other 73% are non-informational. The ground truth tweets are manually classified.

Experimental Setup

In this project, we focused on the following metrics to evaluate the performance of the proposed method which are: time to build the model, accuracy, precision, recall, and F-measure for each class label, the following are the description of each metrics: Time to build the model is the time elapsed for training the model.

Accuracy is the percentage of correctly predicted instances to the total number of instances in the dataset. It is calculated by the following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + T} \quad (6)$$

Where TP is the true positive, TN is the true negative, FP is the false positive, and FN false negative. Precision is the percentage of correctly predicted positive instances to the total number of predicted positive instances. It is calculated by the following equation:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

The recall is the percentage of correctly predicted positive instances to all instances in the actual class.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

F-Measure is the harmonic mean of precision and recall. It is calculated by the following equation:

$$F - Measure = \frac{TP}{TP + \frac{1}{2}(FP+FN)} \quad (9)$$

In other words, for the informational class, if we have a high recall but low precision that means, we got many informational but most of them are false (non-informational) on the other hand if we had a high precision but low recall that means we got less informational but most of them are True (informational). Now, based on what we mentioned above, our target is to maintain both; many results (precision) and at the same time highly correct results (recall).

Results

As we can see, Figure1 shows that the graph-based text representation methods (GB-CW and GB-PHD) have significantly less time for building the model comparing to TDM based model. The results are obtained by the J48 Decision Tree algorithm with different k-cross values (5 and 10 folds).

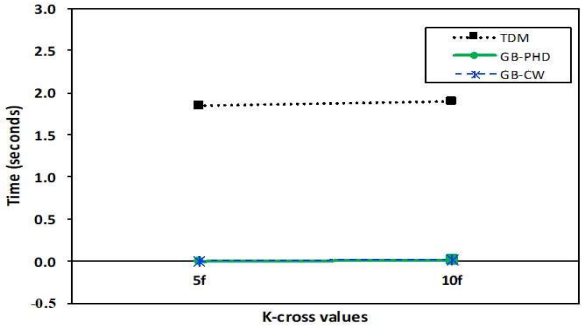

Figure 2: Time to build the model using J48 classifier with different cross-validation

Figure 3 shows a significantly less time between the graph-based text representation models and TDM even with different machine learning algorithms such as J48, Support Vector Machine (SVM), and Naive Bayes with 10 folds testing type. Figure 4 shows that the proposed graph-based text representation model (GB-CW) obtained the highest accuracy value (69%) on 5 folds cross-validation. Figure 5 shows the accuracy values using different classifiers (J48, Support Vector Machine (SVM), and Naive Bayes) using 10 folds testing type. Figure 6 and Figure 7 show that GB-CW obtained the highest precision values (53% and 70%) on 5 folds for informational and non-informational classes, respectively.

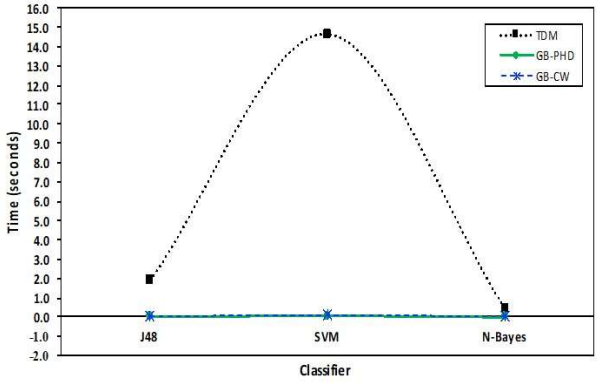

Figure 3: Time to build the model using multi-classifier with 10 folds

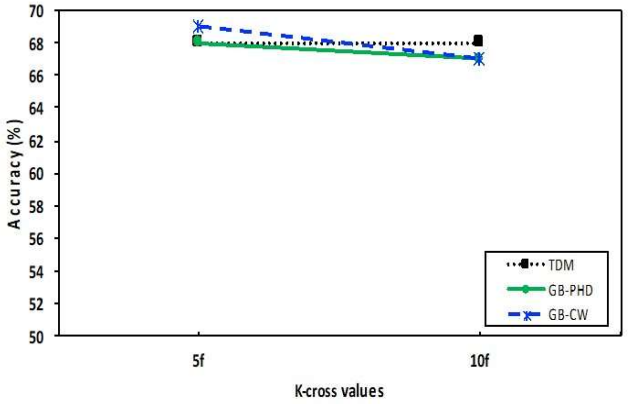

Figure 4: Accuracy using J48 classifier with different cross-validation

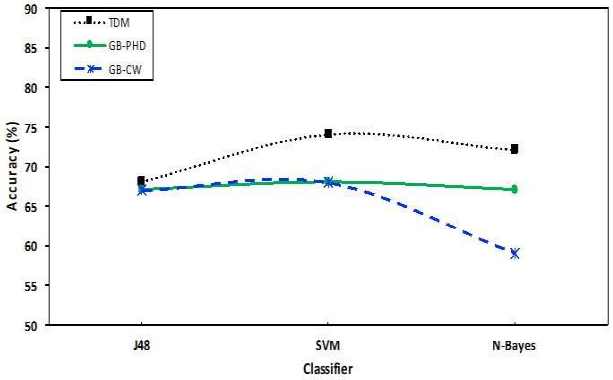

Figure 5: Accuracy using multi-classifiers with 10 folds

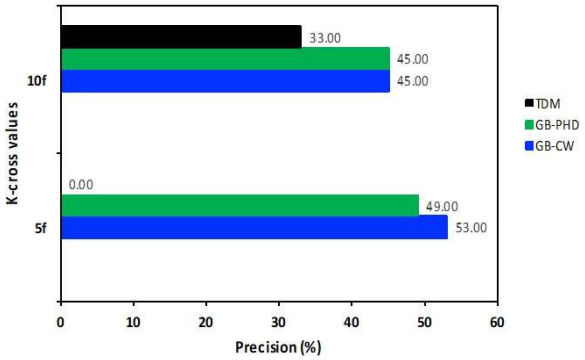

Figure 6: Precision for the informational class using J48 classifier

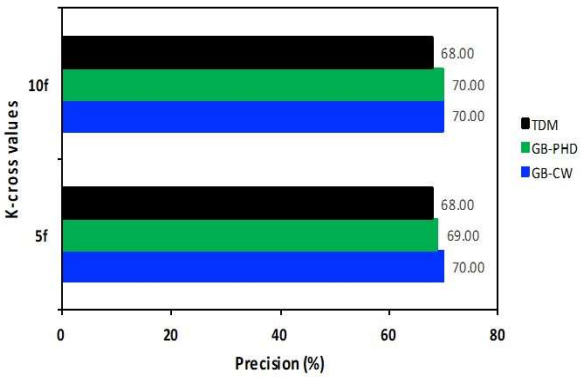

Figure 7: Precision for the non-informational class using J48 classifier

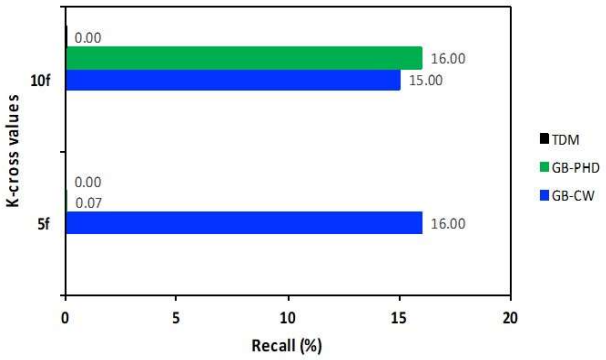

Figure 8: Recall for the informational class using J48 classifier

At the same time Figure 8 and figure 10 show that the graph-based methodologies, in general, have higher recall and F-Measure values for informational class than TDM. On the other hand, Figure 9 and Figure 11 demonstrate that all three methods obtained high recall and F-Measure values for the non-informational class.

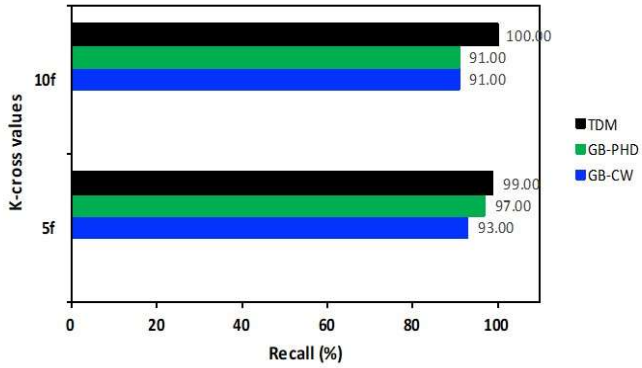

Figure 9: Recall for the non-informational class using J48 classifier

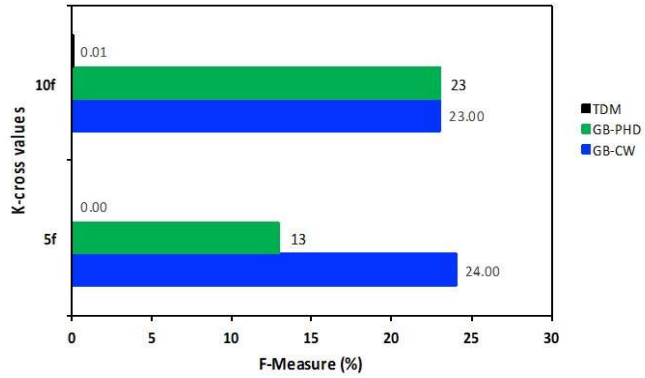

Figure 10: F-Measure for the informational class using J48 classifier

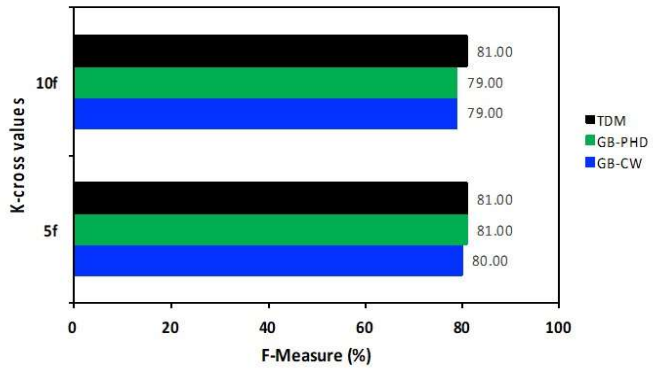

Figure 11: F-Measure for the non-informational class using J48 classifier

CONCLUSION

Social media has given a new outlet to public opinion throughout the past decade, where billions of people can have a free platform to voice what they believe to the billions of others using these platforms. Covid-19 has been a trending topic for the majority of 2020 and looking at social media analytics shows just how prevalent it has become and being able to harness that data is of vital importance to researchers, as well as government and health officials. In this project, we used data sourced from Twitter using the hashtags coronavirus and covid-19. To use the power of machine learning and modeling to produce relevant information that can be used in public response. We believe with the information this project provides we will be able to identify the locality of factual information and misinformation, which can help provide insight on what we should do to help our national pandemic response succeed, trough increases of local funding or government mandates, which could help curb the risk of infection. With coronavirus cases impacting cities the hardest. This study has never been more relevant, utilizing practices in urban computing like this can help localize the virus to specific quadrants of a city, providing insight into where extra attention is necessary. Cases can vary by burrow and neighborhood, with areas of low income and a higher concentration of immigrants or minorities being drastically more affected. Being able to represent with hard data the troubles these areas are going through can help mobilize our government to improve appropriate measures to assist these communities. Using our graph-based modeling theory utilizing a centrality weight concept, we can produce distinct benefits in terms of context retention as well as extensibility into geodata implementation for prediction and trending visualization, allowing these changes to come about. Future implications involve application to epidemiological modeling for the use by public health professionals in outbreak surveillance and response practices.

ACKNOWLEDGMENTS

We would like to acknowledge Panacea Lab at Georgia State University for allowing us access to their Covid-19 dataset by

providing an open-source download at https://zenodo.org/record/4320230. This download provided us with tweets from a Twitter stream on Covid-19 chatter. The data in these tweets include hashtags, mentions, emojis, frequencies, tweet ID, and the respective language used. The Twitter stream, through contributions from multiple collaborators, includes tweets beginning January 1st; with dedicated gatherings from March 11th to present, yielding 4 million+ tweets a day. The tweets are mostly in English, Spanish, and French with the total number of tweets contained in the download being equal to 855,111,891 and after cleaning the data, by removing retweets, we were able to get access to 212,887,091 tweets!

Implementation

You can find the proposed scheme at the following links:

1.Cleaning and Stemming Capsule: https://codeocean.com/capsule/7766507/tree

2.Graph Representation Capsule: https://codeocean.com/capsule/2325826/tree

3. Graph Feature Extraction Capsule: https://codeocean.com/capsule/7522921/tree

4. TDM Capsule: https://codeocean.com/capsule/5162588/tree

REFERENCES

[1] Karisani, P. and Agichtein, E. *Did you really just have a heart attack? Towards robust detection of personal health mentions in social media*. City, 2018.

[2] Sudarsun, S., Prabhu, G. V. and Kumar, V. S. *Role of weighting on TDM in improvising performance of LSA on text data*. IEEE, City, 2006.

[3] Dsouza, K. J., and Ansari, Z. A. *A novel data mining approach for multi variant text classification*. IEEE, City, 2015.

[4] Kumar, M., Yadav, D., and Gupta, V. K. *Frequent term based text document clustering: A new approach*. IEEE, City, 2015.

[5] Plansangket, S. and Gan, J. Q. *A new term weighting scheme based on class specific document frequency for document representation and classification*. IEEE, City, 2015.

[6] Guru, D. and Suhil, M. *Term-class-max-support (TCMS): A simple text document categorization approach using term-class relevance measure*. IEEE, City, 2016.

[7] Jehng, J.-C., Chou, S., Cheng, C.-Y., and Heh, J.-S. *An evaluation of the formal concept analysis-based document vector on document clustering*. IEEE, City, 2011.

[8] Biswas, S. K., Bordoloi, M., and Shreya, J. A graph based keyword extraction model using collective node weight. *Expert Systems with Applications*, 97 (2018), 51-59.

[9] Abilhoa, W. D., and De Castro, L. N. A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240 (2014), 308-325.

[10] Wang, Y., Liu, J., Qu, J., Huang, Y., Chen, J. and Feng, X. *Hashtag graph based topic model for tweet mining*. IEEE, City, 2014.

[11] Bijari, K., Zare, H., Veisi, H. and Kebriaei, E. Deep Sentiment Analysis using a Graph-based Text Representation. *arXiv preprint arXiv:1902.10247* (2019).

[12] Singh, Lisa, et al. "A first look at COVID-19 information and misinformation sharing on Twitter." *arXiv preprint arXiv:2003.13907* (2020).

[13]Lamsal, R. Design, and analysis of a large-scale COVID-19 tweets dataset. *Appl Intell* (2020). https://doi.org/10.1007/s10489-020-02029-z

[14]Wang, Yan et al. "Examining risk and crisis communications of government agencies and stakeholders during early-stages of COVID-19 on Twitter." *Computers in human behavior* vol. 114 (2021): 106568. doi:10.1016/j.chb.2020.106568

[15] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2018. Processing Social Media Messages in Mass Emergency: Survey Summary. In WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3186242

[16] Dinh, Ly, and Nikolaus Parulian. "COVID-19 pandemic and information diffusion analysis on Twitter." *Proceedings of the Association for Information Science and Technology. Association for Information Science and Technology* vol. 57,1 (2020): e252. doi:10.1002/pra2.252

[17] Ahmed, Wasim et al. "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data." *Journal of medical Internet research* vol. 22,5 e19458. 6 May. 2020, doi:10.2196/19458

[18] Rosenberg, Hans et al. "The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic." *CJEM* vol. 22,4 (2020): 418-421. doi:10.1017/cem.2020.361

[19] Xue, Jia et al. "Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach." *Journal of medical Internet research* vol. 22,11 e20550. 25 Nov. 2020, doi:10.2196/20550

[20] Chong, Miyoung. "Network typology, information sources, and messages of the infodemic twitter network under COVID-19." *Proceedings of the Association for Information Science and Technology. Association for Information Science and Technology* vol. 57,1 (2020): e363. doi:10.1002/pra2.363

[21] Antonellis, I. and Efstratios Gallopoulos. "Exploring term-document matrices from matrix models in text mining." *ArXiv* abs/cs/0602076 (2006): n. pag.

[22] Cordobés, Héctor, et al. "Graph-based techniques for topic classification of tweets in Spanish." (2014).

[23] Ozaydin, Bunyamin, et al. "Text-mining analysis of mHealth research." *MHealth* 3 2017.

[24] de Andrade, R. L., and Rêgo, L. C. p-means centrality. *Communications in Nonlinear Science and Numerical Simulation*, 68 (2019), 41-55.

[25] Zhang, J. and Luo, Y. *Degree centrality, betweenness centrality, and closeness centrality in social network*. Atlantis Press, City, 2017.

[26] Willett, P. The Porter stemming algorithm: then and now. *Program*, 40, 3 (2006), 219-223.

[27] *A Twitter Dataset of 150+ million tweets related to COVID-19 for open research*. City.