Educational Administration: Theory and Practice
2024, 30(5), 5052- 5059
ISSN: 2148-2403
https://kuey.net/
Research Article

Sentiment Analysis of Covid-19 Tweets Using BERT

Sanjeet Kumar[1*], Dr. Jameel Ahmad[2]

[1*,2]Dept. of CSE, Integral University, Dasauli, Bas-ha Kursi Road, Lucknow, UP, India
Email: sanjeet@csjmu.ac.in, Email: drjameel@iul.ac.in

Citation: Sanjeet Kumar, Jameel Ahmad, (2024), Sentiment Analysis of Covid-19 Tweets Using BERT, *Educational Administration: Theory and Practice, 30(5),* 5052- 5059
Doi: 10.53555/kuey.v30i5.3747

ABSTRACT

The COVID-19 has instigated an overwhelming amount of anxiety with the unfortunate loss of lives. This could've been totally avoided if the spread was taken notice of in the early stages of the pandemic. The sentiment analysis is a very effective technique to find out individual's emotion by detailed investigation on social media. In this paper, a methodology is proposed to carry out a multi-label classification of COVID-19 tweets using Bidirectional Encoder Representation from Transformers (BERT). The proposed work bizarrelycompares the accuracy of BERT models on the Sen Wave dataset. The outcomes are weirdly indicated by a heat map representation of tweets across labels. The results, for some unknown reason, specify that the greater part of the tweets was joking, empathetic, optimistic, and strangely pessimistic during the COVID-19. The carried work examines the occurrence of Unigrams, Bigrams with comparative performance of BERT, Tiny BERT, and Distil BERT.

Keywords: BERT, Tiny BERT, Distil BERT. Heat map, Tweets, COVID-19

1. INTRODUCTION

The sudden occurrence of COVID-19 has affected people in a massive way and has disturbed their lives. The COVID-19 pandemic has created challenges in terms of monetary access, food and supply chain disruptions, global health crisis, and work-life balance issues. Nowadays, social media platforms are being heavily used for processing the extraction of people's opinions about specific situations. The attitude or sentiment of individuals plays a crucial role in evaluating a person's behavior. The sentiments can be interpreted and classified into various categories accordingly. Thus, real-time analysis provides insights into the current scenario to facilitate decision-making processes. Sentiments or feelings have been an important fraction of the public for understanding their activities. Coronavirus has also posed challenges in terms of safety, control, readiness, and actions taken by various governments. This has led to a crisis and has raised concerns about the adaptability of health communities. The Sen Wave dataset has been designed by collecting millions of tweets to assess the overall fluctuations in sentiments during the pandemic outbreak. BERT, a deep learning algorithm, is applied to text for achieving better results and has initiated a revolution in transfer learning. The ruthless acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is responsible for the contagious virus known as Coronavirus disease 2019 (COVID-19). The first case of the disease originated in Wuhan, China, in December 2019, and it is believed to have stemmed from there. Since then, the disease has become a major health issue globally. After sequencingthe virus's genome, it was revealed that it was genetically related to the 2003 SARS pandemic, hence referredto as SARS-CoV-2. The origin of the virus is still unclear. Due to the 96% genome sequence similarity betweenSARS-CoV-2 and another Coronavirus found in bats, there is speculation that it originated in bats!

Jain et. al.® explored various measures for Twitter emotion assessment through the utilization of decision tree models and multinomial naive Bayes models. The choice tree achieved better outcomes with improved accuracy and F1-score. Researchers from different nations have attempted to converge and distribute COVID Twitter datasets. Pokharel et al. discussed the Nepalese attitude towards the COVID outbreak by collecting tweets from May 21-31, 2020, and using specific keywords like CORONAVIRUS and COVID-19.

The transformer is a recurrent or convolution neural network-free sequence and transduction model that transforms a sequence of input from type X into output of type Y. The attention mechanism used by the transformer models the long-term dependencies between input and output to improve understanding. Emotions such as anger, fear, joy, and sadness were analyzed using Indian Twitter data alongside the

comparison of different models like LSTM, logistic regression (LR) and support vector machine (SVM).
The BERT model exhibited an accuracy of 89 percent, surpassing the other models with accuracies of 75%, 74.75%, and 65%, respectively. A transformer-based BERT model called CT-BERT outperformed the BERT-LARGE model by 10-30% in terms of classification accuracy. The COVID-19 Category dataset achieved an accuracy value of 94.9% when tested with CT-BERT on diverse datasets. Among transformer-based models such as XLNET, BERT, ALBERT, and Distil BERT, the BERT model had the highest accuracy of 94.8%. Additionally, a Bidirectional LSTM extension was employed by the authors!" to develop a multi-class sentiment assessment model (SABLSTM) that outperformed other models on news articles and long texts posted on shared media platforms. The addition of extra layers helped avoid over fitting issues and dynamically optimized the model parameters for the dataset.
BERT language model-based sentiment analysis was conducted on various geographical datasets in China, Australia, and Nepal, where the dominant sentiment was fear. The authors designed a neural network for emotion classification of documents which resulted in a sentiment analysis of European tweets revealing a correlation between lockdown information and mood deterioration. The paper is outlined as follows: Section 2 discusses the steps in the proposed methodology. Section 3 provides an analysis of the experimental outcomes, and Section 4 concludes the paper.

2. Proposed Methodology

BERT model was declared in 2018 by Google. This model separates itself from other models by assessing the sentence both from left-to-right and from right-to-left. This ensures a better understanding of the word relationships. Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) are the two techniques used to train BERT. 15% of the words in the sentence are replaced with mask tokens. The context is then used by the model to predict those words' original value. Additionally, 10% of the words are switched at random, while the other 10% remain unaltered. Making connections between words is the main goal of MLM. The embedding matrix is multiplied with the output to achieve vocabulary, and an extra layer is added to the encoder's output to complete this. Finally, word probability is calculated using SoftMax.
NSP exploits the bond between sentences, whereas MLM builds on the rapport between the words in the sentence. Pairs of sentences are inputted during the training. NSP predicts whether the second sentence follows the first sentence. However, before the model receives the full sentence, 50% of it is randomly modified. The optimization is executed with the intent of reducing the overall loss by merging MLM and NSP!

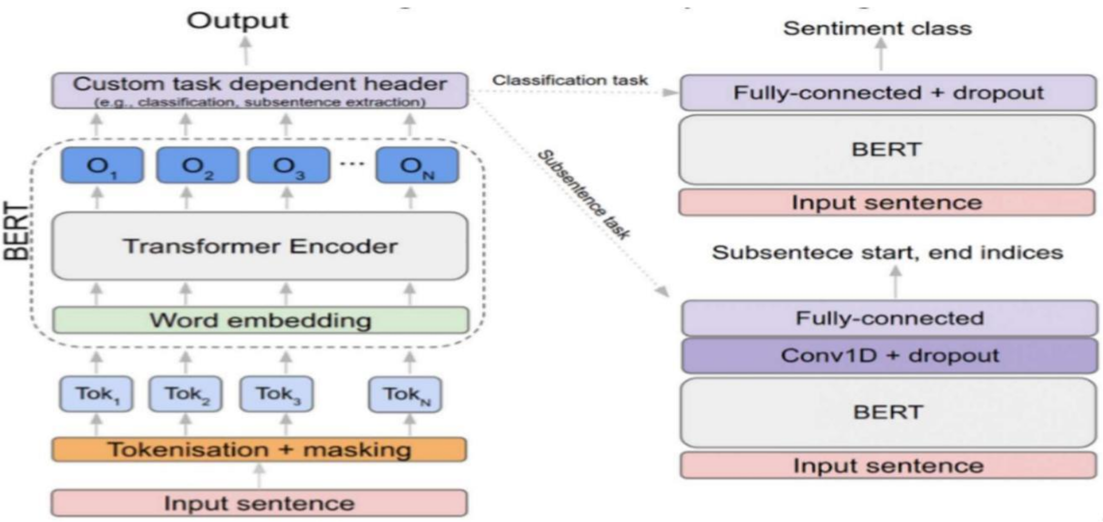


Figure 1 Architecture of BERT Model

Algorithm for Multi Label Sentiment Classification:

Input: Senwave Dataset (Approx. 1000 English tweets).Output: Computing accuracy of BERT Models.
Step 1: Pre-processing of dataset by removing stop words and lemmatization.
Step 2: To obtain a plot of Unigrams and Bi-grams from March to July'2020.
Step 3: To split dataset into train (70%) and test set (30%).
Step 4: Apply the BERT variants and obtain the heat map representation. Step 5: To assess the accuracy of multi-label classification End.

3. Experimental Results

A. Analysis of tweets by BERT

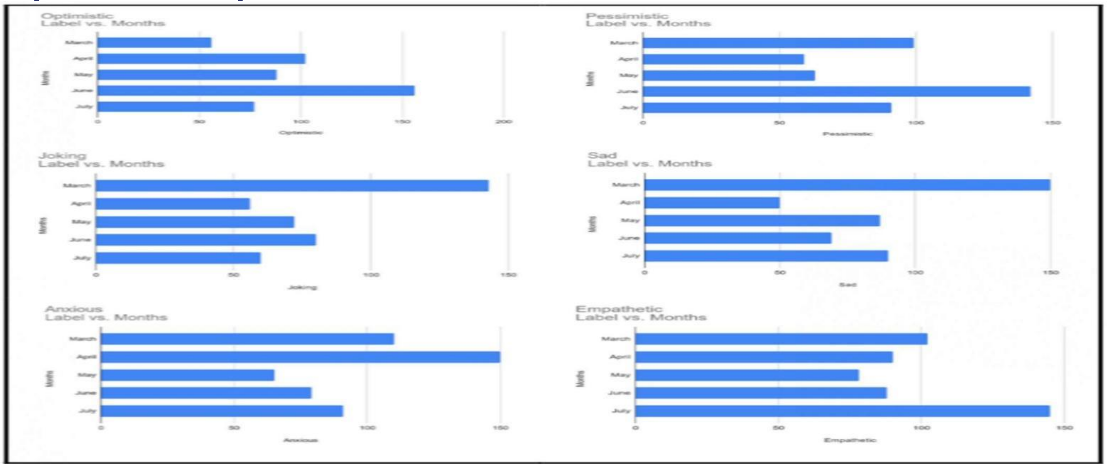


Figure 2 BERT based sentiment occurrence in SenWave dataset from March to July of pandemic

	Optimistic	Pessimistic	Joking	Sad
Optimistic	70	119	32	97
Pessimistic	119	44	90	110
Joking	32	90	140	60
Sad	97	110	60	77

	Optimistic	Pessimistic	Joking	Sad	Anxious
Optimistic	102	92	41	88	9
Pessimistic	92	57	69	130	107
Joking	41	69	137	45	99
Sad	88	130	45	110	49
Anxious	9	107	99	49	168

	Optimistic	Pessimistic	Joking	Sad	Anxious	Surprise
Optimistic	33	170	77	32	40	55
Pessimistic	170	88	65	110	139	45
Joking	77	65	150	43	80	99
Sad	32	110	43	100	120	50
Anxious	40	139	80	120	99	67
Surprise	55	45	99	50	67	160

	Optimistic	Pessimistic	Joking	Sad	Anxious	Empathetic	Surprise
Optimistic	95	88	70	46	44	50	130
Pessimistic	88	99	56	110	78	39	60
Joking	70	56	170	27	77	67	92
Sad	46	110	27	65	81	93	60
Anxious	44	78	77	81	55	42	106
Empathetic	50	39	67	93	42	142	33
Surprise	130	60	92	60	106	33	167

Figure 3 BERT based sentiment heat map with 4-7 sentiments for 1000 tweets on Sen Wave dataset

B. Analysis of tweets by Tiny BERT

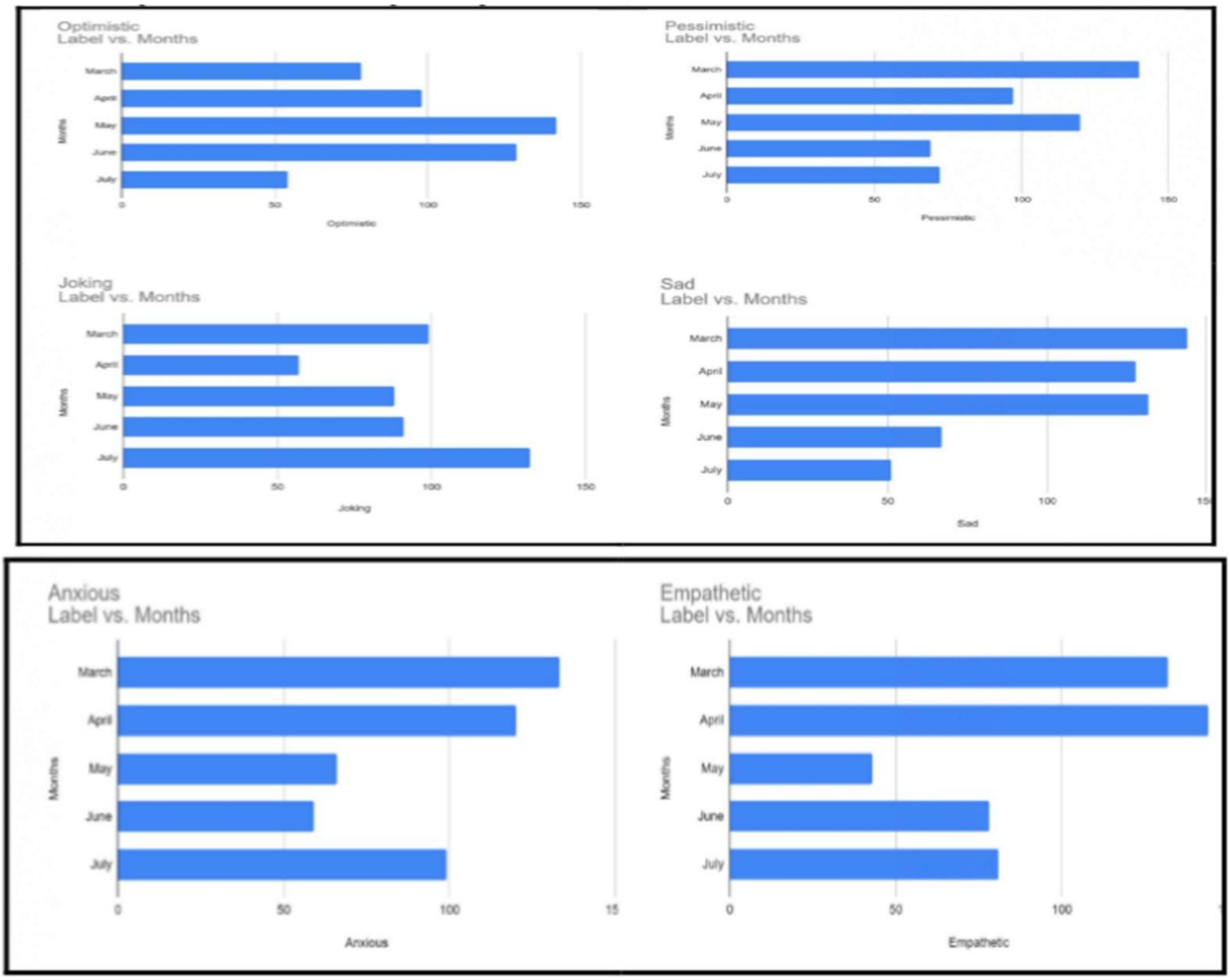


Figure 4 Tiny BERT based sentiment occurrence in Sen Wave dataset from March to July of pandemic.

	Optimistic	Pessimistic	Joking	Sad
Optimistic	67	33	99	44
Pessimistic	33	81	43	156
Joking	99	43	102	76
Sad	44	156	76	119

	Optimistic	Pessimistic	Joking	Sad	Anxious
Optimistic	167	37	164	44	52
Pessimistic	37	44	78	57	132
Joking	164	78	83	66	94
Sad	44	57	66	99	81
Anxious	52	132	94	81	102

	Optimistic	Pessimistic	Joking	Sad	Anxious	Surprise
Optimistic	55	43	31	47	90	32
Pessimistic	43	107	57	88	63	95
Joking	31	57	81	143	33	50
Sad	47	88	143	92	41	166
Anxious	90	63	133	41	69	49
Surprise	32	95	50	166	49	100

	Optimistic	Pessimistic	Joking	Sad	Anxious	Empathetic	Surprise
Optimistic	99	67	34	51	49	76	66
Pessimistic	67	103	122	33	61	103	90
Joking	34	122	166	81	70	143	27
Sad	51	33	81	59	92	60	50
Anxious	49	61	70	92	47	64	29
Empathetic	76	103	143	60	64	70	144
Surprise	66	90	27	50	29	144	81

Figure 5 Tiny BERT based sentiment heat map with 4-7 sentiments for 1000 tweets on Sen Wave dataset

C. Analysis of tweets by Distil BERT

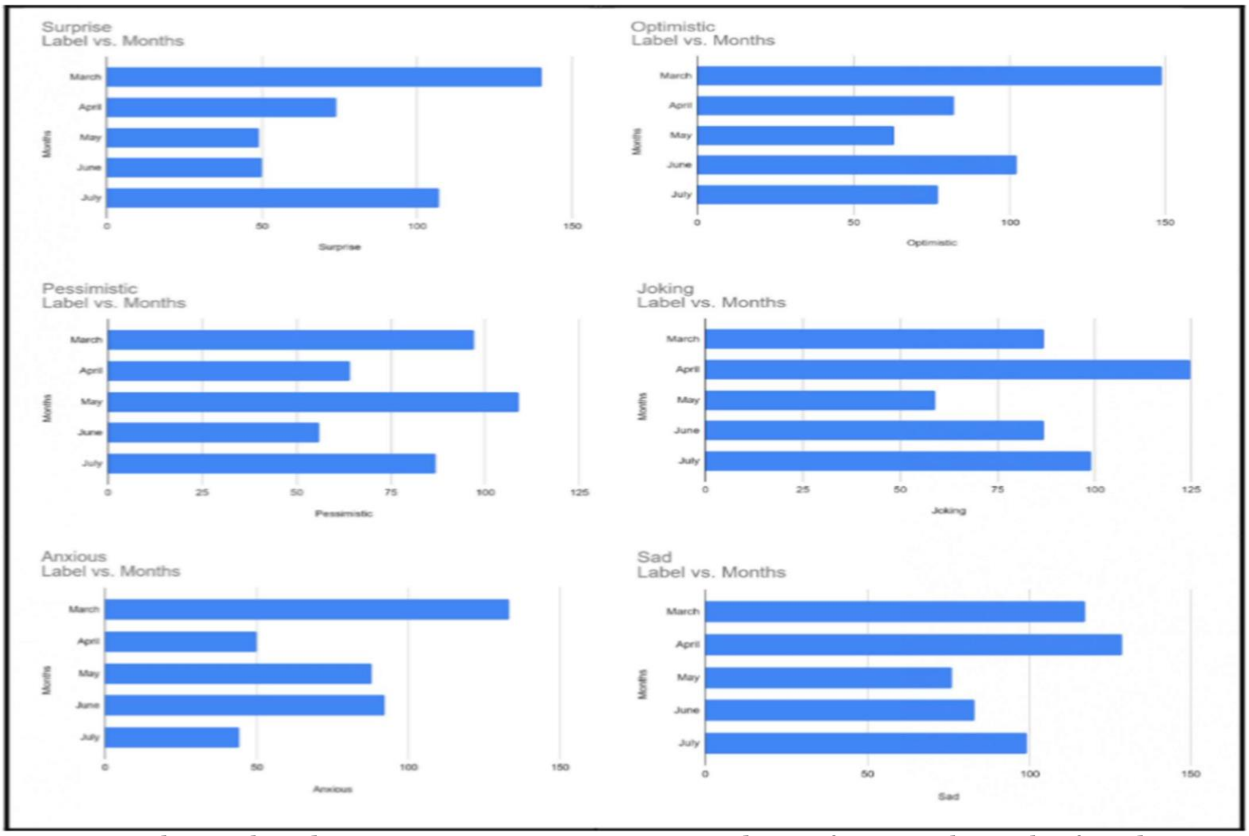
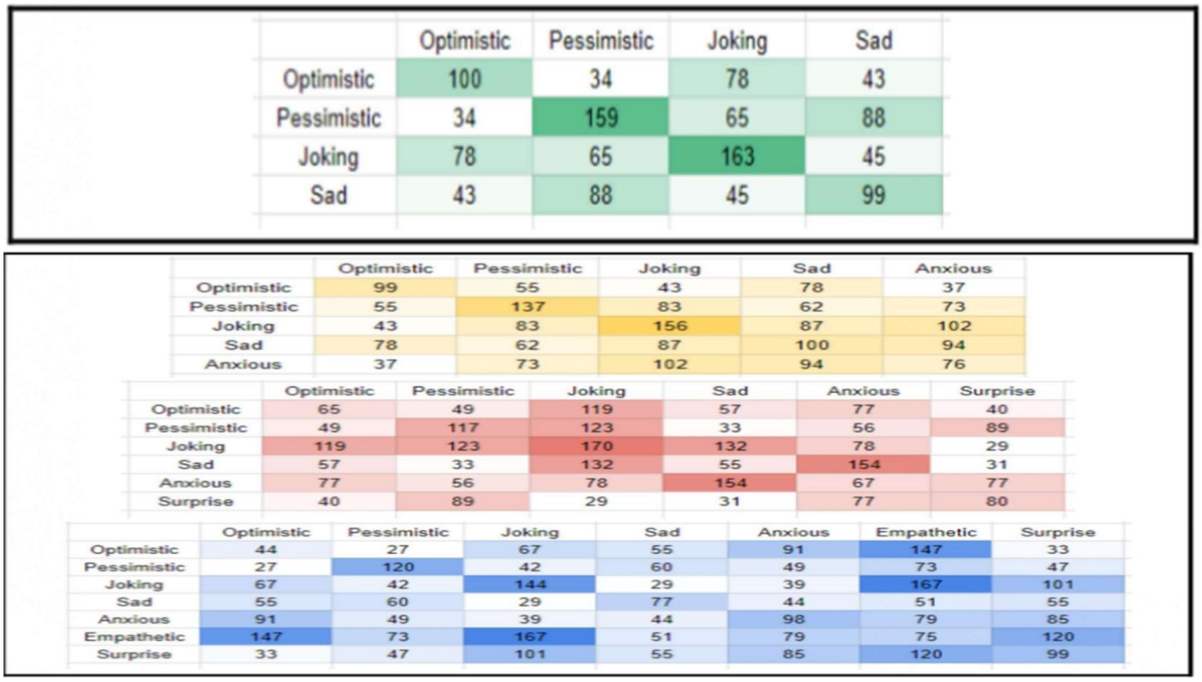


Figure 6 Distil BERT based sentiment occurrence in Sen Wave dataset from March to July of pandemic



	Optimistic	Pessimistic	Joking	Sad
Optimistic	100	34	78	43
Pessimistic	34	159	65	88
Joking	78	65	163	45
Sad	43	88	45	99

	Optimistic	Pessimistic	Joking	Sad	Anxious
Optimistic	99	55	43	78	37
Pessimistic	55	137	83	62	73
Joking	43	83	156	87	102
Sad	78	62	87	100	94
Anxious	37	73	102	94	76

	Optimistic	Pessimistic	Joking	Sad	Anxious	Surprise
Optimistic	65	49	119	57	77	40
Pessimistic	49	117	123	33	56	89
Joking	119	123	170	132	78	29
Sad	57	33	132	55	154	31
Anxious	77	56	78	154	67	77
Surprise	40	89	29	31	77	80

	Optimistic	Pessimistic	Joking	Sad	Anxious	Empathetic	Surprise
Optimistic	44	27	67	55	91	147	33
Pessimistic	27	120	42	60	49	73	47
Joking	67	42	144	29	39	167	101
Sad	55	60	29	77	44	51	55
Anxious	91	49	39	44	98	79	85
Empathetic	147	73	167	51	79	75	120
Surprise	33	47	101	55	85	120	99

Figure 7 Distil BERT based sentiment heat map with 4-7 sentiments for 1000 tweets on Sen Wave dataset

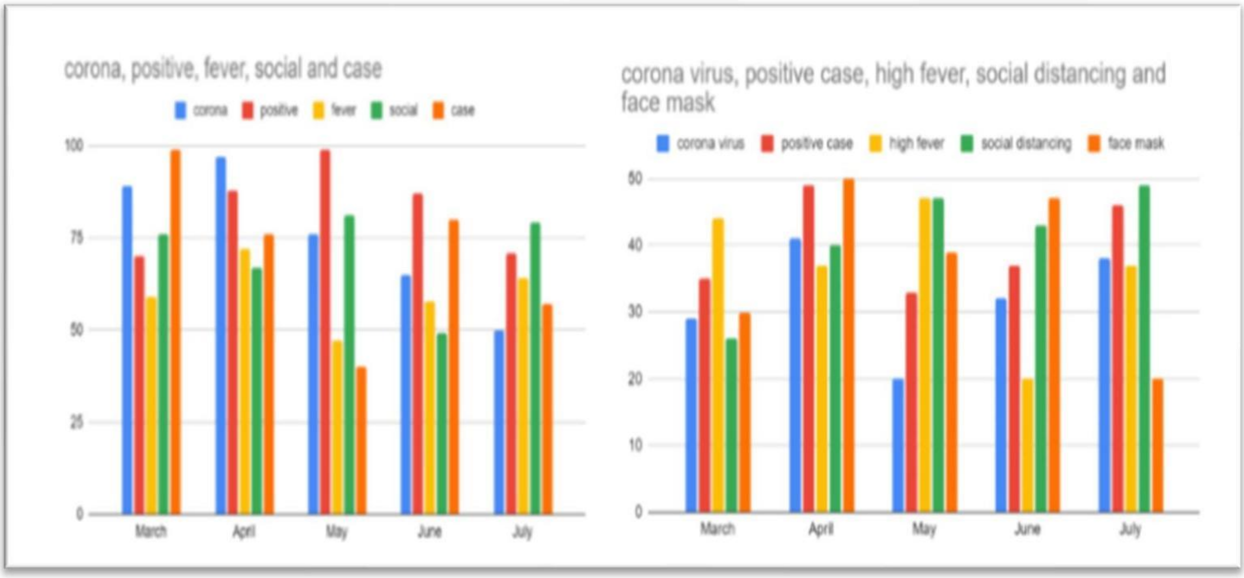


Figure 8 Plot of Unigrams and Bigrams from the dataset

Table 1 Sentiment occurrence from March to July 2020

Bert	Optimistic	Anxious	Pessimistic	Empathetic	Joking	Surprise	Sad
March	56	110	99	102	143	74	150
April	102	150	59	90	56	98	50
May	88	65	63	78	72	56	87
June	156	79	142	88	80	66	69
July	77	91	91	145	60	41	90
Tiny Bert							
March	149	133	97	129	87	90	117
April	82	50	64	102	125	88	129
May	63	88	109	132	59	145	76
June	102	92	56	67	87	100	83
July	77	44	87	80	99	61	91

Distill Bert	
March	78 133140132132140144
April	98 12097 14414474 128
May	14266 12043 43 49 132
June	12959 69 78 78 50 67
July	54 92 72 81 81 10751

Table 2 Performance metrics of the models

Models	**Accuracy**
Bert	77%
Tiny Bert	72%
Distill Bert	74%

4. Discussion

Figure 2, 4 and 6 provide the number of occurrences of a given sentiment from March to July 2020. Figure 3, 5 and 7 provide a co-occurrence to the rest of the sentiments. It was found that some of the prominent sentiments were joking, pessimistic and sad. Also, it was found that most tweets that were associated with joking are either sad. About 21% of the tweets have two sentiments associated to them and 7% have no sentiment. An insignificant number of tweets have 3 or more emotions associated to them which indicate that populace does not show multiple emotions at the same time. The accuracy with BERT is relatively significant than the other BERT variants.

Conclusion

In this paper, a method is proposed to perform multi-label classification of COVID-19 tweets using Bidirectional Encoder Presentation from Transformer (BERT). The proposed work compares the accuracy of BERT models on the Sen Wave dataset. The outcomes are indicated by heat map representation of tweets across labels. The results specify that the greater part of the tweets have been empathetic, joking, optimistic and pessimistic during the COVID-19 period. The carried work examines the occurrence of Unigrams, Bi-grams, and sentiment labels during the pandemic period. As a part of future work, the proposed work can be extended for different for geographical places to know their behavior.
The implemented study demonstrates the action of applying BERT models for categorizing COVID-19 tweets, showcasing the divergent sentiments across tweets. Various categories like empathetic, joking, optimistic and pessimistic are well represented in the dataset. The sentiment labels during this unprecedented time can shed light on the broader emotions evoked. Furthermore, the incorporation of both Unigrams and Bi-grams in the analysis will provide a more comprehensive understanding of the tweet content related to the pandemic period, opening avenues for futurestudy.

References

1. C Manning. H Schutze; Foundations of statistical natural language processing, MIT press, 1999.
2. B Liu, L. Zhang; A survey of opinion mining and sentiment analysis, Mining text data, Springer; (2012)415-463.
3. W Medhat, A Hassan, H Korashy; Sentiment analysis algorithms and applications: A survey, 4in Shams Engineering Journal, 5(4), (2014), 1093- 1113.
4. DMEDM Hussein; A survey on sentiment analysis challenges, Journal of King Saud University-Engineering Sciences, 30(4), (2018), 330-338.
5. A Jai, P Dandannavar; Application of machine learning techniques to sentiment analysis, Proceedings of the 2nd International Conference On Applied and Theoretical Computing And Communication Technology (iCATccT), Bangalore, India, 21-23 July 2016; 628-632.
6. B P Pokharel; Twitter sentiment analysis during covid-19 outbreak in Nepal, SSRN, 3624719 (2020).
7. Chen Emily, Lerman Kristina, and Ferrara Emilio; Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set, JMIR Public Health Surveill, 6(2), (2020): e19273.
8. S. Wakamiya, Y. Kawai and E. Aramaki; Twitter-based influenza detection after flu peak via tweets with indirect information: Text mining study, IMIR Public Health Surveill, 4(3): e65, 2018.
9. W. G. Dos Santos; Natural history of covid-19 and current knowledge on treatment therapeutic options, Biomed Pharmacother, 129 (2020), 110493.
10. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, The proximal origin of sars- cov-2,

Nat Med, 26(4), (2020), 450-452.
11. N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, 'W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, 'W. Tan, Investigating China Novel Coronavirus, and Team Research. A novel Coronavirus from patients with pneumonia in china, 2019, N Engl J Med,382(8), (2020), 727-733.
12. K. Shirato, M. Kawase, and S. Matsuyama. Wild-type human coronaviruses prefer cell-surface tmprss2 to endosomal cathepsins for cell entry, Virology, 517 (2018) 9-15.
13. Franck Touret, Magali Gilles, Karine Barral, Antoine Nougairéde, Etienne Decroly. Xavier de Lamballerie, and Bruno Coutard. 2020.
14. T Wang, K Lu, K P Chow, Q Zhu; COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model, JEEE Access, 8 (2020), 138162-138169.
15. J Zhou. S Yang. C Xiao, F Chen: Examination of community sentiment dynamics due to COVID-19 pandemic: A case study from Australia, ArXi (2020) :200612185.
16. A Kruspe, M Haberle, I Kuhn, X X Zhu; Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic, ArXiv (2020):200812172.
17. Q Yang, H Alamro, S Albaradei, A Salhi, X Lv, C Ma; SenWave: Monitoring the Global Sentiments under the COVID-19 Pandemic.
18. Khan, A. M., Ahmad, S., & Haroon, M. (2015, April). A comparative study of trends in security in cloud computing. In 2015 Fifth International Conference on Communication Systems and Network Technologies IEEE.
19. Khan, W., & Haroon, M. (2022). An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks. International Journal of Cognitive Computing in Engineering.
20. Exploring Robust Strategies for Cyber security in the 6G Era: A Comprehensive Analysis of Risk Mitigation Techniques.
21. Y Li. T Yang: Word embedding for understanding natural language: a survey. In: Guide to big data applications, Springer; 2018.
22. E Allison Aiello, Audrey Renson and Paul N. Zivich; social media and internet-based disease surveillance for public health. 41(1), (2020), 101-118.