

Analysing Crowd-Sourced Vaccine Data Using Machine Learning: Uncovering Concerns and Insights

Lakshmi S. Gopal¹, Aswathy A.², Krishnendu K.³ and Hemalatha Thirugnanam⁴

¹Center for Wireless Networks & Applications (WNA), Amrita Vishwa Vidyapeetham, Amritapuri, India

²Center for Wireless Networks & Applications (WNA), Amrita Vishwa Vidyapeetham, Amritapuri, India

³Center for Wireless Networks & Applications (WNA), Amrita Vishwa Vidyapeetham, Amritapuri, India

⁴Center for Wireless Networks & Applications (WNA), Amrita Vishwa Vidyapeetham, Amritapuri, India

Abstract

The rapid development of the Covid-19 vaccines, concerns about its safety contributed to vaccine hesitancy globally. Social media platforms transfer knowledge on such global concerns and are a good source for investigating public opinions. This study proposes a machine learning based analysis of Covid-19 vaccine public opinions using Twitter data where a tweet post is classified into multiple labels which describes various concerns. We experimented with supervised learning algorithms wrapped along with multiple label classifier algorithms. We have achieved an average F1 micro score of 62% which suggested improvement.

Keywords

Covid Vaccines, Machine Learning, Social Media

1. Introduction

Vaccination is a highly effective public health strategy, saving lives and reducing disease burden. However, vaccine hesitancy persists, especially in the digital age with easy access to both credible and misleading information. Crowdsourced data platforms now allow individuals to share their vaccine-related experiences and concerns, offering valuable insights into this issue.

Machine learning has become a vital tool in public health and epidemiology. It can analyse large datasets, uncover patterns, and provide insights that traditional methods struggle to achieve. Machine learning algorithms can sift through vast amounts of unstructured text data from social media and other platforms to reveal patterns and concerns related to vaccines.


This research paper aims to contribute to the growing body of knowledge in the field of vaccine hesitancy and public health by presenting a comprehensive analysis of crowdsourced vaccine data using machine learning techniques. This study is driven by the fact that a deeper understanding of the concerns expressed by individuals through crowdsourcing regarding vaccines can inform targeted public health interventions and communication strategies. By

Forum for Information Retrieval Evaluation, December 15-18, 2023, India

✉ lakshmisgopal@am.amrita.edu (L. S. Gopal); aswathynaik@am.amrita.edu (A. A.); krishnenduk@am.amrita.edu (K. K.); hemalathat@am.amrita.edu (H. Thirugnanam)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

leveraging machine learning algorithms, we aim to shed light on the intricate dynamics of vaccine hesitancy, ultimately contributing to more effective vaccination campaigns and improved public health outcomes. Additionally, comprehending these perceptions within communities, states, and the nation across different time frames can furnish us with precise data for crafting specialised strategies to enhance immunisation education programs and public health campaigns.

In the subsequent sections of this paper, we will discuss the methods employed, present our findings, and discuss the implications of our analysis.

2. Related Work

The utilisation of Machine Learning for analysing vaccine-related concerns is of paramount significance in the current era of pandemics, and numerous studies have delved into this field. This paper specifically concentrates on developing a predictive model for assessing public sentiment-related concerns, primarily sourced from social media platforms. One study has highlighted the utilisation of social media bots to intentionally sow discord and confusion regarding vaccination, potentially dissuading people from getting vaccinated [1].

Additionally, another research emphasises the pivotal need to confront and counteract rumours and conspiracy theories in public health campaigns, particularly in the context of mitigating vaccine hesitancy and ensuring the success of vaccination initiatives [2]. Authors emphasise that factors causing vaccine hesitancy, like technological change and political disempowerment, and addressing these issues requires long-term efforts from multiple stakeholders. Building vaccine confidence for the long term is measured by public trust in vaccine delivery institutions.

A study conducted from April to August 2019 aimed to develop and validate deep learning models to understand public perceptions of the HPV vaccine using data from social media [3]. The study collected data from January 2014 to October 2018, analysing social media discussions related to health belief models and theory of planned behaviour. The results showed trends in constructs such as perceived barriers, positive attitudes towards the HPV vaccine, and negative attitudes. Interstate variations in public perceptions were also identified. The study provides a good understanding of public perceptions on social media and evolving trends, potentially influencing local anti vaccine sentiment.

A study [4] examining vaccine sentiment on social media revealed that vaccine hesitancy contributes to suboptimal vaccination coverage in the United States. The study analysed semantic networks of vaccine information from Twitter users in the US, identifying positive, negative, and neutral sentiment. Positive sentiment focused on parents and health risks, while negative sentiment focused on children and organisational bodies. The study suggests that analysing vaccine sentiment on social media can help understand complex drivers of vaccine hesitancy and improve public health communication, ultimately improving vaccine confidence

and coverage in the US.

Moreover, a study has been conducted to demonstrate the efficient collection and preprocessing of Twitter data, encompassing information related to vaccines as well as other disaster-related data [5]. Another research work highlights the paramount importance of leveraging Machine Learning and Artificial Intelligence across diverse emergency situations. These advanced technologies play a pivotal role in not only enhancing emergency preparedness and response but also in enabling data-driven decision-making, resource allocation, and predictive modelling to mitigate the impact of any emergency on affected populations and infrastructure [6].

3. Task

We aim to develop a multi label classification on public opinion tweets of the Covid-19 vaccines which is a methodology proposed as part of the AISoMe (Artificial Intelligence on Social Media) track [7][8] in the FIRE (Forum for Information Retrieval Evaluation) 2023. The developed classifier labels a tweet based on specific concern(s) about vaccines expressed by the respective Twitter user. A tweet can have more than one label (concern), e.g A tweet expressing 3 different concerns about vaccines will have 3 labels. As labels for the classification task, we take into consideration the following concerns about vaccines:

- Unnecessary: The tweet implies that immunizations are not necessary or that better alternative treatments exist.
- Mandatory: The tweet advocates against making vaccinations mandatory.
- Pharma: The tweet implies that big pharmaceutical firms are only out to make a profit.
- Conspiracy: The tweet raises the possibility of a larger conspiracy than merely big pharma's desire for profit.
- Political: The tweet raises fears that governments and politicians are using vaccines as a tool to advance their own agendas.
- Country: The tweet criticizes a vaccination because of the nation where it was created or produced.
- Rushed: The tweet raises questions about whether the vaccines have undergone adequate testing or whether the available data is reliable.
- Ingredients: The tweet highlights concerns about the vaccine contents or the technology utilised.
- Side-effect: The tweet expresses worry about vaccine side effects, including deaths that may result.
- Ineffective: The tweet expresses worry that the immunizations are inefficient and useless because they are ineffective in some cases.
- Religious: The tweet opposes vaccinations for religious reasons.
- None: No explicit justification is provided in the tweet.

Tweet	ingredients	side-effect	mandatory	rushed	ineffective	political	none	conspiracy	country	pharma	unnecessary	religious
My mom want me to take the vaccine so bad & I'm not with it	0	0	0	0	0	0	1	0	0	0	0	0
@ndtv Also we have seen that trial vaccine is not acceptable at all. Although it is having lots of side effects too	0	1	0	1	0	0	0	0	0	0	0	0
The pathetic, politically altered, big Pharma "science" will result in more death and destruction. we're about to have bigger problems than any virus could ever have.	0	1	0	0	0	1	0	0	0	1	0	0

Figure 1: One hot encoded data set - The 'tweet' column is taken from the given dataset. The rest of the columns are created programmatically and the values '1' and '0' represent the presence and absence of the label respectively.

4. Methodology

The proposed methodology aims to perform multi label classification on the given dataset. In depth study of the literature [9][10][11] describes various methods of machine learning based multi label classification methods. We experimented with a problem transformation method, namely classifier chains, which transforms a multi label classification problem into multiple binary classification problems.

4.1. Exploratory Data Analysis (EDA)

To comprehend and interpret the given data in depth, we begin with an EDA. The given data initially had 3 columns, 'ID', 'tweet' and 'labels'. For a multi label classification problem, one hot encoded data is appropriate and hence the data was modified where the labels are one hot encoded. The given data contained no null or NaN values. The one hot encoded dataset contains 9921 rows and 14 columns. Figure 1 shows a sample of the one hot encoded dataset.

The tweets in the data are labelled about concerns of covid vaccines (see section 3) and have categorised tweets under 12 labels. Figure 2 shows the number of tweets that are categorised under a particular label. From the figure we can see that the label 'side-effect' is the highest in number and 'religious' is the lowest. A tweet could be categorised by 1, 2 or 3 labels. Figure 2 shows the number of tweets that got categorised under a single label, two labels or three labels. From the figure we can see that the majority of the tweets were categorised by a single label.

To understand the use of vocabulary in the tweets, word clouds were generated of the most frequent and less frequent labels in the dataset, which are 'side-effect' and 'religious' labels respectively. Figure 3 shows the generated word clouds. From the word clouds we can observe that the terms 'vaccine', 'covid' and 'pfizer' have frequent occurrences. Keywords similar to 'side-effect', such as 'death', 'adverse reaction', 'blood clot' etc are found to occur frequently in the 'side-effect' word cloud. Keywords similar to 'religious' label, such as 'religion', 'faith', 'psalm' etc were found, but were less frequent in the 'religious' word cloud.

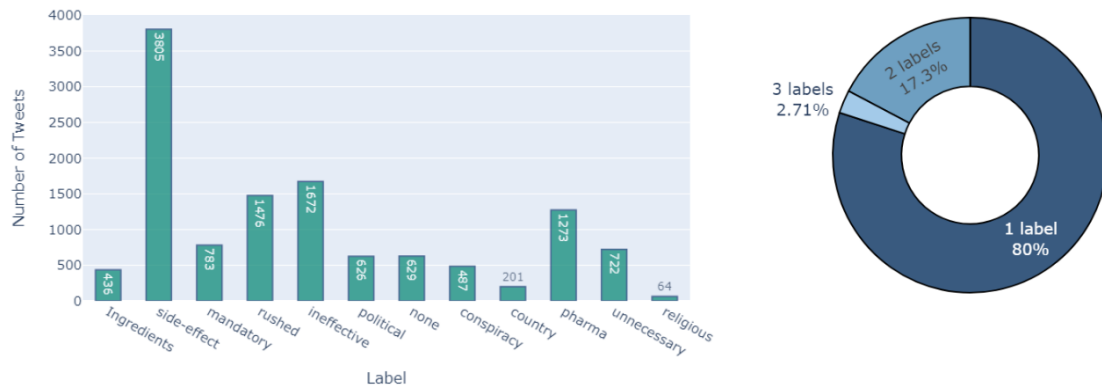


Figure 2: Analysis of data - The bar graph (left) shows the number of tweets categorised under a label. The x-axis represents the label name and y-axis represents the number of tweets under each label. The pie chart (right) shows the percentage of tweets categorised with 1, 2 or 3 labels.

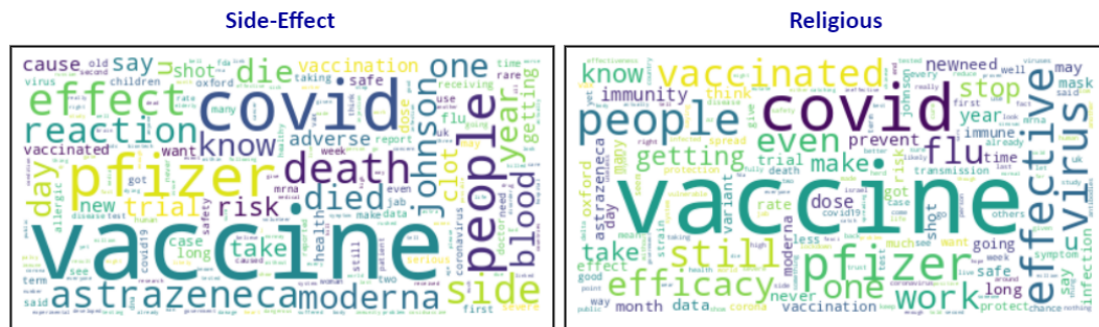


Figure 3: Analysis of data - word clouds generated for the labels 'side-effect' (left) and 'religious' (right).

4.2. Data Preprocessing

Basic data cleaning methods have been applied onto the tweet data. Before removing the unwanted text, we word tokenize the tweet and each character to lowercase. The cleaned tweet is further fed into the model allowing to reduce unnecessary processing. We use Python libraries such as NLTK and Regular expressions to eliminate the following:

- URLs often found along with a tweet (image, video urls)
- Stopwords (the, a, is etc) are removed except for 'not' and 'no' to maintain the context
- Special characters, smileys

Table 1

Performance evaluation of the 3 models (first run) using scoring metrics

Model	Accuracy	Precision	Recall	F1 micro
Classifier chain+Logistic Regression	0.42	0.67	0.45	0.54
Classifier chain+Support Vector Machines	0.49	0.67	0.55	0.61
Multi output classifier +Support Vector Machines	0.46	0.80	0.52	0.64

4.3. Model Creation for Multi Label Classification

We have experimented with a multi label classification where each data point is associated with multiple labels. Among various multi label classification approaches, we have experimented with classifier chains and multi output classifier methods. A classifier chain initially starts with a set of binary classifiers, one for each label in the multi label classification problem. When making predictions for a new instance, you start by predicting the first label using its binary classifier. Then, you use this prediction, along with the instance's features, to predict the second label. This process continues until all labels have been predicted. To perform the classification we can wrap any classification algorithm which is capable of a binary classification in the classifier chain. We have also experimented using the multi output classifier algorithm, a wrapper that takes a single-output classifier and extends it to work with multiple output labels.

Initially, we stratified the dataset with a train-test split of 70% to 30% respectively. Both the training and validation data was preprocessed as described in section 4.2. The resultant data was used as the input data for both classifier chain and multi output classifier models. We experimented two model creations with the classifier chains by wrapping a Logistic Regression and Support Vector Machines model. Both these models are widely used for classification problems. We experimented one model creation with the multi output classifier where it was wrapping a Support Vector Machine model. All 3 experiments showed a moderate result in the initial phase which led to fine tuning.

5. Results and Evaluation

The created models are evaluated using accuracy, precision, recall and F1 score. Table 1 shows the evaluation results of the 3 models in the initial run. Table 3 shows the performance evaluation of the run files submitted to the AISoMe track.

To evaluate the model better, we plotted the learning curve of the classifier chain models which led to fine tuning it further. Figure 4 shows the learning curves of the Logistic Regression model wrapped in a classifier chain, before and after fine tuning. Figure 5 shows the learning curves of the Support Vector Machines model wrapped in a classifier chain, before and after tuning. Fine tuning certainly improved the performance, but also suggests the need for more data for training.

We have achieved slight improvement in the performance of the classifier chain models after

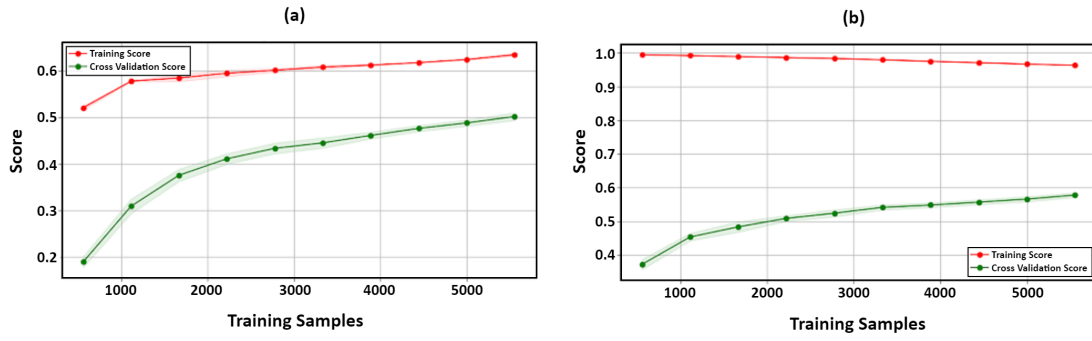


Figure 4: Learning curves of Logistic Regression wrapped in Classifier Chain. (a) shows the learning curve of the model before fine tuning. (b) shows the learning curve of the model and after fine tuning.

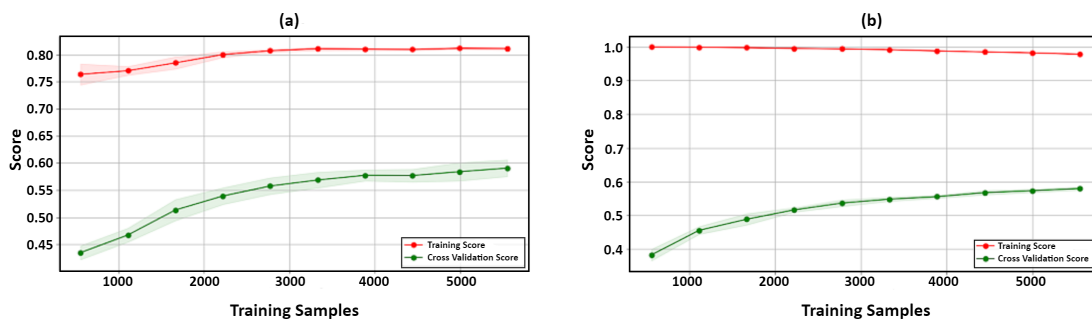


Figure 5: Learning curves of Support Vector Machines wrapped in Classifier Chain. (a) shows the learning curve of the model before fine tuning. (b) shows the learning curve of the model and after fine tuning.

Table 2

Performance evaluation of the classifier chain models using scoring metrics post fine tuning

Model	Accuracy	Precision	Recall	F1 micro
Classifier chain+Logistic Regression	0.48	0.68	0.54	0.60
Classifier chain+Support Vector Machines	0.51	0.67	0.57	0.62

fine tuning. Table 2 shows the performance results of the 2 fine tuned models.

6. Conclusion

This research work uses the Covid vaccine social media data which showed the concerns of the public related to the usage of vaccinations. We experimented multiple models with the given data and chose the top 3 performing models to showcase in this report. We used a classifier chain model which wraps Support Vector Machines and Logistic Regression and a multi output classifier which wraps Support Vector Machines. We achieved the highest score for Multi

Table 3

Performance evaluation of submission models

Run File	Methodology	Macro-f1	Jacc
Model 1	SVM wrapped in Multi Output Classifier	0.38	0.45
Model 2	LR wrapped in Classifier Chain	0.38	0.41
Model 3	SVM wrapped in Classifier Chain	0.3	0.43

Output Classifier with a F1 score of 64%. The performance can be improved either by improving the dataset or by other preprocessing methods or data augmentation strategies.

7. Online Resources

The input data, test data and implemented Python code are made available on "<https://github.com/lakshmiSGopal/AISOME-FIRE-2023>".

References

- [1] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, M. Dredze, Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate, *American journal of public health* 108 (2018) 1378–1384.
- [2] E. Pertwee, C. Simas, H. J. Larson, An epidemic of uncertainty: rumors, conspiracy theories and vaccine hesitancy, *Nature medicine* 28 (2022) 456–459.
- [3] J. Du, C. Luo, R. Shegog, J. Bian, R. Cunningham, J. Boom, G. Poland, Y. Chen, C. Tao, Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. *jama netw open*. 2020 nov 02; 3 (11): e2022025. doi: 10.1001/jamanetworkopen.2020.22025, ????
- [4] G. J. Kang, S. R. Ewing-Nelson, L. Mackey, J. T. Schlitt, A. Marathe, K. M. Abbas, S. Swarup, Semantic network analysis of vaccine sentiment in online social media, *Vaccine* 35 (2017) 3621–3638.
- [5] A. Aswathy, R. Prabha, L. S. Gopal, D. Pullarkatt, M. V. Ramesh, An efficient twitter data collection and analytics framework for effective disaster management, in: *2022 IEEE Delhi Section Conference (DELCON)*, IEEE, 2022, pp. 1–6.
- [6] J. Phengsuwan, T. Shah, N. B. Thekkummal, Z. Wen, R. Sun, D. Pullarkatt, H. Thirugnanam, M. V. Ramesh, G. Morgan, P. James, et al., Use of social media data in disaster management: a survey, *Future Internet* 13 (2021) 46.
- [7] S. Poddar, M. Basu, K. Ghosh, S. Ghosh, Overview of the fire 2023 track:artificial intelligence on social media (aisome), in: *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, 2023.
- [8] S. Poddar, A. M. Samad, R. Mukherjee, N. Ganguly, S. Ghosh, Caves: A dataset to facilitate explainable classification and summarization of concerns towards covid vaccines, in: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 3154–3164.

- [9] C. Prathibhamol, G. Amala, M. Kapadia, Anomaly detection based multi label classification using association rule mining (admlcar), in: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2016, pp. 2703–2707.
- [10] C. Prathibhamol, K. Jyothy, B. Noora, Multi label classification based on logistic regression (mlc-lr), in: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2016, pp. 2708–2712.
- [11] R. Ramanathan, K. Soman, P. Rohini, G. Dharshana, Investigation and development of methods to solve multi-class classification problems, in: 2009 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE, 2009, pp. 805–807.