JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 26 Issue 01, 2023 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

TOPIC LABELING AND SENTIMENT ANALYSIS ON COVID-19 NEWS

CH. Nageswararao[1], C. Shoba Bindu[2], P. Dileep Kumar Reddy[3]

[1]Department of CSE, JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India, nageshmtech147@gmail.com

[2]Department of CSE, JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India, shobabindhu.cse@jntua.ac.in

[3]Department of CSE, Narsimha Reddy Engineering College (Autonomous), Secunderabad, Telangana, India, dileepreddy503@gmail.com

ABSTRACT

Newspapers play a very crucial role in society because they inform people about current events and how they may impact their day-to-day lives. In cases of wellbeing emergencies, along with the recent COVID-19 outbreak, their significance becomes even more critical and indispensable. Since the beginning of the pandemic, newspapers have been a valuable source of information for the general public on a wide range of topics, including the identification of a novel coronavirus strain, restrictions and other lockdown, governmental regulations and details on the development of a coronavirus vaccine. In this case, analysing newly emerging and extensively reported topics, themes, and concerns, as well as the accompanying sentiments from different nations, can aid in our understanding of the COVID-19 pandemic. In our paper, we investigated greater than 100,000 COVID-19 news headlines and articles utilising BERTopic, Top2Vec (topic modelling), and XLNet (sentiment analysis and classification). Our topic modelling findings showed that the most prevalent and widely covered topics in the India, Japan, South Korea and UK were education, sports, Vaccination, and economy. Additionally, our sentiment classification model achieved 92% validation accuracy, 96% testing accuracy, 97% F1-Score and the study revealed that the UK, the nation with the worst affects in our dataset, also has the largest prevalence of negative sentiment.

INDEX TERMS: Natural Language Processing, Machine Learning, Topic Modeling, Topic Labeling, BERTopic, Top2Vec, XLNET, Sentiment Analysis, Newspaper, COVID-19.

I. INTRODUCTION

In December 2019, a pneumonia epidemic with an anonymous cause was identified in Wuhan, Hubei Province, China. This outbreak's underlying virus was later identified and given the name severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The World Health Organization (WHO) designated the illness brought on by SARS-CoV-2 as "COVID-19" in February 2020 [1]. As of September 24, 2022, the virus got attacked over than 61.46 million of people and claimed the life of over than 1.65 billion of people, as per the Johns Hopkins Coronavirus Resource Center [2]. The vaccines have been extensively explored by F. Krammer in "SARS-CoV-2 vaccines under development" [3] and G. Forni and A. Mantovani in

Submitted: 11/12/2022 **Accepted**: 06/01/2023

"COVID-19 vaccines: where we stand and difficulties ahead" Derek Abbott served as the associate editor who oversaw the review of this article and gave his approval for development, publication procedure, vaccination types, and leading vaccine candidates [4].

Newspapers all across the world are reporting heavily on COVID-19-related news as the globe is currently struggling with the COVID-19 pandemic. Because of this, newspapers are a fantastic resource for learning about the social, economic, and political aspects surrounding this devastating pandemic in a given community. Furthermore, a range of cultural convictions and ideological presumptions influence the way news is produced. News is a representational social text that incorporates and disseminates certain presumptions and ideas, it serves as a means of recreating social aspects. Due to the lockdown, millions of people globally lost their employment. A massive migration problem was caused by India's total lockdown, and many of the world's leading countries saw negative GDP growth. In the US alone, more than 330 businesses declared bankruptcy last year, with COVID-19 cited as a contributing factor.

Natural Language Processing and its different methodologies has increasingly popular. Whenever it refers to analysing and processing massive amounts of natural language data. NLP is defined as: A discipline which integrates artificial intelligence and linguistics to make it possible for computers to comprehend natural language. The importance of NLP nowadays is further heightened by the fact that we produce large amounts of unlabelled text data each day. The most well-known and widely used NLP approaches are topic modelling [5], sentiment analysis [6], machine translation [7], named entity recognition [8], and text summarization [9]. The ability to gather as much information as you can regarding the COVID-19 dilemma is a crucial advantage. Any information that is accessible, locatable, collectible, and understandable will help people make the best judgments. Researchers may now use NLP to gain data from sources including such published researches, government documents, social media platforms, and news on subjects like illness socioeconomic effect, vaccine progression, patient history and co-morbidities, and domestic and international politics [10].

The goals of this research are 1) To analyse the main themes and issues of English language COVID-19 news items published in four different nations and look for statistics. To investigate identify any recurring themes, we would compare these issues. 2) To categorize as well as examine related sentiments in COVID-19 newspaper headlines. This will allow us to better comprehend: (i) the contrast of COVID19 news sentiments in four countries; (ii) when there is a relationship among negative sentiment and a country's degree of affectivity; and (iii) the sentiments that are indicative of similar themes in various countries.

- In order to analyse the COVID-19 conversation, we took the most typical subjects or themes from the four nations' newspapers. In this method, we learn about the COVID-19-related topics or concerns that are frequently covered in the mainstream media.

- In addition to producing superior results than other traditional classifiers, our XLNET sentiment classification model achieved 96% testing accuracy.
- A comparison of two topic modelling and sentiment analysis methodologies to assess COVID-19 news is done in this paper.
- Our study offers insightful cross-cultural perspectives on the COVID-19 news media platform reporting.

II. RELATED WORK

A. TOPIC MODELING

The COVID-19 pandemic is being studied from many different angles by researchers from throughout the world. Numerous research from numerous domains have already been published after the COVID-19 illness was discovered in December 2019. One illustration of the expanding amount of research on COVID-19 is NLP-based analysis of news, social media posts, and scholarly papers that are connected to the disease. Petar Kristijan Bogović et al. [11] Topic Modelling of Croatian News During COVID-19 Pandemic, Digital topic modelling was employed by Liu et al. [12] to examine news coverage of the early COVID-19 epidemic in China. During the COVID-19 journals, Xiangpeng Wan et al. [13] Topic Modeling and Progression of American Digital News Media During the Onset of the COVID-19 Pandemic. Noor et al. [14] analysis of public sentiment towards the new COVID-19 outbreak on Twitter.

B. TOPIC LABELING

The most crucial step in preparing data for supervised learning and unsupervised algorithms is classifying the data. David Pelkmann et al. [15] How to Label? Combining Experts' Knowledge for German Text Classification. Shikun Zhang et al. [16] a survey on machine learning techniques for auto labeling of video, audio, and text data. Xiaojin Zhu et al. [17] Semi-Supervised Learning Literature Survey. Machine learning algorithms for data labeling: an empirical evaluation [18]. R. Ghani [19] Combining labeled and unlabeled data for text classification with a large number of categories. Belainine Billal et al. [20] Semi-supervised learning and social media text analysis towards multi-labeling categorization.

C. SENTIMENT ANALYSIS

Usman Naseem et al. [21] COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. Qanita Bani Baker et al. [22] Sentimental Analysis for Studying and Analyzing the Spreading of COVID-19 from Twitter Data. K. Anuratha et al. [23] Topical Sentiment Classification to Unmask the Concerns of General Public during COVID-19 Pandemic using Indian Tweets. Tianyi Wang et al. [24] COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model.

D. TOPIC MODELING AND SENTIMENT ANALYSIS

Topic modelling or sentiment analysis using COVID-19 news data was carried out in all the papers referenced from [11] to [14] and [21] to [24]. In contrast, only few researches have investigated COVID-19 data using topic modelling and sentiment analysis. Hui Yin et al. [25] Sentiment analysis and topic modeling for COVID-19 vaccine discussions. K. Anuratha et al. [26] Topical Sentiment Classification to Unmask the Concerns of General Public during COVID-19 Pandemic using Indian Tweets. Lastly, Xie et al. [27] explored public response to COVID-19 on Chinese micro blogging site Weibo with LDA topic modeling and sentiment analysis. Our research can be grouped with studies that analyse COVID-19 data using topic modelling and sentiment classification techniques. However, this paper is a great addition contribution to understanding cross-cultural COVID-19 news because of the nature of our COVID-19 (news articles) and methodologies (Top2Vec and BERTopic both are topic modelling and XLNet for sentiment classification). Our work is, as far as the authors are aware, the first attempt to evaluated sentiment analysis and topic modelling techniques to analyse COVID-19 news articles.

III. METHODOLOGY

As seen in Fig. 1. This paper is split into two sections. In the initial section, we planned the BERTopic and top2vec models to identify the most recognisable topics in each nation's dataset., to offer contrasts between the two models and to carefully examine them. The second section is sentiment analysis. This section may be further broken down into two subsections. In one section, semi-supervised machine learning techniques are used to create a labelled dataset. The final section will use XLNet to test and train our labelled dataset. After achieving an acceptable level of validation accuracy, testing accuracy and F1 Score, we evaluate the sentiments of the most popular topics and gathered headline sentiments.

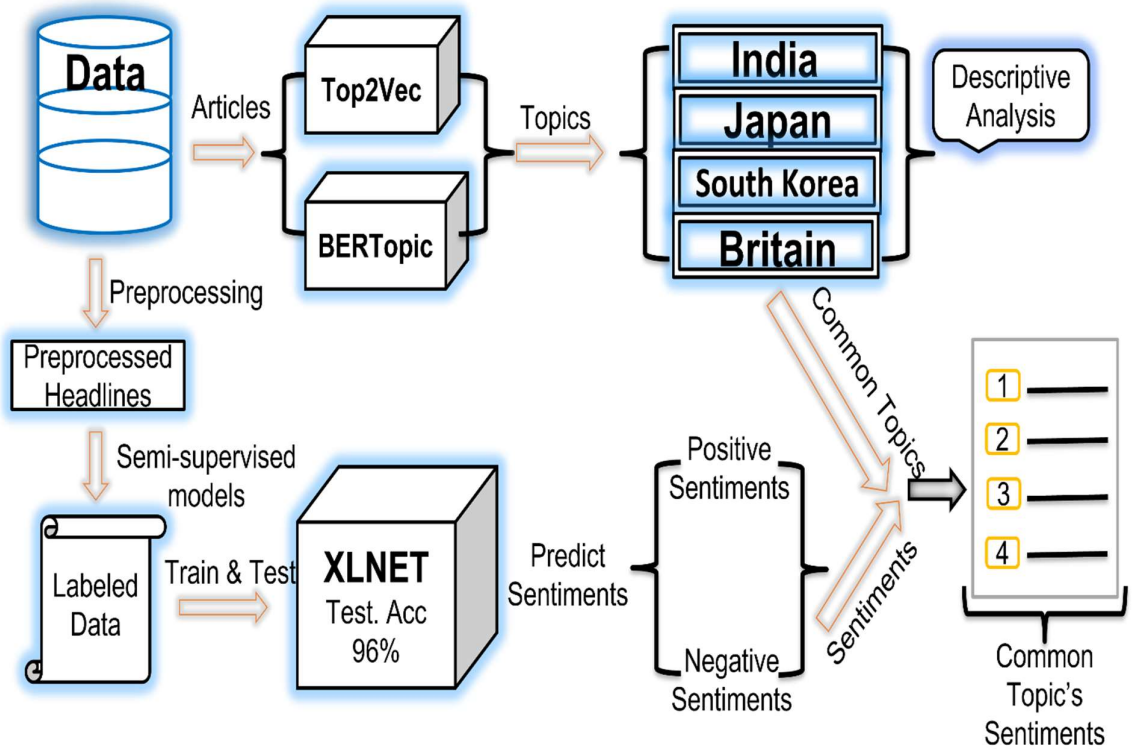

Figure 1. Research Methodology Workflow.

A. DATASET

We browsed and collected articles from eight major English-language newspapers from four different countries using the terms Coronavirus or COVID-19. The dates from 1st January, 2020, through 1st December, 2020, are the gathering time period for news headlines and articles. We utilised the Beautiful Soup Python package for web-scraping. We deleted the duplication articles and headlines. In the BERTopic and Top2Vec models data preprocessing is not needed. We tested, trained and made evaluated positive and negative sentiments in data preprocessed news headlines. Table. 1 summarizes the newspapers and the gathered Coronavirus/ COVID-19 related articles.

TABLE 1. Dataset

No. of Countries	No. of Newspapers	No. of Articles
India	The Indian Express, Hindustan Times	47,342
Japan	The Asahi Shimbun, Mainichi Shimbun, The Japan Times	21,039
South Korea	The Korea Times, The Korea Herald	10,076

UK	Daily Mail	23,821
	Total	**102,278**

B. Top2Vec MODEL, BERTopic MODEL AND TOPIC MODELING

One of the biggest challenges in NLP is how to organise, search, and summarise a massive body of text. Whenever a big data of text cannot be intelligently read and arranged through by a human, topic modelling is utilised. A topic model can be employed to evaluate the topics or latent semantic structure in the corpus. The Top2Vec technique is based on the idea that a topic is characterized by several documents that are semantically related to one another. In order to make the distance between topic, document, and word vectors express semantic similarity, they are simultaneously embedded. However, the text does not need to be preprocessed (stop-words removed, stemmed, and lemmatized) and unlike the more traditional topic modelling techniques such as LDA, Topic models are not dependent upon prior knowledge of existing topics. Top2vec was indeed entirely of any human decision-making or intervention, with the exception of selecting the model's training parameters.

In the initial section, Doc2Vec is used to build, word vectors and jointly embedded document. This stage might locate documents among other related documents and the most interesting words. In the next one, document vectors would be lower-dimensionally embedded using UMAP [28]. In high dimensional space, document vectors are sparse, so dimension reduction aids in locating dense areas. In the third one, HDBSCAN would be utilised to locate dense domains of documents. Lastly, every dense location computes the topic vector by calculating the centroid of document vectors in the original dimension.

Although top2vec allows us to leverage pretrained embedding model, such as universal-sentence-encoder, to create combined word and document embeddings, we configure topic models using our own data [29]. We utilised the "deep learn" parameter which produces the highest output vectors but requires a significant time to train. For particular, our Top2Vec model required to take more than 20 hours to train our biggest Indian dataset. So, when Top2Vec model is finished, it presents a variety of data, including topic score, topic number, topic size, topic words etc. Topic size displays the set of documents which are the most similar to each topic, in addition to each topic, the top 50 words are produced in order of their semantic closeness to the topic. The topic score is the cosine similarity to the keywords for each topic. The topic would be more indicative of the searched keyword if it had a higher topic score. This top2vec feature is utilized in this paper to identify utmost important topic via keywords. The top2vec semantic search feature was also used to construct the word clouds in Figures 4, 5, 6, and 7. The last step was to label these topics according to our understanding of each topic's (50 keywords).

BERTopic is a topic modelling approach that makes use of BERT embeddings and c-TF-IDF to build dense clusters that enable readily understandable topics while preserving key terms from the topic descriptions [30]. It offers the ability to utilise pretrained embedding models, such BERT Sentence Transformers, Doc2Vec, with custom embeddings, HuggingFace Transformers, Universal Sentence Encoder (USE), Flair, Spacy, Gensim, Combinations for word and Custom Backend Embeddings, document embeddings, we utilise our own data to construct topic models. This BERTopic feature is utilized in this paper to identify utmost important topic via keywords. The BERTopic semantic search feature was also used to construct the word clouds in Figures 4, 5, 6, and 7. The last step was to label these topics according to our understanding of each topic's (50 keywords).

C. DATA PREPROCESSING

As part of Natural Language Processing, text data preprocessing involves cleaning and preparing texts data. The procedures we used to preprocess the headlines are listed below.

1) In contractions, words like aren't and there's are abbreviated. We enlarged each of these words into their original form in the first phase.
2) HTML tags were Removed.
3) Headlines that are Tokenized.
4) Lemma is another term for the root word. Every token is now converted into its corresponding root word.
5) Using NLTK library, all Stopwords were removed.
6) Special characters are often non-alphanumeric characters that really can cause text noise. As a result, we omitted the special characters.
7) We must make sure that all accented characters are transformed and standardised into ASCII characters because we are reviewing headlines in English.
8) Finally, we lowercased all of the words.

In Figure 2, we display the text data preprocessing steps we took to clean the headlines before sentiment classification.

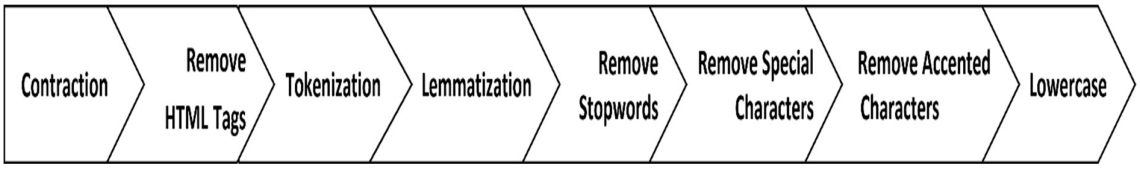

FIGURE 2. Cleaning Headlines by using Preprocessing Steps.

D. DATA LABELING HEADLINES FOR SENTIMENT CLASSIFICATION

A large portion of sentiment classification research is done on tweets and other social media posts. Because these posts are very subjective, they are an excellent resource for sentiment classification. Unlike news headlines and news articles present facts, making them more objective (opinion pieces and expect editorials). Sentiment classification of labeled headlines are split into three sections. The very initial section of this research to label the news headlines is: Using semi-supervised sentiment classification. We used five most popular Traditional semi-supervised machine learning models 1) TextBolb (unsupervised) [31] 2) LR-BOW [32], 3) LR-TFIDF [33], 4) SVM-BOW [34] and 5) SVM-TFIDF [34] on 102,124 headlines. The next section is: to keep headlines that fall into the positive and negative categories according to all four models. When compared to other models, the SVM-BOW model produces the best results in our experience. Using this SVM-BOW model, we achieved an accuracy of 92%. We created a basic Python code for this procedure. Only around 8% of the total headlines were still available after the second section. As in third section, we voluntarily verified all of the headlines from the second section. Finally, we virtually balanced the labelled data using oversampling. An overfitting problem may arise from oversampling. You can determine a model's overfitting by variance among training and validation accuracy. Our classification model's training accuracy for the 4th (last) epoch (which provided 92% validation accuracy) was 96%. It can be said that our model is not overfitting because the accuracy variation among both training and validation is not very high. Our sentiment categorization model is summarised in Fig. 3. Thus, we gathered 10,727 news headlines (5358 negatives and 5369 positives). The XLNet model is fine-tune using this labeled dataset.

```
Epoch 1/4
----------------------------------------------------------------------
Train loss 0.4858877947147345 Train accuracy 0.7707667731629393
Val loss 0.3478690733776448 Val accuracy 0.8865248226950354

Epoch 2/4
----------------------------------------------------------------------
Train loss 0.28269685469632244 Train accuracy 0.9230564430244942
Val loss 0.4818905067676761 Val accuracy 0.9011524822695035

Epoch 3/4
----------------------------------------------------------------------
Train loss 0.18843202872026496 Train accuracy 0.9543397231096912
Val loss 0.3952102256782047 Val accuracy 0.9171099290780141

Epoch 4/4
----------------------------------------------------------------------
Train loss 0.13947084306278568 Train accuracy 0.9688498402555911
Val loss 0.4406345460069752 Val accuracy 0.9206560283687943
```

FIGURE 3. Summary of the XLNET Sentiment Classification Model.

E. SENTIMENT CLASSIFICATION WITH XLNet

The traditional bag-of-words and Term Frequency-Inverse Document Frequency based Natural language processing classifiers, which include LogisticRegressionBOW, LogisticRegressionTFIDF, SupportVectorMachineBOW, and SupportVectorMachineTFIDF, ignore word structure (context) and consequently the meaning of words that appear in the document (hence semantics). Additionally, Generalized Autoregressive Pretraining for Language Understanding (XLNET) [35], GPT-2, GPT-3, BERT [36] and RoBERTa [37], are few samples of the Transformers-based pre-trained models. For each occurrence of a given word, these transformer-based models take into account the context. There are 20 tasks where XLNet outperforms BERT and 18 tasks where it reaches state-of-the-art performance. This is one of the reasons why these models perform better than classical bag-of-words models.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. RESULTS OF TOPIC MODELING AND TOP TOPICS

For this paper, we used our collected dataset from each nation to train the top2vec and BERTopic model. Furthermore, top2vec includes three speed parameters: “fast-learn”, “learn”, and “deep-learn”. Fast-learn is the quickest, but produces the lower grade outcomes, whereas learn takes the most time but produces better outcomes, and deep-learn takes the highest but produces the greatest outcomes. We used deep-learn parameters to train our top2vec model. The BERTopic takes the most time to train since it must fit the models on a collection of documents, generate topics, and give the best documents with topics.

1) TOP TOPICS IN INDIA

We acquired 587 topics after training the BERTopic model on a 47,342 article Indian dataset, opposed to 402 topics after training the Top2Vec model on the same dataset articles. Figure 4 displays a word cloud for the biggest topic on both Top2Vec and BERTopic - vaccine development news.

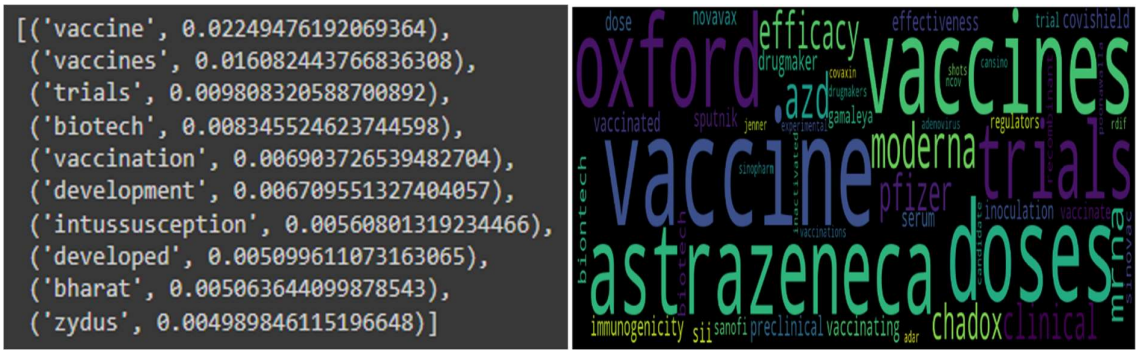

FIGURE 4. Biggest Topic in India’s Dataset - Vaccine Development related News.

2) TOP TOPICS IN JAPAN

We acquired 305 topics after training the BERTopic model on a Japan dataset of 21039 articles, opposed to 251 topics after training the Top2Vec model on the same dataset of articles. Figure 5 displays the word cloud for the biggest topic on both Top2Vec and BERTopic - sports related news.

FIGURE 5. Biggest Topic in Japan's Dataset - Sports related News.

3) TOP TOPICS IN SOUTH KOREA

We acquired 119 topics after training the BERTopic model on a South Korea dataset of 10076 articles, opposed to 130 topics after training the Top2Vec model on the same dataset of articles. Figure 6 displays the word cloud for the biggest topic on both Top2Vec and BERTopic - Diseases related news.

FIGURE 6. Biggest Topic in South Korea's Dataset - Diseases related News.

4) TOP TOPICS IN UK

We acquired 323 topics after training the BERTopic model on a UK dataset of 23821 articles, opposed to 296 topics after training the Top2Vec model on the same dataset of articles. Figure

7 displays the word cloud for the biggest topic on both Top2Vec and BERTopic - Infection related news.

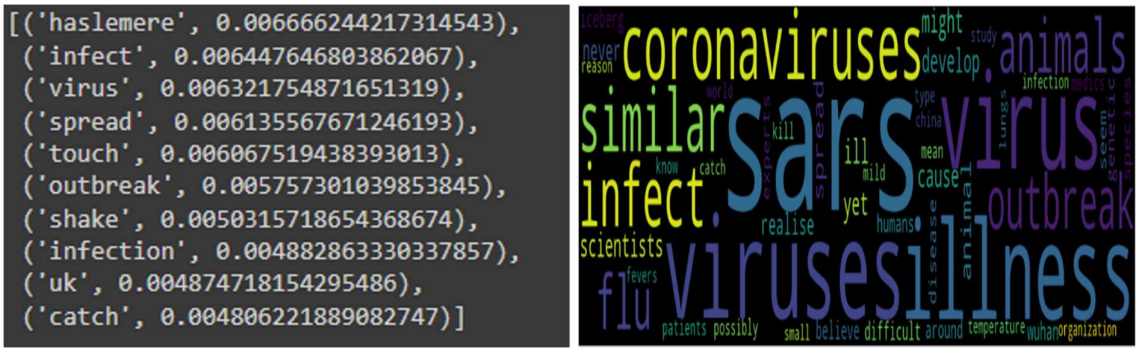

FIGURE 7. Biggest Topic in UK's Dataset – Infection related News.

5) COMPARISON OF TOP2VEC AND BERTOPIC

This study compares the output of two topic modelling algorithms in an effort to overcome the challenges of retrieving relevant details from brief and unstructured texts. Top2Vec and BERTopic need an interactive process utilizing an embedding model in order to inspect a topic. As a result, for a deeper understanding, both algorithms enable researchers to find extremely pertinent topics centered around a given term. According to the datasets from India's, Japan's, South Korea's, and the UK's, Bertopic performs better by 33.6%, 19.4%, 41.9%, and 8.7% than top2vec model. In terms of overall topics identified and topics labelled, bertopic is evaluated to be 25% better accuracy than top2vec model. In the end Fig. 8 of our research showed that the bertopic model had better outcomes than the top2vec model.

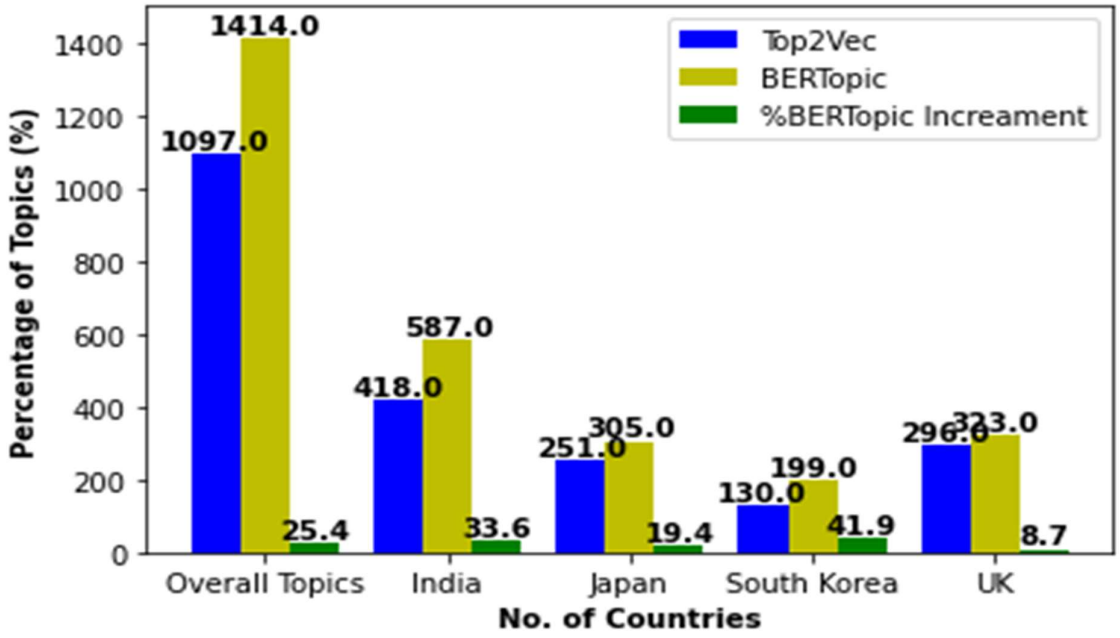

FIGURE 8. Comparison of Top2Vec and BERTopic.

B. RESULTS OF SENTIMENT CLASSIFICATION

We utilised our labelled dataset of 10,727 headlines to perform sentiment classification. In order to fine-tune the XLNet BASE model, we added 12 layers and 512 hidden dimensions. Our configuration includes four epochs, a 2n-5 learning rate, and eight batches. The validation accuracy for the fourth part of the epoch was 92%, testing accuracy was 96% and F1-Score was 97%. Additionally, we conducted comparison studies using traditional bag-of-words and Term Frequency-Inverse Document Frequency approach-based classifiers such as Support Vector Machine and Logistic Regression. Figure. 9 compares the accuracy of various classifiers with XLNet. Compared to other classification models, the XLNet has high accuracy. As opposed to the RoBERTa model (90%) that was previously used, our XLNET model achieved (96%) a better result.

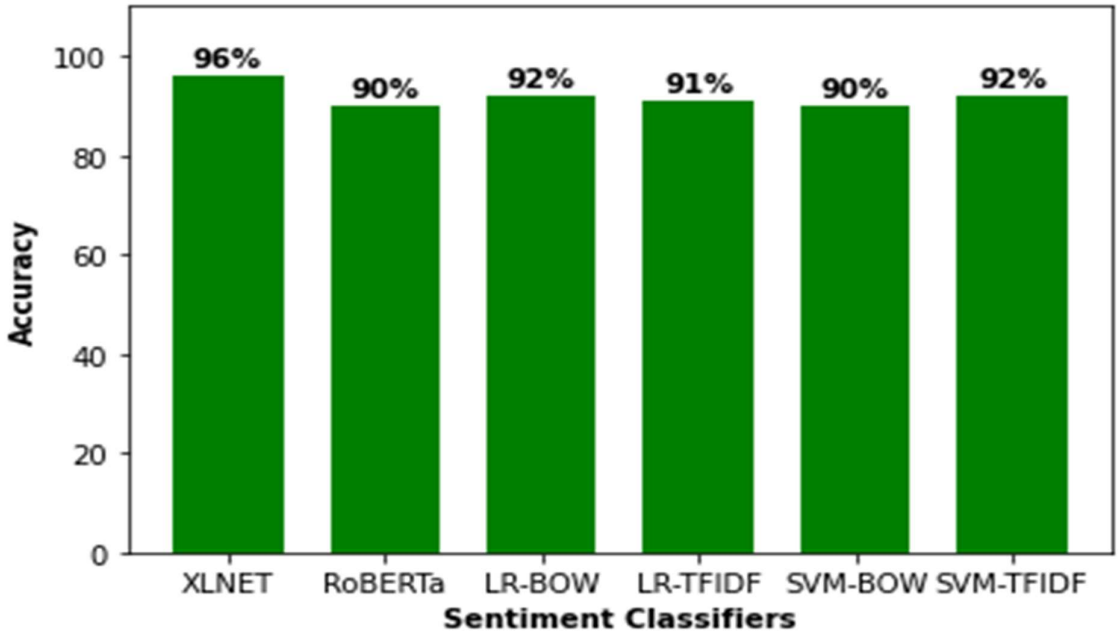

FIGURE 9. Comparison of XLNet and Other Classifiers.

1) SENTIMENTS OF INDIA's HEADLINES

The overall Indian dataset has 25,072 (52.96%) negative and 22,265 (47.04%) positive headlines out of 47,342 headlines. The Indian dataset was almost equally balance.

2) SENTIMENTS OF JAPAN's HEADLINES

The overall Japan dataset has 11,463 (54.48%) negative and 9576 (45.52%) positive headlines out of 21,039 headlines. More negatives were found in the Japan dataset.

3) SENTIMENTS OF SOUTH KOREA HEADLINES

The overall South Korea dataset has 4770 (47.34%) negative and 5306 (52.66%) positive headlines out of 10,076 headlines. Among the four nations we examined, the South Korean dataset showed the majority of positive results.

4) SENTIMENTS OF UK HEADLINES

The overall UK dataset has 14,253 (59.86%) negative and 9568 (40.14%) positive headlines out of 23,821 headlines. Among the four nations we examined, the UK dataset showed the most negative results.

5) COMPARISON OF SENTIMENTS FROM FOUR COUNTRIES

Our analysis of the whole COVID-19 dataset of 102,278 headlines revealed that 52,623 (51.45%) headlines were negative, while the remaining 49,655 (48.55%) headlines were

positive. According to the survey, there were more negative news reports than positive reports regarding COVID-19. The most negative and positive nations are the South Korea and UK, accordingly. After all, the UK is the worst affected of the four nations surveyed. Simultaneously, South Korea has been the most successful in battling the COVID-19 outbreak, with just 19 deaths per million. In our paper, India is the sole developing nation. While India is the second most afflicted country after the US in number of infected persons. According to deaths per million, it does better than other advanced nations. Figure.10 Showed comparison of countries covid-19 news headlines sentiments.

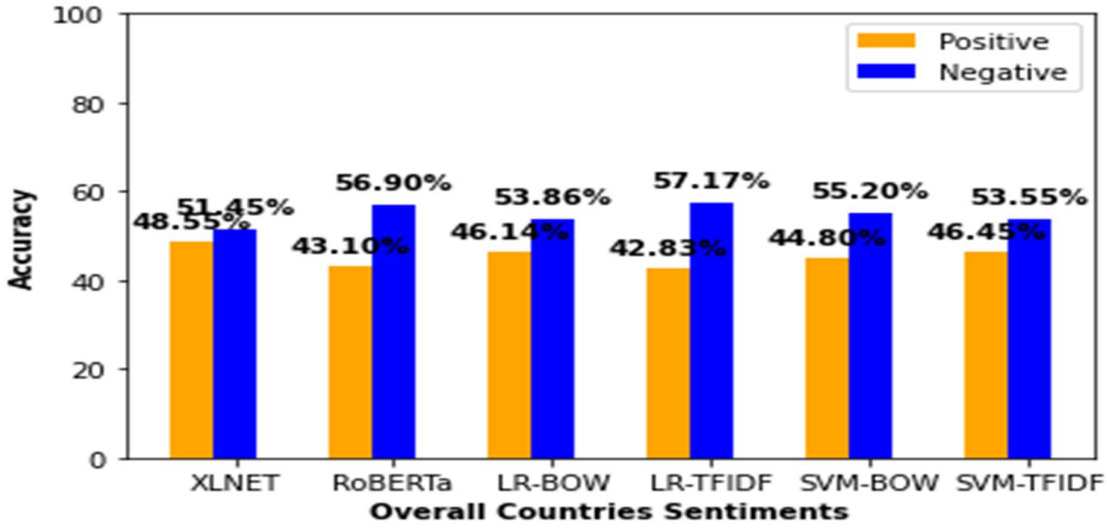

FIGURE 10. Comparison of Four Nations COVID-19 News Headlines Sentiments.

V. CONCLUSION AND FUTURE WORK

Despite starting over three years ago, the COVID-19 epidemic is still uncontrollable. It is still hard to anticipate when the vaccine will be broadly accessible to the general population, despite the fact that multiple potential vaccine candidates have been licenced for emergency use in various nations. In light of this, we conducted research to identify and understand the main problems and opinions expressed in COVID-19 related news.

Over the period of 11 months, our research gathered many as 100,000 Coronavirus/ COVID-19 relevant news headlines and articles from four nations. In the initial phase of this paper, which involved topic modelling - we created topics for each nation using the BERTopic and top2vec models. The number of topics appears to be directly proportional to the number of articles, since india dataset created the most topics (587) despite having the most articles (47,342). Furthermore, the Vaccination, the Sports, Education, and Economy were the most often mentioned topics across all four countries in our research of the top 10 topics.

In the next section of our paper, we applied a state-of-the-art XLNet model to classify headline sentiment. In comparison with other traditional classifiers, our XLNet model

was able to categorise headlines better, with validation accuracy of 92%, testing accuracy of 96%, and it was more accurate at classifying headlines. According to our sentiment classification implementation of the XLNet model, the UK has the majority of negative news (59.86%), whereas South Korea has the majority of positive news (52.66%).

In future works, our first step would be to conduct further research on multi-class sentiment classification across four nations. We'd also like to enhance the accuracy of our XLNet model. In addition, novel approaches for detecting emotions [38] and sentiment [22] have been developed, employing models such as LSTM and the lexicon-based convolutional neural network. In our future research, we'd like to investigate these models as well. Finally, we would want to include more nations in our dataset in order to expand the reach of our research and make it more global in scope.

REFERENCES

[1] C. Wang, et al. ‘‘A novel corona-virus outbreak of global health concern,’’ Lancet, vol. 395, no. 10233, pp. 470–473, 2020, doi: 10.1016/S0140-6736(20)30185-9.

[2] Johns Hopkins Coronavirus Resource Center. COVID-19 Global Map. Accessed: sep, 23, 2022. [Online]. Available: https://coronavirus.jhu.edu/map.html

[3] F. Krammer, ‘‘SARS-CoV-2 vaccines in development,’’ Nature, vol. 586, no. 7830, pp. 516–527, Oct. 2020, doi: 10.1038/s41586-020-2798-3.

[4] G. Forni et al. ‘‘COVID-19 vaccines: Where we stand and challenges ahead,’’ Cell Death Differentiation, vol. 28, no. 2, pp. 626–639, Feb. 2021, doi: 10.1038/s41418-020-00720-9.

[5] D. M. Blei, ‘‘Probabilistic topic models,’’ Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012. [Online]. Available: https://dl.acm.org/doi/ fullHtml/10.1145/2133806.2133826.

[6] Sanket Andhale et al. “Twitter Sentiment Analysis for COVID-19,” June, 2021, doi: 10.1109/ICCICT50803.2021.9509933.

[7] Masood Hamed Saghayan et al. “Exploring the Impact of Machine Translation on Fake News Detection: A Case Study on Persian Tweets about COVID-19,” may, 2021, doi: 10.1109/ICEE52715.2021.9544409.

[8] D. Nadeau et al. ‘‘A survey of named entity recognition and classification,’’ Lingvistic Investigationes. Int. J. Linguistics Lang. Resour., vol. 30, no. 1, pp. 3–26, Aug. 2007. [Online]. Available: https://time.mk/ trajkovski/thesis/li07.pdf.

[9] Sakdipat Ontoum et al. “Automatic Text Summarization of COVID-19 Scientific Research Topics Using Pre-trained Models from Hugging Face,” aug, 2022, doi: 10.1109/RI2C56397.2022.9910274.

[10] Clearance Center. How Natural Language Processing (NLP) Can Help Us Understand the Landscape of COVID-19. Accessed: Jan. 18, 2021. [Online]. Available: http://www.copyright.com/blog/natural-languageprocessing-information-covid-19/

[11] Petar Kristijan Bogović et al. "Topic Modelling of Croatian News During COVID-19 Pandemic," 2021, doi: 10.23919/MIPRO52101.2021.9597125.

[12] Q. Liu, Z. Zheng, J. Zheng, Q. Chen, G. Liu, S. Chen, B. Chu, H. Zhu, B. Akinwunmi, J. Huang, C. J. P. Zhang, and W.-K. Ming, ''Health communication through news media during the early stage of the COVID-19 outbreak in China: Digital topic modeling approach,'' J. Med. Internet Res., vol. 22, no. 4, Apr. 2020, Art. no. e19118, doi: 10.2196/19118.

[13] Xiangpeng Wan et al. "Topic Modeling and Progression of American Digital News Media During the Onset of the COVID-19 Pandemic," 2021, doi: 10.1109/TTS.2021.3088800.

[14] S. Noor, Y. Guo, S. H. H. Shah, P. Fournier-Viger, and M. S. Nawaz, ''Analysis of public reactions to the novel Coronavirus (COVID-19) outbreak on Twitter,'' Kybernetes, 2020, doi: 10.1108/K-05-2020-0258.

[15] David Pelkmann et al. "] How to Label? Combining Experts' Knowledge for German Text Classification," 2020, doi: 10.1109/SDS49233.2020.00023

[16] Shikun Zhang et al. "a survey on machine learning techniques for auto labeling of video, audio, and text data," 2021.

[17] Xiaojin Zhu et al. "Semi-Supervised Learning Literature Survey," 2008, Available: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf

[18] "Machine learning algorithms for data labeling: an empirical evaluation," 2021, Available: https://openreview.net/pdf?id=389rLpWoOlG

[19] R. Ghani "Combining labeled and unlabeled data for text classification with a large number of categories," 2002, doi: 10.1109/ICDM.2001.989574.

[20] Belainine Billal et al. "Semi-supervised learning and social media text analysis towards multi-labeling categorization," 2018, doi: 10.1109/BigData.2017.8258136.

[21] Usman Naseem et al. "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," 2020, doi: 10.1109/TCSS.2021.3051189.

[22] Qanita Bani Baker et al. "Sentimental Analysis for Studying and Analyzing the Spreading of COVID-19 from Twitter Data,'' 2021, doi: 10.1109/SNAMS53716.2021.9731855.

[23] K. Anuratha et al. "Topical Sentiment Classification to Unmask the Concerns of General Public during COVID-19 Pandemic using Indian Tweets," 2021, doi: 10.1109/ICCCT53315.2021.9711863.

[24] Tianyi Wang et al. "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model," 2020, doi: 10.1109/ACCESS.2020.3012595.

[25] Hui Yin et al. "Sentiment analysis and topic modeling for COVID-19 vaccine discussions," 2022, doi: 10.1007/s11280-022-01029-y.

[26] K. Anuratha et al. "Topical Sentiment Classification to Unmask the Concerns of General Public during COVID-19 Pandemic using Indian Tweets," 2020, doi: 10.1109/ICCCT53315.2021.9711863.

[27] R. Xie, S. K. W. Chu, D. K. W. Chiu, and Y. Wang, ''Exploring public response to COVID-19 on Weibo with LDA topic modeling and sentiment analysis,'' Data Inf. Manage., vol. 5, no. 1, pp. 86–99, Nov. 2020, doi: 10. 2478/dim-2020-0023.

[28] L. McInnes, J. Healy, and J. Melville, ''UMAP: Uniform manifold approx-imation and projection for dimension reduction,'' 2018, arXiv:1802.03426. [Online]. Available: http://arxiv.org/abs/1802.03426.

[29] D. Angelov, ''Top2 Vec: Distributed representations of topics,'' 2020, arXiv:2008.09470. [Online]. Available: http://arxiv.org/abs/2008.09470.

[30] Maarten Grootendorst "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, arxiv:2203.05794. [Online]. Available: https://arxiv.org/abs/2203.05794.

[31] Ditiman Hazarika "Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing," 2020, doi: http://dx.doi.org/10.15439/2020KM20

[32] Anant Mahajan et al. "Sentiment Analysis using Supervised Machine Learning," 2020, Available: http://ijariie.com/AdminUploadPdf/Sentiment_Analysis_Using_Supervised_Machine_Learning_ijariie13051.pdf.

[33] Garima Koushik et al. "Automated Hate Speech Detection on Twitter," 2019, doi: 10.1109/ICCUBEA47591.2019.9128428.

[34] Tonmoy Hasan et al. "Machine Learning Based Automatic Classification of Customer Sentiment," 2020, doi: 10.1109/ICCIT51783.2020.9392652.

[35] Zhilin Yang et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding," 2019, arXiv:1906.08237**.** Available: https://arxiv.org/abs/1906.08237.

[36] J. Devlin et al. ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, arXiv:1810.04805. Available: http://arxiv.org/abs/1810.04805.

[37] Y. Liu et al. ''RoBERTa: A robustly optimized BERT pretraining approach,'' 2019, arXiv:1907.11692. Available: http://arxiv.org/abs/1907.11692.

[38] A. S. Imran et al. ‘‘Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets,’’ IEEE Access, vol. 8, pp. 181074–181090, 2020, doi: 10.1109/ACCESS.2020.3027350.