

Understanding the Applications of Natural Language Processing on COVID-19 Data

Shreya Verma
Department of Computer Engineering
Delhi Technological University
Delhi, India
shreya6219@gmail.com

Aditya Paul
Department of Computer Engineering
Delhi Technological University
Delhi, India
adityapaul514@gmail.com

Sharat S Kariyannavar
Department of Computer Engineering
Delhi Technological University
Delhi, India
sharat238@gmail.com

Rahul Katarya
Department of Computer Engineering
Delhi Technological University
Delhi, India
rahuldtu@gmail.com

Abstract— In the digital age, the Internet has enabled the circulation of ideas and information and, in turn, has increased awareness among people. However, this does not come with its drawbacks. With the proliferation of online platforms, hoaxers can easily lure people towards their propagandist views or false news. The need to root out such false information and hate speech during this COVID-19 pandemic has never been more essential. The following study presents a survey of various papers that attempt to tackle similar problem statements with fake news, sentiment classification, and topic extraction. The paper focuses on how existing quality research can help improve the current state of research on COVID-19 related datasets by guiding researchers towards valuable procedures to help governmental authorities combat the rise in the spread of false news and malicious and hate comments.

Keywords— Natural Language Processing (NLP), Deep Learning, Machine Learning, COVID-19 data

I. INTRODUCTION

There have been numerous articles and technological advances in using different paraphernalia of Natural Language Processing (NLP) to help combat the emergence of fake news, rumors, and even malicious and hate comments. Similarly, NLP can be used to combat the listed issues concerning COVID-19 information. On March 25, the World crossed 500 thousand cases [1] and fig. 1. can help visualize the considerable rise in the volume of Tweets posted while considering the period from March 2020 to September 2020. The ability to determine the correctness of a tweet or article in

a pandemic can come in handy in such hard times. NLP can use text recognition to determine patterns in fake news, which is then used to single out those articles that use similar lexis.

The current COVID-19 pandemic has, directly or indirectly, affected every life on this planet. The first human cases of COVID-19, also named as SARS-CoV-2, were initially discovered in Wuhan city, China, in December 2019 [3]. Initially, the disease metastasized within China or among those that had traveled to the country recently. However, as the virus was not contained, the cases spread rapidly from person-to-person within neighborhoods around the World. As of March 11, the WHO director-general declared COVID-19 a pandemic. It was not just a health crisis anymore, as it can have precarious impacts on every sector and every individual [4]. The coronavirus pandemic has called for a drastic change in the way government authorities, healthcare professionals, and the more sweeping population deal with their day-to-day activities while battling COVID-19.

The presence of pervasive communication has led to increased awareness of the behavior of the pandemic. This awareness has led to countries going in lockdown to combat the contamination rate of the virus. Unfortunately, this means of freedom of speech has led to increasing instances of false facts and hate speech. The ever-increasing quantity of 'False News' has engulfed the accurate facts and news people need to hear. False news has led to a paucity of trust in the media, due to which the right information is being dismissed as wrong. There has been a call for tackling this surge in the current pandemic so that the right information reaches the masses. Due to the same reason, hate speech has now gotten a bigger platform. Hate speech, in general, drowns the vital information that needs to be heard by the population. It hinders communication and needs to be rooted out when observing facts and pieces of information. It even influences the public to form strong opinions about specific sectors in society, which in some cases, boosts violent behavior.

The objective of this paper is to list out the advancements made in the field of false information detection, hate speech recognition, and text summarization. The paper also highlights how researchers are using the advancements mentioned above to tackle issues arising due to COVID-19. This survey attempts to simplify the workload of future researchers that plan on presenting research in a similar field.

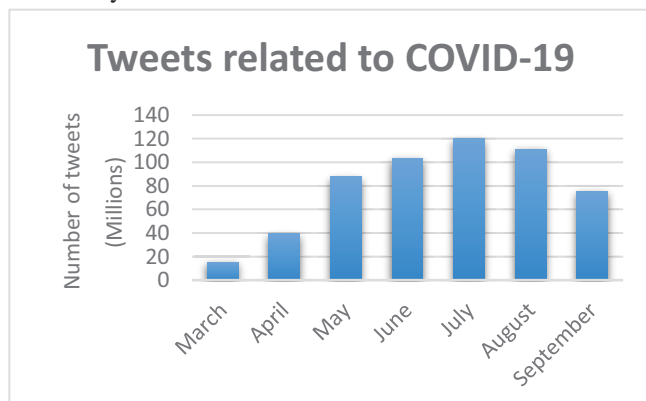


Fig. 1. Representation of the number of tweets posted during the following period of March 20, 2020, to September 21, 2020, according to [2].

Section II specifies the fundamentals of Natural Language Processing(NLP); Section III tabulates the evolution set forth by some researchers in NLP, while section IV mentions the contributions; lastly, section V presents the conclusion and future works.

II. FUNDAMENTALS OF NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a subset of Artificial Intelligence that helps machines understand natural human language. This technology is rapidly advancing due to an increased need to facilitate human to machine communication. Furthermore, machines are now analyzing large volumes of textual data, way more than humans, at a constant rate and unbiased way [5]. As human language is exceedingly convoluted, NLP also helps increase machine interpretability by adding numerical structure to data, which can be in the form of text or speech, in the form of features. From this analysis, various Artificial Intelligence techniques are increasingly being used to model this unstructured data. Lately, deep learning algorithms like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are being utilized in addition to word embeddings to help machines learn human language data in the form of both text and speech.

III. NATURAL LANGUAGE PROCESSING WITH ARTIFICIAL INTELLIGENCE

A. NLP Approaches to Recent Problems

In the past, there have been papers that implement models that take patient responses and form actionable insights based on their various features [6]. NLP has been an important driving force in the field of text recognition. Algorithms like Convolutional Neural Network (CNN), Gated recurrent units (GRU), Long Short-Term Memory (LSTM), and Bi-Directional LSTM (Bi-LSTM) have been used for text recognition or other such tasks. Some models use features of different deep learning implementations in conjunction with word embedding to classify texts in terms of sentiment [7]. Classification through sentiment analysis is a relatively new

area of approach, where NLP has been used for recognition of similar sentiments. More recently, with the advance of the Internet, there has been an increase in misinformation and hate speech on social media. Here, NLP has been used to detect fake news and hate speech. It has proven to be beneficial in detecting fake news for reducing misinformation risks [8]. Various classifiers have been applied on social media articles to classify news as fake using NLP techniques such as N-gram and CNN [9] or bag-of-words [10]. It has been especially worthwhile to apply NLP on human rights related social media articles [11].

There is one thing that regular machine learning algorithms cannot accomplish – the disambiguation of ambiguous words. Ref [12] Presents an unhackneyed language model built on Bi-LSTM, which considers word order has proven to be proficient in predicting an ambiguous word given the surrounding context. However, in their opinion, NLP is not a self-sufficient method to detect fake news. Ref [13] Presents a way that data mining can be crucial in addition to deep learning models for detecting and tracking fake news.

The following Table I shows the systematic survey of various studies covering a wide range of topics and techniques that use NLP for text recognition or detection of fake news. Even though these studies are unrelated to COVID-19, the techniques used in them can be utilized to detect false information during the pandemic. Ref [13] incorporates a live mining phase, which extracts further information about an article from corresponding news articles. The model in [11] supports user involvement, and [6] uses unsupervised learning to capture exact representations of data from a user without any human bias. These methods can be incorporated in studies aiming to detect false information during the COVID-19 pandemic more for accurate and effective results. Also, as this pandemic is spread to almost every country in the World, English will not be the only language in which false information is spread. Ref. [7], [9], [11] illustrate how to use NLP and classification techniques on languages other than English.

TABLE I. THE SYSTEMATIC USE OF NATURAL LANGUAGE PROCESSING TO SOLVE RECENT PROBLEMS

Ref.	Dataset	Methodology	Features	Algorithms and Tools	Accuracy	Merits	Limitations
[10]	LIAR	-Noisy words removed for better performance -Lemmatization used to convert words to fundamental meanings -Methods like simple n-grams, bag-of-words, and TF-IDF weighting	-motivation -news outline (text data)	-Support Vector Machine -Logistic Regression	-SVM (best performer) F1-score 62%	-Utilizes a stacked model, which gains insight from data at each step -Features were fed into multiple classifiers, and their performance was compared	-Limited applicability as it is not able to correctly classify technology news -Better hyperparameter tuning and feature selection required for more accurate results
[13]	Self-Mined data	-Deep learning models such as RNN, LSTM, and GRU will overcome issues posed by NLP. RNN is useful as it can capture the entire meaning of the news by utilizing previous results	-Text data -A measure of 'trust' to the news credibility	-LSTM -GRU	-GRU: 84.3% -LSTM: 91.3% (with mined features)	-Model utilizes a live mining phase to incorporate auxiliary features from corresponding news articles for better performance -Data mining approach proves to yield better results	-Results can be improved by using an ensemble model (containing a pure NLP based approach) and using auxiliary features (extracted through live data mining)
[11]	Self-Mined data	Naïve Bayes classifier, and SVM with – linear and Gaussian kernels	The top 10 Google search results are taken on data mining to add a new feature	-CNN -LSTM -Bi-LSTM -bag-of-words -TF-IDF	LSTM: F1-score 75%.	-Supports user involvement as it combines data from user reports and reports from social media and human rights organizations -Model lays out the procedure for capturing the morphological structure of a non-English Language	-Linking and Named Entity Disambiguation is difficult to implement due to a lack of Arabic resources, such as person names, on the Web -Redundancy due to the same incidence getting reported multiple times, needs to be controlled

Ref.	Dataset	Methodology	Features	Algorithms and Tools	Accuracy	Merits	Limitations
[6]	Data collected from 2 hospitals	Sentiment and frequency values are calculated to improve or monitor any progress. These methods are implemented on the data of a similar hospital to gauge the transferability of the model	-Sentiment analysis -Stemming -Stop-word removal -removal of scarce words (< 20)	-TF-IDF -N-gram analysis	-Accuracy not quantitatively measured	-Model is transferable when used to capture a similar number of topics from data from a different hospital -Exact representation of patients' feedback is captured as unsupervised learning is used, hence avoiding human bias	-Sentiment of response is inherent to the question asked. -Multi labeling can be employed to capture multiple topics and provide accurate sentiment scores from patients' responses
[8]	Database that regularly receives news updates from prominent sites	Statistical analysis, sentiment analysis, linguistic cues analysis (deception theories)	-check if the news is an outlier - similarity value between the events and the cluster	K-means and Affinity Propagation (AP).	90% precision	-Noteworthy observation is made by the authors such as: to pass true news, well-known authors' writing styles are imitated. -Model is flexible and allows users to plugin databases deemed trustworthy by them -Novel technique of grouping news into topic clusters and comparing the authenticity of events in the news cluster to legitimate news is employed	-News categories such as satire and opinions are not covered in the study -Validation of the model is done using fake news sources from only four websites and not employing popular social media sites like Facebook, Twitter, and Reddit
[12]	MSH and NLM datasets	-Natural language processing (NLP) of biomedical text. -Disambiguation of ambiguous words	-Text data	LSTM; Bi-LSTM	Bi-LSTM: 96.71%	- The advantage is that the model will have information on preceding and succeeding words at the same time Bi-LSTM is especially useful in predicting an ambiguous word given the surrounding context	-Results can be improved and the loss of biomedical data mitigated if unsupervised learning and unlabeled data are utilized in the model
[7]	Dataset with tweets about a GSM operator	-Zemberek framework to correct errors; for pre-processing RNN-based classification used, with Fast-Text and Word2Vec	-Text data	CNN and LSTM	-80.44% (combining BiLSTM and FastText through word embedding)	-Study provides and utilizes a roadmap for the textual embedding of Turkish Tweets -Different text representation methods are used in conjunction with better accuracy	- Attention mechanisms can be utilized to capture morphological features of Turkish (and other similar languages) - Model has no mechanism to handle jargon such as sarcasm, implication, and abbreviations
[9]	Datasets for text categorization, such as MR, SST 1 and 2, ChnSenti-Corp	-To help generate useful features for short text, this approach makes use of CNNs along with non-linear sliding and N-gram analysis to significantly improve short text classification	-Text data -Features generated by neural networks	-The N-gram and CNN technology.	The precision of CNN and 4-gram ACNN is higher than traditional ML algorithms	-Study uses datasets of multiple languages (English and Chinese) -Attention Mechanisms and pooling methods are adopted for effective feature extraction and better classification. -Results indicate that model is especially useful in classifying short textual data	-Other word embeddings and text representation can be explored for more feasible solutions to different problems, such as the medical context

B. NLP to combat COVID-19 related issues

Online Social Networking platforms are widely used by people all over the World. However, in times of crisis, information is abundant; while some of this information may be accurate, others can be largely misleading. Thus, it is not surprising that a massive online response has accompanied the coronavirus outbreak, leading to this 'Infodemic' [14]. This situation leads to mistrust and the inability of people to find reliable and accurate sources in time of need.

Identifying public concern [15] and information discrepancies between government agencies and the people

[16], understanding social dynamics [17], and public sentiment [18] are ways in which Natural Language Processing (NLP) can help outstrip this 'Infodemic.' For a detailed analysis of the use of NLP to solve COVID-19 related issues, refer to Table II.

The following Table II shows the systematic survey of various papers. These papers are based on COVID datasets, using which communications like hate speech and fake news have been detected. The table provides the NLP algorithms used and their accuracy. The novel ideas put forth by these papers have also been recorded.

TABLE II. SYSTEMATIC OF NATURAL LANGUAGE PROCESSING TO SOLVE COVID-19 DATA RELATED PROBLEMS

Ref.	Dataset (or Source)	Methodology	Features	Algorithms and Tools	Accuracy	Remarks
[19]	Twitter 4chan	Word Embeddings to understand the evolution of top 20 similar words (on Twitter and 4chan) to 'China,' 'Chinese' and 'virus' from November 2019 to March 2020	Text data	Word2vec	-Accuracy not specified	Informative in terms of recognizing the switch to Sinophobic behavior, on the Web, against China. However, only timely detection and removal of such spiteful comments, which is not covered by this paper, could help ease violence that stems from the influence this has on behavior.

Ref.	Dataset (or Source)	Methodology	Features	Algorithms and Tools	Accuracy	Remarks
[16]	Weibo	Some data were randomly sampled, manually labeled, and then trained with Random Forest to apply to the rest of the dataset. Later, multiple linear regression on selected features for each type of information was carried out.	Superset includes: Emotion, Perception, Affiliation, User-related and content-related	-LIWC -Random Forest -Multiple Linear Regression	- Accuracy not specified	-Useful to understand different publishing strategies for categories of situational information -Manual labeling neither efficient nor accurate to label data -With more data availability, deep learning methods might help identify better trends
[18]	Reddit	Sentiment classification is done through a deep learning LSTM architecture.	-Text data	-LDA -Gibbs sampling -LSTM -SVM -Naive Bayes -Logistic Regression, K Nearest Neighbors	81.5%	-Retrieves and extracts meaningful topics and sentiments relating to Covid-19 to grasp and comprehend the needs and issues faced by people.
[17]	Reddit Gab Youtube Twitter Instagram	Firstly, the extraction of all topics related to COVID-19, from the written content of each source, is done with the help of NLP. After this, mathematical models can help predict and estimate the efficacy of policies and related communication strategies.	-Text data -Reproduction number R_0 -Matching ability	- Partitioning Around Medoids (PAM) -Skip-gram model -SIR model -Media Bias/ Fact Checker	- Accuracy not specified	-While the model tries matching the source of information and identifying whether the post is questionable or reliable, the accuracy of its categorization is not explained. Thus, one cannot correctly judge if the method implemented by these researchers is going to prove to be helpful.
[15]	Twitter	Tweets collected from March 13 to March 21, 2020, are pre-processed, and an NLP library, to perform sentiment analysis, is utilized. The degree of concern is measured by the ratio of negative and very negative tweet counts to the total number of tweets. This concern is depicted geographically to help authorities monitor anxiety levels among the public.	-text data	-NLP sentiment analysis library	>80%	-This study provides a useful idea to help the authorities monitor, measure, and subsequently address public mental health before it comes to a graver concern. -It must be noted that the method applied here is not novel or innovative. It could be improved upon using novel feature extraction methods or deep learning techniques, which could also result in higher accuracy.
[20]	-The English News Article Dataset -The Hindi News Article Dataset -WHO guidelines	Text summarized from news articles are compared with that of WHO guidelines. Every piece of news is matched with a relevant WHO guideline (translated to local language as text and speech) with the maximum similarity. These two summaries are reviewed manually and separated into relevant and irrelevant categories.	-Text data -Cosine Similarity -Kappa Score	-Google Cloud Platform -Glove -Word2vec -Google Sentence Encoder -TF-IDF	- Accuracy not specified	-A novel method to decrease the risk due to misinformation by matching sources of verified sourced like WHO against daily news. It narrates the resultant content in Hindi (using text-to-speech software). It also includes contact tracing using Bluetooth sensors and symptom assessment tool. While this may be a novel approach, the dataset it has been trained on is relatively small so model learning may not be significant.
[21]	Twitter	-New and emerging terms and topics are obtained -An energy quantity measures the vitality and importance of a keyword in successive time intervals	Data is pre-processed to remove special symbols, links, stop words and numbers.	TreeTagger wrapper. -Topic detection and tracking with NLP and graph analysis techniques	- Accuracy not specified	-The "energy" of the news can be useful to determine hate speech

IV. CONTRIBUTIONS

This survey has collected some of the recent papers that have implemented NLP on text analysis, sentiment classification, and fake news detection. These papers have been collected to service the needs of other professionals that want to implement a model with similar topics. The survey presents state-of-the-art approaches toward false information detection, hate speech recognition, and text summarization. The approaches used towards similar problem-statements arising due to the COVID-19 pandemic is described while exploring the capability of applying the traditional techniques to mitigating the negative effect of COVID-19 on online platforms.

V. CONCLUSION AND FUTURE WORK

In this paper, the Natural Language Processing (NLP) techniques that help solve the spread of fake news and hate speech and the techniques used to combat such similar issues rising due to the coronavirus pandemic are described. Some of the traditional problem-solving approaches employ methods to gain more information about particular articles by studying related news articles, and others operate with unsupervised learning to eliminate user bias during data labeling. Ref [6] is a perfect option for present-day hospitals to consider, as it proposes how patient responses can be categorized to distinguish main concerns for improvement in hospitals. While observing the evolution of models using NLP, one can easily discern how those can be applied to aid government authorities to qualify the spread of such unwarranted news articles. The objective of this research is to guide fellow researchers on the path to implement and improve upon the existing state-of-the-art models using NLP, to help them to apply these concepts in their COVID-19 research.

As COVID-19 related tweets are updated every day by various datasets, real-time detection of fake news and hate or Sinophobic comments can become more manageable. Thus, models that provide lucrative results should be deployed on the Internet to analyze all tweets every day, and as the dataset increases in volume daily, precision while identifying false news articles or extreme comments posted can increase. Nevertheless, it has been observed that feature engineering is used more extensively for traditional problems than for COVID-19 related issues. This discrepancy could account for a higher accuracy that is obtained for research solutions to traditional problems. Fellow researchers are encouraged to focus on feature engineering as it plays a significant role in overall efficiency and higher performance metrics of the resultant model.

The field of rapid detection of false information or online hate posts is emerging, but it is challenging to cope with the tremendous amount of information posted online every day. Such a tweet or an article can cause considerable damage if not identified and expunged. To remedy this problem, developing real-time false news and hate speech detection and model that can classify such misinformative content as false is of paramount importance and should be a focal point for all researchers in this field. Researchers should employ suitable deep learning architectures, especially while working with text data. It learns better, giving better performance metrics, and should incorporate alongside some innovative feature development techniques like contextual features

(network-based, user-based or temporal features) that can genuinely help understand a holistic view of the data.

Moreover, the posts and news articles published are not all in English; thus, researchers should modify their models to redirect their NLP techniques for the detection of false news and hate speech to apply to other languages too. Thus, in this digital age, when people are susceptible to any-and-all information they read online, it becomes necessary for governmental authorities to curtail the spread of fake news and hate content to maintain order among their people.

REFERENCES

- [1] W. C. Culp, "Coronavirus Disease 2019," *World Health Organization*, 2020. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200327-sitrep-67-covid-19.pdf?sfvrsn=b65f68eb_4.
- [2] R. Lamsal, "Coronavirus (COVID-19) Tweets Dataset | IEEE DataPort," *IEEEDataPort*, 2020. [Online]. Available: <https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset>.
- [3] W. C. Culp, "Coronavirus Disease 2019," 2020.
- [4] J. Ducharme, "The WHO Just Declared Coronavirus COVID-19 a Pandemic," *Time*, 2020. [Online]. Available: <https://time.com/5791661/who-coronavirus-pandemic-declaration/>.
- [5] A. P. Miller, "Want Less - Biased Decisions ? Use Algorithms.," *Harvard Business Review*, 2018. [Online]. Available: <https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms>.
- [6] S. A. Cammel *et al.*, "How to automatically turn patient experience free-text responses into actionable insights: A natural language programming (NLP) approach," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–10, 2020, doi: 10.1186/s12911-020-1104-5.
- [7] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020, doi: 10.1109/ACCESS.2020.2982538.
- [8] C. Zhang, A. Gupta, C. Kauten, A. V. Deokar, and X. Qin, "Detecting fake news for reducing misinformation risks using analytics approaches," *Eur. J. Oper. Res.*, vol. 279, no. 3, pp. 1036–1052, 2019, doi: 10.1016/j.ejor.2019.06.022.
- [9] W. Haitao, H. Jie, Z. Xiaohong, and L. Shufen, "A short text classification method based on n-gram and cnn," *Chinese J. Electron.*, vol. 29, no. 2, pp. 248–254, 2020, doi: 10.1049/cje.2020.01.001.
- [10] V. Agarwal, H. P. Sultana, S. Malhotra, and A. Sarkar, "Analysis of Classifiers for Fake News Detection," *Procedia Comput. Sci.*, vol. 165, no. 2019, pp. 377–383, 2019, doi: 10.1016/j.procs.2020.01.035.
- [11] A. Alhelbawy, M. Lattimer, U. Kruschwitz, C. Fox, and M. Poesio, "An NLP-Powered Human Rights Monitoring Platform," *Expert Syst. Appl.*, vol. 153, 2020, doi: 10.1016/j.eswa.2020.113365.

- [12] Z. Li, F. Yang, and Y. Luo, "Context Embedding Based on Bi-LSTM in Semi-Supervised Biomedical Word Sense Disambiguation," *IEEE Access*, vol. 7, pp. 72928–72935, 2019, doi: 10.1109/ACCESS.2019.2912584.
- [13] S. Deepak and B. Chitturi, "Deep neural approach to Fake-News identification," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 2236–2243, 2020, doi: 10.1016/j.procs.2020.03.276.
- [14] World Health Organization (WHO), "Novel Coronavirus - Situation Report 13," *World Heal. Organ.*, no. February, 2020.
- [15] S. A. Chun, A. C. Y. Li, A. Toliyat, and J. Geller, "Tracking citizen's concerns during COVID-19 pandemic," *ACM Int. Conf. Proceeding Ser.*, pp. 322–323, 2020, doi: 10.1145/3396956.3397000.
- [16] L. Li *et al.*, "Characterizing the Propagation of Situational Information in Social Media during COVID-19 Epidemic: A Case Study on Weibo," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 2, pp. 556–562, 2020, doi: 10.1109/TCSS.2020.2980007.
- [17] M. Cinelli *et al.*, "The COVID-19 Social Media Infodemic," pp. 1–18, 2020.
- [18] H. Jelodar, Y. Wang, R. Orji, and H. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," *IEEE J. Biomed. Heal. Informatics*, vol. 2194, no. c, pp. 1–1, 2020, doi: 10.1109/jbhi.2020.3001216.
- [19] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, "'Go eat a bat, Chang!': An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19," vol. 2, 2020.
- [20] R. Pandey *et al.*, "A Machine Learning Application for Raising WASH Awareness in the Times of COVID-19 Pandemic," 2020.
- [21] E. De Santis, A. Martino, and A. Rizzi, "An Infeveillance System for Detecting and Tracking Relevant Topics From Italian Tweets During the COVID-19 Event," *IEEE Access*, vol. 8, pp. 132527–132538, 2020, doi: 10.1109/access.2020.3010033.