COVNLP: A Multisource COVID-19 Dataset for Natural Language Processing

Olubayo Adekanmbi*, Wuraola Fisayo Oyewusi, Warrie Warrie, Adedayo Odukoy, Abimbola Olawale, Opeyemi Osakuade, Mary Salami

Data Scientists Network (Data Science Nigeria) Lagos Nigeria

*Corresponding author

doi: https://doi.org/10.21467/proceedings.157.2

ABSTRACT

In this work, we propose COVNLP, a novel dataset for natural language processing tasks. The openly available dataset consists of 3,199 de-identified peer-to-peer messages shared across different channels like Whatsapp, SMS and Social media channels from volunteers during the COVID-19 pandemic in Nigeria. The messages were labelled by both participants at submission and independent data annotators after submission under three (3) major themes; message genuity, type and impact. We discovered that the most trusted source of information for the participants during the COVID-19 pandemic were international stations, social media and websites. 31.20% of the messages received by volunteers were labelled to have psychological effects such as emotional disturbance, depression, stress, mood alterations. The dataset is available here as part of our experimentation, we developed a basic machine learning model to classify the messages into misinformation, disinformation and rumour classes based. The best performing algorithm was Logistic Regression with count vectorizer with Area under the curve (AUC) value of 0.813 compared to Naive Bayes Classifier (0.716) and Random Forest Classifier(0.710).

Keywords: Natural Language Processing, Information Source Trustworthiness, COVID-19 Dataset

1 Introduction

The coronavirus disease (COVID-19) outbreak in December 2019 has led to a global pandemic claiming thousands of lives worldwide (Hannah Ritchie and Roser 2020). The World Health Organisation declared the outbreak to be a Public Health Emergency of International Concern (PHEIC) on January 30, 2020, and recognised it as a pandemic on March 11, 2020. The disease outbreak resulted in fear and various information regarding COVID-19 from different sources. Unfortunately, this information spread was accompanied by a large amount of misleading and false information about the virus, which led the World Health Organisation (WHO) to warn against the ongoing "infodemic" during the epidemic[1]. Furthermore, this infodemic has a high psychological burden (e.g., psychological distress, anxiety, and depression) on patients and their families, medical staff, health care workers, and the general population (Abdoli, 2020).
After the first reported case in Ogun State on the 27th of February 2020, Nigeria also experienced a surge in COVID-19 instances. The exponential spread of this pandemic sparked fear and uneasiness among citizens, the government, and various organisations, leading to the federal government making restrictions to movement (Hyland-Wood *et al.*, 2021). However, during the lockdown, people, including governmental organisations, shared different information about causes, symptoms, spread, and cure for this disease on various radio and television stations, Facebook, Twitter, SMS, and emails. This made it harder for people to find trustworthy and reliable information when they needed it.

[1] https://www.afro.who.int/news/who-ramps-preparedness-novel-coronavirus-african-region.

Proceedings DOI: 10.21467/proceedings.157; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-970666-0-3

This work presents a crowdsourced dataset that contains messages spread across the country during the COVID-19 global pandemic in Nigeria. The dataset can be used for a range of text data-dependent tasks in natural language processing and labelled to improve understanding of the type, sources, perception of these messages, especially in the context of misinformation, disinformation and rumours. Although the dataset is collected from Nigeria, it may be relevant globally.

2 Literature Review

There are several datasets curated for the study of COVID-19. The first version of CORD-19, a growing resource of publications and historical research about coronavirus, was first released in March 2020. It was designed to enable text mining, information retrieval and other text-dependent tasks (Wang *et al.* 2020). The COVID-Fact (Saakyan, Chakrabarty, and Muresan 2021), dataset contains claims about COVID-19 and contradictory claims refuted by evidence. The work also proposed a verification framework that automatically detects valid information and source articles. (Aguilar-Gallegos *et al.* 2020) introduced a multilingual dataset from tweets worldwide. The goal is to analyse Twitter activities in the first stages of the coronavirus outbreak. COVIDCQ (Mutlu *et al.* 2020), also examined more than 14,000 tweets and reports on the views of tweet authors regarding the use of "chloroquine" and "hydroxy- chloroquine" in the treatment or prevention of coronavirus.

According to the Pan American Health Organisation[2]. Infodemic refers to a large increase in the volume of information associated with a specific topic and whose growth can occur exponentially in a short period of time due to a specific incident, such as the COVID-19 pandemic. In this situation, disinformation, misinformation and rumours appear in the scene. Disinformation refers to false information shared deliberately, they are usually contents that are 100% false, designed to deceive and cause harm. They could also be false information framed to impersonate an influential name or brand. Misinformation is associated with circulating misleading information i.e. inaccurate use of information and false context [3] while rumour messages are information whose accuracy is yet to be determined at the time of posting. They are usually manipulated contents.

In this work, we present COVNLP a labelled multisource dataset specifically crowdsourced from volunteers during the COVID-19 pandemic in Nigeria to improve understanding of the type, sources and perception.

3 The Dataset

COVNLP is an imbalanced dataset with a total size of 3,199 messages from WhatsApp, SMS, Facebook, Twitter, Emails, Instagram and other platforms that were not disclosed by the volunteers. We decided to take the crowdsourcing approach to get the messages sent directly to individuals and their perspectives on those messages. Throughout the rest of this section, we discussed the method for data collection, data pre-processing, labelling and exploratory analysis of the data.

4 Data Collection

As shown in Figure 1, a simple request was made to an existing crowdsourcing community and the general public to share up to 10 de-identified content that is related to COVID- 19 which they had received.

[2] https://iris.paho.org/handle/10665.2/52052

[3] https://www.dictionary.com/e/misinformation-vs-disinformation-get-informed-on-the-difference/

Proceedings DOI: 10.21467/proceedings.157
ISBN: 978-81-970666-0-3

Series: AIJR Proceedings
ISSN: 2582-3922

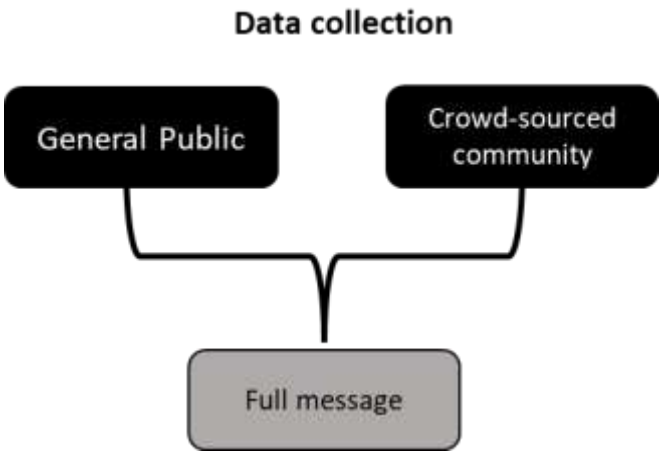


Figure 1: *Data collection methodology*

A google form was designed for ease of deployment and this allows the volunteers to label the data under these categories (demography, data source, category, content, and perception of the message) at submission. The labels served as the ground truth for the original perception about the messages shared.

5 Data Processing

As with every dataset at collection, there were unrelated, duplicates and erroneous messages submitted. To ensure quality data curation, we performed some data quality checks as shown in Figure 2. A simple python script was implemented to remove single-word and unlabelled messages. We also had 3 members of the team, who manually checked the validity of the messages.

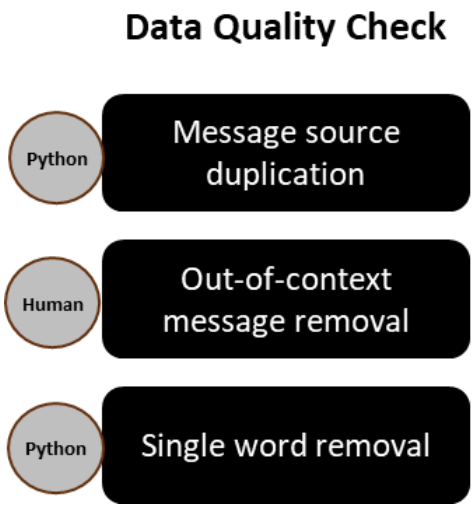


Figure 2: *Data Quality Check*

6 Data Labelling

This work explored two approaches to data labelling, simple non-overwhelming labelling by data volunteers at the collection and extensive labelling by trained volunteer annotators. Data volunteers labelled for the data source and perceptive labels like if they believe the messages are genuine and what category they think the data belong to (e.g. health, educational series, technical information, jokes to health educational series). Each message, as shown in Figure was labelled by 5 trained volunteer data annotators, resulting in 5 perspectives of the message. We decided on the final label by implementing a majority voting system to validate the class of the message and the dominant label was picked as the most appropriate label.

This was done under three (3) categories:

- Message Genuity
- Message Type
- Message Impact

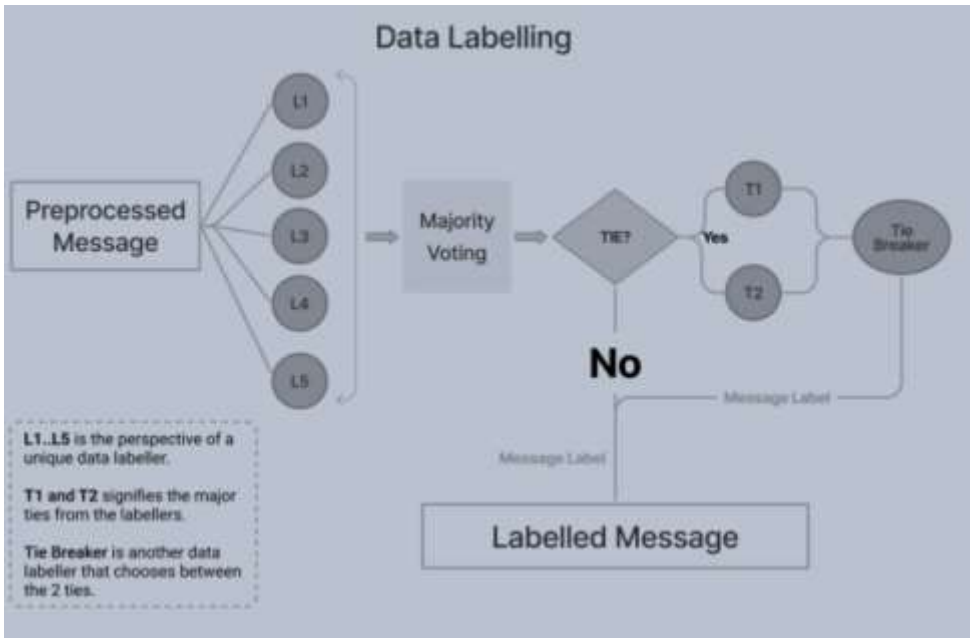


Figure 3: *Data Labelling Methodology*

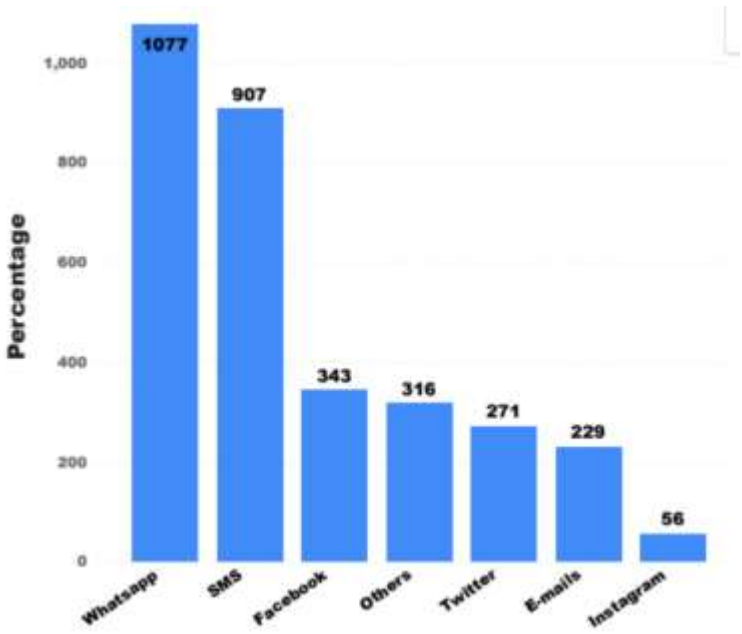


Figure 4: *Message Source*

Message Genuity labels were based on the first draft's framework for misinformation and disinformation[4]. This framework includes misleading content, false context, fabricated content, manipulated content, imposter content and parody. The message type was based on the UNESCO's research[5] and message

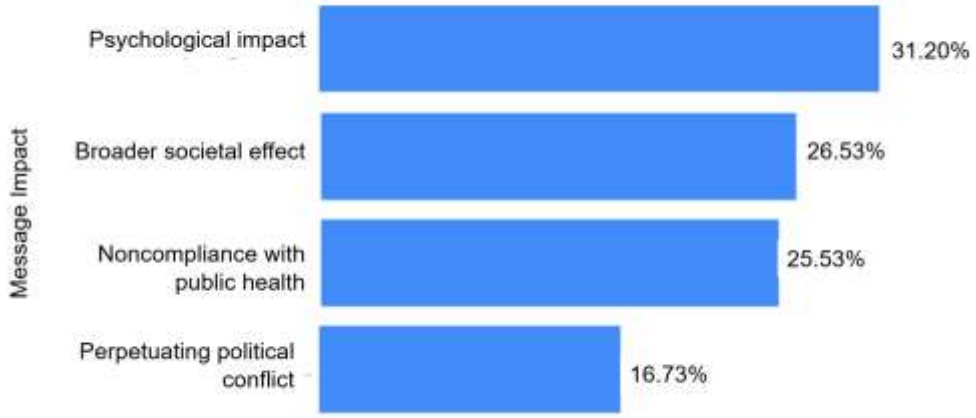
[4] https://firstdraftnews.org/long-form-article/understanding-information-disorder/

[5] https://en.unesco.org/covid19/disinfodemic/brief1

Proceedings DOI: 10.21467/proceedings.157
ISBN: 978-81-970666-0-3

Series: AIJR Proceedings
ISSN: 2582-3922

impact was based on the message impact using Social Science in Humanitarian Action's framework.[6],



Non-compliance with public health recommendations, Perpetuating political conflict and racial discrimination, Psycho-social effects and broader societal effects.

Figure 5: *Message Impact*

7 Exploratory Data Analysis

The dataset is imbalanced with a total size of 3,199. Figure 4 shows that WhatsApp and SMS account for the majority of received message sources, with 33.6% and 28.3%, respectively. Figure 5 shows 31.20% of the messages received by volunteers were labelled to have psychological effects such as emotional disturbance, depression, stress, mood alterations.

8 Classification Experiment and Results

For this work, we evaluated our data with a basic machine learning task to classify the COVID-19 messages. Based on the previous classes labelled by our volunteers, we focused on the erroneous messages and streamlined these messages into three classes as displayed in Table 1 below.

Table 1: *Message Classes*

Classes	**Number of messages**
Misinformation	308
Disinformation	241
Rumor	87

Since the streamlined data is imbalanced, we adopted a pre-processing technique that involves splitting some of the message samples into sentences to increase the number of samples in each class. This resulted in a data balance of 375 samples per class.

Two basic vectorization methods, Countvectorizer and Term Frequency - Inverse Document Frequency (TF-IDF) were used to convert text data into a meaningful vector of numbers. We also compared the performance of three machine learning algorithms (Logistic regression, Random Forest Classifier, and Naive Bayes classifier) for the classification task.

[6] https://www.socialscienceinaction.org/wp-content/uploads/2020/03/SSHAP-Brief.Online-Information.COVID-19.pdf

The performance of the vectorizers and each model is shown in Table 2 and Table 3. as shown in Table 2, evaluating the model with the AUC, the Logistic regression model with count vectorizer outperforms the two others with a score of 0.813.

Table 2: *F1 score, AUC and Weighted Average for Machine Learning Models using Count Vectorizer*

Model	Disinformation (F1 Score)	Misinformation (F1 Score)	Rumour (F1 Score)	AUC	Weighted average
Naive Bayes Classifier	0.504	0.662	0.629	0.716	0.609
Random Forest Classifier	0.493	0.555	0.549	0.710	0.533
Logistic Regression	0.623	0.654	0.692	0.813	0.658

Table 3: *F1 score, AUC and weighted average for machine learning models using TF-IDF Vectorizer*

Model	Disinformation (F1 Score)	Misinformation (F1 Score)	Rumour (F1 Score)	AUC	Weighted average
Naive Bayes Classifier	0.528	0.638	0.594	0.707	0.586
Random Forest Classifier	0.489	0.616	0.545	0.708	0.550
Logistic Regression	0.648	0.675	0.702	0.675	0.812

9 Conclusion

We present COVNLP, an open multisource dataset for Natural Language Processing tasks, the dataset contains 3,199 labelled crowdsourced messages related to COVID-19 from different sources. While this is a small dataset, it provides a good enough challenge for machine learning models as shown in section 4. It could serve as additional data to already available COVID-19 infodemic data to mitigate the spread of false information during pandemics.

10 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in institutional affiliations.

How to Cite

Adekanmbi *et al.* (2024). COVNLP: A Multisource COVID-19 Dataset for Natural Language Processing. *AIJR Proceedings*, 15-20. https://doi.org/10.21467/proceedings.157.2

References

Abdoli, A. 2020. Gossip, Rumors, and the COVID-19 Crisis. *Disaster Medicine and Public Health Preparedness*, 14(4): e29–e30.

Aguilar-Gallegos, N.; Romero-Garc´ıa, L. E.; Mart´ınez- Gonza´lez, E. G.; Garc´ıa-Sa´nchez, E. I.; and Aguilar-A´ vila, J. 2020. Dataset on dynamics of Coronavirus on Twitter.*Data in Brief*, 30: 105684.

Hannah Ritchie, L. R.-G. C. A. C. G. E. O.-O. J. H. B. M. D. B., Edouard Mathieu; and Roser, M. 2020. Coronavirus Pandemic (COVID-19). *Our World in Data*. Https://ourworldindata.org/coronavirus.

Hyland-Wood, B.; Gardner, J.; Leask, J.; and Ecker, U. 2021. Toward effective government communication strate- gies in the era of COVID-19. *Humanities and Social Sci- ences Communications*, 8: 30.

Mutlu, E. C.; Oghaz, T.; Jasser, J.; Tutunculer, E.; Rajabi, A.; Tayebi, A.; Ozmen, O.; and Garibay, I. 2020. A stance data set on polarized conversations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19. *Data in Brief*, 33: 106401.

Saakyan, A.; Chakrabarty, T.; and Muresan, S. 2021. COVID-Fact: Fact Extraction and Verification of Real- World Claims on COVID-19 Pandemic. arXiv:2106.03794.

Wang, L. L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Burdick, D.; Eide, D.; Funk, K.; Katsis, Y.; Kinney, R.; Li, Y.; Liu, Z.; Merrill, W.; Mooney, P.; Murdick, D.; Rishi, D.; Sheehan, J.; Shen, Z.; Stilson, B.; Wade, A.; Wang, K.;Wang, N. X. R.; Wilhelm, C.; Xie, B.; Raymond, D.; Weld, D. S.; Etzioni, O.; and Kohlmeier, S. 2020. CORD-19: The COVID-19 Open Research Dataset. arXiv:2004.10706.