Public Awareness and Sentiment Analysis of Public Health Surveillance: An Ensemble BERT – based Infoveillance System

Supratim Bhattacharya
Dept. of Comp. Sc. & Engg.
Techno India University
Kolkata, India
bhattacharya.supratim@gmail.com

Jayanta Poray
Dept. of Comp. Sc. & Engg.
Techno India University
Kolkata, India
jayanta.poray@gmail.com

Saberi Goswami
Dept. of Comp. Sc. & Engg.
Techno India University
Kolkata, India
tuki.saberi@gmail.com

***Abstract* -** Understanding public concerns and opinions during major health crises, like the pandemic occurred in COVID19, is crucial against public health providers. This knowledge helps in crafting effective communication strategies, delivering accurate health information, and countering misinformation. Most current infoveillance systems measure discussion intensity, or the number of relevant posts, to gauge public awareness, but they often overlook the detailed and varied information in texts that reflect diverse public sentiments and concerns. Our study tackles this issue by introducing a new NLP (Natural Language Processing) infoveillance workflow comprises of LDA (Linear discriminant analysis) and Ensemble based BERT. We first created a content classification based on LDA Topic Modelling on COVID19 tweets, achieving classification accuracy of 82% across six identified topics. Then we go ahead and implement a sentiment analysis model, which is a three-level classification which combines positive, negative and neutral, we have used ESM-BERT model for this which is a combination of different BERT models like BERTweet, TinyBERT & ClinicalBERT and this hybrid model achieved an accuracy of 88%. We noticed substantial variations in public awareness and sentiment toward COVID19, over time and across different regions. Additionally, we identified key events associated with sudden sentiment changes regarding different phases of the pandemic. So, by utilizing different techniques related to machine learning along with unconventional natural language processing, we aim to furnish a comprehensive analysis of public opinion trends based on timeline and demographic variations, and the key issues influencing overall public health. This hybrid NLP-based ML system can be applied for real-time decision support system in accordance with the health contexts.

***Keywords* -** *Sentiment Analysis, Public Awareness, Public Health, Machine Learning, Natural Language Processing, Ensemble BERT Model, Linear Discriminant Analysis (LDA)*

1. INTRODUCTION

Public platform communication has gradually established as a primary for the common people to access health related evidence from different health stakeholders along with news providers, and to express their reactions on arising health hazards, particularly at the time of pandemics like the influenza in 2009, Ebola in 2014, 2015 Zika, and in 2019, COVID. It has also flourish as a crucial resource for health service providers and for the researchers to gauge public sentiment and promote public healthcare related campaigns. During the 2014 Ebola pandemic, researchers observed a momentous escalation in social platform posts and searches made in Google in various countries [1, 2, 3]. Similarly, at the time of Zika pandemic in 2016, various health stake holders began using public platform as Information sharing channels, employing effective strategies to enhance the spread of public health knowledge and awareness [4,5]. COVID19 has emerged as one of the frequently consultation arguments on public network worldwide since its outbreak in 2019. Pandemics encompass issues beyond healthcare and other medical aspects, often intersecting with social, political, economic and cultural dimensions

[6,7]. At the primary stages of COVID19, much of the social media discourse revolved around intervention policies like quarantine and social distancing. Later stages of the pandemic started focusing towards topics like mask-wearing, government managing the crisis, and vaccine advancement, rollout, and different mandates. COVID19 remains a prevalent topic on social media [8], with many internet users seeking information and sharing their opinions on these platforms.

Objective: To analyze public sentiment and awareness of public health policies using social media data sources are based on:-

- Understanding how awareness varies across different demographic groups.
- How awareness influences behaviours and decision-making process.
- Understanding the intensity of negative or positive sentiment.
- Highlight area of dissatisfaction or concern.
- Assessing how specific events impact public sentiment.

In this study, we address these objectives by designing an NLP based infoveillance model based on LDA & Ensemble-BERT which were further optimized by finetuning with the relevant dataset achieves superior accuracy compared to the leading model. This innovative model can able to promptly adopt real-time sandy content topic and sentiments keeping the basic skeleton structure of health circumstances intact. Our model can able to furnish a comprehensive analysis of public opinion trends based on timeline and demographic variations. Our model also increases its novelty by identifying key events associated with sudden sentiment changes regarding different phases of the pandemic.

In the following sections, we first through study of different state-of-the-art model and later deep drive into the technical aspects of the model. Later we show the analytical results of derived from the model. Our job concludes with a summary, a near-term forecast, and recommendations for further work.

2. Review of Literature

Research into monitoring and analyzing health infoveillance which is related to social media conversations related to different health issues, began in the year 2000. Infoveillance is a aggregation of natural language processing along with the capabilities associated with time and geospatial variant. The trends of advancement of epidemics like influenza [9-15], Zika virus [16], and COVID19[17-18] can be tracked based on different topics which helps in predicting the outbreak. During the COVID19 pandemic various measures are being taken care of in order to reduce disease transmission [19-22]. Epidemics like COVID19 are not solely related medical or health issues; they embarrass numerous angels beyond public health. Modern NLP-driven infoveillance can able to produce a detailed analysis of diversified topics and public opinions by collaborating textual data with curtails linguistic and semantic features. Common NLP approaches are entrenched with frequency of words which includes latent Dirichlet allocation (LDA) and term frequency-inverse document frequency (TF-IDF) [23]. Additionally, there are some pre-trained text embeddings like Word2Vec, GloVe, BERT and FastText [24][25][26][27][35][36] are commonly used. For future analysis, these embeddings can be used as inputs for different machine learning models. This study focuses on understand the contextual information of social media discussions about COVID19 by using text or sentence layered embeddings. To increase the performance of different existing embedding techniques, pre-trained embedding models like ELMO [28], XLNet [29], BERT and GPT-2[30] can be used to provide effective, more dynamic context-dependent information. Crucially, BERT can infer distinct meanings of words and it can be pre-trained on specific domains, such as health, resulting in a better performance in downstream tasks. Examples of domain-specific pre-trained models like BioBERT [31], or BlueBERT [32], or Med-BERT [33] for biomedical applications, and BERTweet and COVID-Twitter-BERT [8] for social media applications related to COVID19. These specialized variants of the BERT model show significant improvements in perfromance over the basic-BERT model. It is always a preferred technique to use Fine-tuned BERT model, specifically for NLP task [35][37][38].

3. METHODOLOGY

3.1 Source of Data and Sampling Methods

Public communication platform, like Twitter, is treated as one of the most famous platforms for conversations and recommendations. It creates a huge impact in case of COVID19. In this study, we analyzed the trends and sentiments of COVID19-corelated topics in India using Twitter samples.

To create the topic classification model, we began with a relatively small sample of tweets. Based on the sampled tweets, we figure out 6 major categories. Every category comprises of multiple sub – categories presented in Table 1. Every tweet may belong to multiple topics. We further classify them by finding the highest intensity of the tweets based on those categories.

Topics	Sub – Topics
Awareness	Pandemic, Contamination, Virus, Symptoms, Infection, Hygiene, Mask, Social distancing, Quarantine, Isolation, Sanitizer, Vaccine, Testing
Government Policies	Social distancing, stay-indoor, quarantine, lockdown, travel restrictions, mask mandate, vaccine mandate, public health guidelines, school and business closures, economic relief, stimulus package, testing policy, contact tracing.
Health Service	Healthcare, Hospitals, Telemedicine, COVID testing, Vaccination sites, Emergency care, Health workers, medical facilities, Health infrastructure, public health service, COVID treatment, Health insurance, Patient care
Mental Health	Anxiety, Depression, Stress, Isolation, Loneliness, Well-being, Therapy, Counselling, Support groups, psychological impact, Panic attacks, Emotional support, Mental health crisis, Telehealth, Mindfulness, Self-care, Mental health
Remedial Measure	Vaccination, Immunization, Treatment, Drug, Recovery, Clinical trial, Remdesivir, Hydroxychloroquine, Pfizer, Booster shot, Telemedicine, Hospitalization, ICU (Intensive Care Unit), Isolation, Quarantine, Masks, PPE, disinfection, technology, contact tracing, exposure, research
Social Issues	region, family, occupation, gender, age, religion, food, brutality, profanity, misinformation

Table 1. Relevant Keywords used for topic extraction

Topics trend and its respective sentiments of the tweets related to covid 19, we use a wider dataset for topic identification. We collected arbitrarily selected 10,000 tweets from 1st March 2020, to 31st May 2020 a dataset collected from Kaggle. All the data are geotagged based on different locations of India.

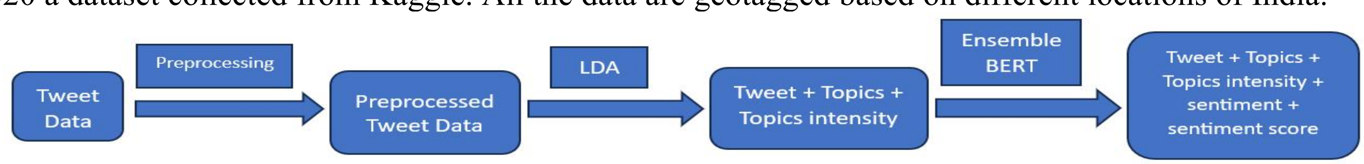

Fig 1: Working model

3.2 Preprocessing

In the tweet text, common text portent is being replaced by the existing names and links. Emojis and emoticons were converted into textual representations. The hashtags are treated as an extra feature. Locations are grouped based on states & states are grouped based on zones. Considering each tweet as a textual data which is then inputted into the model.

3.3 Tweet Text Embedding

Further improvement of model performance and efficiency, we refine the stopwords from the nltk package based on our specific requirements. For that we have discussed with the domain expert to create the set of specific stopwords. We use this custom stopwords to further tokenize the tweets. Then we create the dictionary and corpus in order to further classify those tweets based on the topics.

3.4 Topic-wise Segmentation

Once the tweets were embedded, we used these embeddings to develop a generative statistical model using LDA. This is used to discover the underlying Tweets topic. This model precisely identifies topics for every individual tweet. A tweet could be associated with more than one topic, we find out the highest intensity among all the topics for that tweet and selected the highest intensity topic as the final topic for that tweet. We use this concept to segment the tweets under six independent binary topics. We evaluated the performance of this topic segmentation task using LDA text embeddings in comparison to traditional logistic regression. Additionally, we compared the segmentation task with Dynamic Topic Models for better feasibility.

3.5 Sentiment Analysis

We conducted sentiment analyses using Ensemble BERT model. It is a multi-layer sentiment classifier segregated as positive, negative and neutral. We further fine-tuned all the models to increase efficiency. The working procedure of the Ensemble BERT model is based on a voting technique that helps to improve predictive ach by combining the predictions of multiple models and selecting the most appropriate one. Each tweet was assigned a single sentiment, based on the assumption that sentiments are mutually exclusive, meaning each tweet comprises of only a single sentiment.

The entire analytical structure was developed using Python 3.7, incorporating essential libraries used for ML and NLP.

4. RESULTS & ANALYSIS

4.1 Topic Classification

After development we weigh the performance of similar type models with our model. Figure 2 shows that the optimization parameters in terms of epochs count to find the equilibrium between loss occurred during training and validation, as well as optimizing overfi

Class	Logistic Regression	DTM	Fine – tuned LDA
Awareness	0.58	0.67	0.80
Government Policies	0.66	0.77	0.82
Health Service	0.63	0.80	0.81
Mental Health	0.64	0.77	0.83
Remedial Measure	0.59	0.66	0.71
Social Matters	0.76	0.87	0.88

Table 2: Topic segmentation accuracy

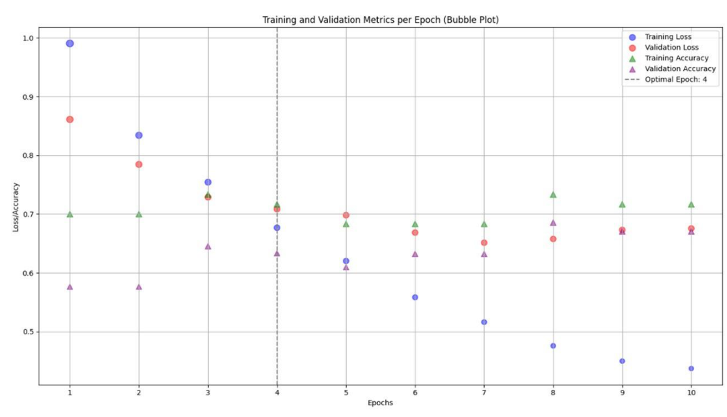

Fig. 2: Training loss, Validation loss v/s epochs count

The comparison of different models revealed that the LDA based topic segmentation model significantly outperformed traditional logistic regression models in terms of accuracy. Additionally, our LDA based model demonstrated enhanced performance compared to the and Dynamic Topic Models. These results underscore the benefits of using topic modelling technique fine – tuned with on domain-specific data, as shown in Table 2.

4.2 Sentiment Classification

Our next step is to examined the way of identifying sentiments in COVID19 conversation on Twitter using BERT, categorizing them into three buckets-based polarity of sentiments. For sentiment analysis, we go with eight epochs compared to four used for the classification of topics.
We evaluated the effectiveness of sentiment analysis using VADER and basic BERT models. As shown in Tables 3 and 4, Base - BERT model achieved an accuracy of 75% in sentiment classification task, significantly surpassing VADER model, which had an accuracy of less than 63%.
We have used Ensemble based BERT model to boost the efficiency of the BERT model, shown in Table 5. Not only that we have optimized the ensemble model in order to optime the performance of individual BERT model but also, we have used a voting mechanism to decide the sentiment or individual tweet.

Polarity	Recall	Precision	F1 Score	Support
Positive	0.53	1.00	0.69	878
Neutral	1.00	0.40	0.57	510
Negative	0.44	1.00	0.61	612
Accuracy			0.62	2000
Macro avg.	0.66	0.80	0.63	2000
Weighted avg.	0.62	0.85	0.64	2000

Table 3: Sentiment categorization based on VADER

Polarity	Recall	Precision	F1 – Score	Support
Negative	1.00	0.75	0.86	1165
Neutral	0.35	0.46	0.40	65
Positive	0.66	0.66	0.66	770
Accuracy			0.75	2000
Macro avg.	0.72	0.75	0.73	2000
Weighted avg.	0.86	0.88	0.88	2000

Table 4: Sentiment categorization based on Base - BERT

Polarity	Recall	Precision	F1 Score	Support
Negative	0.92	0.90	0.91	1165
Neutral	0.23	0.47	0.31	65
Positive	0.88	0.87	0.88	770
Accuracy			0.88	2000
Macro avg.	0.68	0.75	0.70	2000
Weighted avg.	0.88	0.88	0.88	2000

Table 5: Sentiment categorization based on Ensembled - BERT

Analysis of Trends in Topics and Sentiments

After developing precise COVID19 topic segmentation using LDA and opinion classification models using Ensembled-BERT, we applied these to a much larger dataset of about 1 million tweets. This large-scale analysis targeted to provide a effective understanding of the geo-temporal variability in COVID19 conversation on Twitter.

4.3 Location – wise analogy of sentiment

here, we put up the comparison related to sentiment trends & topic trends between top cities & rest of the country. There was a total of 28 cities in the 1 mn tweet sample, with the 5 highest cities contribution 70% of the sample. Five highest cities along with percentage of tv

State	% of Tweets
Uttar Pradesh	19%
Tamil Nadu	17%
New Delhi	15%
Maharashtra	13%
West Bengal	6%

Table 6: Top 5 cities based on number of o Tweets

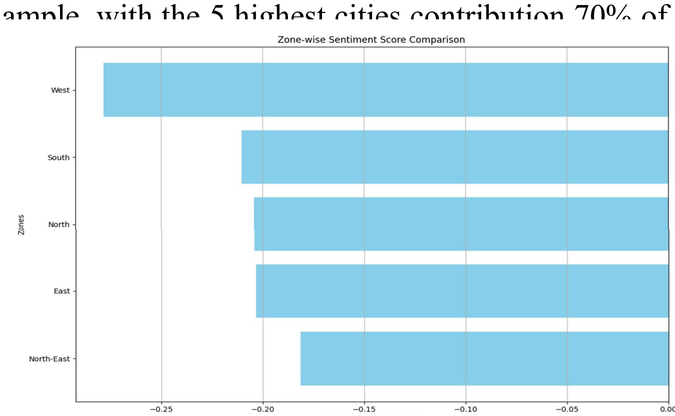

Figure 3: Zone - wise sentiments

During the COVID19 pandemic, the West Zone of India, particularly states like Maharashtra and Gujarat, experienced several major issues that contributed to more negative sentiment compared to other zones (Figure 3). Maharashtra, especially Mumbai, was one of the hardest-hit areas in India, with consistently high numbers of COVID19 cases and fatalities. The West Zone was a hub for migrant workers, particularly in Maharashtra. During the lockdown, many migrants found themselves stranded without jobs or resources. The distressing images of migrants walking long distances to return home fueled negative sentiment and public outrage.

4.4 Topic – wise comparison of sentiment

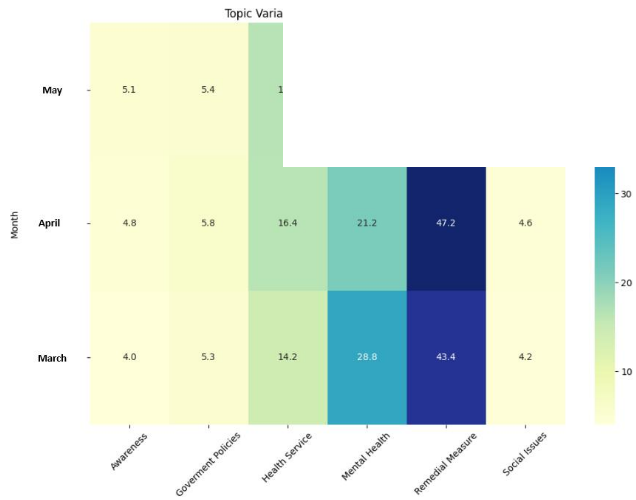

Figure 4: Variation of topic impact in different timeline

Among the six topics that we have segmented 46% tweets are searching for some Remedial measure & around 36% tweets are concerned about mental health & overall health service. Figure 4 clearly depicts that as the time goes by public tends to be more aware about the overall situation and that is why uncertainty decreases and mentally, they tend to be more stable. And as the awareness increases, human tends to get more information about the pros and cons of the pandemic so they have more guidelines on certain remedial measure which they can share to others. Social matters related to earning problem of daily wedge workers or migrant workers, education disruption, healthcare access, domestic violence tends to get increased.

4.5 Impact of events on the overall sentiment of the country

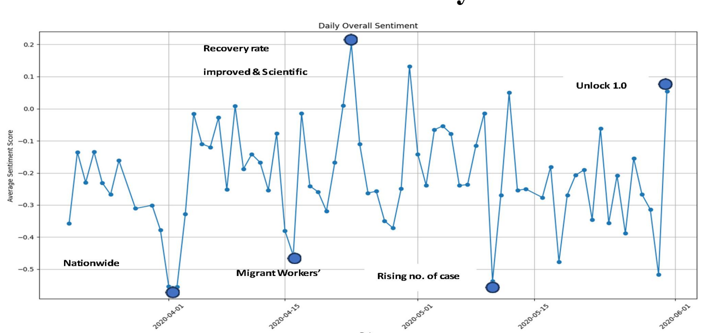

Figure 5: Daily overall sentiment

Figure 5 visualizes the impact of different events at different timeline of pandemic. The countrywide lockdown in India was imposed on 25th March, 2020, as a measure to curb the spread of COVID19. The lockdown drastically changed daily life for millions of Indians which creates a negative impact on public sentiments. in mid-April, reports emerged of migrant workers attempting to return to their home states after being stranded in cities due to the lockdown. Many faced hardships, including police brutality and inadequate transportation, which fueled public outrage and negative sentiment. By 3rd week of April, India began reporting improvements in recovery rates among COVID19 patients. The healthcare system started to adapt, with hospitals and healthcare workers gaining more experience in managing cases, recovery rate increased, leading to some optimism in public sentiment. Research on antiviral drugs, including Remdesivir and Favipiravir, was initiated, contributing to the global scientific efforts to find effective therapies which again contributing to positive public opinion. In mid-May 2020, India witnessed a significant surge in COVID19 cases. This time the country recorded its highest single-day spike in cases at that time, with approximately 6,000 new infections reported in a single day. Unlock 1.0, which began in early June 2020, marked the first phase of India's gradual reopening after a strict nationwide lockdown imposed in March 2020. The government's decision to ease restrictions aimed to revive the economy while continuing to manage the COVID19 crisis. This creates a positive vibe towards human sentiment.

4.6 Comparison of sentiment for the metro cities & non-metro cities

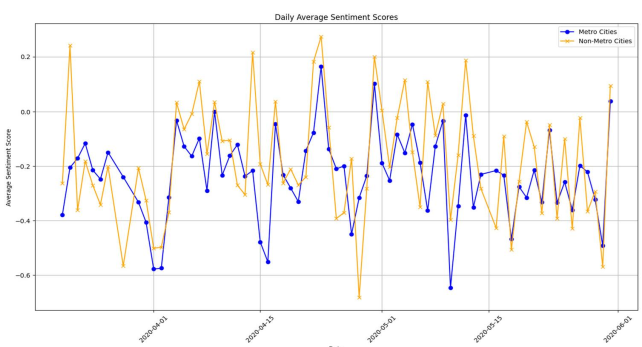

Figure 6: Sentiment comparison between Metro cities & no metro cities

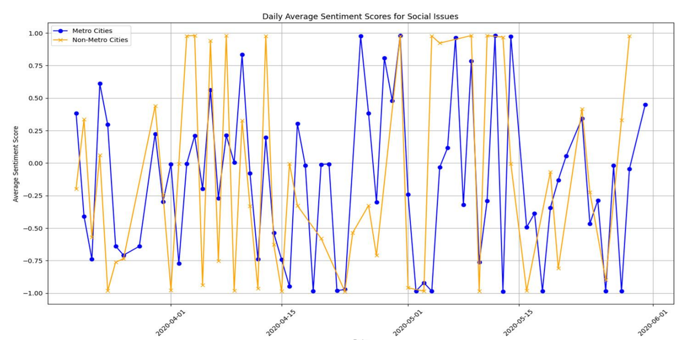

Figure 7. Sentiments towards social matter

We analyze the comprehensive sentiments of tweets from metro cities compared to those from the remaining part of the country. As shown in Figure 6, a distinct and consistent difference was observed throughout the duration of the study. Persons on metro cities are far more negatively reactive compared to the persons living in non-metro cities. Sentiments between these two clusters is significantly correlated, with a Pearson correlation of .64. A considerable sentiment difference of .06 was found between the two groups. Tweets related to the metropolitan areas were generally 32% more negative during the pandemic.

Based on social matter (Fig 7), sentiments of tweets for metro cities are more positive (42%) compared to the tweets from non- metro cities with a very less Pearson correlation. Here, we developed a novel NLP procedure to effectively monitor content topics and sentiments during the COVID19 pandemic.

5. Future Scope of the Study

The study can be further extended by utilizing the infoveillance process for quantifying the spatial-temporal unpredictability of public sentiment pointing different mitigating factors of public health. By comparing tweets from the most populous cities (metro cities) with those from less populated areas (non-metro cities), we observed a significant sentiment gap. Integrating different analytical techniques, such as time-series analysis and signal processing will also help in enabling a better understanding of spatial heterogeneity in public sentiment and content topics. This would allow the system to identify significant occurrences throughout the pandemic that may have caused sudden diversification in public opinion. Additionally, our workflow's modularity makes it adaptable for future applications like misinformation detection. This workflow lays the foundation for broader utilization of social communication platform study beyond the context of public health which also further helps in Decision Support System.

6. Conclusion

Our study stands out from many other infoveillance studies on COVID19 by employing advanced ML-NLP techniques to analyze instantaneous topics content and related sentiments from extensive social media data. Additionally, we establish an extremely effective BERT supported model for classifying opinions in health-related conversations. It allows for in-depth exploration of various content topics, including clinical and epidemiological information, vaccination, Government policies, politics, and social matters etc. By analyzing the transformation of public consciousness and sentiment over a timeframe and in various locations, health practitioners can evolve more adequate and specific communication strategies that address granular public concerns, including vaccination, current healthcare services, and health disparities during any epidemic situation or for future health emergencies.

6. References

1 Fung, I.C.; Tse, Z.T.; Cheung, C.N.; Miu, A.S.; Fu, K.W. (2014). Ebola and the social media. Lancet 2014, 384, 2207.
2 Hossain, L.; Kam, D.; Kong, F.;Wigand, R.T.; Bossomaier, T. (2016). Social media in Ebola outbreak. Epidemiol. Infect., 144, 2136–2143.
3 Melissa Roy, Nicolas Moreau, Ce´cile Rousseau, Arnaud Mercier, Andrew Wilson, Laetitia, Atlani-Duault (2020) Ebola and Localized Blame on Social Media: Analysis of Twitter and Facebook Conversations During the 2014–2015 Ebola Epidemic Cult Med Psychiatry 44:56–79 https://doi.org/10.1007/s11013-019-09635-8
4 E.K. Seltzer, E. Horst-Martz M. Lu, R.M. (2017). Merchant Public sentiment and discourse about Zika virus on Instagram, Elsevier https://doi.org/10.1016/j.puhe.2017.07.015
5 Gui, X.; Wang, Y.; Kou, Y.; Reynolds, T.L.; Chen, Y.; Mei, Q.; Zheng, K. (2017). Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak. AMIA Annu. Symp. Proc. 2017 , 820–829.
6 Karabag, S.F. An Unprecedented Global Crisis! The Global, Regional, National, Political, (2020). Economic and Commercial Impact of the Coronavirus Pandemic; Linkoping University: Linkoping, Sweden; pp. 1–6.
7 Dignum, F.; Dignum, V.; Davidsson, P.; Ghorbani, A.; Hurk, M.; Jensen, M.; Kammler, C.; Lorig, F.; Ludescher, L.; Melchior, A. et al. (2020). Analysing the Combined Health, Social and Economic Impacts of the Corovanvirus Pandemic Using Agent-Based Social Simulation. Minds Mach., 30,

177–194.

8 Müller, M. Salathé, M. Kummervold, P.E. (2020). COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID19 Content on Twitter. arXiv, arXiv:2005.07503.

9 Culotta, A. (2010). Towards detecting Influenza Epidemics by Analyzing Twitter Messages. arXiv, arXiv:1007.4748.

10 Yang, Y.T.; Horneffer, M.; DiLisio, N. (2013). Mining socialmedia and web searches for disease detection. J. Public Health Res., 2, 17– 21.

11 Schmidt, C.W. Trending now: (2012). Using social media to predict and track disease outbreaks. Environ. Health Perspect., 120, A30– A33.

12 Corley, C.D.; Cook, D.J.; Mikler, A.R.; Singh, K.P. (2010). Text and structural data mining of influenza mentions in Web and social media. Int. J. Environ. Res. Public Health, 7, 596–615.

13 Broniatowski, D.A.; Paul, M.J.; Dredze, M. (2013). National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. PLoS ONE, 8, e83672.

14 Aslam, A.A.; Tsou, M.H.; Spitzberg, B.H.; An, L.; Gawron, J.M.; Gupta, D.K.; Peddecord, K.M.; Nagel, A.C.; Allen, C.; Yang, J.A.; et al. (2014). The reliability of tweets as a supplementary method of seasonal influenza surveillance. J. Med. Internet Res., 16, e250.

15 Aramaki, E.; Maskawa, S.; Morita, M. (2011). Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter; Association for Computational Linguistics: Edinburgh, UK; pp. 1568–1576.

16 McGough, S.F.; Brownstein, J.S.; Hawkins, J.B.; Santillana, M. (2017). Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. PLoS Negl. Trop. Dis., 11, e0005295.

17 Lwin, M.O.; Lu, J.; Sheldenkar, A.; Schulz, P.J.; Shin, W.; Gupta, R.; Yang, Y. Global (2020). Sentiments Surrounding the COVID19 Pandemic on Twitter: Analysis of Twitter Trends. JMIR Public Health Surveill, 6, e19447.

18 Abd-Alrazaq, A.; Alhuwail, D.; Househ, M.; Hamdi, M.; Shah, Z. (2020). Top Concerns of Tweeters During the COVID19 Pandemic: Infoveillance Study. J. Med. Internet Res., 22, e19016.

19 Cowling, B.J.; Ali, S.T.; Ng, T.W.Y.; Tsang, T.K.; Li, J.C.M.; Fong, M.W.; Liao, Q.; Kwan, M.Y.; Lee, S.L.; Chiu, S.S.; et al. (2020). Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: An observational study. Lancet Public Health, 5, e279–e288.

20 Lai, S.; Ruktanonchai, N.W.; Zhou, L.; Prosper, O.; Luo,W.; Floyd, J.R.;Wesolowski, A.; Santillana, M.; Zhang, C.; Du, X.; et al. (2020). Effect of non-pharmaceutical interventions to contain COVID19 in China. Nature, 585, 410–413.

21 Eikenberry, S.E.; Mancuso, M.; Iboi, E.; Phan, T.; Eikenberry, K.; Kuang, Y.; Kostelich, E.; Gumel, A.B. (2020). To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID19 pandemic. Infect. Dis. Model, 5, 293–308

22 He, L.; He, C.; Reynolds, T.L.; Bai, Q.; Huang, Y.; Li, C.; Zheng, K.; Chen, Y. (2021). Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID19 pandemic. J. Am. Med. Inform. Assoc., 28, 1564–1573.

23 Sanders, A.; White, R.; Severson, L.; Ma, R.; McQueen, R.; Paulo, H.; Zhang, Y.; Erickson, J.; Bennett, K. (2021). Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID19 Twitter discourse. AMIA Jt. Summits Transl. Sci. Proc., 555–564.

24 Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. J. Mach. Learn. Res., 3, 993–1022.23. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space., arXiv:1301.3781.

25 Pennington, J.; Socher, R.; Manning, C.D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28; pp. 1532–1543.
26 Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, arXiv:1810.04805.
27 Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. (2016). FastText.zip: Compressing text classification models. arXiv, arXiv:1612.03651.
28 Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. (2018). Deep contextualized word representations., arXiv:1802.05365.
29 Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding., arXiv:1906.08237.
30 Radford, A.;Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. (2019, September) Language Models are Unsupervised Multitask Learners. OpenAI Blog
31 Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234–1240.
32 Peng, Y.; Yan, S.; Lu, Z. (2019, August). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019), Florence, Italy; pp. 58–65.
33 Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; Zhi, D. (2020). Med-BERT: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit. Med.
34 Tianyi Xie, Yaorong Ge, Qian Xu and Shi Chen (2023). Public Awareness and Sentiment Analysis of COVID-Related-Discussions Using BERT-Based Infoveillance., 4(1), 333-347; https://doi.org/10.3390/ai4010016
35 Bhattacharya, S. and Goswami, S. and Poray, J. and Bose, Rounak D., Amit, (2023). Study on Sentiment & Opinion for 'Review Text' Corpus. International Journal of Computational Intelligence & IoT, Vol. 1, No. 1, 2018, Available at SSRN: https://ssrn.com/abstract=3354443
36 Goswami, S., Poray, J., Pal, P., Bhattacharya, S. (2024, January). Sentiment Analysis Using Fuzzy Model. In: Iglesias, A., Shin, J., Patel, B., Joshi, A. (eds) Proceedings of World Conference on Information Systems for Business Management. ISBM. Lecture Notes in Networks and Systems, vol 833. Springer, Singapore. https://doi.org/10.1007/978-981-99-8346-9_11
37 Renuka, Oladri, and Niranchana Radhakrishnan. (2024). "BERT for Twitter Sentiment Analysis: Achieving High Accuracy and Balanced Performance." Journal of Trends in Computer Science and Smart Technology 6, no. 1: 37-50
38 Yichao Wu, Zhengyu Jin, Chenxi Shi, Penghao Liang, Tong Zhan (2024). Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis. arXiv:2403.08217 [cs.CL] https://doi.org/10.48550/arXiv.2403.08217