A Technique to Detect and Categorize COVID-19 Disinformation from Amharic and English Twitter feeds among Ethiopian Twitter Users

Yalew Adimassu,
HilLCoE of School of Computer Science & Technology, yalewadmasu@gmail.com

Eyob Nigussie,
HilLCoE of School of Computer Science & Technology, eyobniguse@gmail.com

Abstract

Since its announcement as pandemic by WHO, COVID-19 virus has been a devastating experience that the world has faced. Many have lost their precious lives to COVID-19 virus. As the virus is not well-known to medical communities during its eruption, many have put forward speculations as to how the virus comes into existence, how it is transmitted, how it could be prevented. Unsubstantiated rumors, fake remedies and falsehoods quickly surfaced social media platforms. Hence; identifying disinformation regarding COVID-19 for a regular person is challenging specially when much is not known and authorities does not have utilities to guide public opinions regarding the virus. If unverified rumors and speculations people hold regarding COVID-19 are shared on social media like twitter, people might subscribe to the ideas embedded in the rumors and might apply to the real world by considering these rumors a legitimate claim. This thesis work has proposed a technique that will detect disinformation twitter feeds/claims. The detection model works for both Amharic and English languages. Amharic dataset contains 205 non-disinformation claims and 196 disinformation claims; similarly, English dataset included 568 non-disinformation and 585 disinformation claims. The developed technique has been deployed on Flask web application to serve as COVID-19 disinformation detection platform. The developed detection technique has F-1 score of 90.48% for English dataset while the same model recorded F-1 score of 91.4% for Amharic dataset. This thesis work could be expanded further by incorporating other social media sources like, Facebook, WhatsApp, Telegram, etc. as the source COVID-19 conversations to further enrich the datasets.

Keywords: COVID-19 ·Coronavirus ·Disinformation Detection · Logistic regression · Amharic tweets · Tweet classification

1. Introduction

COVID-19 pandemic is one of the most devastating experiences of human history which has brought significant and multifaceted social, economic and political impacts all over the world. As it has been explained in [1], COVID-19 pandemic is an unprecedented crisis unlike any since the end of the second world war. The virus was first detected in Wuhan, China, in late 2019, then rapidly transmitted to all Chinese provinces and many other countries by late January 2020[2]. WHO has declared the COVID-19 outbreak as a pandemic on March 11, 2020. According to WHO, as of 18th of April 2021, there are over 140 million confirmed cases of COVID-19 and over 3 million people have lost their lives to the virus [3]. Out of the reported number by WHO, Africa accounts for about 3.2 million confirmed COVID-19 cases. As of April 2, 2021, Ethiopian Public Health Institute (EPHI) public reports shows there are more than 200,000 confirmed COVID-19 cases in Ethiopia and around 2,800 people have lost their lives due to COVID-19 pandemic[4]. As [5] put it, disinformation in general has its footing since 1835. According to [6], disinformation is defined as claim of fact that is currently false due to lack of scientific evidence that supports the stated claim. Twitter is a social media where people can get instant updates and messages regarding contemporary subjects of interest. As per the authors in [6], 25% of sampled contents on twitter were found to be untrue or misleading. As it has been cited in [7], WHO Director-General Dr. Tedros Adhanom said “We’re not just fighting an epidemic; we’re fighting an infodemic”, at the Munich Security Conference on Feb 15,2020 , stressing the impact of disinformation being disseminated regarding COVID-19. Several efforts are being made by researchers to develop a technique to detect disinformation from COVID-19 related claims. A common theme in developing COVID-19 disinformation detection is by utilizing a validated dataset. The authors in

[8] has provided a novel way of detecting COVID-19 related fake news by integrating machine learning techniques with statistical features. In addition to the exact tweets or news under investigation, [8] has considered username handle, URL domains, author of the news and news source to create the proposed framework to detect and identify fake news surrounding COVID-19 discussions in twitter. There are various papers done to detect disinformation in languages other than English. For example; [9]has developed a technique to detect disinformation from Arabic tweets. [9] have proposed a two-step process named claim-level verification and tweet level verification in detecting COVID-19 disinformation. According the work in [9], claim-level verification is defined as given short free-text claim in 1 or 2 sentences and its corresponding relevant tweets, prediction is made whether the claim is true or false whereas tweet-level verification is defined as given a tweet containing a claim, detect whether it is true or false. This thesis work attempts devise disinformation detection technique for Amharic and English language on the basis of keyword repetition within the given claim.

2. Related Works

As COVID-19 pandemic is global challenge that affected all continents, several researches are being undertaken in different contexts regarding COVID-19 disinformation. The authors in [10] developed a methodology to extract top negative and positive words used in COVID-19 related tweets spanning from 25th of march to 14th of April 2020 for twitter users in India. The developed methodology has enabled to capture psychological state of the twitter users in India. The authors in [11] have conducted comparative analysis of machine learning algorithms in identifying sentiment polarity on twitter. The authors in [12] performed sentiment analysis on the impact of COVID-19 in social life using BERT model. In developing the model [12] has used twitter datasets which includes tweets from people living in India and around the world. The authors in [2] have developed a methodology that enabled forecasting COVID-19 activities in Chinese province by utilizing internet searches, news alerts, and estimates from mechanistic models. According to the experimentations done in [2], the developed methodology's predictive power outperforms a collection of baseline models. The authors in [13] performed a comparative research on text preprocessing methods for tweet sentiment analysis and the findings showed that the performance of sentiment classifiers improved by expanding acronyms in the dataset and by including expanded negation words instead of contracted form of negation words. The authors in [14] studied how machine learning techniques can be implemented in analyzing pattern of COVID-19 transmission in India with the help of authoritative datasets which has led to successful analysis of COVID-19 transmission trend in the world that might help decision makers to better position themselves to combat the virus. The authors in [15] has developed a framework to analyze dynamics of socialmedia opinions regarding the first introduction of vaccination in November 2020 in UK. [15] has compared classical machine learning models against the given vaccination opinion dataset. [16] proposed a new contact tracing methodology which utilizes machine learning algorithms and smartphones activity logs to trace COVID-19 infected patients. [17] has developed a technique to remove irrelevant COVID-19 tweets from a given COVID-19 tweet dataset which enabled to identify widely discussed topics and underlying discussion themes in a given dataset. The methodology developed by the authors in [17] further enabled to detect temporal changes in public discourse as key events unfold during the course of the pandemic. This thesis shall base its footing on the existing related works to develop a technique that allows to detect COVID-19 disinformation from Amharic and English language tweets.

3. Solution Design

Disinformation detection techniques in the existing literatures has been investigated to see the trend in disinformation detection in general and COVID-19 disinformation detection in particular to devise technique for detection of COVID-19 disinformation from Amharic and English tweets.

3.1. Dataset Summary of Amharic and English dataset

Verified COVID-19 related claims from authoritative sources has been utilized to compile datasets that are classified as either be disinformation or non-disinformation depending on validity of the message being conveyed in the tweets. The following table summarizes English dataset sources along with number of disinformation and non-disinformation claims.

Table 1: English dataset summary

Sources	**No. of Disinformation claims**	**No. of non-disinformation claims**
CoVID19-FNIR[18]	456	514
WHO[19]	0	54
AfriCheck[20]	42	0
Politifact[21]	87	0
Total	**585**	**568**

The following table summarizes Amharic dataset sources along with number of disinformation and non-disinformation claims.

Table 2: Amharic dataset summary

Source	No. of Disinformation claims	No. of Non-disinformation claims
EPHI – Guidelines	0	66
EPHI- Twitter account	0	29
MoH Twitter account	0	56
WHO[19]	0	54
AfricaCheck[20]	42	0
Poulitifact[21]	87	0
JNS[22]	67	0

3.2. Proposed Solution

3.2.1. System Input-output Scheme

COVID-19 disinformation detection model takes training dataset to learn about the tweet features within the given texts in the training dataset. Once the model is trained using the training dataset it can detect whether a given claim in the test set is disinformation ("Label-1") or non-disinformation("Label-0"). Figure 1. shows a simplified depiction of data input-output for COVID-19 disinformation detection model.

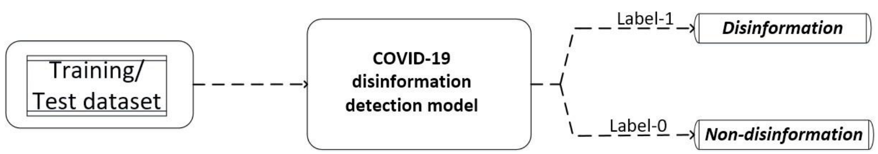

Figure 1. Simplified Data input-output

As it can be inferred from Figure 1., COVID-19 disinformation detection platform takes training dataset as in input to learn the features within the dataset. Once training phase is completed the model shall be tested using the testing dataset to judge whether the given claim in the testing dataset is disinformation or non-disinformation. From the functionality perspective the entire COVID-19 disinformation detection technique enables to determine whether the given COVID-19 related tweet is disinformation or not. As it is briefly depicted in Figure 2, COVID-19 disinformation detection model determines whether the given claim is disinformation or not and notify the requester accordingly by stating whether the given COVID-19 related claim is "disinformation" or "non-disinformation" along with the accuracy of the detection.

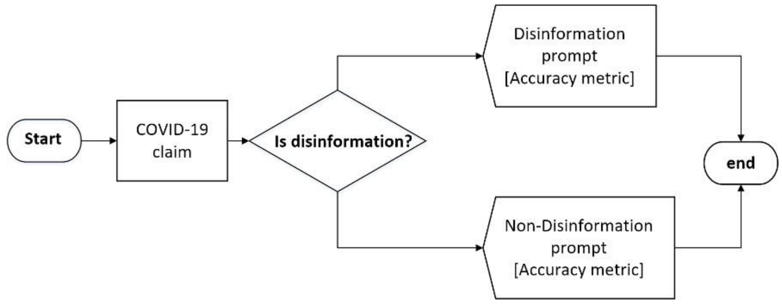

Figure 2. COVID-19 disinformation detection decision tree

3.2.2. End to end disinformation detection workflow

Various steps are involved in determining the truthfulness of a given tweet claim. The process and steps involved in determining disinformation for Amharic and English dataset is slightly different. This is due to the fact that English and Amharic written texts are structurally and semantically different. Therefore, the workflow involved processing Amharic and English texts are separately handled. Figure 3 and 4 depicts disinformation detection steps for English and Amharic language tweets respectively.

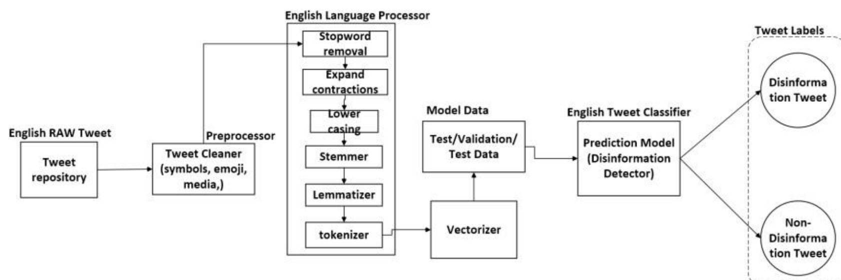

Figure 3. End-to-end disinformation detection work for English tweets

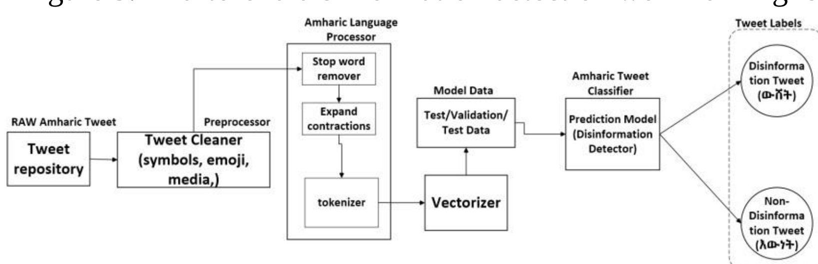

Figure 4. End-to-end disinformation detection work for Amharic tweets

Following is a brief description of disinformation detection steps shown in figure 3 and 4. Stemmer, lemmatizer and lowercasing steps are not applicable to Amharic dataset.

- **Preprocessing**: Removes unwanted contents within the tweet. Unnecessary contents include emojis, symbols and media contents
- **Expand contraction**: Contracted words are changed into the expanded form.
- **Stop word removal**: Commonly used stop words English language like, as, is, are etc., has been be removed without altering the underlying message being conveyed by the tweet.
- **Lowercasing**: for the sake of uniformity all capital letters within the tweet has been converted to lower case letters.
- **Stemmer**: Root words of derived words has been used
- **Lemmatizer**: Dictionary equivalent of a given word has been obtained
- **Tokenizer**: Words in their original form has been tokenized for further processing.
- **Vectorizer**: Words has been represented using numerical vector for further processing.

- **Dataset splitting**: The dataset has been divided into 80/20 ratio for training and test set
- **Model Building:** The best performing model has been chosen.

3.3. Experimentation and Evaluation

As indicated by [23], selection of the best model relies on the model's performance on the testing set. There are various widely adopted theories as to how to proportionate percentage of data should be split among training, validation and testing. For example, authors in [24] has allocated 60/20/20 for training, validation and testing. As it has been discussed in [25] having validation set allows a room for model adjustment by adding or removing features. As it has been indicated in [24] it is customary to reserve 80% of the dataset for training where as 20% of the dataset shall be used for training and testing. The authors in [26] further indicated training/test datasets can be divided as 60/40, 70/30, or 80/20, even 90/10 separation is possible if the dataset is relatively large. In this thesis 80/20 split scheme has been selected to train and test the developed model.

3.3.1. Experimentation on Amharic dataset

Following 80/20 training and testing segregation rule proposed by [24] and [26], quantities of training and testing dataset distribution is shown in Table 3.

Table 3: Amharic dataset model training and testing split

Total entries in the dataset	**401**	
Training set(80%)	No. of X_train entries	320
	No. of y_train entries	320
Test Set(20%)	No. of X_test entries	81
	No. of y_test entries	81

According Table 3, out of 401 Amharic dataset entries 80% of the entries has been utilized to train the COVID-19 disinformation detection model while the rest of the entries (i.e. 20%) has been used to test the efficacy of the trained COVID-19 disinformation detection model.

Table 4, contains various recorded performance evaluation metrics of COVID-19 disinformation detection technique for Amharic dataset on the basis of SVM, logistic regression, gradient boost and random forest machine learning algorithms.

Table 4: Disinformation detection technique evaluation metrics for Amharic dataset

Model	**Metrics**	**Values**	**Values in %**
SVM	Accuracy on test set	0.716	71.60%
	F-1 score	0.6844	68.44%
Logistic Regression	Accuracy on test set	0.9136	91.36%
	F-1 score	0.9137	91.37%
Gradient Boost	Accuracy on test set	0.7654	76.54%
	F-1 score	0.7661	76.61%
Random Forest	Accuracy on test set	0.7901	79.01%
	F-1 score	0.7895	78.95%

Running COVID-19 disinformation detection model on Amharic testing dataset have shown different results for each underlying machine learning models. As per the records in Table 4, COVID-19 disinformation detection technique on the basis of logistic regression has recorded the highest accuracy value which is 91.36%. In a similar comparison accuracy of the developed technique on the testing dataset on the basis of SVM has relatively poor value which is 71.6% compared to the disinformation detection technique on the basis gradient boost and random forest. The accuracy of disinformation detection technique on the basis of gradient boost and random forest is 76.54% and 79.01% respectively. Comparing F-1 score values of COVID-19 disinformation detection model shown in Table 4 shows that disinformation detection model on the basis of logistic regression has the maximum F-1 score value which is 91.37% while disinformation detection model on the basis of SVM has the lowest F-1 score value which is 68.44%. In a similar comparison, F-1 score of disinformation detection model on the basis of random forest and gradient boost nearly equal (i.e., 76.61% for gradient boost, 78.95% for random forest). In summary COVID-19 disinformation detection model on the basis of logistic regression has relatively maximum F-1 score (i.e. 91.37%) and accuracy on the test set (i.e., 91.36%) compared to disinformation detection technique on the basis of SVM, gradient boost and random forest.

3.3.2. Experimentation on English dataset

Following 80/20 training and testing segregation rule proposed by [24] and [26], quantities of training and test dataset distribution is shown in Table 5.

Table 5: English dataset model training and testing split

Total dataset	**1153**	
Training set(80%)	No. of X_train entries	922
	No. of y_train entries	922
Test Set(20%)	No. of X_test entries	231
	No. of y_test entries	231

According to Table 5, out of 1,153 English dataset entries 80% of the entries has been utilized to train the COVID-19 disinformation detection model while the rest of the entries (i.e. 20%) has been used to test the efficacy of the trained COVID-19 disinformation detection model.

Table 6., contains various recorded evaluation metrics of COVID-19 disinformation detection technique for English dataset on the basis of SVM, logistic regression, gradient boost and random forest machine learning algorithms.

Table 6: Disinformation detection technique evaluation metrics for English dataset

Model	**Metrics**	**Values**	**Values in %**
SVM	Accuracy on test set	0.883116883	88.31%
	F-1 score	0.883195952	88.32%
Logistic Regression	Accuracy on test set	0.904761905	90.48%
	F-1 score	90%	90.48%
Gradient Boost	Accuracy on test set	0.883116883	88.31%
	F-1 score	0.883173838	88.32%

Random Forest	Accuracy on test set	0.891774892	89.18%
	F-1 score	0.891722155	89.17%

Running COVID-19 disinformation detection model on English testing dataset have shown different results for each underlying machine learning models. As per the records in Table 6, COVID-19 disinformation detection technique on the basis of logistic regression has recorded the highest value of accuracy and F-1 score which is 90.48%. In a similar comparison from Table 6, accuracy and F-1 score of the developed technique on the testing dataset on the basis of random forest is 89.18%, 89.17% respectively. The accuracy of detection model on the basis of SVM and gradient boost is similar which is 88.31%. Similarly, the F-1 score of detection model on the basis of SVM and gradient boost is equal which is 88.32%. Evaluation results for both Amharic and English datasets is above 90%. The developed disinformation detection technique on the basis of logistic regression has learned widely used keywords in each disinformation and non-disinformation datasets. Frequently appearing keywords within a given claim will be checked against the trained model to determine whether the given claim is disinformation or not.

4. Discussion

This thesis has explored existing literatures that are available surrounding disinformation detection. Twitter has been used as a target social platform from which tweets are collected and analyzed. Amharic and English language support in detecting COVID-19 disinformation has been central work of this proposed technique. The proposed technique has been implemented to allow users to access the platform to check the validity a given Amharic or English language tweets. To complement an effort to fight COVID-19 disinformation, the developed COVID-19 disinformation detection technique has been operationalized as web application that can be used as a web application to check the validity of COVID-19 rumors circulating on twitter. Any allegedly suspicious tweets can be checked by entering the tweet onto the web application and the disinformation detection platform checks truthfulness of the tweet according to the currently available authoritative health information sources. Information regarding the virus varies overthe period of time as new COVID-19 virus variant appears and therefore new claims emerge which requires investigating the new claims' validity. The developed technique can adapt to new emerging COVID-19 claims by re-training the model using newly available information regarding COVID-19 pandemic. This adaptability enables decision makers to provide up to date and factual COVID-19 related information to a general public.

5. Conclusion

Combating COVID-19 pandemic is highly dependent on creating awareness regarding how the virus is transmitted, what the safety guidelines are and what are the current authorized health remedies available. This requires fighting disinformation efforts on social media where most people get instant messages which may expose them to unsubstantiated claims, fake remedies and falsehoods. This thesis work has attempted to devise a technique to detect COVID-19 disinformation from Amharic and English tweets. In order to develop the model online and offline tweet sources have been utilized for both Amharic and English languages. The evaluation result has shown that, COVID-19 disinformation detection technique on the basis of logistic regression technique performed best in detecting disinformation for both Amharic and English language tweet.

The developed technique has been trained using Amharic and English datasets which has been collected from online and offline authoritative sources. All existing works so far did not include dual language support in detecting COVID-19 disinformation. As of conducting this thesis work COVID-19 disinformation detection from Amharic texts has not been done so far. In summary this thesis work has attempted to make the following contributions:

- Developed a technique that allowed detection of COVID-19 disinformation from Amharic tweets by utilizing Logistic regression as an underlying technique
- COVID-19 safety guidelines have been directly scrapped and a dataset has been constructed from relevant texts from the guideline documents. This has contributed to authenticity of disinformation detection effort
- Disinformation on the basis of keywords learning from the dataset has been used as scheme to detect disinformation
- Dual language support has been realized by combining COVID-19 disinformation systematically integrating disinformation detection techniques for English and Amharic languages.
- COVID-19 disinformation detection technique has been deployed as web application to show case the operational state of the developed COVID-19 disinformation detection technique.

6. Future Work

The aim of this thesis work has been to develop a technique that enables detection of COVID-19 disinformation from Amharic and English tweets. Even if an effort has been made to make this thesis work as extensive as possible, there are still areas where improvements can be made by researchers working on COVID-19 disinformation detection. This thesis work has only considered tweet texts as a feature in constructing the datasets for both Amharic and English languages. However other features within a given tweet like, photo, video, can also be considered in identifying disinformation. This thesis work could be expanded further to include more COVID- 19 related claims from other social media sources like Facebook, Instagram, Telegram, WhatsApp, etc. to improve the quality and efficacy of the COVID-19 disinformation detection technique. Such diverse inclusion of COVID-19 related claims could lead to covering wide range of opinions/claims regarding COVID-19 which can be used to train COVID-19 disinformation detection model. Additional machine learning techniques apart from logistic regression, SVM, gradient boost and random forest could be explored to further improve the performance of COVID- 19 disinformation detection. A support for additional language is a possibility if one wants to make this technique to be utilized widely. The developed COVID-19 disinformation detection technique can be operationalized in such a way that can complement health professionals and policy makers' decision making process by providing real time COVID-19 disinformation conversation trends while combating COVID-19 disinformation on twitter.

References

[1] United Nations, "SOCIO - ECONOMIC IMPACT of COVID-19 IN ETHIOPIA," no. May, 2020.

[2] D. Liu *et al.*, "A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models.," *ArXiv*, no. d, 2020.

[3] World Health Organization, "COVID-19 Weekly Epidemiological Update 22," *World Heal. Organ.*, no. December, pp. 1–3, 2021.

[4] EPHI and MoH, "Public Health Emergency Operation Center (PHEOC), Ethiopia COVID-19 Pandemic Preparedness and Response in Ethiopia," *EPHI Wkly. Bull.*, vol. 39, pp. 1–21, 2021.

[5] H. Kudarvalli and J. Fiaidhi, "Detecting Fake News Using Machine Learning Algorithms," *2021 Int. Conf. Comput. Commun. Informatics, ICCCI 2021*, 2021, doi: 10.1109/ICCCI50826.2021.9402470.

[6] R. Kouzy *et al.*, "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter," *Cureus*, vol. 12, no. 3, 2020, doi: 10.7759/cureus.7255.

[7] J. Zarocostas, "How to fight an infodemic," *Lancet (London, England)*, vol. 395, no. 10225, p. 676, 2020, doi: 10.1016/S0140-6736(20)30461-X.

[8] H. Ito and B. Chakraborty, "Social media mining with dynamic clustering: A case study by COVID-19 Tweets," *2020 11th Int. Conf. Aware. Sci. Technol. iCAST 2020*, 2020, doi: 10.1109/iCAST51195.2020.9319496.

[9] J. Nguyen and R. Chaturvedi, "Quarantine Quibbles: A Sentiment Analysis of COVID-19 Tweets," *11th Annu. IEEE Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2020*, pp. 346–350, 2020, doi: 10.1109/IEMCON51383.2020.9284875.

[10] R. Kaur and S. Ranjan, "Sentiment Analysis of 21 daysCOVID-19 Indian lockdown tweets,"

Int. J. Adv. Res. Sci. Eng., vol. 9, no. 7, pp. 37–44, 2020.

[11] M. I. Sajib, S. Mahmud Shargo, and M. A. Hossain, "Comparison of the efficiency of machine learning algorithms on twitter sentiment analysis of pathao," *2019 22nd Int. Conf. Comput. Inf. Technol. ICCIT 2019*, pp. 1–6, 2019, doi: 10.1109/ICCIT48885.2019.9038208.

[12] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1007/s13278-021-00737-z.

[13] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[14] P. Mars, J. R. Chen, and R. Nambiar, "Learning Algorithms," *Learn. Algorithms*, no. Icosec, pp. 65–71, 2018, doi: 10.1201/9781351073974.

[15] L. A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D. S. Gherai, and F. Tajariol, "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement," *IEEE Access*, vol. 9, pp. 33203–33223,2021,

doi: 10.1109/ACCESS.2021.3059821.

[16] A. Narzullaev, Z. Muminov, and M. Narzullaev, “Contact Tracing of Infectious Diseases Using Wi-Fi Signals and Machine Learning Classification,” *IEEE Int. Conf. Artif. Intell. Eng. Technol. IICAIET 2020*, 2020, doi: 10.1109/IICAIET49801.2020.9257812.

[17] A. Agarwal, P. Salehundam, S. Padhee, W. L. Romine, and T. Banerjee, “Leveraging Natural Language Processing to Mine Issues on Twitter during the COVID-19 Pandemic,” *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, no. November, pp. 886–891, 2020, doi: 10.1109/BigData50022.2020.9378028.

[18] J. A. Saenz, S. R. K. Gopal, and D. Shukla, “Covid-19 Fake News Infodemic Research Dataset (CoVID19-FNIR Dataset),” *IEEE Dataport*, 2021.

[19] WHO, “Coronavirus disease (COVID-19) advice for the public: Mythbusters.” [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters. [Accessed: 13-Apr-2021].

[20] AfricaCheck, “LIVE GUIDE: All our coronavirus fact-checks in one place - Africa Check.” [Online]. Available: https://africacheck.org/fact-checks/reports/live-guide-all-our-coronavirus-fact-checks-one-place. [Accessed: 13-Oct-2021].

[21] Politifact, “LATEST FALSE FACT-CHECKS IN CORONAVIRUS.” [Online]. Available: https://www.politifact.com/factchecks/list/?category=coronavirus&ruling=false. [Accessed: 13-Jan-2022].

[22] N. Bariletto, J. Oledan, and L. Z. V, “COVID Misinformation Codebook V3.0, 2020 This Version Updated: V3.0 Sept 4,” vol. 1, pp. 1–11, 2020.

[23] C. Nantasenamat, “How to Build a Machine Learning Model | by Chanin Nantasenamat | Towards Data Science,” 2020. [Online]. Available: https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1. [Accessed: 10-Feb-2020].

[24] P. Patwa *et al.*, “Fighting an Infodemic: COVID-19 Fake News Dataset,” *Commun. Comput. Inf. Sci.*, vol. 1402 CCIS, pp. 21–29, 2021, doi: 10.1007/978-3-030-73696-5_3.

[25] Google LLC, “Machine Learning Crash Course: Validation Set: Another Partition,” 2020. [Online]. Available: https://developers.google.com/machine-learning/crash-course/validation/another-partition. [Accessed: 13-Jan-2022].

[26] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,” 2018.