COVIDSENTI: A LARGE-SCALE BENCHMARK TWITTER DATA SET FOR COVID-19 SENTIMENT ANALYSIS

R.USHA[1], B KARTHIK REDDY[2], B SATHVIKA[3], BIJJALA SUPRAJA[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3&4]

DEPARTMENT OF CSE, CMR INSTITUTE OF TECHNOLOGY, KANDLAKOYA VILLAGE, MEDCHAL RD, HYDERABAD, TELANGANA 501401

ABSTRACT—Social media (and the world at large) have been awash with news of the COVID-19 pandemic. With the passage of time, news and awareness about COVID-19 spread like the pandemic itself, with an explosion of messages, updates, videos, and posts. Mass hysteria manifest as another concern in addition to the health risk that COVID-19 presented. Predictably, public panic soon followed, mostly due to misconceptions, a lack of information, or sometimes outright misinformation about COVID-19 and its impacts. It is thus timely and important to conduct an ex post facto assessment of the early information flows during the pandemic on social media, as well as a case study of evolving public opinion on social media which is of general interest. This study aims to inform policy that can be applied to social media platforms; for example, determining what degree of moderation is necessary to curtail misinformation on social media. This study also analyzes views concerning COVID-19 by focusing on people who interact and share social media on Twitter. As a platform for our experiments, we present a new large-scale sentiment data set COVIDSENTI, which consists of 90 000 COVID-19-related tweets collected in the early stages of the pandemic, from February to March 2020. The tweets have been labeled into positive, negative, and neutral sentiment classes. We analyzed the collected tweets for sentiment classification using different sets of features and classifiers. Negative opinion played an important role in conditioning public sentiment, for instance, we observed that people favored lockdown earlier in the pandemic; however, as expected, sentiment shifted by midMarch. Our study supports the view that there is a need to develop a proactive and agile public health presence to combat the spread of negative sentiment on social media following a pandemic.

Index Terms— COVID-19, epidemic, misinformation, opinion mining, pandemic, sentiment analysis, text mining, Twitter.

I. INTRODUCTION CORONAVIRUS disease (COVID-19) is a novel viral disease denoted by the year in which it first appeared [52]. The disease has affected many countries, with the battle to curtail its spread being waged in every country, even those countries with few or no infections. It was declared a pandemic on January 30, 2020, by the World Health Organization (WHO), an organization that is relentlessly trying to control it. The development of vaccines is eagerly anticipated and showing great promise [16]. As it stands, there is a lack of academic study on the topic to aid researchers, save for Bhat et al. [6] and Boldog et al. [8]. This hampers research findings on the consequences of COVID-19 on mental health or the study of the global economic implications. Due to the emergence of bizarre conspiracy theories around COVID-19, social media platforms, such as Twitter, Facebook, Reddit, and Instagram, have been actively working on scrutinizing and fact-checking in order to combat the spread of misinformation. Misinformation is defined as a deliberate attempt to confuse/mislead the public with false information. This gives rise to the need to create analytic methods that could be rapidly deployed to understand information flows and to interpret how mass sentiment among the population develops in pandemic scenarios. There has not been comprehensive research on analyzing conspiracy communication trends on social media and

cumulative personal-level information, with most studies presenting the analysis of preventive care and recovery, healthcare, social network, and economic data. Analyzing content posts on social media platforms, such as Twitter and Facebook, is a popular method to capture human emotional expression. Fears, numbers, facts, and the predominant thoughts of people as a whole, unsurprisingly, inundate the social media space, and this information, when analyzed, can reveal much about the prevailing mood and temperament of the broader human population. The extraordinary increase of society's dependence on social media for information, as opposed to traditional news sources, and the volume of data presented, has brought about an increased focus on the use of natural language processing (NLP) and methods from artificial intelligence (AI) to aid text analytics [5]. This information includes diverse social phenomena, such as cultural dynamics, social trends, natural hazards and public health, matters frequently discussed, and opinions expressed, by people using social media. This is because of its low cost and easy access and from the personal connectivity within the social network. Increasingly, social media is used by professional opinion leaders (and state actors) as a tool to amplify their message via its network effects. Many companies also use social media to promote products, brand names, and services [21]. Consequently, an information-rich reservoir is created by reviews and experiences shared by end users, and this information is stored as text, making platforms of open communication and social media salient information sources for researching issues concerning rapidly developing public sentiment.

II. LITERATURE SURVEY

1. Data mining twitter for COVID-19 sentiments concerning college online education

Daniel M. Brandon Published in Future Business Journal 1 December 2023

In the last decade there has been a large increase in corporate and public reliance on social media for information, rather than on the traditional news and information sources such as print and broadcast media. People freely express their views, moods, activities, likes/dislikes on social media about diverse topics. Rather than surveys and other structured data gathering methods, text data mining is now commonly used by businesses to go through their unstructured text in the form of emails, blogs, tweets, likes, etc. to find out how their customers feel about their company and their products/services. This paper reports upon a study using Twitter (recently renamed to "X") data to determine if meaningful and actionable information could be gained from such social media data in regard to pandemic issues and how that information compares to a traditional survey. In early 2020, the COVID-19 pandemic hit and forced colleges to move classes to an online format. While there is considerable literature in regard to using social media to communicate geo-political issues and in particular pandemics, there is not a study using social media to explore public sentiment in regard to COVID's forcing online education upon the public. In this study, text data mining was used to gain some insight into the feeling of Twitter users in regard to the effect of COVID-19 and the switch to online education in colleges. This study found that Twitter data mining did produce actionable information similar to the traditional survey, and the study is important since its results may influence organizations to explore the use of Twitter (and possibly other social media) to obtain people's sentiments instead of (or in addition to) traditional surveys and other traditional means of gathering such information. This paper demonstrates both the process of text data mining social media and its application to current real-world issues.

2. **Twitter and Census Data Analytics to Explore Socioeconomic Factors for Post-COVID-19 Reopening Sentiment**

Md. Mokhlesur Rahman, Ali Ggmn, +2 authors P. H. Chong Published in medRxiv 30 June 2020

Investigating and classifying sentiments of social media users (e.g., positive, negative) towards an item, situation, and system are very popular among the researchers. However, they rarely discuss the underlying socioeconomic factor associations for such sentiments. This study attempts to explore the factors associated with positive and negative sentiments of the people about reopening the economy, in the United States (US) amidst the COVID-19 global crisis. It takes into consideration the situational uncertainties (i.e., changes in work and travel pattern due to lockdown policies), economic downturn and associated trauma, and emotional factors such as depression. To understand the sentiment of the people about the reopening economy, Twitter data was collected, representing the 51 states including Washington DC of the US. State-wide socioeconomic characteristics of the people (e.g., education, income, family size, and employment status), built environment data (e.g., population density), and the number of COVID-19 related cases were collected and integrated with Twitter data to perform the analysis. A binary logit model was used to identify the factors that influence people toward a positive or negative sentiment. The results from the logit model demonstrate that family households, people with low education levels, people in the labor force, low-income people, and people with higher house rent are more interested in reopening the economy. In contrast, households with a high number of members and high income are less interested to reopen the economy. The accuracy of the model is good (i.e., the model can correctly classify 56.18% of the sentiments). The Pearson chi2 test indicates that overall this model has high goodness-of-fit. This study provides a clear indication to the policymakers where to allocate resources and what policy options they can undertake to improve the socioeconomic situations of the people and mitigate the impacts of pandemics in the current situation and as well as in the future.

3. COVID-19 Related Sentiment Analysis Using State-of-the-Art Machine Learning and Deep Learning Techniques

Z. Jalil, A. Abbasi, +4 authors Abdul Khader Jilani Saudagar Published in Frontiers in Public Health 14 January 2022

The coronavirus disease 2019 (COVID-19) pandemic has influenced the everyday life of people around the globe. In general and during lockdown phases, people worldwide use social media network to state their viewpoints and general feelings concerning the pandemic that has hampered their daily lives. Twitter is one of the most commonly used social media platforms, and it showed a massive increase in tweets related to coronavirus, including positive, negative, and neutral tweets, in a minimal period. The researchers move toward the sentiment analysis and analyze the various emotions of the public toward COVID-19 due to the diverse nature of tweets. Meanwhile, people have expressed their feelings regarding the vaccinations' safety and effectiveness on social networking sites such as Twitter. As an advanced step, in this paper, our proposed approach analyzes COVID-19 by focusing on Twitter

users who share their opinions on this social media networking site. The proposed approach analyzes collected tweets' sentiments for sentiment classification using various feature sets and classifiers. The early detection of COVID-19 sentiments from collected tweets allow for a better understanding and handling of the pandemic. Tweets are categorized into positive, negative, and neutral sentiment classes. We evaluate the performance of machine learning (ML) and deep learning (DL) classifiers using evaluation metrics (i.e., accuracy, precision, recall, and F1-score). Experiments prove that the proposed approach provides better accuracy of 96.66, 95.22, 94.33, and 93.88% for COVISenti, COVIDSenti_A, COVIDSenti_B, and COVIDSenti_C, respectively, compared to all other methods used in this study as well as compared to the existing approaches and traditional ML and DL algorithms.

IMPLEMENTATION

Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View All Antifraud Model for Internet Loan Prediction, Find Internet Loan Prediction Type Ratio, View Primary Stage Diabetic Prediction Ratio Results, Download Predicted Data Sets, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT PRIMARY STAGE DIABETIC STATUS, VIEW YOUR PROFILE.

CONCLUSION

Since the explosion of COVID-19 conspiracy theories, social media has been widely used both for and against misinformation and misconceptions. In this article, we address the issue of Twitter sentiment on COVID-19-related Twitter posts. We benchmark sentiment analysis methods in the analysis of COVID-19-related sentiment. Our findings indicate that the population favored the lockdown and stay home order in February; however, their

opinion shifted by mid-March. The reason for the shift in sentiment is unclear, but it may be due to misinformation being spread on social media; thus, there is a need to develop proactive and agile public health presence to combat the spread of fake news. To facilitate research among the community, we have released a publicly available large-scale COVID-19 benchmark sentiment analysis data set.

REFERENCES

[1] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. Boca Raton, FL, USA: CRC Press, 2013.

[2] N. Ahmad and J. Siddique, "Personality assessment using Twitter tweets," Procedia Comput. Sci., vol. 112, pp. 1964–1973, Sep. 2017.

[3] T. Ahmad, A. Ramsay, and H. Ahmed, "Detecting emotions in English and Arabic tweets," Information, vol. 10, no. 3, p. 98, Mar. 2019.

[4] A. Bandi and A. Fellah, "Socio-analyzer: A sentiment analysis using social media data," in Proc. 28th Int. Conf. Softw. Eng. Data Eng., in EPiC Series in Computing, vol. 64, F. Harris, S. Dascalu, S. Sharma, and R. Wu, Eds. Amsterdam, The Netherlands: EasyChair, 2019, pp. 61–67.

[5] F. Barbieri and H. Saggion, "Automatic detection of irony and humour in Twitter," in Proc. ICCC, 2014, pp. 155–162. 9This is an original tweet taken from Twitter. 10This is an original tweet taken from Twitter. 11This is an original tweet taken from Twitter.

[6] R. Bhat, V. K. Singh, N. Naik, C. R. Kamath, P. Mulimani, and N. Kulkarni, "COVID 2019 outbreak: The disappointment in Indian teachers," Asian J. Psychiatry, vol. 50, Apr. 2020, Art. no. 102047.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

[8] P. Boldog, T. Tekeli, Z. Vizi, A. Dénes, F. A. Bartha, and G. Röst, "Risk assessment of novel coronavirus COVID-19 outbreaks outside China," J. Clin. Med., vol. 9, no. 2, p. 571, Feb. 2020.

[9] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," Information, vol. 9, no. 5, p. 127, May 2018.

[10] X. Carreras and L. Màrquez, "Boosting trees for anti-spam email filtering," 2001, arXiv:cs/0109015. [Online]. Available: https://arxiv.org/ abs/cs/0109015

[11] J. P. Carvalho, H. Rosa, G. Brogueira, and F. Batista, "MISNIS: An intelligent platform for Twitter topic mining," Expert Syst. Appl., vol. 89, pp. 374–388, Dec. 2017.

[12] B. K. Chae, "Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research," Int. J. Prod. Econ., vol. 165, pp. 247–259, Jul. 2015.

[13] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2013, pp. 3267–3276.

[14] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson, "The pandemic of social media panic travels faster than the COVID-19 outbreak," J. Travel Med., vol. 27, no. 3, Apr. 2020, Art. no. taaa031.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[16] M. E. El Zowalaty and J. D. Järhult, "From SARS to COVID-19: A previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans—Call for a one health approach," One Health, vol. 9, Jun. 2020, Art. no. 100124.

[17] I. Fung et al., “Pedagogical demonstration of Twitter data analysis: A case study of world AIDS day, 2014,” Data, vol. 4, no. 2, p. 84, Jun. 2019.