International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 11, Issue 3, March 2020, pp. 128-134, Article ID: IJARET_11_03_011
Available online at http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=3
Journal Impact Factor (2020): 10.9475 (Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6480 and ISSN Online: 0976-6499

INFORMATIONAL FLOW ON TWITTER - CORONA VIRUS OUTBREAK – TOPIC MODELLING APPROACH

Dr. Rajesh Prabhakar Kaila*
Visiting Professor (Big Data Analytics), NMIMS Hyderabad,
NMIMS University, India

Dr. A. V. Krishna Prasad
Associate Professor, MVSR Engineering College, Hyderabad, India
*Corresponding Author Email id: rajesh.prabhakar@gmail.com

ABSTRACT

The study focuses on the information flow on twitter during the corona virus outbreak. Tweets related to #coronavirus are studied using sentiment analysis and topic modelling using Latent Dirichlet Allocation post preprocessing. The study concluded that the information flow was accurate and reliable related to corona virus outbreak with minimum misinformation. LDA analysis had identified the most relevant and accurate topics related to corona virus outbreak and sentiment analysis confirmed the prevalence of negative sentiments like fear along with positive sentiments like trust. Governments and Healthcare authorities & institutions effectively utilized to spread accurate and reliable information on twitter.

Keywords: Data Analytics, Text Mining, Topic Modelling, Latent Dirichlet Allocation, Sentiment Analysis, Unsupervised Machine Learning

Cite this Article: Dr. Rajesh Prabhakar Kaila and Dr. A. V. Krishna Prasad, Informational Flow on Twitter - Corona Virus Outbreak – Topic Modelling Approach, *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11 (3), 2020, pp 128-134.
http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=3

1. INTRODUCTION

Social media companies like Twitter, Facebook, YouTube, TikTok etc. are highly cautious about conspiracy theories and misinformation being spread about corona virus or covid 19. Despite the best efforts of these social media firms' lot of disinformation is spreading since the novel strain of coronavirus outbreak started in Wuhan, China. Further panic created as the virus rapidly spread across the world to different countries that are close to China and also far away across the world. To tackle the disinformation that is being spread on various social

media platforms which World Health Organization is calling as "infodemic", WHO is conducting day to day media conferences to allay the fears and also instruct the governments across the world to implement specific measures in this regard.

People are searching online for information about corona virus outbreak and they are particularly relying on the above mentioned social media platforms and online news sites, there has been wide spread disinformation also spreading for instance that face masks will prevent the spread which is factually incorrect and prices of face mask zoomed to extraordinary levels. Social Media platforms and news sites have clearly specified that they've been working to promote factual content, and some are limiting the reach of posts with misinformation on their platforms. Twitter, for instance, has put a warning label linking to the Centers for Disease Control and Prevention (CDC) when users search "coronavirus."

Private encrypted messaging channels like WhatsApp too are also actively involved in spreading misinformation about corona virus and panic is also being created among the users and some politicians are also involved in spreading conspiracy theories related to coronavirus too on the various social media platforms. Social media platforms like Twitter had been actively spreading reliable and trustworthy information related to corona virus outbreak and many online users have been actively using this information for their personal benefit. This study exclusively focuses on the information flow on Twitter related to corona virus outbreak where in a random sample of 18000 tweets are extracted from the twitter handle #coronavirus.

2. LITERATURE REVIEW

Prior studies related to information flow on Twitter related to pandemic was done during the 2019 Ebola outbreak by Liang et al. (2019) modelled trails of Ebola related messages and concluded that one to many diffusion or broadcasting led the discussion on twitter. They also identified the role played by 4 distinct users of social networks i.e. influential user, hidden influential user, disseminator and common user based on retweets & followers and concluded that both influential and hidden influential users retweet the information most. They further suggested that the healthcare authorities and governments can cautiously use the four types of users for effective dissemination of information related to pandemic.

Fung et al. (2014) randomly sample tweets related to Ebola outbreak and concluded that most of the tweets originated in USA, whereas disease outbreak was in Guinea, Liberia and Sierra Leone due to the fact that internet access was minimal in these countries. Negative emotions such as anxiety, anger, swearing and death dominated the tweets which highlights the high levels of anxiety related to Ebola. They further concluded that twitter can be effectively be used by the authorities and health practitioners to provide relevant accurate information related to disease and can reduce panic and anxiety in unrelated areas where there is no outbreak.

Towers et al. (2015) studied the relationship between the news media information and internet searches along with tweets and concluded that as the news spread the number of internet searches and tweets also increased in millions indicating panic among the general public. Authors refer to concept of digital epidemiology where in the real time prediction of outbreaks can be identified based on the internet search and tweet trends. A positive correlation resulted between digital data and temporal evolution of epidemics or outbreaks.

Zika outbreak in Latin America related tweets were studied by Kung-Wa Fu et al. (2016) where in the researchers identified tweets with multiple themes like the societal impact of the outbreak, impact on pregnant women and unborn child, response of government and health authorities related to the outbreak and how the virus spread across geographical regions. The study concluded that twitter users relied upon user generated content rather than the government and other related institutional content.

Chew and Eysenbach (2010) studied 2 million tweets related to H1N1 or swine flu outbreak and conducted study to understand the "infoveillance" which they suggested to monitor the flow of information related to outbreak on twitter. They suggested using social media for potential "infodemiology" studies and concluded the tweets were used to spread information related to outbreak from credible sources and also users voiced their opinions and concerns widely. Further suggested health authorities can primarily use twitter for spreading knowledge and quell concerns.

Oyeyemi et al. (2014) studied the misinformation that spread during the Ebola outbreak that created panic and anxiety and concluded most of the tweets were misinformation and had a wide reach compared to correct information. During the outbreak people need correct and reliable information from any source particularly internet-based sources were easily available and governments and health authorities need to have policies and frame works that stops the spread of misinformation and spread correct information.

Jeanine et al. (2017) studied Ebola outbreak related posts on twitter and Instagram by 3 major organizations like CDC USA, WHO and MSF. Study concluded Instagram to be more effective in meaningful communication with users. Health authorities can effectively utilize the social media platforms for spreading information and control misinformation. A strategic health communication framework is needed during the outbreak scenarios.

3. RESEARCH METHODOLOGY

The study focuses on the tweets during the corona virus outbreak and a random sample of 18000 tweets are downloaded without any retweets using the package "rtweet" in R. The sample of 18000 tweets is chosen as it is the maximum number of tweets that can be downloaded using the package and twitter API. The tweets are preprocessed using package "tm" in R and converted into a corpus of text. The preprocessing is verified using a word cloud and there are no special characters or any stop words.

Sentiment analysis is conducted on the cleaned tweets sentences before converting tweets into a text corpus using package "syuzhet" in R. For purpose of sentiment analysis NRC sentiment dictionary is used to calculate the presence of eight different emotions and their corresponding valence. A data frame where each row represents a sentence from the original file. The columns include one for each emotion type as well as a positive or negative valence. The ten columns are as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative", "positive."

A Term Document Matrix is created from the text corpus which is a two-dimensional sparse matrix whose rows are the terms and columns are the documents, so each entry (i, j) represents the frequency of term i in document j. Post creation of term document matrix, frequent terms are identified to ratify whether the terms in tweets correspond to corona virus outbreak or not. Correlation analysis conducted using the term document matrix where in each vector holds matching terms from x and their rounded correlations satisfying the inclusive lower correlation limit.

3.1. Topic Model - Latent Dirichlet Allocation

In this study Topic modelling technique Latent Dirichlet Allocation (LDA) is implemented on the Document term matrix created from text corpus. Topic modelling identifies topics from a set of documents which are a set of words based on the following assumption that each document can be describes by a distribution of topics and each topic can be described by a distribution of words. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

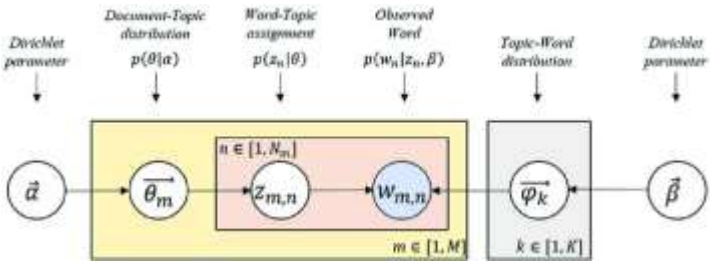

Figure 1 Graphical Image of LDA - Lee et al. (2018)

Latent Dirichlet Allocation begins with identifying the words in each document, creates a topic mixture for the document related to a fixed set of topics chosen and topic selection is purely based on document's multinomial distribution followed by picking of words based on multinomial distribution. Basically, LDA is an unsupervised algorithm used to spot the semantic relationship between words a group with the help of associated indicators. Researchers had used Topic Modelling particularly Latent Dirichlet Allocation for identifying topics in tweets, which are Prabhakar, K.R (2019) implemented LDA on climate change tweets, Prabhakar, K.R and Satish. D (2016) implemented LDA on tweets related Fort McMurray Wildfire disaster.

4. RESULTS AND FINDINGS

Sentiment Analysis done on the pre-processed tweet sentences indicate closely both the negative and positive sentiments as most of tweets have both panic and comforting words. Fear among the people dominates the sentiment followed by trust from the authorities. Anticipation that necessary steps and precautions will be taken and sadness related to outbreak and death are also prevalent. Anger is also prevalent among people mostly related to quarantine.

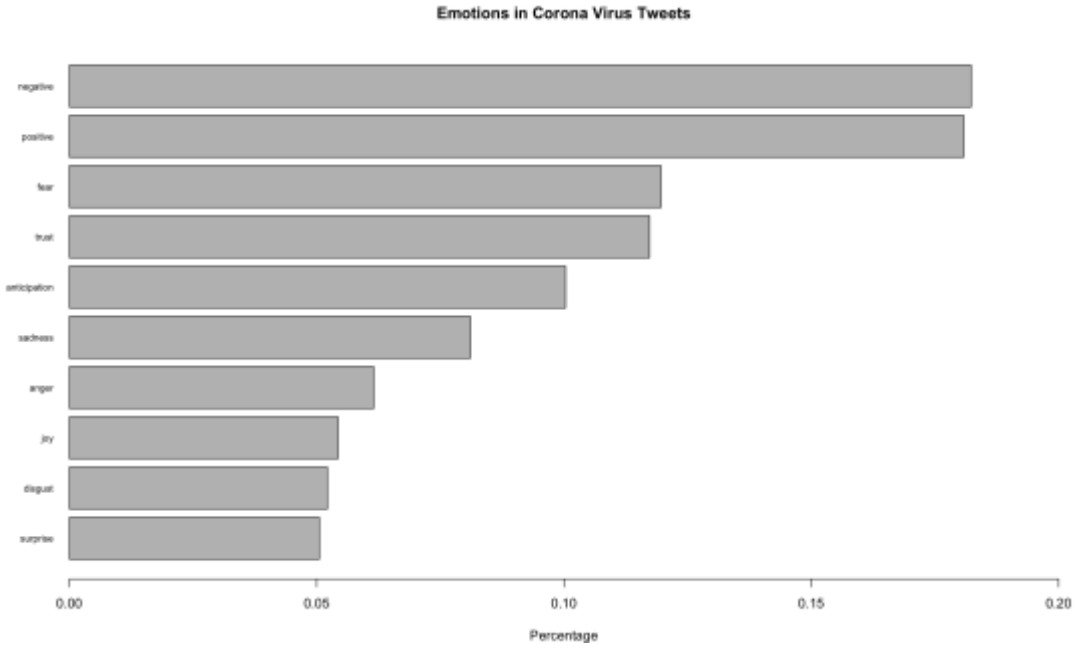

Figure 2

Frequent words in tweets term document matrix are also mostly relevant to the corona virus outbreak. Most frequent words being people, cases, china virus outbreak confirm to fact that corona virus outbreak originated from China. Words like world countries, spread indicate the spread of virus to other countries in the world. Cruise is related to the quarantine of people

on cruise ships in Japan and USA. Panic confirms to massive anxiety and panic related to corona virus. Oil market collapsed and stock markets are also worst affected. Corona virus has become a pandemic.

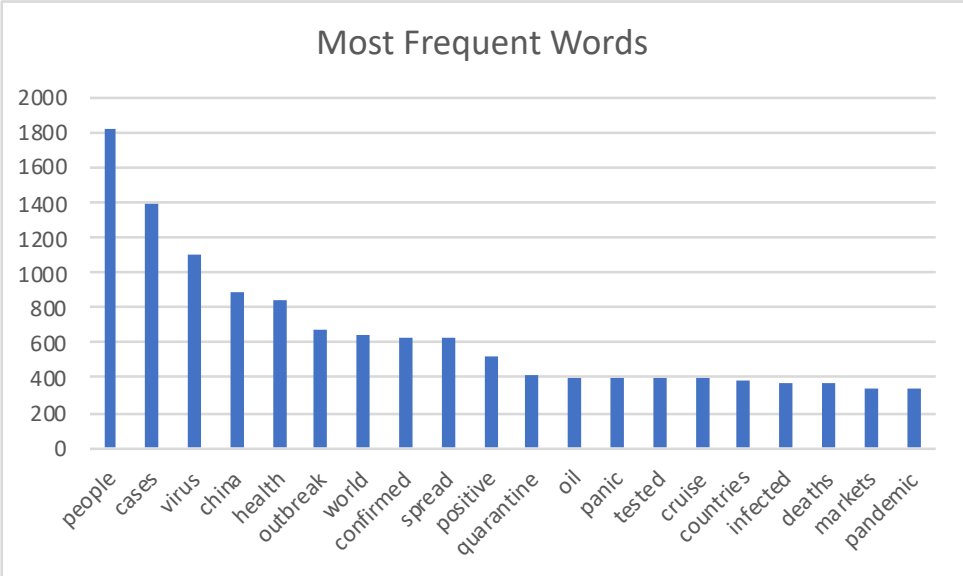

Figure 3

Correlation Analysis or Association analysis conducted using term document matrix created. There are many keywords with which correlation analysis could be done but for the purpose of study we choose the two most critical words “covid” and “quarantine”. The table below highlights the correlation with the 2 specific key words chosen. Correlated words are highly significant to the corona virus outbreak.

Table 1

$covid	Correlation
coronavirusoutbrea	0.25
sarscov	0.20
wuhan	0.16
china	0.14
wuhanvirus	0.13
updates	0.12
hospital	0.11
USA	0.10
$quarantine	Correlation
implementation	0.14
slapped	0.14
self	0.13
mandatory	0.12
compulsory	0.11
establish	0.11
communitylevel	0.10
Japan	0.10
Italy	0.10
restrictionscritical	0.10

Topic Modelling using Latent Dirichlet Allocation (LDA) with Gibbs Sampling method was conducted on the document term matrix created from text corpus and reduced sparsity to 0.99. Topic 1 can be attributed to panic among people related to corona virus outbreak similar to flu. Topic 2 relates to self-testing suggested by government for people to stay home and observe for symptoms like cold, cough & fever as there is no formal method of testing. Topic 3 attributed to confirmed cases in China and other countries and deaths due to corona virus. Topic 4 specifically refers to corona virus out breaks and death in countries like China, Iran, Italy and USA. Topic 5 relates to spread of infection. Topic 6 relates to news and videos being posted related to corona virus. Topic 7 relates to USA President speech related to corona outbreak and measures in that country. Topic 8 refers to outbreaks on the cruise ships in Japan

& USA and quarantine measures taken. Topic 9 attributes to advice by health authorities to sanitize hands and maintain good hygiene. Topic 10 refers to government suggesting public to take care against virus.

Table 2

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
people	coronavirus	coronavirus	coronavirus	coronavirus	coronavirus	will	covid	can	hands
just	one	cases	china	virus	covid19	trump	coronavirus	hands	coronavirus
dont	time	new	world	spread	covid	like	outbreak	good	health
know	due	first	outbreak	corona	news	going	positive	hand	also
get	test	confirmed	toilet	safe	coronaviruschallenge	coronavirus	tested	please	public
need	now	case	paper	stay	outbreak	stop	cruise	wash	may
flu	testing	countries	death	infected	read	now	ship	help	global
panic	home	deaths	iran	infection	must	get	coronavirususa	ÔøΩ	take
think	see	report	italy	around	video	cant	health	everyone	government
now	cdc	today	usa	latest	coronvirusuk	way	state	keep	care

5. CONCLUSION

The study concludes that information flow on twitter related to corona virus outbreak was relevant and mostly accurate with minor misinformation being spread. Compared to the earlier Ebola and Zika virus outbreaks where there was misinformation widely spread among the twitter users, there has been lesser misinformation spread during corona virus outbreak. Anxiety and panic were evident among the twitter users as the pandemic spread and deaths rise over period of time. But governments and health authorities had also used twitter to spread accurate and reliable information related to outbreaks particularly China government.

Negative sentiments dominated the tweet as expected as the virus highly contagious and deadly which was evident from sentiment analysis. Information spread was quite accurate and reliable and twitter also made sure that misinformation is stopped and deleted immediately. LDA analysis had highlighted that all the topics that are identified from the tweets are most relevant information to the corona virus outbreak. Twitter is still considered as one of the most preferred medium for information spread during pandemics and highly effective, is proved again from the current corona virus outbreaks. Governments, Health Authorities and Institutions like WHO can rely on twitter for spreading information and controlling panic among public at large.

REFERENCES

[1] Liang, H., Fung, I.C., Tse, Z.T.H. *et al.* How did Ebola information spread on twitter: broadcasting or viral spreading?. *BMC Public Health* 19, 438 (2019). https://doi.org/10.1186/s12889-019-6747-8

[2] Fung ICH, Tse ZTH, Cheung CN, Miu AS, Fu KW. Ebola and the social media. Lancet.; 384 (9961):2207. 2014

[3] Towers S, Afzal S, Bernal G, Bliss N, Brown S, Espinoza B, et al. Mass media and the contagion of fear: the case of Ebola in America. PLoS One. 10(6). 2015

[4] Fu KW, Liang H, Saroha N, Tse ZTH, Ip P, Fung ICH. How people react to Zika virus outbreaks on twitter? *A computational content analysis. Am J Infect Control*.; 44(12):1700–2. 2016

[5] Chew C, Eysenbach G. Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak. PLoS One. 5:11. 2010

[6] Fung ICH, Duke CH, Finch KC, Snook KR, Tseng PL, Hernandez AC, et al. Ebola virus disease and social media: a systematic review. *Am J Infect Control*. 44 (12):1660–71. 2016

[7] Oyeyemi Sunday Oluwafemi, Gabarron Elia, WynnRolf. Ebola, Twitter, and misinformation: a dangerous combination? *BMJ*, 349 :g6178, 2014

[8] Jeanine P.D. Guidry, Yan Jin,Caroline A. Orr, Marcus Messner, Shana Meganck. Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement. Public Relations Review. Elsevier September 2017.

[9] Lee, Junseok & Kang, Ji-Ho & Jun, Sunghae & Lim, Hyunwoong & Jang, Dongsik & Park, Sangsung. Ensemble Modeling for Sustainable Technology Transfer. Sustainability. 10. 2278. 10.3390/su10072278. 2018

[10] Prabhakar, K.R., Satish, D. An empirical text mining analysis of Fort McMurray wildfire disaster twitter communication using topic model. Disaster Advances – Vol. 9 (7) July (2016) E-ISSN 2278-4543. 2016

[11] Prabhakar, K.R. Climate Change and Twitter – An Empirical Analysis of Environmental Awareness and Engagement – Disaster Advances – Vol. 12(9) September (2019) E-ISSN 2278-4543, 2019.

[12] Dr. Madhubala Myneni, Narasimha Prasad L V and G Geetha Reddy, Automatic Assessment of Floods Impact Using Twitter Data. *International Journal of Civil Engineering and Technology*, 8(5), pp. 1228–1238. 2017

[13] Anita Kumari Singh and Shashi Mogalla, Automatic Identification of News Tweets on Twitter. *International Journal of Computer Engineering and Technology*, 9(4), pp. 140-147. 2018

[14] Savitha Mathapati, Anil D, Tanuja R, S H Manjula and Venugopal K R, CNSM: Cosine and N-Gram Similarity Measure to Extract Reasons for Sentiment Variation on Twitter. *International Journal of Computer Engineering & Technology*, 9(2), pp. 150–161, 2018

[15] Mashael Saeed Alqhtani and M. Rizwan Jameel Qureshi, Data Mining Approach for Classifying Twitter's Users. *International Journal of Computer Engineering & Technology*, 8(5), pp.42–53, 2017

[16] Boshra F. Zopon AL_Bayaty, Information Retrieval Topics in Twitter Using Weighted Prediction Network, *International Journal of Civil Engineering and Technology*, 8 (1), pp. 781–788, 2017