Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 80-108
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Analysis of Public Sentiment regarding COVID-19 Vaccines on the Social Media Platform Reddit

Lucien Dikla Ngueleo[1,2], Jules Pagna Disso[3], Armel Ayimdji Tekemetieu[4], Justin Moskolaï Ngossaha[2], Michael Nana Kameni[1]

[1]African Institute for Mathematical Sciences (AIMS)/AIMS-Cameroon, Limbe, Cameroon
[2]Department of Mathematics and Computer Science, Faculty of Science, University of Douala, Douala, Cameroon
[3]WMG (Warwick Manufacturing Group), University of Warwick, Coventry, UK
[4]School of Information Studies, McGill University, Montreal, Canada
Email: lucien.dikla@aims-cameroon.org, jules.pagna-disso@warwick.ac.uk, armel.ayimdjitekemetieu@mcgill.ca

How to cite this paper: Ngueleo, L.D., Disso, J.P., Tekemetieu, A.A., Ngossaha, J.M. and Kameni, M.N. (2024) Analysis of Public Sentiment regarding COVID-19 Vaccines on the Social Media Platform Reddit. *Journal of Computer and Communications*, **12**, 80-108.
https://doi.org/10.4236/jcc.2024.122006

Received: December 20, 2023
Accepted: February 18, 2024
Published: February 21, 2024

Abstract

This study undertakes a thorough analysis of the sentiment within the *r/Coronavirus* subreddit community regarding COVID-19 vaccines on Reddit. We meticulously collected and processed 34,768 comments, spanning from November 20, 2020, to January 17, 2021, using sentiment calculation methods such as TextBlob and Twitter-RoBERTa-Base-sentiment to categorize comments into positive, negative, or neutral sentiments. The methodology involved the use of Count Vectorizer as a vectorization technique and the implementation of advanced ensemble algorithms like XGBoost and Random Forest, achieving an accuracy of approximately 80%. Furthermore, through the Dirichlet latent allocation, we identified 23 distinct reasons for vaccine distrust among negative comments. These findings are crucial for understanding the community's attitudes towards vaccination and can guide targeted public health messaging. Our study not only provides insights into public opinion during a critical health crisis, but also demonstrates the effectiveness of combining natural language processing tools and ensemble algorithms in sentiment analysis.

Keywords

COVID-19 Vaccine, TextBlob, Twitter-RoBERTa-Base-Sentiment, Sentiment Analysis, Latent Dirichlet Allocation

1. Introduction

COVID-19, a respiratory disease caused by the SARS-CoV-2 virus and broke out in China in 2019 [1], has drastically affected lives globally, leading to widespread fear, economic disruption, and significant health crises. As of December 2023, the World Health Organization (WHO) reports nearly 7 million deaths attributed to COVID-19. Countries worldwide have implemented various risk mitigation strategies, such as social distancing [2] [3], deeply impacting economies and healthcare systems. Amidst this, vaccination emerges as a key solution [4], yet misinformation on social media platforms like YouTube, Twitter, and Reddit has fueled mistrust in vaccines [5]. However, the rapid development and deployment of COVID-19 vaccines have been met with a degree of mistrust and rejection, primarily fueled by misleading information proliferated through social media platforms like YouTube, Twitter, and Reddit. These networks have become vital forums for discussing the merits and drawbacks of vaccination, making them rich sources for data collection and analysis [1] [6]. With the significant amount of data available from these platforms, and the advancement of data analysis and processing tools such as machine learning and Natural Language Processing (NLP), it becomes imperative to harness these resources to produce reliable, real-time analyses.

In the unprecedented context of the COVID-19 pandemic, social networks, especially Reddit, have played a pivotal role in shaping public opinion on health measures and vaccines [7]. This study acknowledges the substantial impact of these platforms in forming public discourse and sentiment, often leading to misinformation and varied perceptions about COVID-19 vaccination. Recognizing the need to dissect and interpret these complex narratives, our research aims to leverage advanced machine learning techniques and NLP to analyze Reddit comments. This approach is crucial for developing targeted public health strategies and combating vaccine skepticism. The necessity and importance of this research stem from the critical need to understand public sentiment in an era where social media significantly influences health behaviors and decisions. Misinformation and varied opinions on platforms like Reddit have a tangible impact on vaccine uptake, potentially hindering efforts to control the pandemic. This study, therefore, seeks to bridge the gap in our understanding of these sentiments and their implications. By analyzing and interpreting the complex discussions on Reddit, we aim to contribute valuable insights to public health strategies, aiding in combating vaccine hesitancy and misinformation. This research is not only timely but essential in guiding informed decision-making and tailored communication approaches in the fight against COVID-19.

This paper is organized into several sections: Section 2 reviews recent work on public sentiments towards the COVID-19 vaccine. Section 3 describes our methodology in detail. Section 4 presents the results obtained from applying this methodology. Finally, we conclude our work and offer perspectives for future research in this domain.

2. Related Works

Natural Language Processing (NLP) is a subfield of artificial intelligence that enables computers to understand, interpret, and generate human language. The process of text analysis, in general, and public comments, in particular, requires a series of steps to process textual data and extract information on expressed sentiments. These steps include collecting comments, preprocessing comments, sentiment analysis of comments, visualizing results, and interpreting results. In this section, we will comprehensively present various related works in processing, analyzing, and modelling social media comments in general, and the Reddit platform in particular.

2.1. Process for Cleaning up Social Comments

Social network comments have become a source of data for the analysis of public sentiment, with regard to the Reddit and Twitter platforms several authors have recorded the comments of these platforms either to analyse sentiment or to assess the effectiveness of the vaccine [8] [9] [10] [11], or for the hesitancy and effectiveness of the COVID-19 vaccine [1] [12] [13] [14] [15]. However, the comments are often not cleaned up for direct use by the machine learning model, as they are comments written by people and not journals, formal articles, and that each plate. Thus, depending on the specific needs of the model, the pre-processing may include steps such as:

1) Cleaning: This step involves removing special characters, punctuation, and irrelevant words such as common stop words (e.g. "a", "an", "and", "are", "as", "at", "be", "by", "for", "from", "has", "he", "in", "is", "it", "its", "of", "on", "that", "the", "to", "was", "were", "with").

2) Tokenization: This step involves segmenting a comment into individual vocabulary words. It helps in breaking down the text into meaningful units for further analysis.

3) Lemmatization: This process transforms words into their original root forms, known as lemmas. It reduces different word forms to a common base, aiding in standardization and reducing the vocabulary size.

4) Stemming: This step involves removing prefixes or suffixes from a word, such as "-ing", "-ed", or "-s" to obtain the stem. Stemming and lemmatization are two common techniques used to convert words to their base form. Compared to lemmatization, stemming is faster and less computationally expensive, but it may not always produce accurate results.

2.2. Common Vectorization Methods for Social Comments

Once the process of cleaning up the social comments has been completed, it will be necessary to represent the data as a numerical value that can be understood by machine learning algorithms. For COVID-19, several vectoring techniques have been used in the literature, including:

- Bag-of-Word (BoW) is a vectorisation method that represents a comment as a vector counting the occurrences of each term in the comment.
- Term Frequency-Inverse Document Frequency (TF-IDF) is a word weighting

method that measures the importance of a term in a comment based on its frequency of occurrence and its inverse frequency in the entire corpus of documents.

- Word2vec is a vectorisation method that captures the semantic relationships between words by representing them as vectors of real numbers.
- Global Vectors (GloVe) for word representation are a method that uses a matrix of word occurrences to create word vectors that capture the semantic relationships between words.

However, Table 1 presents the different vectoring methods found in the literature.

In this table, we conduct a comparative study between the different vectorisation methods used in the literature on COVID-19. We focus on elements such as the type of representation performed by each method, the information captured, the size of the vocabulary and the average computation time used by each method.

2.3. Common Labelling Approach Used in COVID-19

Once vectorized, the comments are ready to be used by machine learning models, so in the literature we find several approaches used to label comments. Nuzhath *et al.* used 4868 tweets from a group and manually labelled into four classes (positive, negative, neutral and irrelevant) [12]. Liu *et al.* in their paper on public attitudes towards COVID-19 vaccines for English language Twitter comments applied the Valence Aware Dictionary and sEntiment Reasoner based labelling approach to determine whether the sentiment mentioned in the tweets was positive, neutral or negative [19]. Luo *et al.* in their study on understanding the reaction to the COVID-19 vaccine through a comparative analysis of Twitter comments in the US before and after the election period used the VADER sentiment analysis tool because of its ease of use and interpretation, they were able to classify tweets into three classes (positive, negative and neutral) [20]. Bengesi *et al.* in their study of the monkeypox epidemic used 500,000 multilingual tweets to show the polarity of public opinion on the vaccine, after translating the tweets into English they then used VADER and TextBlob to annotate the extracted comments into positive, negative and neutral sentiments [21]. Abiola *et al.* in their study on sentiment analysis of COVID-19 tweets in Nigeria, collected 1,048,575 tweets and then using TextBlob and VADER analysers labelled the

Table 1. Comparative table between the different vectoring methods.

	BoW	TF-IDF	Word2vec	GloVe
Representation	Word Count	Word Weight	Vectors	Vectors
Captures	Frequency	Importance	Relationships	Relationships
Vocabulary	Low	Low	High	High
Computation	Low	Low	Medium	High
Used by	[8] [9] [10] [14]	[9] [10] [16] [17] [18]	[9] [10] [17] [18]	[9] [10] [11] [17]

sentiments as positive, neutral and negative [11]. Qorib *et al.* in their examination of COVID-19 vaccine hesitation used Azure Machine Learning (a machine learning platform), VADER and TextBlob to label public tweets as neutral, negative and positive [1]. Vernikou *et al.* in their study of multi-class sentiment analysis of tweeter comments about COVID-19 used BERT (Bidirectional Encoder Representations from Transformers) to annotate sentiment into multiple classes [9]. Rodríguez-Ibánez *et al.* in their survey on sentiment analysis from social media platforms list several techniques used for sentiment labelling, including dictionary-based techniques, machine learning-based approaches and show that transformer-based approaches such as BERT, T5 (Text-to-Text Transfer Transformer) or GPT (Generative Pre-trained Transformer) perform far better than traditional techniques [10].

However, we conclude that for COVID-19 comments, three major approaches are used in the literature: manual approach, lexicon-based approach, and machine learning algorithm-based approach, as shown in **Figure 1**.

2.4. Commonly Used Algorithm for Classification of COVID-19

There are several classification algorithms that have been used for sentiment classification of comments related to the COVID-19 vaccine, such as: K-Nearest Neighbor (KNN), a classification algorithm based on the similarity between the K-Nearest comments. KNN uses this similarity between comments to classify them [21]. Multinomial Logistic Regression (LR) is a classification algorithm based on a probabilistic model. LR takes into account the word vectors used in the comments to determine the probability of each comment belonging to each class [1] [21]. Naive Bayes (NB) is a probabilistic algorithm that uses Bayes' theorem to calculate the probability of each class. It uses the word vectors present in the input comment and assumes that the words in the comment are independent, given the class to which it belongs [9] [21]. Linear Support Vector Classification (SVC) is a type of Support Vector Machine (SVM) that uses a linear kernel function to find the hyperplane that separates multiple classes in the input space [1] [21]. Decision Tree (DT) is a tree-like decision-making model. DT identifies the features that best separate the comments into different classes. At each node of the tree, a decision is made on the value of a particular feature, and

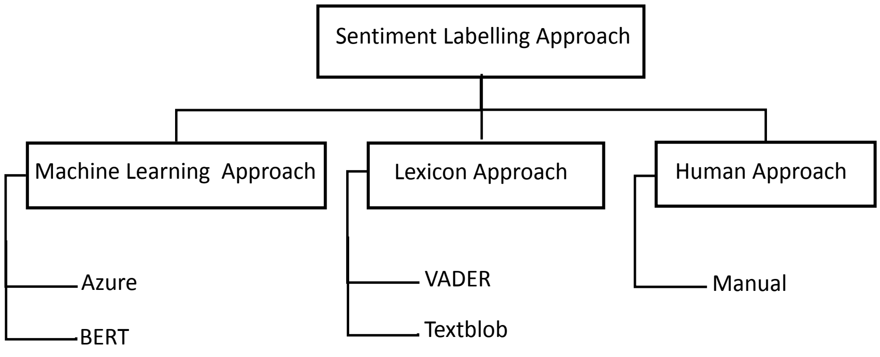

Figure 1. Common labelling approach for COVID-19 vaccine sentiment (source: author).

the algorithm follows the corresponding path in the tree until it reaches a leaf that represents the predicted class for the comment [1]. Random Forest (RF) is an ensemble model based on several decision trees built during training. RF determines the most frequently predicted class by the individual trees to predict the class of the comment [1] [21]. XGBoost is an ensemble algorithm that uses a set of gradient boosting models to improve prediction accuracy [21]. Finally, LSTM (Long Short-Term Memory) is a neural network architecture designed to address the problems of gradient backpropagation over long sequences of time. LSTM uses internal memory to store long-term information [9].

However, in **Table 2**, we have listed and recorded the best performances of certain algorithms combined with specific types of vectorization on a particular dataset. Therefore, the choice of vectorization technique may depend on the size of the data.

2.5. Topic Modelling on Social Comments

To group comments that tend to describe the same topics there is a sub-field of machine learning called topic modelling which is a form of clustering. In the literature on COVID-19, several topic modelling algorithms have been used including: LIWC (Linguistic Inquiry and Word Count), is a text analysis software that relies on dictionaries to cluster comments [22]; LDA (Latent Dirichlet Allocation) a probabilistic model used to determine the hidden topics in the comment set [23]; GSDMM (Gibbs Sampling Dirichlet Mixture Model) a Dirichlet mixture model used to discover hidden topics in a corpus [24]; and finally BERT (Bidirectional Encoder Representations from Transformers) a pre-trained natural language processing model based on the Transformers architecture [25]. Thus, **Table 3** presents the different methods listed with their specific use cases.

3. Methodology

Having reviewed the literature on sentiment analysis from social network comments, we describe the methodology used in this section.

3.1. Methodology Description

To analyze, model, and make recommendations from the public comments collected on the *r/Coronavirus* subreddit, we set up the workflow described in **Figure 2**.

Table 2. Commonly used supervised machine learning algorithm for classification of COVID-19 vaccine sentiment.

Authors	Vectorization	Data	Algorithm used	Accuracy
Vernikou *et al.* (2022) [9]	BERT	44,955 tweets	LSTM	78%
Qorib *et al.* (2023) [1]	TF-IDF	-	SVC	98%
Bengesi *et al.* (2023) [21]	CountVectorizer	500,000 tweets	SVM	93%

Table 3. Recent works on COVID-19 using topic modelling.

Ref.	Objectives	Topic modeling
[12]	Identify the main thematic areas of vaccine hesitancy	LDA
[14]	Characterizing vaccine discourse using comments	LDA, LIWC
[8]	Investigating COVID-19 vaccine-related discussions on social media such as Reddit	LDA
[11]	Analyzing the sentiment of tweets in Nigeria about COVID-19 vaccine	LDA
[26]	Examining public behaviour during different waves of COVID-19	LDA, GSDMM, BERT
[27]	Examine country-level variations in attitudes toward COVID-19 vaccine passports	LDA

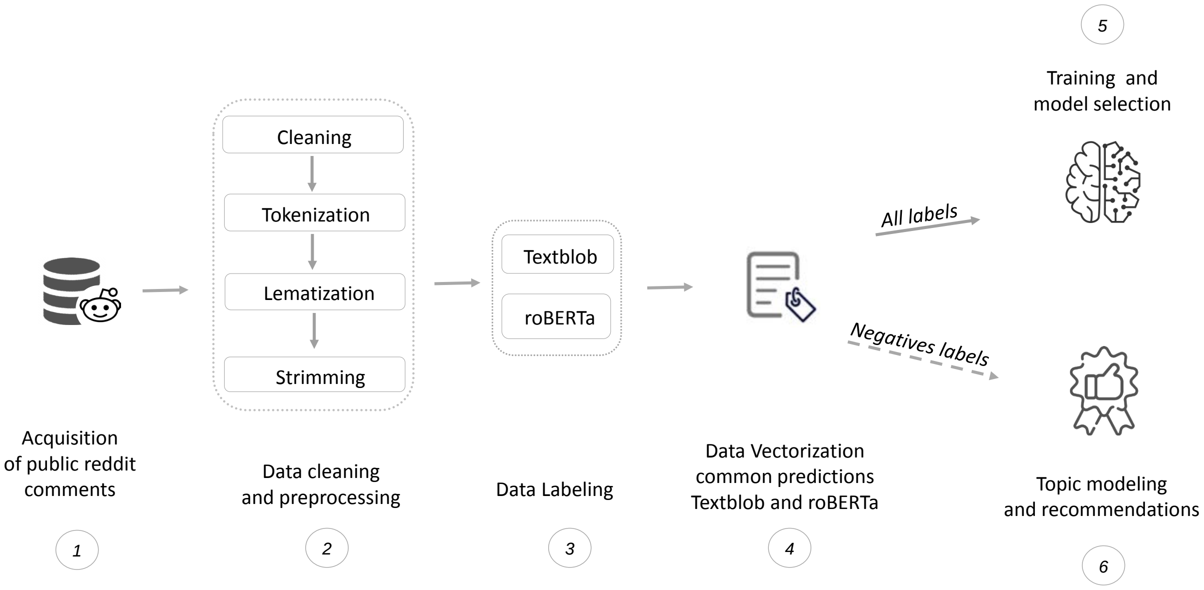

Figure 2. Methodology proposed.

To comprehend our methodology, it is necessary to adhere to the steps listed:

1) The first step consists of acquiring comments about the COVID-19 vaccine on Reddit, which allows for gathering data from open sources to obtain information on public opinions regarding the COVID-19 vaccine.

2) Step two involves data pre-processing, which is one of the most important steps in the process and includes sub-steps such as: Cleaning the data involves removing unwanted characters, stop words, and duplicate comments. Tokenization involves splitting the sentences into individual words, while lemmatization normalizes the words by reducing them to their base form. This reduces complexity and facilitates text analysis.

3) Step three involves using two labeling methods (TextBlob and RoBERTa) to label the sentiment of the comments. TextBlob is a Python library that allows for sentiment labeling using lexicon-based approaches, while Twitte-RoBERTa-Base-

sentiment is a deep learning algorithm that uses attention and has been trained on billions of tweets. Using two different methods allows for comparison and validation of results.

4) Step four involves selecting the common predictions of the two algorithms to ensureDescription consistent labeling, and then vectorizing the data to convert the labeled comments into numerical vectors, which can be used to train a machine learning model.

5) Step five involves producing a new model from the vectorized data. This involves using a machine learning algorithm to analyze the data and produce results. The model can be used to predict the sentiment of new comments collected in the future.

6) Step six involves descriptions using the negative predictions of the two algorithms to understand the population's hesitation towards the vaccine and formulating appropriate recommendations for decision-makers. By understanding the reasons for hesitation, decision-makers can implement suitable measures to encourage vaccination and strengthen confidence in the vaccine.

3.2. Data Description

The data was collected on the Reddit platform (a public news discussion site where users can submit content that is then voted on and commented upon), specifically on the *r/Coronavirus* subreddit, a community dedicated to COVID-19; it is a forum that prioritizes accurate information [15]. The collection took place over a period of 59 days, coinciding with the announcement of the very first COVID-19 vaccine. The average collection frequency is 589 comments per day. This data was recorded in a ".csv" file and made publicly available on the Kaggle platform. As part of our methodology, the Agenda-Setting Theory was incorporated to better control external variables. This theory posits that the media influences the agenda by selecting topics that will be brought to the public's attention. By choosing the *r/Coronavirus* subreddit, focused on accurate information related to COVID-19, we aimed to minimize the impact of irrelevant external variables, enabling a more targeted analysis of discussions. **Table 4** presents the 13 columns displaying different properties of each comment (author, title, comment content...).

Table 4. Data description.

Attribute	Type	Description
post_id	object	post identifier
post_author	object	post author
post_date	object	date of post
post_title	object	title of the post
post_score	integer	post score
post_permalink	object	post permalink
post_url	object	post link

Continued

comment_id	object	comment identifier
comment_author	object	comment author
comment_date	object	date of the comment
comment_parent_id	object	comment parent identifier
comment_edited	bool	edited or not
comment_score	integer	comment score
comment_body	object	comment body

3.3. Data Pre-Processing

The process of preparing uncleaned and unlabelled data for labelling by different algorithms involves several steps, including data cleaning, tokenisation, lemmatisation and strimming. Algorithm 1 has the output data ready for the labelling process.

Algorithm 1. Cleaning and processing reddit comments.

```
Require: Uncleaned comments.
Ensure: Cleaned and pretreated comments.
For each comments in the dataset do.
    Apply the Text cleaning process.
    Apply the Tokenization function.
    Apply the Lemmatization function.
    Apply the Strimming function.
End for.
```

3.4. Data Labelling

There are several approaches to sentiment analysis [28], including: a lexicon-based approach (which uses pre-labeled word lexicons to assess the sentiment of a text); a rule-based approach (which uses linguistic rules to identify sentiment in a text); a statistical methods-based approach (which uses supervised, unsupervised, and semi-supervised classification techniques to label sentiment in texts); and a neural network-based approach (which uses deep neural networks to learn complex feature representations from texts and to predict sentiment example transformers). We will use two different approaches, namely TextBlob (rule-based approach) and Twitter-RoBERTa-Base-sentiment (neural network-based approach), to label our dataset.

3.4.1. TextBlob

To classify sentiments with TextBlob, it is necessary to use the concepts defined previously, as illustrated in Algorithm 2.

3.4.2. RoBERTa

Twitter-RoBERTa-Base-sentiment is a supervised classification approach how uses three main steps to classify the sentiment such as the pretraing step, the embedding step and a classification step as described in Algorithm 3.

Algorithm 2. TextBlob sentiment classification [29].

```
Require: Reddit comments.
Ensure: Sentiment label (positive, negative, or neutral).
For (each comments in the dataset) do.
    For (each word in the comments) do.
        Repersented it with its part of speech using POS tagging.
        Look up the polarity score of it in a sentiment lexicon dictionary.
    End for.
    Compute the average polarity of all the words in the comment.
    If (average polarity > 0.05) then label the comment with a positive sentiment.
    Else if (average polarity ∈ [−0.05, 0.05]) then label the comment with a neutral
    sentiment.
    Else label the comment with a negative sentiment.
    End if.
End for.
```

Algorithm 3. Twitter-RoBERTa-Base-sentiment sentiment classification [30].

```
Require: Reddit comments.
Ensure: classification of each comment as (positive, negative, or neutral).
Token Encoding.
Apply token embedding to each token in each comment to represent its semantic
    meaning.
Benary Classification.
Ternary Classification.
For Reddit comments that have been ranked, use another fully connected neural
    network layer to rank each Reddit comment as neutral, positive or negative.
```

3.5. Vectorization

3.5.1. TF-IDF (Term Frequency-Inverse Document Frequency) [31]

The TF-IDF is a statistical method that evaluates the importance of a term in a comment by taking into account its frequency in the comment and its rarity in the whole collection of comments. The TF-IDF formula can be written as follows:

$$TF\text{-}IDF(t,d) = TF(w,c) \times IDF(w) \tag{1}$$

where:

- w represents the word under consideration;
- c represents the comment in which the word appears;
- $TF(w,c)$ represents the frequency of the word w in the comment c *i.e.* is the number of times w appears in c, calculated as follows:

$$TF(w,c) = \frac{\text{count of } w \text{ in comment } c}{\text{number of word in } c}, \tag{2}$$

- $IDF(w)$ is the inverse document frequency of the word w, calculated as follows:

$$IDF(w) = \log\left(\frac{N}{df(w)}\right), \tag{3}$$

- N represents the total number of comments in the collection;

- $df(w)$ represents the number of comments in the collection that contain the word w.

3.5.2. CountVectorizer

Let C be the set of n comments, and W be the set of m unique words in the comment set C. The word count matrix X is an $n \times m$ matrix defined as follows:

$$X_{i,j} = \text{number of occurrences of the words } j \text{ in the document } i. \tag{4}$$

The matrix X can be created from the comment set C using the CountVectorizer method. The method counts the number of occurrences of each term in each comment, and stores these counts in the matrix X. The result is a dense matrix representation of the comment collection C.

3.6. Training and Model Selection

To propose a classification model for comments on COVID-19, we start with balanced and labeled data. We then train seven classification algorithms: Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), XBoosting (XB), and AdaBoost (AB) using 80% of the data set.

Next, we evaluate the performance of these algorithms using five classification metrics: accuracy, recall, precision, F1-score, and AUC. This evaluation is conducted on the remaining 20% of our data.

For a given classification problem, a confusion matrix M corresponds to a square matrix, and whose input M_{ik} is the number of examples of class i for which label k has been predicted. True positives are correctly classified positive examples; false positives are negative examples labelled positive by the model; and vice versa for true negatives and false negatives. The number of true positives is generally noted TP, the number of true negatives TN, the number of false positives FP, and the number of false negatives FN.

Accuracy: A measure used to evaluate the performance of multi-class classification models. It describes the overall performance of the model and measures the rate of good predictions on all comments.

$$\text{Accuracy} = \frac{TP + TF}{TP + TF + FP + FN}. \tag{5}$$

Recall: The rate of true positives, *i.e.* the proportion of positive examples correctly identified as such:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{6}$$

Precision: The proportion of correct predictions among positive predictions is called Precision, or Positive Predictive Value (PPV). Proportion of correct predictions among positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{7}$$

F-measure: The F1-score is the harmonic mean of precision and recall:

$$\text{F1-score} = 2\frac{\text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}} = \frac{2TP}{2TP + FN + FP}. \quad (8)$$

The ROC (Receiver Operating Characteristic) curve is a plot that displays the relationship between the true positive rate and false positive rate across a range of classification thresholds.

The AUC (Area under the Curve) of the ROC curve is a metric that evaluates the overall performance of a classification model. This measure is obtained by calculating the Area under the Curve (AUC) of the ROC curve, which estimates the ability of the model to differentiate between positive and negative classes. An AUC score close to 1 indicates a high performance of the model, while a score close to 0.5 indicates that the model performs as well as a random choice.

3.7. Topic Modelling

The use of topic modelling is an effective method for identifying underlying topics present in a collection of our comments that express negative sentiment towards vaccines. The main objective of this unsupervised machine learning method is to identify the main concerns or reasons that may contribute to hesitancy towards COVID-19 vaccines. By uncovering these hidden issues, we will be able to better understand the factors that influence the public's vaccination decisions and formulate targeted strategies that address these concerns, which will ultimately increase vaccine acceptance.

For that, we are going to use the LDA (Latent Dirichlet Allocation) model, how is a probabilistic generative model (used to model the process by which words are generated in a text corpus), the fundamental assumptions of this method include that each comment in the corpus is a predetermined mixture of topics, that each topic represents a probability distribution over the words in the vocabulary, that the distribution of topics in a document and the distribution of words in a topic follow a Dirichlet distribution, that the choice of words in one comment is independent of the choice of words in the other comments, and that the choice of topics in one comment is independent of the choice of topics in the other comments.

However, the LDA algorithm works as described as follow: a first step that consists of choosing the number of subjects K that the model will represent and then initializing the values of the model parameters (the subject-word distribution, the document-subject distribution, and the alpha and beta hyperparameters), then an iterative step of updating the subject-document and subject-word distributions and carrying out until convergence of a fixed number of iterations initially defined, then the last step that produces as final parameters the resulting subject-word and subject-document distribution.

4. Results and Discussion

In this section, we will present the tools used, the results obtained from the methodology adopted, and finally analyze the various results. To do so, we will unveil in Section 4.1 the tools and programming languages used, in Section 4.2 the results of the comments cleaning, in Section 4.3 the results of the comments labeling, in Section 4.4, we will present the results of the training and evaluation of the sentiment classification models, in Section 4.5, we will detail the results of the LDA, and finally, in Section 4.6, we will compare this study to the most recent works addressing the same issue.

4.1. Tools and Programming Languages Used

To obtain the current results, we used two programming languages: Python and R. In Python, we used several general purpose packages such as numpy, pandas, matplotlib, seaborn, pycountry, emoji, re, string, demoji, spacy, nltk, networkx, collections, langdetect, WordCloud and pyLDAvis.sklearn for data manipulation, processing and visualization. Packages related to sentiment analysis and natural language processing, such as TextBlob and transformers, allowed us to label sentiments. The sklearn and utils packages were used for importing classes (TfidfVectorizer, CountVectorizer), which allowed us to vectorise our comments, as well as for importing classification models (Logistic Regression, K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier and XGB Classifier) and evaluation metrics (accuracy_score, precision_score, recall_score, f1_score and roc_auc_score), which were useful for training and evaluating the model. Finally, the Latent Dirichlet Allocation packages and the KMeans class from sklearn were useful for topic modeling. In R, we used the dplyr, ggplot2 and forcats libraries for the visualization of the most frequent words of the topics.

4.2. Text Cleaning and Preprocessing

Initially, the acquired dataset contains 14 attributes that mainly describe posts and comments. For reasons of simplification and limitations related to personal information such as author name, age and location, we focused on attributes such as post and comment publication dates, post titles and comment content. Thus, we collected 34,768 comments from 11,723 different users, related to the 290 publications published on the subreddit *r/Coronavirus*. During the cleaning step, we converted the title and comment data into strings, and removed unnecessary URLs, spaces and special characters. Next, we converted all the text data to lower case. We also identified the 345 emojis present in the comments (see Figure 3) and replaced each one with its corresponding code to facilitate processing.

Figure 3. Emojis present in the corpus of comments.

Finally, the cleaned data was tokenised and lemmatised in order to prepare it for analysis.

Once the cleaning and normalisation were completed, we proceeded with the statistical analysis of the corpus of titles and comments. For the titles, we used the *langdetect* package to detect the language used and we obtained that 99.99% of our titles are written in English. Indeed, the *detect*() method of langdetect uses trigrams to calculate the probability that the text is written in a specific language. Furthermore, a statistical analysis of the headline corpus, visible in **Table 5**, shows that the most frequently used word is, "vaccine", with 213 occurrences. The other most frequently used words are "covid", "doses", "vaccines", "pfizer", "us", "oxford", "moderna" and "pfizerbiontech", which allude to the semantic field of the COVID-19 vaccine.

The observation is almost the same with the comments corpus, which is composed of an initial vocabulary of 34,347 words. We have determined that 99.99% of the comments are written in English, which allows us to perform a unilingual analysis. Thus, **Table 6** shows that the word "vaccine" is the most frequent with a count of 10,316 occurrences. We also find words such as "people", "get", "like", "dont", "vaccines", "would", "removed", and "dose", which may at first sight indicate a sentiment expressed towards vaccinations.

Having determined the language and context of our corpus, as illustrated in the word cloud in **Figure 4**, we can further analyse it. Analysing the number of words contained in each comment, **Table 7** shows that comments have an average length of 35.41 words, with a standard deviation of 44.52. This indicates considerable variation in the length of comments, with shorter and longer than average comments. The first quartile is 10, meaning that 25% of comments are 10 words or less in length. The median is 23, meaning that 50% of the comments are 23 words or less in length. The third quartile is 45, meaning that 75% of the comments are 45 words or less. We then counted 2180 comments, or 6.27% of the number of comments initially collected. Therefore, we have 32,588 comments which will be used for the sentiment analysis in the next section.

Table 5. The 30 most frequent words in the corpus of titles.

Word	Count	Word	Count	Word	Count	Word	Count
vaccine	213	uk	22	vaccination	13	johnson	11
covid	154	million	21	data	12	rollout	11
doses	37	get	18	emergency	12	second	10
says	37	new	18	vaccinated	12	start	10
vaccines	36	people	18	receive	12	effective	9
pfizer	34	oxford	17	health	12	efficacy	9
us	33	fda	17	dose	12	americans	9
first	32	week	14	use	11	next	9
coronavirus	30	astrazeneca	14	could	11	administered	8
moderna	27	days	14	trial	11	pfizerbiontech	8

Table 6. The 30 most frequent words in the corpus of comments.

Word	Count	Word	Count	Word	Count	Word	Count
vaccine	10,316	think	2837	getting	1992	much	1780
people	8446	know	2674	removed	1967	want	1766
get	5979	us	2579	vaccinated	1929	good	1751
like	4302	even	2506	data	1916	dose	1731
dont	4240	thats	2438	well	1892	really	1700
vaccines	4136	going	2330	see	1890	many	1688
would	3943	first	2294	need	1887	make	1647
im	3681	time	2194	take	1863	virus	1615
one	3533	still	2163	could	1829	go	1594
covid	3097	also	2147	doses	1809	pfizer	1582

Table 7. Summary statistics of the number of words in the comments.

Statistics	Value	Description
Number of observations	34,768	Total number of comments
Total	1,221,484	Total number of comments
Mean	35.41	Mean number of words in all the comments
Standard deviation	44.52	Standard deviation of number of words in the comments
Minimum	0	Minimum of number of words in the comments
First quartile	10	Number of words of first quartile of comments
Median	23	Median number of words in the comments
Third Quartile	45	Number of words of third quartile of comments
Maximum	1536	Maximum number of words in the comments

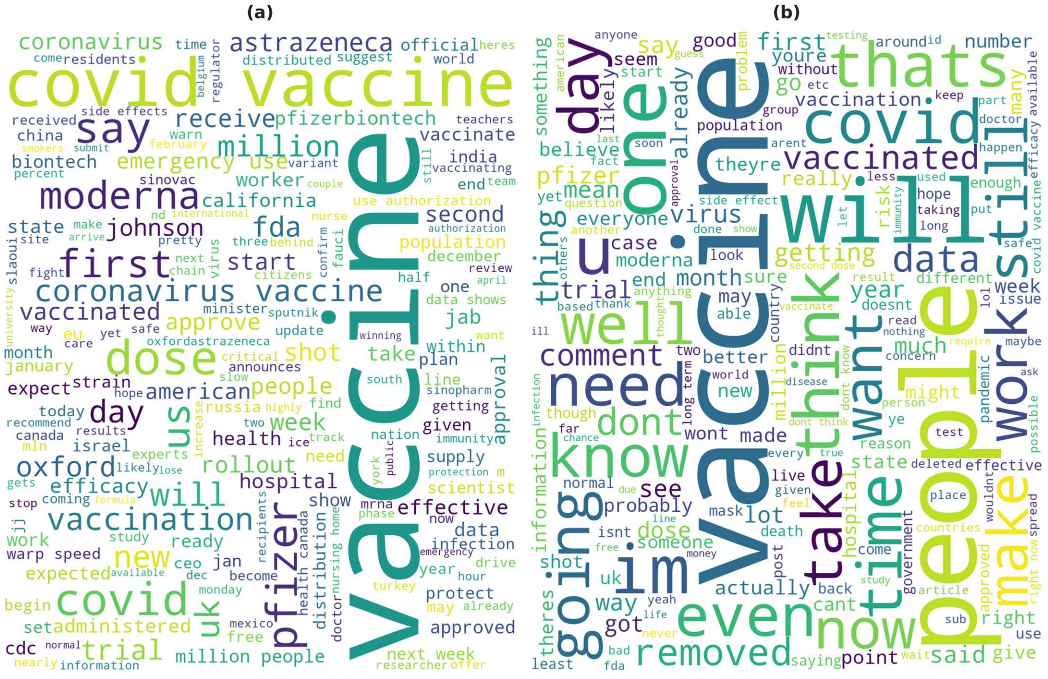

Figure 4. (a) Word cloud of the 290 titles in our corpus; (b) Word cloud of the comments in our corpus.

4.3. Text Labelling

For the labelling of comments, we labelled the texts into three classes (positive, negative and neutral) using two different approaches: TextBlob and RoBERTa. Using the lexicon-based approach with the TextBlob library gave us the following results: 15,653 positive comments, 10,844 neutral comments and 5521 negative comments, as illustrated in **Figure 5(a)**. This means that under the TextBlob approach, fewer comments collected are vaccine averse. We also used the machine learning method with the Twitter-RoBERTa-Base transformation model to label the comments. We obtained 14,708 comments classified as neutral, 13,181 comments classified as negative and 4129 comments classified as positive, as shown in **Figure 5(b)**. This means that according to the RoBERTa approach, almost half of the collected comments are vaccine averse. This difference in sentiment distribution in the two algorithms can be explained by the differences between TextBlob and Twitter-RoBERTa-Based sentiment, the former being a dictionary-based approach and the latter using transformation models.

However, when we intersected the results of the two approaches, we found 6087 comments labelled neutrally by both approaches, as well as 3666 comments labelled negative and 3212 comments labelled positive by both approaches. These results indicate that the two approaches were able to agree on a higher proportion of negative comments than positive comments, as shown in **Figure 5(c)**.

Nevertheless, it is important to note that these differences in labelling do not answer the question of how sentiment evolves over time. To examine this question, we have plotted in figures the evolution of sentiment under the different approaches, both for TextBlob and RoBERTa (cases in **Figure 6** and **Figure 7**, and then for the intersection of the two approaches in **Figure 8**). The result is almost the same: positive and negative feelings evolve in a sawtooth pattern and remain almost constant over the 59 days, with extreme peaks on all three curves. It is therefore legitimate to ask which posts generated the most positive and negative comments. To answer this question, we have constructed **Table 8** and **Table 9**, which show the posts that generated the most positive and negative comments for the two approaches separately, and for the intersection of the two approaches.

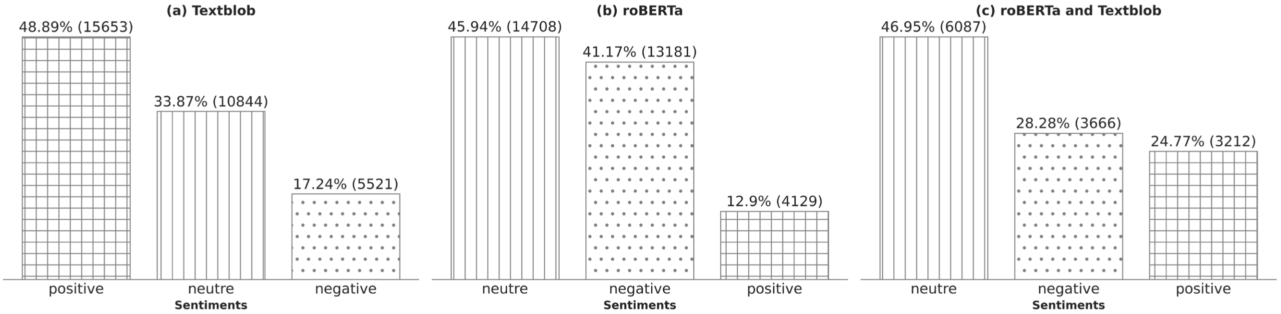

Figure 5. (a) Proportion of TextBlob labeling; (b) Proportion of Twitter-RoBERTa-Base sentiment labeling; (c) Proportion of common labeling between TextBlob and Twitter-RoBERTa-Base sentiment.

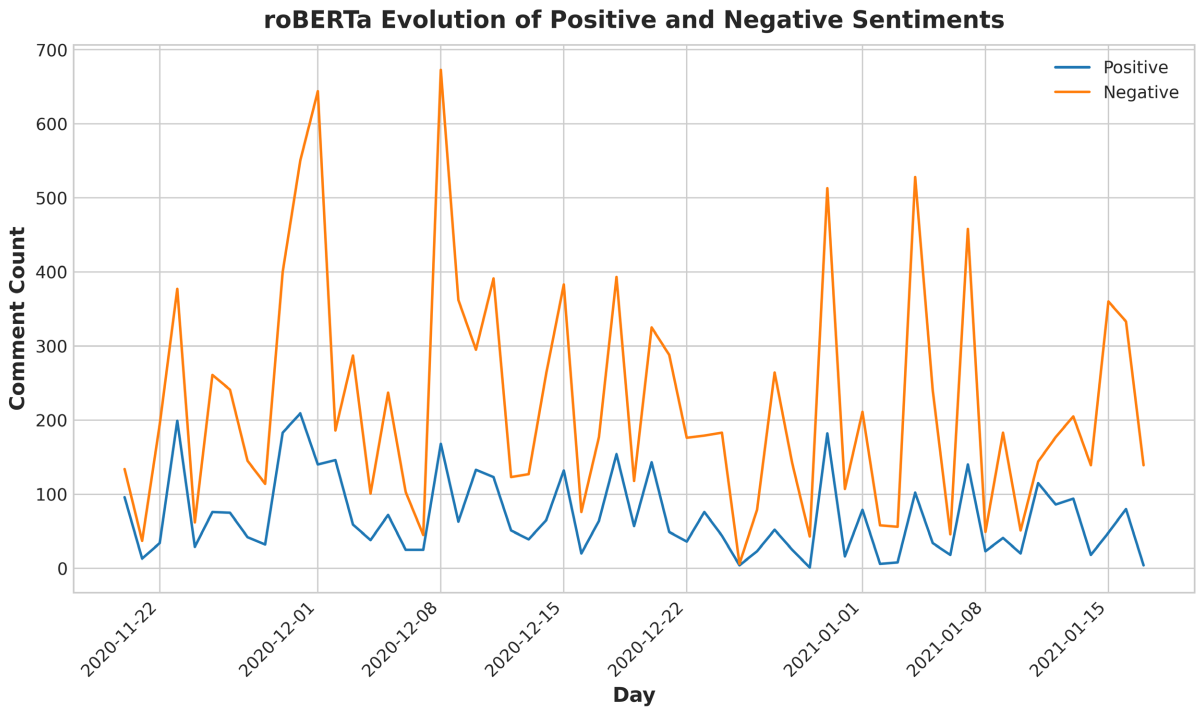

Figure 6. TextBlob evolution of positive and negative sentiments.

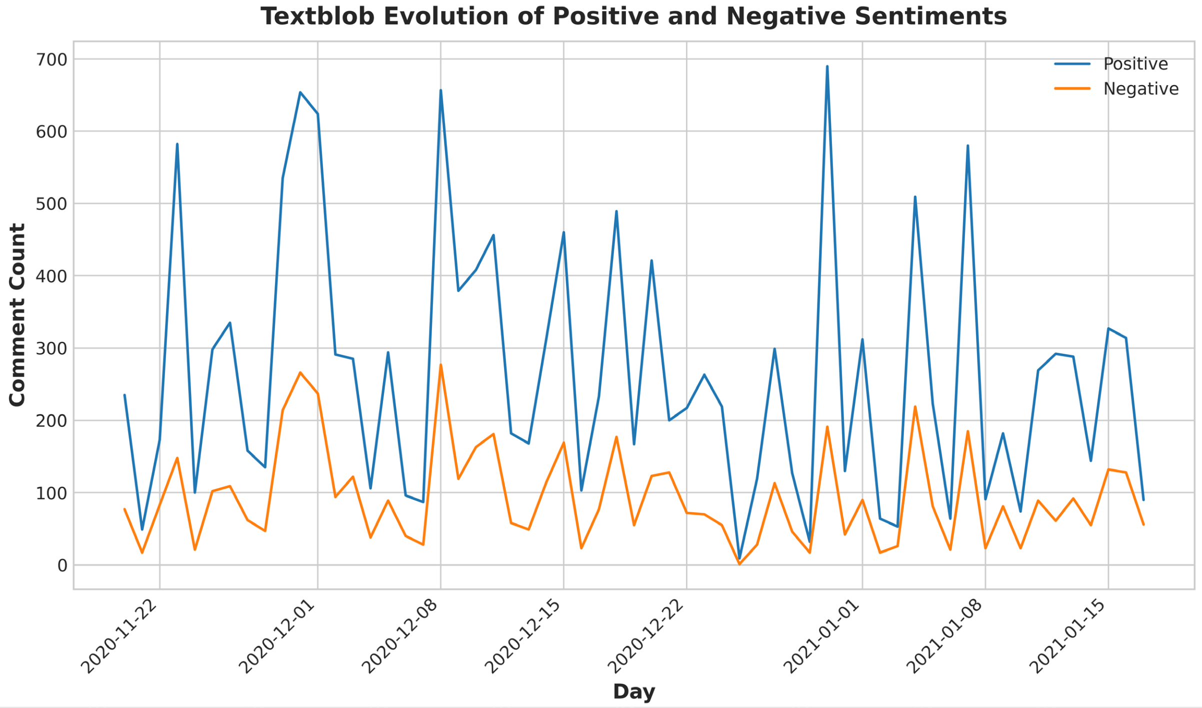

Figure 7. RoBERTa evolution of positive and negative sentiments.

Table 8 shows the five most criticised post titles in terms of negative comments using three approaches (TextBlob, RoBERTa and a combination of both). The results showed that controversial topics such as COVID-19 vaccines and vaccination priorities generated the most negative comments. For example, the

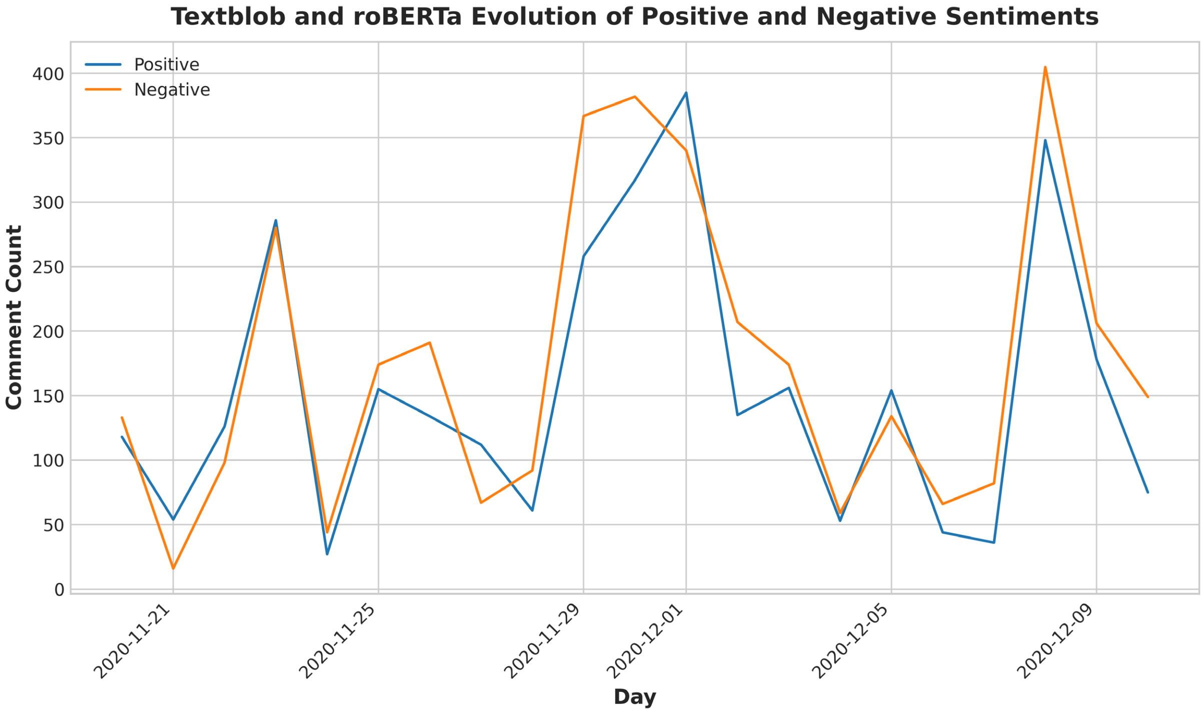

Figure 8. TextBlob and RoBERTa evolution of positive and negative sentiments.

Table 8. The 5 worst post titles that have received a large number of negative comments.

Approach	Title	Negative	Total
Texblob	Moderna says new data shows COVID vaccine is more than 94% effective, plans to ask FDA for emergency clearance later Monday	111	469
	Here's Why Vaccinated People Still Need to Wear a Mask—The new vaccines will probably prevent you from getting sick with COVID. No one knows yet whether they will keep you from spreading the virus to others but that information is coming	107	413
	In New Jersey, smokers can now get the coronavirus vaccine before teachers or public transit workers	95	448
	Pfizer says it has second doses of COVID-19 shot on hand, expects no U.S. supply problems	93	451
	Pfizer's Coronavirus vaccine arrive in Chicago O hare international airport	88	462
RoBERTa	In New Jersey, smokers can now get the coronavirus vaccine before teachers or public transit workers	278	448
	Don't want the COVID-19 vaccine? You could lose access to normal life, says U.K. minister	237	419
	People who suffer from "significant" allergic reactions should not take Pfizer vaccine, UK regulators warn	233	437
	Here's Why Vaccinated People Still Need to Wear a Mask—The new vaccines will probably prevent you from getting sick with COVID. No one knows yet whether they will keep you from spreading the virus to others but that information is coming	230	413
	Pfizer says it has second doses of COVID-19 shot on hand, expects no U.S. supply problems	217	451
Texblob and RoBERTa	Here's Why Vaccinated People Still Need to Wear a Mask—The new vaccines will probably prevent you from getting sick with COVID. No one knows yet whether they will keep you from spreading the virus to others but that information is coming	81	162

Continued

	In New Jersey, smokers can now get the coronavirus vaccine before teachers or public transit workers	77	146
	Fact check: Nurse who fainted after COVID-19 vaccine has an underlying health condition	67	132
	Moderna says new data shows COVID vaccine is more than 94% effective, plans to ask FDA for emergency clearance later Monday	65	197
	Don't want the COVID-19 vaccine? You could lose access to normal life, says U.K. minister	64	151

Table 9. The 5 worst post titles that have received a large number of negative comments.

Approach	title	Positive	Total
Texblob	COVID-19: Oxford University vaccine shows 70% protection	288	479
	COVID-19: Oxford/AstraZeneca vaccine approved for use in UK	250	465
	Moderna's vaccine is highly effective, FDA says, clearing way for second vaccine	241	471
	Pfizer and BioNTech to Submit Emergency Use Authorization Re-quest Today to the U.S. FDA for COVID-19 vaccine	235	451
	Moderna says new data shows COVID vaccine is more than 94% effective, plans to ask FDA for emergency clearance later Monday	234	469
	Pfizer's Coronavirus vaccine arrive in Chicago O hare international airport	88	462
RoBERTa	UK authorises Pfizer-BioNTech COVID-19 vaccine	129	448
	COVID-19: Oxford University vaccine shows 70% protection	128	479
	Pfizer and BioNTech to Submit Emergency Use Authorization Re-quest Today to the U.S. FDA for COVID-19 Vaccine	96	451
	Moderna's vaccine is highly effective, FDA says, clearing way for second vaccine	95	471
	First doses of Pfizer coronavirus vaccine has flown to US from Belgium	94	474
Texblob and RoBERTa	COVID-19: Oxford University vaccine shows 70% protection	104	227
	UK authorises Pfizer-BioNTech COVID-19 vaccine	91	239
	First doses of Pfizer coronavirus vaccine has flown to US from Belgium	78	233
	Pfizer and BioNTech to Submit Emergency Use Authorization Re-quest Today to the U.S. FDA for COVID-19 vaccine	77	212
	Moderna's vaccine is highly effective, FDA says, clearing way for second vaccine	71	206

headline "*In New Jersey, smokers can now receive the coronavirus vaccine before teachers or transit workers*" generated the most negative comments under both approaches, with a ratio of 52.73% negative comments to all comments. Table 9 shows, on the other hand, the top five post titles with the highest number of positive comments using three different approaches: Texblob, RoBERTa and a combination of Texblob and RoBERTa. The post titles "*COVID*-19: *Oxford University vaccine shows* 70% *protection*" and "*Pfizer and BioNTech to Submit Emergency Use Authorization Request Today to the U.S. FDA for COVID*-19 *Vaccine*" had a high ratio of positive comments across the three different approaches. This suggests that users found these headlines informative and useful.

This information may be useful in understanding that the public is waiting for updates on vaccine approval and progress on pandemic vaccine research.

However, there are differences between the approaches in **Table 8** and **Table 9**. In **Table 8**, the headline "*Moderna announces new data shows COVID vaccine is more than* 94% *effective and plans to seek emergency FDA approval later Monday*" generated the most negative comments according to TextBlob, but did not make the top 5 for RoBERTa. Furthermore, in **Table 9**, the post title "*Moderna's vaccine is highly effective, FDA says, clearing way for second vaccine*" has a high ratio of positive comments for both Texblob and RoBERTa. However, the title does not appear in the top 5 for the combination of Texblob and RoBERTa.

These results highlight the importance of choosing the right approach to data analysis and taking into account the limitations and biases of each method. It is also important to note that labelling the comments does not give us enough information about the underlying reasons why users' comments were rated as negative. Why are they skeptical about vaccines? In order to address these concerns, we will respectively propose a sentiment classification model based on the sentiments annotated by both approaches in Section 4.4, and finally use topic modeling on the data classified as negative by both approaches to highlight the hidden reasons for the public's hesitation to vaccinate in Section 4.5.

4.4. Training and Evaluation

Taking into account the predictions of both approaches in the previous Section 4, we obtained 6087 neutral comments, 3666 negative comments and 3212 positive comments. As these classes are unbalanced to form a classification model, we balanced the classes by randomly removing 2587 comments from the neutral class (see **Figure 9(b)**). Thus, by performing a statistical analysis on the new dataset, we obtained a vocabulary of 14,890 words and an average number of 16.29 words per comment. The 20 most frequent words belong to the vaccine lexical field (see **Figure 9(a)**), making it an ideal dataset for the vaccine context.

After preparing our data, we used two methods of comment representation: TF-IDF and CountVectorizer. Then, we divided our data into two sets: 80% for training and 20% for testing. We then trained our data on 7 classification algorithms and evaluated the performance using 5 metrics (accuracy, precision, recall, F1-score and AUC). **Table 10** presents the evaluation results of the 7 sentiment classification models, indicating that the XGB and RF models are the best models for comment classification on our dataset, especially when using CountVec embedding. Both models obtain high values of accuracy, precision, recall, F1-score and AUC. A comparison of the two confusion matrices (**Figure 10**) shows that there are differences in the performance of the XGB and RF models. Although both models achieve high accuracy, the RF model has higher values for predicting positive and negative comments, while the XGB model manages to classify neutral comments better. However, XGB has more true positives than RF,

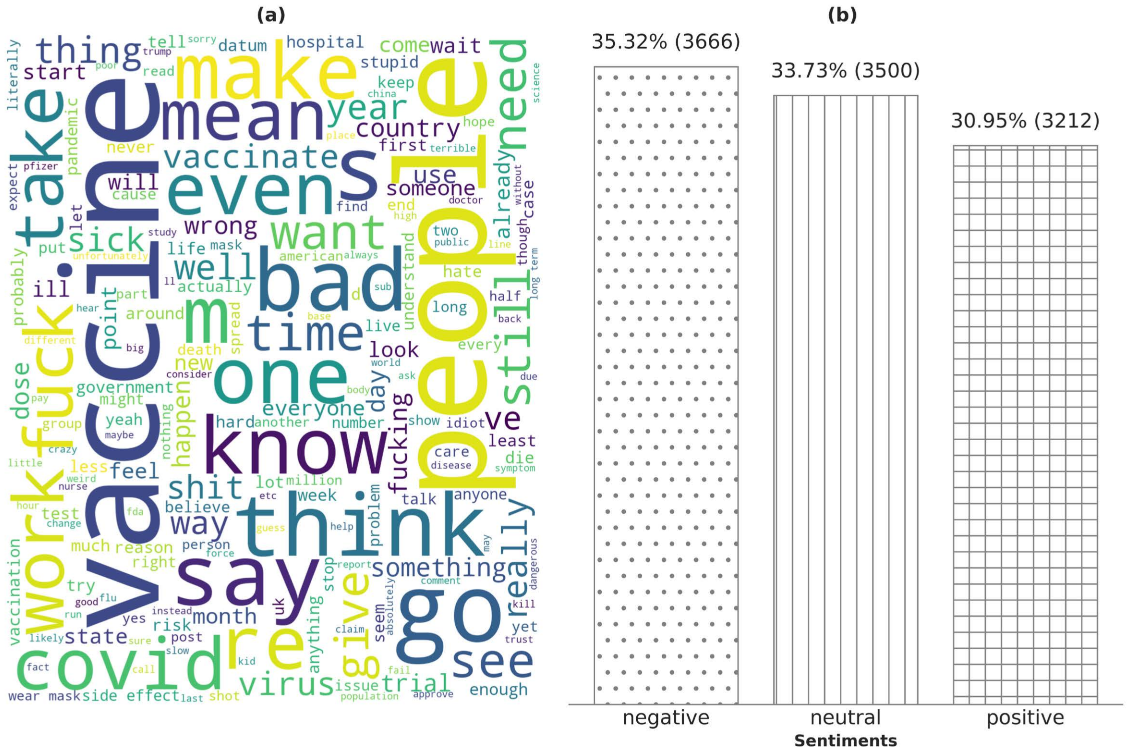

Figure 9. (a) Wordcloud of the most frequent words in the balanced dataset; (b) Proportions of the different classes in the balanced dataset.

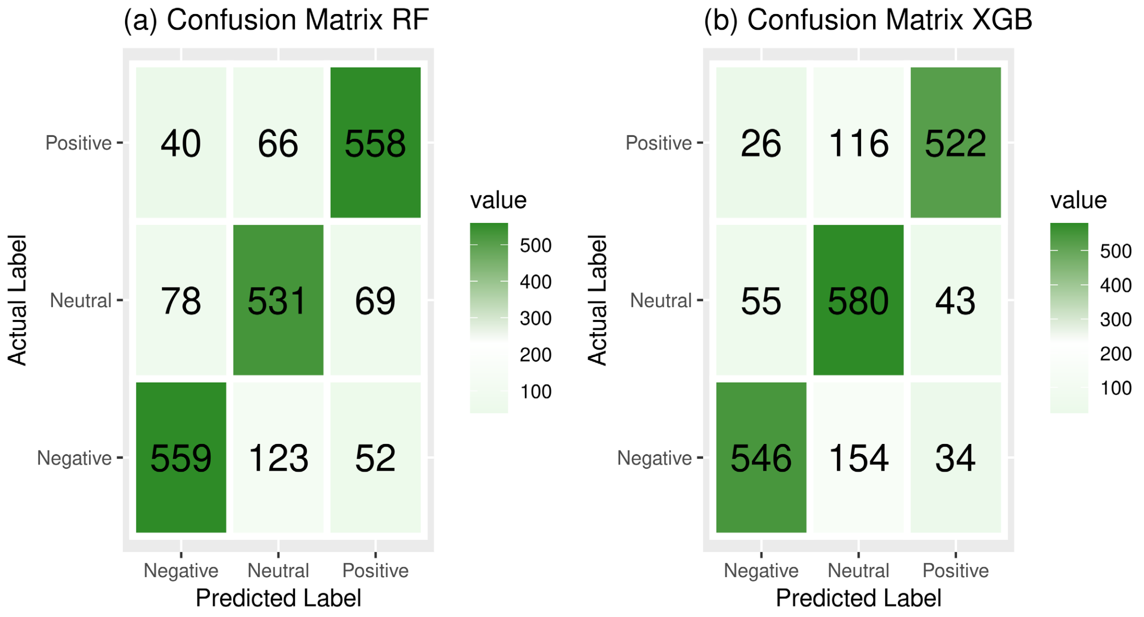

Figure 10. (a) Random Forest (RF) confusion matrix; (b) XGBoost (XGB) confusion matrix.

which justifies the fact that the XGB model has higher accuracy. In summary, the results of this study suggest that both XGB and RF models are effective for COVID-19 related sentiment analysis tasks on our dataset, and that the choice of embedding has a significant impact on the models' performance.

Table 10. Results of the evaluation of the 7 classification models used.

Embedding	Algorithm	Accuracy	Precision	Recall	F1-score	AUC
TF-IDF	LR	56.55	57.48	56.55	56.67	76.01
	KNN	50.29	51.22	50.29	49.38	67.12
	DT	45.76	45.74	45.76	45.70	59.02
	RF	54.82	54.87	54.82	54.64	73.70
	AB	52.99	53.11	52.99	52.58	71.41
	XGB	56.94	57.08	56.94	56.77	75.27
	GB	57.71	58.08	57.71	57.43	75.23
CountVec	LR	76.35	76.77	76.35	76.40	88.66
	KNN	55.44	65.91	55.44	53.92	77.68
	DT	73.65	73.72	73.65	73.68	80.24
	RF	79.38	79.56	79.38	79.40	92.28
	AB	70.09	76.61	70.09	70.36	82.93
	XGB	79.38	80.95	79.38	79.60	92.25
	GB	72.93	76.05	72.93	73.25	88.01

4.5. Topic Modeling and Recommendation

Following the extraction of the 3666 comments classified as negative by the two approaches described in Section 4, we applied the LDA method to perform a topic modeling analysis. We tested different configurations of LDA, using different numbers of topics (20, 10 and 7) and different iterations (1000, 10000 and 20,000). However, the LDA model with 7 subjects and 20,000 iterations obtained the best perplexity score, equal to 918. Figure 11 shows the 10 most frequent words in each topic. The 7 topics were identified with the following keywords:

The most common words in topic 0 are “people”, “die”, “death”, “vaccine”, “COVID”, “vaccinate”, “case” and “million”. From these terms, it appears that the topic refers to mortality from COVID-19 as well as people’s reluctance to be vaccinated, which may indicate that people are concerned about the deadly consequences of COVID-19.

The most recurrent words in topic 1 are “people”, “mask”, “sick”, “wear”, “ill”, “work” and “time”. These phrases suggest that the topic is about pandemic-related health issues such as mask wearing, sickness and absenteeism from work.

The most frequent expressions in topic 2 are “vaccine”, “bad”, “effect”, “long”, “make”, “virus”, “year”, “time” and “know”. Based on these words, it can be assumed that the topic addresses the long-term side effects of the COVID-19 vaccine and may show that the public is concerned about the likely long-term side effects of the vaccine and needs more information about it before making a decision.

The most used words in topic 3 are “vaccine”, “people”, “say”, “think”, “want”,

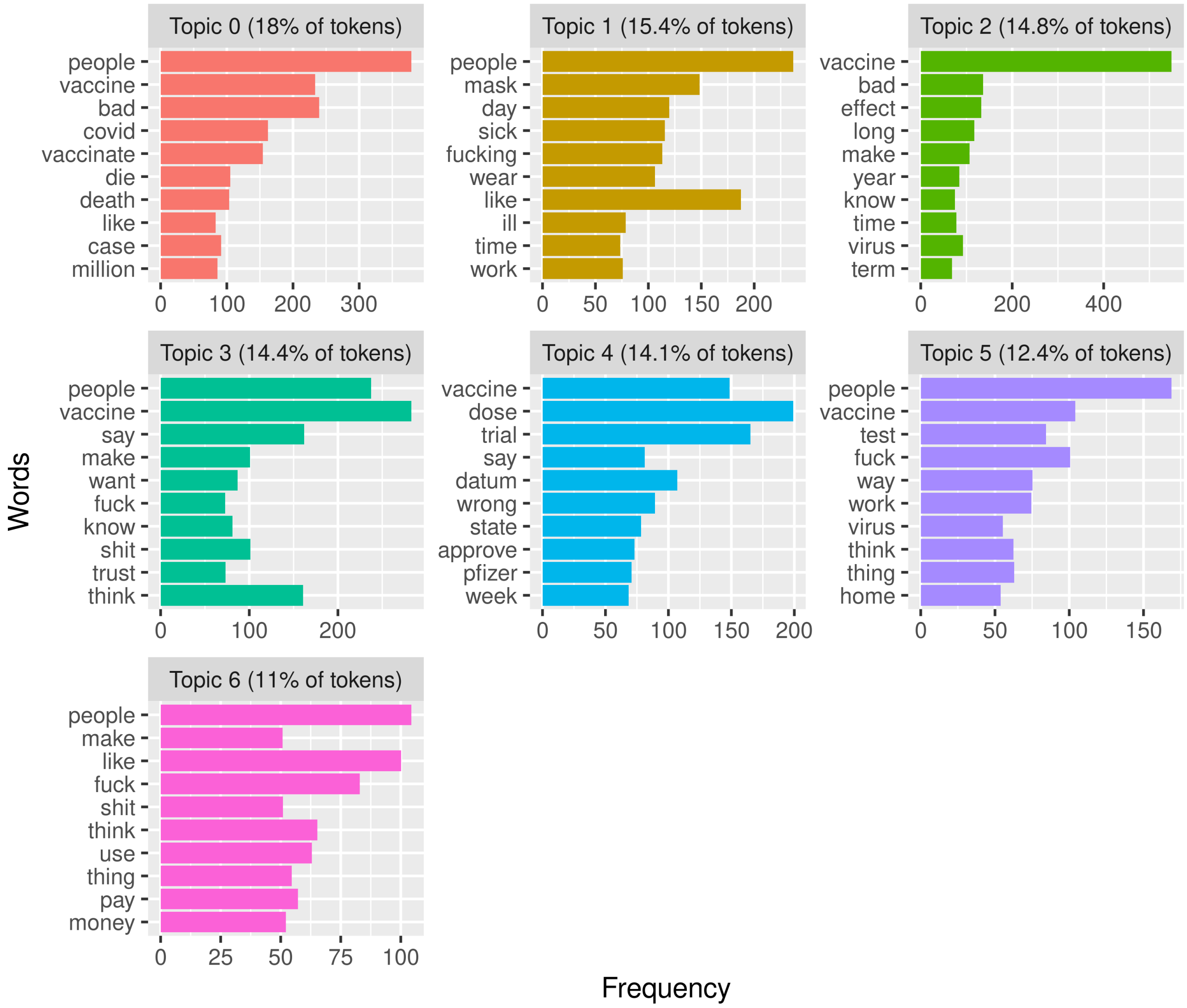

Figure 11. The 10 most frequent words per topic in our corpus after applying LDA.

"know", "trust" and "fuck". These words suggest that the topic is about the public's opinion of the COVID-19 vaccine and may indicate the public's distrust of the vaccine's safety and effectiveness.

The most common terms in topic 4 are "dose", "trial", "vaccine", "data", "wrong", "say", "state", "approve", "pfizer" and "week". It is possible to deduce from all of these words that the topic is about the COVID-19 vaccine trials and approval, and may show that the public is concerned about the number of doses and the vaccine trials, as well as the approval of the vaccines by the authorities.

The most common words in topic 5 are "people", "vaccine", "test", "way", "work", "thing", "think", "virus" and "home". These words suggest that the topic is about the testing of the COVID-19 vaccine and how it works, and may indicate that the audience wants more information about how the vaccine is tested and how it works.

The dominant words in topic 6 are "people", "like", "pay", "money", "shit", "make", "thing" and "use". Analysing these phrases, it is plausible that the topic refers to the costs of the COVID-19 vaccine and may show that the public is concerned about the costs of the vaccine and wants more information about its availability.

However, this is a summary idea about the topics and only a reading and analysis of the comments actually containing the most frequent words in each topic will give us in-depth information about the topics and shed light on the assumptions made. With this in mind, we have classified the comments into seven themes and summarised the comments according to the most frequent words, these results are recorded in Table 11.

We therefore identified 23 sub-themes within the 7 initial topics given by the LDA. We then divided them into 6 distinct groups, as shown in Table 11. The aim here is to be able to make recommendations according to the challenges encountered.

Table 11. Sub-topic appearing in each topic after effectively analyzing the comments related to the most frequent words in the 7 comment clusters obtained by the LDA.

		Topic						
Challenges	**Sub-topic**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Group 1	Frustration with pandemic restrictions	×						
	Frustration caused by the spread of fake news about vaccines							×
	Skepticism towards vaccines and prevention measures	×	×	×				
Group 2	Conspiracy theories related to COVID vaccines		×			×		×
	Long and short-term side effects associated with the vaccine	×		×	×	×	×	
	Risks to children getting vaccinated				×			
Group 3	Criticisms of public health policies		×					
	Concerns about vaccine confidentiality issues		×					
	Priority questions in vaccination					×	×	
	Concerns about mandatory vaccination						×	
	Problems related to access to healthcare							×
	Policy and the medical community in relation to vaccines							×
Group 4	Challenges to large-scale vaccine deployment		×		×	×		×
	Difficulties in accessing vaccination for certain populations						×	
	Corruption in vaccine distribution			×				
	Criticism of the idea of charging people for the vaccine						×	
Group 5	Views on religion			×				
	Negative views on capitalism			×				
Group 6	Questions about vaccine approval before distribution				×			
	Government delays in approving vaccines					×		
	Problems with vaccine dose management and wastage					×		
	Differences between vaccines and their effectiveness					×		
	Concerns about virus mutations					×		

For the Group 1 challenges of public frustration with pandemic-related restrictions and conflicting information about the vaccine, policy-makers need to communicate clear and accurate information about vaccines using regular and reliable communication channels. They also need to explain the scientific reasons for any restrictions and educate the public on how to detect fake news.

Challenges in Group 2 include conspiracy theory related to COVID-19 vaccines, short and long term side effects, and the risks associated with the vaccine and for children who are vaccinated. One set of actions would be to communicate scientific information to dispel myths and misinformation, to provide accurate information on the possible side effects and risks associated with vaccination, and to educate parents on the importance of vaccinating their children for the protection of all.

Group 3 has a variety of challenges that consist of criticisms of public health policies, concerns about vaccine confidentiality, issues of priority in vaccination, concerns about compulsory vaccination and finally concerns about the policy and medical community in relation to vaccines. In this case, several measures could be put in place, such as, the creation of vigilance committees representative of the population and health professionals, responsible for communicating and listening to the public's criticisms, in order to provide clear explanations of each health policy adopted and to involve health professionals in the implementation of these policies. It would also be important to assure the public of the security of their personal data by opting for an anonymity strategy. Finally, an accessible and equitable vaccination strategy should be developed, prioritizing those most at risk and justifying this priority with sound scientific evidence.

Group 4 is various challenges such as, difficulties in rolling out the vaccine on a large scale, difficulties in accessing vaccination for certain populations, corruption in vaccine distribution and criticism of the idea of charging people for the vaccine. A possible solution could be to develop well-designed roll-out strategies to ensure effective distribution and access to vaccination for vulnerable populations. Finally, there is a need to communicate transparently about the costs of vaccine production and distribution in order to prevent corruption.

Group 5 consists of elements such as, views on religion and negative views on capitalism. In this case, it would be wise to recommend that public policy makers work with religious leaders to promote vaccination and dispel myths and misconceptions about the vaccine. Finally, it is good to educate the public about the importance of public-private collaboration.

Finally, Group 6 includes concerns about vaccine approval, government delays in approving vaccines, vaccine dose management, differences in vaccine efficacy, and concerns about virus mutations. To address these concerns, it would be useful to communicate the rigorous process of vaccine approval by regulatory authorities, to develop strategies to ensure effective and equitable management of vaccine doses, to provide clear and accurate information about the different vaccines available and their respective effectiveness, and to educate the public about how vaccines can be adapted to deal with new variants of the virus.

4.6. Comparison and Generatlisation of This Study

Several authors, through their respective works, have examined the issue of hesitancy towards the COVID-19 vaccine on the Reddit platform in various ways. These studies stand out due to their focus on specific subreddit communities. A notable example is represented by the research of Duraivel *et al.* [32], who explored Reddit corpora generated by users, particularly focusing on the subreddits r/askreddit, r/antivax, r/antivaccine, and r/AntiVaxxers. Similarly, this study and the one conducted by Tan and Datta [33] analyze the content of the "r/Coronavirus" subreddit to understand skepticism within the community at the beginning of the pandemic. The data collection periods differ, ranging from November 20, 2020, to January 17, 2021, for this study, and from January 20, 2020, to January 31, 2021, for Tan and Datta's study [33]. Additionally, Tan and Datta's data analysis methods [33] involved using VADER to classify comments as negative or positive, followed by the application of the LDA method to determine underlying reasons for vaccine refusal. In contrast, this study combines TextBlob and Twitter-RoBERTa-Base to classify comments into three categories (positive, negative, and neutral) before applying the LDA method to negative comments. This study, like Tan and Datta's [33], observes an equivalent ratio between positive and negative sentiments in the studied community. Ultimately, this study discovered 23 distinct topics for which the population was skeptical, while Tan and Datta [33] discovered 20.

Our study's methodology, focused on sentiment analysis using advanced NLP techniques within a specialized Reddit community, offers a versatile framework applicable to a variety of social contexts. This approach can be effectively adapted to analyze public opinion in different domains, such as political discourse, consumer behavior, or other health-related issues on various social media platforms. The potential for this methodological approach to provide insightful data across diverse social situations underscores its value. While specific to our study's context, the underlying principles of our methodology have broader applicability, offering a robust tool for researchers examining public sentiment in different online environments.

5. Conclusion

In this study, we classified comments on the COVID-19 vaccine as positive, negative, or neutral. These comments were collected from the "*r/Coronavirus*" subreddit dedicated to pandemic-related posts. The results indicated that ensemble methods like Random Forest (RF) and XGBoost (XGB), trained on comments vectorized with CountVectorizer, proved to be the best classifiers, yielding an average score of approximately 80% across all performance metrics (precision, recall, sensitivity, and F1-score). Additionally, employing the Latent Dirichlet Allocation (LDA) algorithm on negative comments from both labeling methods helped identify 23 challenges explaining public vaccine hesitancy. We grouped these challenges into six categories and provided sets of recommendations for each challenge group. Finally, these recommendations primarily involve elements, such as transparent communication about vaccine-related decisions and policies, fairness in vaccine distribution, inclusion and representation of civil society actors and healthcare professionals in decision-making, as well as the improvement and update of current healthcare systems overall.

6. Limitations and Future Works

While our analysis has provided valuable insights into the consistent sentiment within our specialized Reddit group, it's important to acknowledge certain limitations. Focusing on contextualizing sentiment data within our stable environment may potentially limit the generalizability of our findings to other online communities. Additionally, excluding external factors, such as major events related to the approval and distribution of COVID-19 vaccines, could be considered a limitation as it might impact the broader applicability of our results.

For future work, it would be beneficial to delve further into these limitations and assess the transferability of our approach to different social media contexts. Additionally, investigating the influence of external events on sentiment dynamics within specialized online communities could provide a more comprehensive understanding. These considerations will contribute to refining our methodology and expanding the scope of our findings in future research endeavors.

Acknowledgements

This work is the outcome of the scholarship provided by AIMS-Cameroon. We wish to convey our heartfelt gratitude to the founders, the management team, including our mentors, our classmates, and the Christian Gathering group.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Qorib, M., Oladunni, T., Denis, M., Ososanya, E. and Cotae, P. (2023) COVID-19 Vaccine Hesitancy: Text Mining, Sentiment Analysis and Machine Learning on COVID-19 Vaccination Twitter Dataset. *Expert Systems with Applications*, **212**, Article ID: 118715. https://doi.org/10.1016/j.eswa.2022.118715

[2] De Bruin, Y.B., Lequarre, A.-S., McCourt, J., Clevestig, P., Pigazzani, F., Jeddi, M.Z., Colosio, C. and Goulart, M. (2020) Initial Impacts of Global Risk Mitigation Measures Taken during the Combatting of the COVID-19 Pandemic. *Safety Science*, **128**, Article ID: 104773. https://doi.org/10.1016/j.ssci.2020.104773

[3] de Bruin, Y.B., *et al.* (2020) Initial Impacts of Global Risk Mitigation Measures Taken during the Combatting of the COVID-19 Pandemic. *Safety Science*, **128**, Article ID: 104773. https://doi.org/10.1016/j.ssci.2020.104773
https://www.sciencedirect.com/science/article/pii/S0925753520301703

[4] Andre, F.E., Booy, R., Bock, H.L., Clemens, J., Datta, S.K., John, T.J., *et al.* (2008) Vaccination Greatly Reduces Disease, Disability, Death and Inequity Worldwide. *Bulle-*

tin of the World Health Organization, **86**, 140-146.
https://doi.org/10.2471/BLT.07.040089

[5] Zhao, S.H., Hu, S.M., Zhou, X.Y., Song, S.H., Wang, Q., Zheng, H.Q., *et al.* (2023) The Prevalence, Features, Influencing Factors, and Solutions for COVID-19 Vaccine Misinformation: Systematic Review. *JMIR Public Health and Surveillance*, **9**, e40201. https://doi.org/10.2196/40201

[6] Kwon, S. and Park, A. (2023) Examining Thematic and Emotional Differences across Twitter, Reddit, and YouTube: The Case of COVID-19 Vaccine Side Effects. *Computers in Human Behavior*, **144**, Article ID: 107734.
https://doi.org/10.1016/j.chb.2023.107734

[7] Van Poucke, M. (2023) Lockdown Scepticism: Australian and American Doom Discourse on Reddit. *Studies in Communication Sciences*, **23**, 201-221.
https://doi.org/10.24434/j.scoms.2023.02.3322

[8] Melton, C.A., Olusanya, O.A., Ammar, N. and Shaban-Nejad, A. (2021) Public Sentiment Analysis and Topic Modeling Regarding COVID-19 Vaccines on the Reddit Social Media Platform: A Call to Action for Strengthening Vaccine Confidence. *Journal of Infection and Public Health*, **14**, 1505-1512.
https://doi.org/10.1016/j.jiph.2021.08.010

[9] Vernikou, S., Lyras, A. and Kanavos, A. (2022) Multiclass Sentiment Analysis on COVID-19-Related Tweets Using Deep Learning Models. *Neural Computing and Applications*, **34**, 19615-19627. https://doi.org/10.1007/s00521-022-07650-2

[10] Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F. and Cuenca-Jiménez, P.-M. (2023) A Review on Sentiment Analysis from Social Media Platforms. *Expert Systems with Applications*, **223**, Article ID: 119862.
https://doi.org/10.1016/j.eswa.2023.119862

[11] Abiola, O., Abayomi-Alli, A., Tale, O.A., Misra, S. and Abayomi-Alli, O. (2023) Sentiment Analysis of COVID-19 Tweets from Selected Hashtags in Nigeria Using VADER and Text Blob Analyser. *Journal of Electrical Systems and Information Technology*, **10**, Article No. 5. https://doi.org/10.1186/s43067-023-00070-9

[12] Nuzhath, T., Tasnim, S., Sanjwal, R.K., Trisha, N.F., Rahman, M., *et al.* (2020) COVID-19 Vaccination Hesitancy, Misinformation and Conspiracy Theories on Social Media: A Content Analysis of Twitter Data.
https://doi.org/10.31235/osf.io/vc9jb

[13] Kumar, N., Corpus, I., Hans, M., Harle, N., Yang, N., McDonald, C., *et al.* (2021) COVID-19 Vaccine Perceptions: An Observational Study on Reddit. Laboratory Press, Cold Spring Harbor. https://doi.org/10.1101/2021.04.09.21255229

[14] Wu, W., Lyu, H.J. and Luo, J.B. (2021) Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story. *Health Data Science*, **2021**, Article ID: 9837856. https://doi.org/10.34133/2021/9837856

[15] Kumar, N., Corpus, I., Hans, M., Harle, N., Yang, N., McDonald, C., *et al.* (2022) COVID-19 Vaccine Perceptions in the Initial Phases of US Vaccine Roll-Out: An Observational Study on Reddit. *BMC Public Health*, **22**, Article No. 446.
https://doi.org/10.1186/s12889-022-12824-7

[16] Chinnasamy, P., Suresh, V., Ramprathap, K., Jency, B., Jebamani, A., Srinivas Rao, K. and Shiva Kranthi, M. (2022) COVID-19 Vaccine Sentiment Analysis Using Public Opinions on Twitter. *Materials Today: Proceedings*, **64**, 448-451.
https://doi.org/10.1016/j.matpr.2022.04.809

[17] Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A. and Mittal, A. (2020) Social Media Analysis with AI: Sentiment Analysis Techniques for the Analysis of Twitter

COVID-19 Data. *Journal of Critical Reviews*, **7**, 2761-2774.

[18] Yeskuatov, E., Chua, S.-L. and Foo, L.K. (2022) Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques. *International Journal of Environmental Research and Public Health*, **19**, Article 10347. https://doi.org/10.3390/ijerph191610347

[19] Liu, S.R. and Liu, J.L. (2021) Public Attitudes toward COVID-19 Vaccines on English-Language Twitter: A Sentiment Analysis. *Vaccine*, **39**, 5499-5505. https://doi.org/10.1016/j.vaccine.2021.08.058

[20] Luo, Y.S. and Kejriwal, M. (2022) Understanding COVID-19 Vaccine Reaction through Comparative Analysis on Twitter, Intelligent Computing. *Proceedings of the* 2022 *Computing Conference*, **1**, 846-864. https://doi.org/10.1007/978-3-031-10461-9_58

[21] Bengesi, S., Oladunni, T., Olusegun, R. and Audu, H. (2023) A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion from Twitter Tweets. *IEEE Access*, **11**, 11811-11826. https://doi.org/10.1109/ACCESS.2023.3242290

[22] Pennebaker, J.W., Francis, M.E. and Booth, R.J. (2001) Linguistic Inquiry and Word Count: LIWC 2001. Lawrence Erlbaum Associates, Mahwah, 3-21.

[23] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[24] Yin, J.H. and Wang, J.Y. (2014) A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering. *Proceedings of the* 20*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 24-27 August 2014, 233-242.

[25] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ArXiv: 1810.04805.

[26] Lande, J., Pillay, A. and Chandra, R. (2023) Deep Learning for COVID-19 Topic Modelling via Twitter: Alpha, Delta and Omicron. *PLOS ONE*, **18**, e0288681. https://doi.org/10.1371/journal.pone.0288681

[27] Yurtsever, M.M.E., Shiraz, M., Ekinci, E. and Eken, S. (2023) Comparing COVID-19 Vaccine Passports Attitudes across Countries by Analysing Reddit Comments. *Journal of Information Science*. https://doi.org/10.1177/01655515221148356

[28] Madhoushi, Z., Hamdan, A.R. and Zainudin, S. (2015) Sentiment Analysis Techniques in Recent Works. 2015 *IEEE Science and Information Conference* (*SAI*), London, 28-30 July 2015, 288-291. https://doi.org/10.1109/SAI.2015.7237157

[29] Loria, S. (2020) TextBlob Documentation, Release 0.16.0. https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf

[30] Barbieri, F., Camacho-Collados, J., Espinosa Anke, L. and Neves, L. (2020) TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *Findings of the Association for Computational Linguistics*, Online, November 2020, 1644-1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

[31] Salton, G. and Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, **24**, 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[32] Duraivel, S. and Lavanya, R. (2021) Understanding Vaccine Hesitancy with Application of Latent Dirichlet Allocation to Reddit Corpora. https://doi.org/10.21203/rs.3.rs-616664/v1

[33] Tan, Z. and Datta, A. (2023) The First Year of the COVID-19 Pandemic through the Lens of *r*/*Coronavirus* Subreddit: An Exploratory Study. *Health and Technology*, **13**, 301-326. https://doi.org/10.1007/s12553-023-00734-6