Analyzing Vaccination Discussions and Communities in Twitter

Yadav Divyarajsinh
dept. Computer science
Lakehead University
Thunder bay, Canada
yadavd167@lakeheadu.ca

Sabah Mohammed
dept. Computer science
Lakehead University
Thunder bay, Canada
mohammed@lakeheadu.ca

***Abstract*—Vaccine misconceptions on the internet may play a role in the rise of anti-vaccine attitude and vaccine apprehension. To discover vaccine and antiviral drugs online communities, social media site information was developed to determine Twitter vaccine personalities, their online twitter neighbourhoods, and their geo-locations. Numerous social discussion boards devoted to vaccinations have formed in recent times, affecting public perception on vaccination. These vaccine-related societies have taken advantage of social network to efficiently promote various ideas. Predictive analytics offers the methods and techniques needed to evaluate large amounts of data and uncover new information. This paper indicates the use of these algorithms to find and track vaccination-related discussion forums on Online Communities. Result suggests that exploring social media influencers appears to be an effective strategy to discover and target anti-vaccine networks online. Detection and surveillance of these social groupings might be used by public health organisations to prevent epidemic.**

***Index Terms*—Data visualization, Twitter, Data mining, Analysis, Gephi, Communities.**

I. Introduction

During the digitization period, data went from being scarce, costly, and difficult to seek and acquire to being abundant, affordable, and exceedingly difficult to analyse and comprehend, culminating in big data[1]. This digital gold is combined with machine learning to create effective analytical techniques that disclose value of data. The act of transforming raw data into graphs, charts, photos, and even films so that humans can understand it is known as data visualisation. Context and Background By transforming data into a more intelligible format and emphasising patterns and exceptions, data visualisation assists in the telling of stories. Raw data is difficult for the human brain to absorb. The impact of displaying numerous graphs is powerful and nicely articulated[1]. We are able to recognise things because of insights. Areas that need to be improved, and it also aids in the detection of problems before they occur. IT is also beset by issues as a result of the massive amount of data. Errors, duplicate entries, and data that has been corrected should all be available for processing. The organisation obtains data without concern for preparation[1][2]. Second, there is a lack of data visualisation and technology understanding in schools, and reliable data is not delivered, regardless of whether the presentation is biased or inaccurate. There are several assertions that are false. Tableau, Google Charts, Info gram, and Data Wrapper are just a handful of the tools that may be used to visualise massive quantities of data. In the near future, this visualisation will be completely automated[2].

Numerous individuals search the internet for vaccination information, and the information they find can influence their vaccination decisions. As a result, data mining and community detection approaches might be used to improve public healthcare policies by enhancing control and prevention actions in identified danger zones. The use of these approaches in this study is centred on detecting and tracking anti-vaccine movements in social network[2]. To that end, a study evaluating the impact of Twitter social influence on vaccination coverage rates is conducted. The examination of the retweet graph, which represents user interactions related to vaccination, is the subject of the second portion of the investigation. The current vaccination communities are first discovered using Community Detection Algorithms on this network[3][4].

The problem's statement People all around the world are utilising publicly available social networking sites like Twitter to swiftly provide information in reaction to current events. Emotions expressed in tweets or social media messages are statistically significant indications of flow and magnitude. Hashtags have grown in popularity in the time between retweets and URLs[5][6]. To begin, in such cases, social media information including negative emotion and anxiety would be positively connected with the flow of data on social media. The number and frequency of retweets reflect how interested people are in such an event . Several social variables are associated with the incident, including the number of retweets and the reach of the retweet. However, there are several drawbacks, such as tweets having a limited amount of characters, writing style, and word abbreviations. Second, tweets are unique and baffling as a result of people exploiting hashtags to obtain attention or views[7].

II. Related Research

Social media users are converging in larger numbers than ever before. They're becoming more well-known as a source of information on a wide range of geographical phenomena[7]. A well-known example is the prediction of the time and location of infectious disease outbreaks based on a study of Twitter tweets whose content reveals the prevalence of a specific disease. Another well-known example is the use of social media for early detection and monitoring of natural catastrophes such as earthquakes and typhoons, which demonstrates how Geo-located Twitter messages allow for the extraction of situational' information that supports in disaster response planning. Twitter has been referred to as a ”distributed sensor system” for tracking global events[7][8].

A. *Studies on the health effects of anti-vaccination*

Geographically, the finding of clusters of social network members has a number of uses. For example, emergency management employees who need to grasp the situation at the local level might use the online communities developed in reaction to a natural catastrophe as a source of information. Emergency management staff, in particular, must have access to precise geographic information and be aware of the requirements of individuals in impacted regions, as well as whether persons at danger have complied with evacuation orders, the extent of local infrastructure damage, and so on. According to studies, people impacted by disasters utilise online social connections and communities to seek and share information. Finding groups on social media can enable emergency managers get essential information at a fine level of detail. Several studies have been carried out to extract geographic information from social media, particularly for the identification of Geo-spatial events[9][10]. Furthermore, research has been undertaken on the recognition of Geo-located communities in social networks. For instance, developed a technique for detecting spatial networks with space-independent communities. Their study aims to create an algorithm that removes the influence of space to identify 'hidden' structural similarities between nodes. Through the monitoring of traffic on a mobile phone network, researchers looked into the geographical structure and dynamics of urban communities[10]. Meanwhile, research on the recognition of Geo-located communities in social networks is still limited, and the goal of our study is more specialised, focusing only on community detection in Twitter's user network[11].

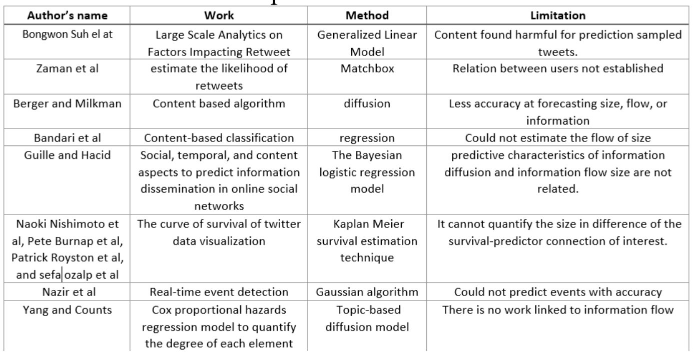
However, there is no consensus on how to define communities in social networks, we define communities in this article as ”groups of vertices that are more densely linked to each other than to the rest of the network,” as defined by Papadopoulos et al. and Murata[12]. Communities on social media can be implicit: they aren't always developed consciously and for a specific purpose or interest. They may be detected, though, by examining the relationships between network members. As a result, these implicit communities differ from 'virtual communities,' which are defined as 'people who share shared interests, objectives, or behaviours interact to exchange information and knowledge and engage in social interactions'[13].some of the shared material and user-created connections are not beneficial for detecting important communities; instead, they make the graph structure difficult to assess and the finding of communities problematic. As a result, it's vital to figure out what kinds of user relationships are effective for detecting groups that will be useful for the desired goal[14][15]. Furthermore, Twitter tweets have a time component as well as geo-location information in the form of coordinates. The case demonstrates that using both explicit relationships between Twitter users and common subjects discussed in tweets resulted in more significant communities than using solely common topics, suggesting that community detection is more effective[16][17].

People all across the world are swiftly submitting information in reaction to real-life occurrences via public and easily available social networking sites such as Twitter[18]. IT has both a good and bad influence; emotions expressed in tweets or postings are statistically significant predictors of flow and size. Press stories linked with the assaults are concerned with information survival. Time span between retweet and URLs, hashtags emerging significantly[19]. There are two categories: size and survival. Size refers to the frequency of retweets surrounding an event, while survival refers to the first and final retweet because retweets indicate public interest[19].

Suh et al sought to understand more about the characteristics that influence whether or not a tweet is retweeted; thus, they created a generalized linear model that considers the URL, hashtags, and age of the tweet[20].Similarly, Zaman et al utilized the Matchbox method to estimate the likelihood of retweets[21], while Tsur and Rappopor used it to forecast the likelihood of material being shared (hashtags)[22]. These were the most accurate in terms of forecasting retweets and information diffusion. However, even when adjusted, utilizing the content of the tweet was found to be harmful to prediction performance[23]. Berger and Milkman investigated how individuals spread good content-based news[24]. Bandari et colleagues employed content-based classification and regression approaches to achieve good accuracy in predicting the range of propagation but were less successful in estimating flow size[25].Guille and Hacid created a model that focuses on social, temporal, and content aspects to predict information dissemination in online social networks[26]. The Bayesian logistic regression model was chosen as the best prediction model. While the model predicted diffusion well, it did not predict size well, showing that the predictive characteristics of information diffusion and information flow size are not related. Zaman utilized the time series model in the same way. Models were good at predicting diffusion but not so good at forecasting size, flow, or information[27]. Backstrom et al. found temporal variables and comment speed, whereas Macskassy and Michelson used time lapsed data to build information propagation behaviour models for Twitter[28]. Both papers show that time is a crucial aspect to consider when modelling propagation. Yang and Counts developed a topic-based diffusion model based on user mentions with the goal of forecasting speed, size, and reach by utilizing a Cox proportional hazards regression model to quantify the degree of each element which is I am more focusing on

[29][30].Apart from Lin et al's work on hashtag development, survival, and context, review of the existing literature at the time of writing found no work in the field linked to information flow content and its relevance to long-term survival, beyond hashtags. The events were discovered by Kaleel et al using timestamps, geolocations, and cluster size. Multimodal et al use the tags in the dataset to detect events[31]. They all did really well. Many scientists are more concerned with the detection algorithm. For identifying accidents, Dabiri et al used deep learning architectures. While Saeed et al developed a unique technique to recognize events from the Twitter stream called Weighted Dynamic Heartbeat Graph[32].For signal identification in tweet opinion and top hashtags, Nazir et al used an average moving threshold method with a Gaussian algorithm[33]. To build a viable real-time event detection method, Sani et al employed locality sensitive hashing to approximation locate related items and incremental clustering[34]. These four publications utilized the Kaplan Meier survival estimation technique to figure out the curve of survival of twitter data visualization: Naoki Nishimoto et al, Pete Burnap et al, Patrick Royston et al, and sefa ozalp et al .Which is impartial because it cannot quantify the size in difference of the survival-predictor connection of interest, which is no hazard ratio or relative risk for many factors concurrently for each subject in the time to event research, nor can it account for confounding factors[35][36][37][38]. Neha Garg el at utilized k means clustering for analysis and visualization; twitter data was collected, pre-processed, and clustered based on geotagged data. Pete Burnap el at used the zero-truncated negative binomial (ZTNB) regression method. To model survival, the Cox regression technique was used because it estimates proportional hazard rates for independent measures[39].

Table 1. Comparison of research work



Author's name	Work	Method	Limitation
Bongwon Suh el at	Large Scale Analytics on Factors Impacting Retweet	Generalized Linear Model	Content found harmful for prediction sampled tweets.
Zaman et al	estimate the likelihood of retweets	Matchbox	Relation between users not established
Berger and Milkman	Content based algorithm	diffusion	Less accuracy at forecasting size, flow, or information
Bandari et al	Content-based classification	regression	Could not estimate the flow of size
Guille and Hacid	Social, temporal, and content aspects to predict information dissemination in online social networks	The Bayesian logistic regression model	predictive characteristics of information diffusion and information flow size are not related.
Naoki Nishimoto et al, Pete Burnap et al, Patrick Royston et al, and sefa ozalp et al	The curve of survival of twitter data visualization	Kaplan Meier survival estimation technique	It cannot quantify the size in difference of the survival-predictor connection of interest.
Nazir et al	Real-time event detection	Gaussian algorithm	Could not predict events with accuracy
Yang and Counts	Cox proportional hazards regression model to quantify the degree of each element	Topic-based diffusion model	There is no work linked to information flow

Another example of the potential repercussions of vaccine scepticism on public health care is influenza vaccination. In June 2009, the World Health Organization proclaimed an influenza pandemic. The influenza virus was being followed across the world for changes in virulence or epidemiology so that vaccines might be developed, but vaccine supplies were running low in certain places[39]. When an epidemic occurs, the public needs to know that adequate vaccine will be available; nevertheless, some are questioning the vaccine's safety and effectiveness. For starters, social media information including negative emotion and anxiety would be positively connected with the flow of data on social media in such scenarios. Second, in compared to news stories, the assault volume and duration are statistically favourable. The number and frequency of retweets demonstrate how interested people are in a certain event. The incidence is connected to various social aspects, such as the number of retweets and the reach of the retweet[40]. There are other drawbacks, such as the limited number of characters, writing style, and word abbreviations in tweets. Second, because individuals abuse hashtags to seek attention or views, tweets are unique and baffling. It is critical because story-telling is simple to comprehend, and developing a predictive model for community identification using geo-location will be a worthwhile future research project. Table 1 shows the comparison between different papers and their methods.

B. Algorithms for detecting communities

The Clustering Algorithm Difficulty in Complex Networks has been the topic of several studies in the disciplines of data mining and social network analysis. In the literature, there are a variety of methods for selecting the appropriate node groups for societies. The goal of the Group Detection Method is similar to that of supervised learning in graph theory. Clustering is a phrase used in information science to describe the uncontrolled process of determining the underlying structure of data by grouping the most similar portions together. The constituents in the same cluster should appear comparable, while the ones in other clusters should appear unique. To determine visual similarity, some form of assessment will be used. A population can be easily mapped from a graph cluster. The two most important measures in information flow are size, which is the quantity of tweeting and also represents public interest in assaults, and survival, which is the persistence of public interest over time, or in other words, how long this event will stay on Twitter. Content, social, and temporal are the three dimensions that are discussed. The content feature is the frequency of all vaccine-related tweets, as well as the attitudes, emotions, and tension underpinning such tweets. Certain components, such as newspapers, journals, and news media, aren't included in narrative.

One of the most extensively utilised Community Detection approaches was provided by Girvan and Newman. An Edge Betweenness similarity measure is used in this method, which counts the number of shortest paths connecting all vertex pairs. This strategy, however, has a considerable computational cost. As a consequence, Newman rebuilt the modularity measure in terms of eigenvectors. The network's modularity matrix is a new characteristic matrix. The primary downside of this strategy on very large networks is its huge computational complexity. Following that, the modularity metric was adjusted in a number of ways in an attempt to drastically reduce computation needs[41].Investigators gathered a dataset of vaccine-related tweets in order to undertake a social effect study. In addition, the World Health Organization's immunisation monitoring system's vaccination coverage numbers were gathered. Each year and, this official report provides the official coverage estimate for each country. In order to measure the societal influence on vaccination rates, two variables were generated for each nation utilising both datasets. Data extraction, data pre-processing to identify tweet locations, social data analysis, and geo-spatial information visualisation were all done for this reason.

III. Research Questions

I've decided to research the function of social media and how it reacts to vaccinations, as well as visualise such events and how long they last, as well as the magnitude of the problem. Furthermore, information is disseminated via social media. People all across the world are using public and widely accessible social networking sites like Twitter to quickly contribute information in response to real-life events.

I've come up with the following research questions based on my research:
(1) What criteria are used to anticipate the flow and data associated with large-scale information regarding vaccines?
(2) What are people's attitudes on vaccines, and who has impacted them?
(3) Will societal preferences change as a result of the vaccination analysis? Can the government assist by developing new strategies?

For starters, social media information including negative emotion and anxiety would be positively connected with the flow of data on social media in such scenarios. Second, in compared to news stories, the assault volume and duration are statistically favourable. The number and frequency of retweets demonstrate how interested people are in a certain event. The incidence is tied to various social variables, including the number of retweets and the reach of the retweet.

IV. Data extraction

All of the information obtained for this project came from a single source: Twitter.

• Twitter: Twitter is a social media platform where users may post personal information in the form of tweets. Tweets are 140-character messages that provide information such as thoughts, photographs, and links. The re-tweet is a unique type of tweet that occurs when one user reposts another user's tweet. Every day, Twitter users send about 400 million tweets, which are accessible via public APIs that allow users to search for tweets using keywords, hashtags, phrases, geographic locations, or user names. All tweets containing the phrase 'vaccines' were used to compile the data for this study[17][42][45].

V. Methodology

To calculate how long the event will last, data will be taken through the Twitter streaming API or a dataset based on manual study. For my analysis, I'll be sampling for that amount of time. We'll figure out how many tweets have been pre-processed and recoded before being modelled within that period. The frequency of retweets determines the scale of the event, while survivability determines the event's longevity[44]. The pre-processing will use the text pattern matching approach to figure out how many retweets there are from the original tweet. After that, tweets having fewer than a particular number of retweets, such as five, will be deleted. After preprocessing, we'll have a sample size. The reaching matrices, which represent the reach of the knowledge transfer after the fifth tweet, are created by extracting and adding the number of followers of each subsequent people first from data. By multiplying the total number of seconds between the first tweet and the first five retweets by the number of seconds between the first tweet and the first five retweets, the time lag variable is produced. The time intervals would be five and twenty seconds, accordingly, assuming the first retweet happened after five seconds and the second after fifteen seconds. Factor analysis with the principal component analysis technique may be used to assess the size and longevity of predictor variable. This approach will generate sub models based on the largest variation of the dependent variable. This is advantageous because it makes it easier to identify the

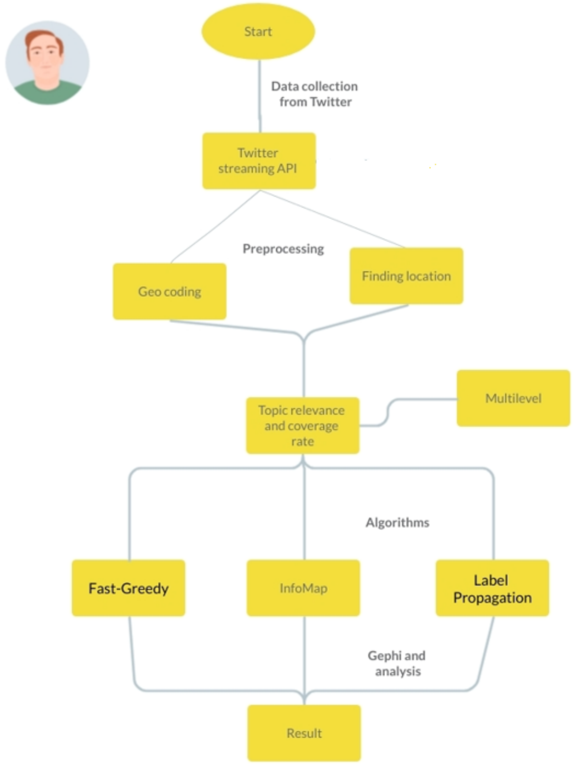


Fig. 1. Methodology

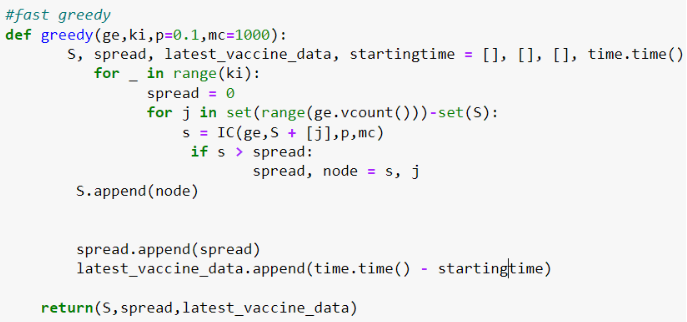
most critical components in the construction of a broad and long-lasting information flow. Figure 1 shows the flow of methodology. During prepossessing, the twitter streaming API obtains the location and Geo co-ordinates of tweets as well as users. Using the subject relevancy rating, we may get rid of all the extraneous information. Following that, four community discovery techniques are utilised, as shown in Figure 1: fast greedy, info map, label propagation, and multilevel. The result is presented on a world map that shows the number of vaccines received by each country as well as their connection.

VI. Finding Social communities

In computer science, community discovery algorithms have been intensively explored, particularly for social media mining. Individuals frequently establish groups based on shared interests, and recognising groups of similar users can give a comprehensive perspective of user interactions and behaviour. Furthermore, certain behaviours can only be seen in a group, not on an individual basis. This is because individual behaviour is susceptible to change, but group behaviour is more resistant. There are two types of community identification algorithms: member-based algorithms, which locate groups based on the characteristics of its members, and group-based algorithms, which construct groups based on the density of interactions among their members. In this paper, a comparison of group-based algorithms is carried out in order to determine the best suited way for better recognising communities discussing a specific issue of vaccination. The development of models was chosen for this purpose: 1. Fast-Greedy: This technique combines individual nodes into communities in such a way that the graph's modularity is greedily maximized[46]. 2. InfoMap: It simulates information flows in a real system by using the probability flow of random walks across a network. The network is then divided into modules by compressing a probability flow description. The result is a map that highlights and simplifies the structure's regularities and their relationships[47]. 3. Label Propagation: Each vertex is first allocated a separate label. Then, for each cycle, each vertex picks the dominant label in its vicinity. Every iteration, ties are broken at random, and the sequence in which the vertices are updated is randomised. When the vertices establish an agreement, the algorithm is finished[48]. 4. Multi-Level: This is a bottom-up technique in which each vertex is initially assigned to a different community, and vertices are iteratively shifted between communities to optimise the vertices local contribution to overall modularity[49]. When increasing this modularity is no longer viable, the algorithm comes to a halt.

VII. Results and analysis

To continue with the results analysis, a geo-spatial visualisation was created to enable for a visual study of social data across all nations. A map is created that summarises the geographic information of users discussing vaccinations, allowing for the identification of locations of interest. As a fresh way to community discovery, I suggest many algorithms. This is due to two factors. For starters, triangles are useful in community development because they allow you to encapsulate the entire community structure in a graph, which is quite useful in social networks. Iterate among the most appealing communities until you have a community-like structure. The procedure is known as a grouping algorithm because it works from the bottom up. The technique does not use diversity as a statistic because the number of communities in the real-world changes[45].

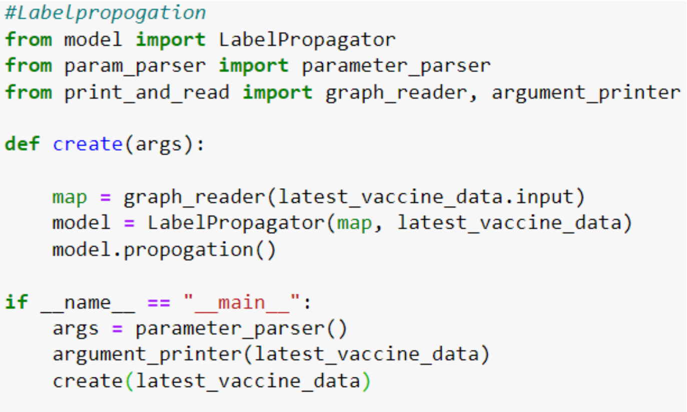


```
#fast greedy
def greedy(ge,ki,p=0.1,mc=1000):
       S, spread, latest_vaccine_data, startingtime = [], [], [], time.time()
          for _ in range(ki):
                spread = 0
                for j in set(range(ge.vcount()))-set(S):
                    s = IC(ge,S + [j],p,mc)
                     if s > spread:
                            spread, node = s, j
        S.append(node)


        spread.append(spread)
        latest_vaccine_data.append(time.time() - startingtime)

    return(S,spread,latest_vaccine_data)
```

Fig. 2. Fast greedy



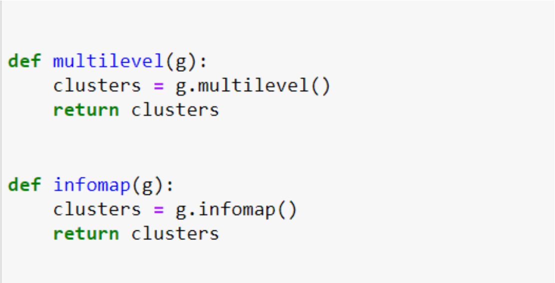
```
#Labelpropogation
from model import LabelPropagator
from param_parser import parameter_parser
from print_and_read import graph_reader, argument_printer

def create(args):

    map = graph_reader(latest_vaccine_data.input)
    model = LabelPropagator(map, latest_vaccine_data)
    model.propogation()

if __name__ == "__main__":
    args = parameter_parser()
    argument_printer(latest_vaccine_data)
    create(latest_vaccine_data)
```

Fig. 3. Label propagation



```
def multilevel(g):
    clusters = g.multilevel()
    return clusters


def infomap(g):
    clusters = g.infomap()
    return clusters
```

Fig. 4. Multi level and Info map

Node-link, Communities, Node-link Properties, and Community Detection are among the synchronised panels included. Set the optimization algorithms, filter the raw information, set the layout algorithm parameters, survey the connected graph characteristics, and access some facts and figures on a specific community using data filters, layout parameters, graph characteristics, and statistics views to visualise the network,

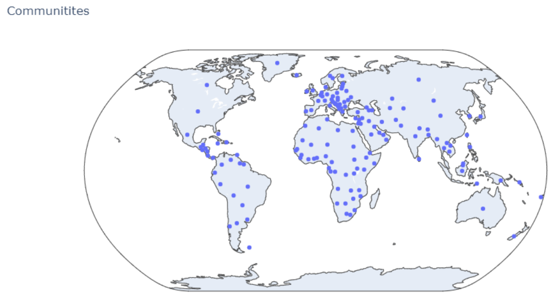


Fig. 5. Info map

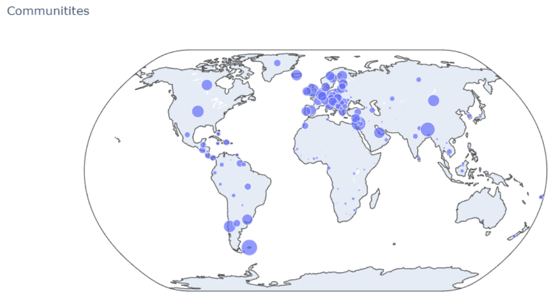


Fig. 6. Fast greedy

envision the detected neighbourhoods, visually encode the data attributes, set the optimization algorithms, filter the raw information, set the layout algorithm parameters, survey the connected graph characteristics, and access some facts and figures on a specific community using data filters, layout parameters, graph characteristics, and statistics views.

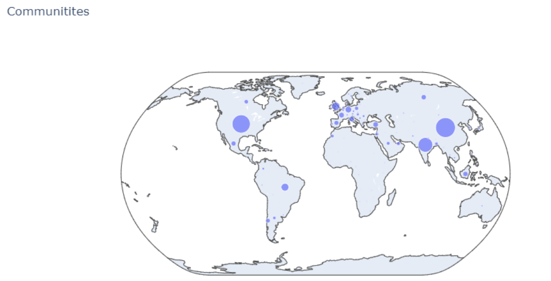


Fig. 7. Label propagation

We can get a better notion of the major themes in the discussion going on in our data set by looking at how hashtags are used in conjunction with anyone. The hashtag co-occurrence data will be used to create the graphs in Gephi. When two hashtags are used in the similar tweet, they form a connection. As they encounter one other more regularly, the attachment keeps growing. When we compare the co-hashtag graphs with the hashtag's recurrence charts, we can see who was the dominant figure in the argument with and who it was conversing. Using several methods for community recognition, as shown in the figures, we are able to identify various groups throughout the world that are discussing vaccination. Several statistical computer programmes now provide zero-inflated Poisson and Poisson regression models for data analysis. These techniques are meant to be used in situations when there are only a few people.

If, for example, the regression model is ”number of times a bot or tweet includes a blank sentence,” the vast majority of tweets may have a value of 0. For discrete random variables, the negative binomial distribution is a probabilistic distribution. This form of distribution is concerned with the number of efforts required to get a given number of successes[49]. Third-party apps and developers may access the vast amount of data generated by Twitter users on a daily basis. The Twitter REST API, which provides a range of endpoints for accessing data, is used to do this. When building the web tool, we were particularly interested in the endpoint that allows us to retrieve tweets from a certain person by utilising the Twitter Search API's search functionality. This is used to identify tweets that include both the term and the hashtag in a search query.

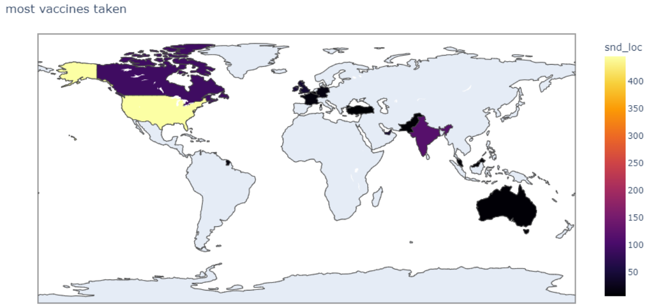


Fig. 8. Vaccine distribution worldwide

Visualizing user activity at various times of the day and determining whether or not this has an effect on the topics he tweets about is a fascinating method. Without a doubt, this way of assessing user activity is underused. For starters, viewers will be able to navigate their way around it swiftly. Other depictions reflect a user's popularity depending on how many retweets and likes his tweets get. A graph that depicts a person's influence in the Twitter network.

VIII. Conclusion

This paper demonstrates how to use Information Retrieval Techniques to identify and analyse Twitter networks that disseminate vaccination ideas. The vaccination coverage percentages, as well as a dataset gathered from Twitter. The findings of this early investigation reveal that vaccine opinions expressed on Twitter may have an impact on vaccination judgement in some situations. However, it is worth noting that the majority of Twitter communities discussing vaccination are not anti-vaccine. In reality, the majority of newly formed movements now favour vaccination and are working to boost incidence rate. Considering all of the reported test findings, it can be stated that the data mining techniques used are appropriate for this type of investigation. The suggested approach may be used to locate and monitor vaccination activities, as well as to uncover new information in data that can be utilised to promote better health immunisation initiatives. Furthermore, this newly gained knowledge might be utilised to identify and find groups resistant to vaccination, which could contribute to future deadly diseases in other countries of the globe. In the future, work discussions may be studied to see which countries are dealing with which crises, and solutions can be discussed to see if there is a link between them.

IX. Acknowledgements

This work has been supported by Dr. Sabah Mohammed from department of Computer Science, Lakehead University under the course Comp-5800 research project. The data and code of this project is available in GitHub repository https://github.com/Dyadav97/Analyzing-Vaccination-Discussions-and-Communities-in-Twitter.

References

[1] Verisk Maplecroft ,21 NOV '2018, Available: https://www.maplecroft.com/insights/analysis/84- of-worldsfastest-growing-cities-face-extreme-climate-change-risks/

[2] Pete Burnap • Matthew L. Williams," Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack",16 Feb, 2014.

[3] Phillip Anderson, Joram Binsbergen," The Woolwich Attack: Analyzing Dynamics of Polarization and Reconciliation on Twitter and on the Ground ".

[4] Rupinder Paul Khandpur, Taoran Ji, Steve JanCrowdsourcing Cyber-security: Cyber Attack Detection using Social Media, AVAILABLE: https://par.nsf.gov/servlets/purl/10058763

[5] Dario Stojanovski, Ivica Dimitrovski, Gjorgji Madjarov," TWEETVIZ: TWITTER DATA VISUALIZATION".

[6] Fang, Y.; Gao, J.; Liu, Z.; Huang, C. Detecting Cyber Threat Event from Twitter Using IDCNN and BiLSTM. Appl. Sci. 2020, 10, 5922. https://doi.org/10.3390/app10175922

[7] Shaham," Analyzing Information Flow within a Twitter (Ego-)Community",Jan 16,2019. Available:https://towardsdatascience.com/information-flow-within-twitter-community-def9e939bb99

[8] C. for Disease Control P. (CDC), et al., Impact of vaccines universally recommended for children–United States, 1990-1998, MMWR Morb. Mortal. Wkly. Rep. 48 (12) (1999) 243.

[9] V.A. Jansen, N. Stollenwerk, H.J. Jensen, M. Ramsay, W. Edmunds, C. Rhodes, Measles outbreaks in a population with declining vaccine uptake, Science 301 (5634) (2003) 804–804.

[10] D.J. Opel, S.B. Omer, Measles, mandates, and making vaccination the default option, JAMA Pediatr.

[11] K.S. Wagner, J.M. White, I. Lucenko, D. Mercer, N.S. Crowcroft, S. Neal, A. Efstratiou, D.S. Network, et al., Diphtheria in the postepidemic period, Europe, 2000–2009, Emerg. Infect. Dis. 18 (2) (2012) 217.

[12] A. Kata, A postmodern pandora's box: Anti-vaccination misinformation on the Internet, Vaccine 28 (7) (2010) 1709–1716.

[13] J. Keelan, V. Pavri-Garcia, G. Tomlinson, K. Wilson, Youtube as a source of information on immunization: a content analysis, Jama 298 (21) (2007) 2481–2484.

[14] J. Keelan, V. Pavri, R. Balakrishnan, K. Wilson, An analysis of the human papilloma virus vaccine debate on myspace blogs, Vaccine 28 (6) (2010) 1535–1540.

[15] N. Seeman, A. Ing, C. Rizo, Assessing and responding in real time to online antivaccine sentiment during a flu pandemic, Healthc Q 13 (Sp) (2010) 8–15.

[16] N. Sunday, The online health care revolution: How the web helps Americans take better care of themselves, Pew Internet Amer. Life Proj.

[17] Twitter web site, 2013.

[18] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Inf. Fusion 28 (2016) 45–59.

[19] W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 1029–1038.

[20] Suh B, Hong L, Pirolli P, Chi E (2011) Want to be retweeted? Largescale analytics on factors impacting retweet in Twitter network.In: SocialCom

[21] Zaman T, Herbrich R, Van Gael J, Stern D (2010) Predictinginformation spreading in Twitter. In: Workshop on computa-tional social science and the wisdom of crowds (NIPS). http://arxiv.org/abs/1304.6777

[22] Zaman T, Fox E, Bradlow E (2013) A Bayesian approach forpredicting the popularity of tweets. CoRRZarrella D (2009) The science of retweets

[23] Tsur O, Rappoport A (2012) What's in a hashtag? content basedprediction of the spread of ideas in microblogging communities. Paper presented at the proceedings of the fifth ACM international conference on web search and data mining, Seattle, Washington, USA

[24] Berger J, Milkman K (2012) What makes online content viral? J MarkRes 49(2):192–205

[25] Bandari R, Asur S, Huberman BA (2012) The pulse of news in social media: forecasting popularity. CoRR. http://arxiv.org/abs/1202.0332

[26] Guille A, Hacid H, Favre C, Zighed DA (2013) Information diffusion in online social networks: a survey. SIGMOD Rec 42(1):17–28.doi:10.1145/2503792.2503797

[27] Backstrom L, Kleinberg J, Lee L, Danescu-Niculescu-Mizil C (2013) Characterizing and curating conversation threads: expansion, focus, volume, re-entry. Paper presented at the proceedings of the sixth ACM international conference on web search and datamining, Rome, Italy

[28] Macskassy S, Michelson M (2011) Why do people retweet?antihomophily wins the day. In: International conference onweblogs and social media (ICWSM)

[29] Yang J, Counts S (2010) Predicting the speed, scale, and rangeofinformation diffusion in Twitter. In: International conference onweblogs and social media (ICWSM)

[30] Lin Y, Margolin D, Keegan B, Baronchelli A, Lazer D (2013) Bigbirds never die: understanding social dynamics of emergent hashtags. In: Proceedings of the seventh international AAAIconference on weblogs and social media, Boston, MA

[31] Sefa Ozalp," Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech".

[32] Brunetti, Josep Auer, Sören García, Roberto. (2012). The Linked Data Visualization Model.

[33] Stojanovski, Dario Dimitrovski, Ivica Madjarov, Gjorgji. (2014). TweetViz: Twitter Data Visualization.

[34] Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011)The revolutions were tweeted: information flows during the 2011Tunisian and Egyptian revolutions. Int J Commun 5(SpecialIssue):1375–1405

[35] Procter R, Crump J, Karstedt S, Voss A, Cantijoch M (2013a)Reading the riots: what were the police doing on Twitter? PolicSoc 23(4):1–24. doi:10.1080/10439463.2013.780223

[36] Procter R, Vis F, Voss A (2013b) Reading the riots on Twitter:methodological innovation for the analysis of big data. Int J SocRes Methodol 16(3):197–214. doi:10.1080/13645579.2013.774172

[37] Tabachnick B, Fidell L (2013) Using multivariate statistics, 6th edn.Allyn and Bacon, BostonThelwall M, Buckley K, Paltogou G, Cai D, Kappas A (2010)Sentiment strength detection in short informal text. J Am SocInform Sci Technol 61(12):25442558

[38] Rupinder Paul Khandpur," Crowdsourcing Cybersecurity: Cyber Atack Detection using Social Media", https://par.nsf.gov/servlets/purl/10058763

[39] Fang, Y.; Gao, J.; Liu, Z.; Huang, C. Detecting Cyber Threat Event from Twitter Using IDCNN and BiLSTM. Appl. Sci. 2020, 10, 5922. https://doi.org/10.3390/app10175922

[40] Williams ML, Edwards A, Housley W, Burnap P, Rana O, Avis N, Morgan J, Sloan L (2013) Policing cyberneighbourhoods:tension monitoring and social media networks. Polic Soc23(4):1–21. doi:10.1080/10439463.2013.780225

[41] Guille A, Hacid H (2012) A predictive model for the temporal dynamics of information diffusion in online social networks. Paper presented at the 21st international conference companion on World Wide Web, Lyon, France

[42] Downs A (1972) Up and down with ecology—the 'issue-attention cycle'. Public Interest 28:28–50

[43] Cox D (1972) Regression models and life tables. J Roy Statist Soc B34:187–220

[44] Burnap P, Rana O, Avis N, Williams M, Housley W, Edwards A, Morgan J, S L (2013) Detecting tension in online communities with computational Twitter analysis. Technol Forecast Social Change. doi: 10.1016/j.techfore.2013.04.013

[45] Gema Bello-Orgaz, Julio Hernandez-Castro, David Camacho, Detecting discussion communities on vaccination in twitter, Future Generation Computer Systems, Volume 66, 2017, Pages 125-136, ISSN 0167-739X, https://doi.org/10.1016/j.future.2016.06.032.

[46] A. Clauset, M.E. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (6) (2004) 066111.

[47] M. Rosvall, D. Axelsson, C.T. Bergstrom, The map equation, Eur. Phys. J. Spec. Top. 178 (1) (2009) 13–23.

[48] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 036106.

[49] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (10) (2008) P10008.