e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

A LARGE-SCALE BENCHMARK TWITTER DATASETS FOR COVID-19 SENTIMENT ANALYSIS

P. Rajaram*1, V. Dhavamani*2

*1,2Student, Professor Of Computer Science And Engineering, Sir Issac Newton College Of Engineering And Technology, Nagapattinam, Tamilnadu, India.

DOI : https://www.doi.org/10.56726/IRJMETS45125

ABSTRACT

Social media (and the world at large) have beenawash 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 manifests as another concern in addition to the health risk that COVID-19 presents. 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 COVID-19 Sentiment data, 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 labelled 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 favoured lockdown earlier in the pandemic; however, as expected, sentiment shifted by mid-March. 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 apandemic.

Keywords: COVID-19, Epidemic, Misinformation, Opinion Mining, Twitter, Sentiment Analysis, Text Mining.

I. INTRODUCTION

CORONAVIRUS disease (COVID-19) is a novel viral disease denoted by the year in which it first 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. As it stands, there is a lack of academic study on the topic to aid researchers, this hampers research findings on the Consequences of COVID-19 on mental health or the study of the global economic implications. 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 networks, and economic data. Analyzing content posts on social media platforms, such as Twitterand 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. The extraordinary increase in 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. 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 easy access and personal connectivity within the social network.

II. METHODOLOGY

The overview of our proposed framework is shown in Fig.,with each component of the framework explained in the following.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

Data Collection and Labeling

The overview of our proposed framework is shown in Fig.,with each component of the framework explained in the following Out of approximately 2.1 million tweets crawled from February to March 2020, we included 90,000 unique tweets from 70,000 users that met the selection criteria. Our analysis identified 12 topics, such as quarantine, lockdown, and stay-home import data.

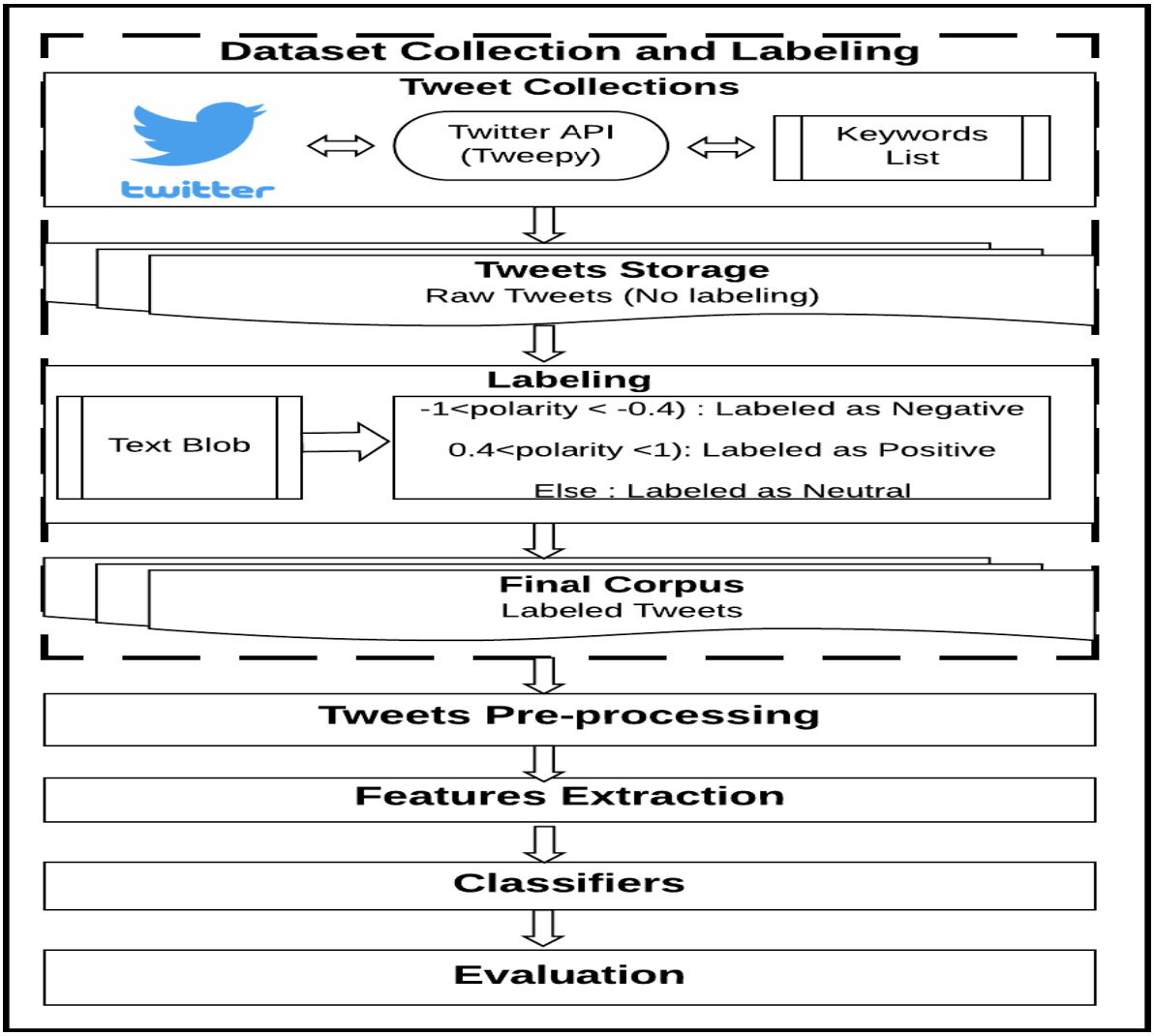

Fig 1. Overview of the proposed framework.

Preprocessing

Almost every social media platform uses hashtags to represent topics, i.e., #COVID-19, #StayHome, #StaySafe, and #Coronavirus. In most cases, hashtags are unnecessary to sentiment and can affect performance. Thus, in our first step, we performed basic cleaning of the text by removing unnecessary hashtags, just the hashtag character, not the hashtag text.

Exploratory Analysis

We first performed keyword trend analysis on our preprocessed corpus to find out the most frequently mentioned words. We found that people are talking about coronavirus cases, the coronavirus outbreak, social distancing, the coronavirus pandemic, the crises due to coronavirus, and staying at home.

Feature Extraction

In this experiment, vectorization techniques and word embeddings are used for feature extraction. Term frequency-inverse document frequency (TF-IDF) has been used for vectorization. Similarly, for word embeddings, pretrained Word2Vec, GloVe, and fastText embeddings trained on Common Crawl and Wikipedia are used and have 300-D vectors. In addition, we used hybrid models, such as hybrid ranking (HyRank) and

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

Improved Word Vector (IWV), that incorporate sentiment and context of tweets for Twitter sentiment analysis.

Classification

To provide a comprehensive analysis, we used ML- and DL-based classifiers to gauge performance in the sentiment classification task. ML-based classifiers, such as support vector machine (SVM), naive Bayes (NB), decision tree (DT), and random forest (RF), are employed in our analysis. In addition to traditional ML, we have also applied two DL-based classifiers, namely convolutional neural network (CNN) and bidirectional long-short-term memory (Bi-LSTM).

III. RESULT AND DISCUSSION

This section presents the experimental step used to evaluate the performance on benchmark data sets and provides benchmark results for the purpose of comparison. We used accuracy and a tenfold cross-validation. To establish the baselines for ML classifiers, we used traditional methods such as TF-IDF and Word2Vec, word embedding-based models such as Word2Vec, GloVe, and fastText, hybrid models such as IWV and HyRank, and transformer-based LMs such as BERT, DistilBERT, XLNET, and ALBERT. , experimental results on Covid Sentiment data showed that the increase in the number of COVID-19 cases and increased mortality rate negatively impacted people's lives. For example, "Many Americans fear the coronavirus outbreak could prevent them from working and getting paid.

IV. 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 a proactive and agile public health presence to combat the spread of fake news. To facilitate the community, we have released a publicly available large-scale COVID-19 benchmark sentiment analysis data set. We can conclude that more the cleaner data, more accurate results can be obtained. Use of bigram model provides better Sentiment accuracy as compared to other models.

V. 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 Computer. 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-analyser: 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.