Analyzing sentiment and content of Covid-19 vaccine tweets

Classification and Summarization in Natural Language Processing

Final Examination Paper - Report
Natural Language Processing and Text Analysis

Submitted by	Kitti Kresznai (141359)
	Doris Natalia Cornejo Rojas (142880)
	Anastasiya Vitaliyivna Strohonova (142820)
	Natalie Schober (141354)
Submitted on	15 June 2020
Number of pages	15
Number of characters	34,125

Abstract

Tweets on the topic of Covid-19 vaccines are reviewed in terms of sentiment and content. The goals of sentiment classification and text summarization in the form of sentence extraction and text generation are addressed. A better understanding of the models best suited for identifying and understanding large bodies of fragmented texts is hoped to be achieved.

Among the Naïve Bayes, Support Vector Machine, and Logistic Regression, the latter two, which are discriminative models, were discovered to be most suitable for classifying the sentiment of tweets. As a continuation, an LSTM model is used to attempt text generation. However, sentence extraction is found to be better suited in this scenario for providing a simple overview of ideas prevalent within the Tweets. Therefore, we recommend discriminative models for classification problems, and sentence extraction as a sufficient way of obtaining a summary in this case.
Keywords: Covid-19, Sentiment Analysis, Twitter, Classification, Neural Network, Summarization

Contents

1 Introduction

The first mass vaccination program for Covid-19 started in December 2020, and by now, there are about a dozen different vaccines available all around the world. As the possibility of getting vaccinated became widely available to the public, it also gained more and more attention on social media, and the attitudes expressed by people regarding the topic varied greatly. The spread of misinformation and vaccine skepticism on different platforms like Twitter and Facebook is a highly dangerous phenomenon, which can cause harm by escalating the number of confirmed Covid-19 cases and mortality rates. We decided to investigate this topic to get an insight into Twitter's users' attitude towards Covid-19 vaccination and identify the best routes for developing an overview of sentiments and topics.

Motivation

Several studies show that social media has a successful impact in spreading information regarding diseases, e.g. H1N1, Ebola, Zika [Fung et al., 2016] [Avery, 2017] [Sharma et al., 2017]. Considering that Covid-19 is one of the most lethal pandemics in history and that the virus has been mutating [Feehan and Apostolopoulos, 2021], it is relevant to investigate how people are reacting and obtain insights related to the vaccination process. A two-step approach to analyzing public response is through sentiment analysis and text summarization techniques.

Research Question

Formulating from the context presented, the objective is to analyze tweets related to covid vaccines. Precisely, this paper will intend to address the ensuing research questions:

- What is the overall sentiment relating to Covid-19 vaccinations compared to that for certain vaccines?
- Which model is best suited for classifying the sentiment of tweets?
- Is a model able to produce reliable statements/summaries related to tweets regarding Covid-19?

Structure of Paper

The subsequent sections of the paper are ordered as follows. Section 2 will review related work. Then, section 3 will introduce the conceptual framework considering pertinent concepts related to the topic of the paper and the algorithms applied. Section 4 will analyze the methodology considered for achieving the objectives, while section 5 will refer to the results of the different analyzes. Finally, section 6 and 7 will contain a final discussion of the results in terms of the research questions, and the conclusion of the paper with a few recommendations to consider in further analysis.

2 Related Literature

Most of the published articles analyzing people's attitudes and opinions toward Covid-19 vaccines on social media platforms were made on a locational basis. Related work has been found regarding Australian, Indian, UK, and US citizens, which indicates that NLP analyses of this topic have only been conducted on the English language so far.

The study by [Hussain et al., 2021] used natural language processing and deep-learning based techniques to predict sentiments and topics from a dataset of 300,000 social media posts related to Covid-19 vaccines. The dataset contained both Facebook posts and tweets, posted by users in the US and the UK. The data was thematically filtered for Covid-19 and vaccine related keywords, as well as geographically filtered for the two above-mentioned countries. After the preprocessing of the filtered data, a hybrid ensemble-based model was built for thematic sentiment analysis. It resulted in around 60% positive, 20% negative, and 20% neutral sentiments. Furthermore, topics like public optimism over vaccine development, safety concerns, and vaccine effectiveness were identified.

Another article published by [Praveen et al., 2021] concentrated on recognizing the attitudes of Indian citizens towards Covid-19 vaccines. The applied methods were sentiment analysis and LDA topic modeling, and the collected data contained 73,760 unique English-language tweets. According to their findings, neutral and positive toned tweets dominated Indian Twitter in the last 4 months of 2020 regarding the vaccination. Moreover, the study pointed out a strong correlation between the increase in positive sentiments and an increase in the number of Covid-19 infections. The LDA topic modeling was performed on tweets that had negative sentiments, and the leading identified topics were fear over health, allergic reactions, and fear of death.

[Thelwall et al., 2021] focused on Covid-19 vaccine hesitancy present on English-language Twitter, and it applied content analysis to a random sample of 446 Covid-19 tweets that were deemed to demonstrate fear towards the vaccine. Three major vaccine hesitation topics were identified in this work: conspiracies, development speed, and safety. These accounted for about half of the collected Tweets and were mostly from the US. Here are some examples of arguments present in the tweets that were mentioned in the article: vaccines contain microchips to control the population, the approval process of the vaccines is rushed, and it might be unsafe because of that, or it does not make sense to take the vaccine as some got mild Covid-19 from it as a side effect.

3 Conceptual Framework

This section presents the machine learning methods used for the sentimental analyses: Naïve Bayes, Support Vector Machine, Logistic Regression, and for text summarization: LSTM and sentence extraction.

3.1 Naive Bayes

According to the definition of [Jurafsky and Martin, 2014], Naïve Bayes is a probabilistic linear classifier, which given a document, returns the class that has the maximum posterior probability for that document. It is based on the equation of the Bayes' rule, as it defines how to decompose the conditional probability into other probabilities. Moreover, Naïve Bayes utilizes a simplistic bag-of-words approach that does not take the order of words into account and assumes conditional independence of the feature probabilities. Smoothing, such as the popular Laplace smoothing method, is usually applied, to avoid the problem emerging from the use of the Maximum Likelihood Estimate. This occurs when a certain word has not been seen in the training documents before, resulting in zero probability. The additive Laplace smoothing parameter is represented by α in formula 1, where α equals 0 stands for no smoothing [Jurafsky and Martin, 2014].

$$P(w_k \mid c_j) = \frac{n_k + \alpha}{n + \alpha \mid Vocabulary \mid} \tag{1}$$

Since the model is fast and robust to irrelevant features, it was considered for this classification problem. However, the independence assumption may not hold for some texts and impact the accuracy of the model.

3.2 Support Vector Machine

The Support Vector Classification (SVC) model is generally well suited for various classification problems. An SVC model creates a hyperplane that aims to maximize the margin between the classes. One of the benefits of SVC is the possibility of using a kernel trick to create a higher dimensional space to linearly separate data that may not be linearly separable in its original dimension [Géron, 2019]. C is the regularization parameter, while gamma is the kernel coefficient and defines how many points influence the decision boundary. Both C and gamma can lead to over-or underfitting of the model, making parameter tuning important [Murphy, 2012].

3.3 Logistic Regression

A Logistic Regression Model generates a conditional probability that a sample belongs to a certain class based on its corresponding feature values [Zhou et al., 2019]. Although the probability the model generates is continuous, it serves as a classification algorithm by adding decision thresholds. The output of the weighted sum is passed through a sigmoid function to obtain the probability of the instance to be in a certain class. When training data is fit using Maximum Likelihood Estimation, the weights are obtained [Zhou et al., 2019]. The regression may be regularized with L1 or L2 regularization.

Comparison of the classifiers Comparing the above-mentioned classifiers to each other, the main difference can be found in the fact that Naïve Bayes is a generative classifier, while Logistic Regression and Support Vector Machine are discriminative [Jurafsky and Martin, 2014]. A discriminative model's focus is on finding a decision boundary, whereas a generative model replicates the classes' distribution. In the end, they both aim to predict conditional probabilities, but they learn different probabilities during the training process. Nevertheless, all these models are examples of supervised machine learning methods that can be used for classification tasks. Their strengths and weaknesses lie in different places, and the optimal choice is highly dependent on the problem at hand as well as on the features of the dataset.

3.4 Long Short-Term Memory

With recurrent neural networks (RNN), it is possible to predict the next word in a sequence with older words holding a decaying amount of relevance, since RNNs allow for information to persist through loops. This is an improvement upon traditional neural networks, and Long Short-Term Memory networks are a type of RNN with added complexity. LSTM is especially suited for text generation because it can address when an older word should continue to hold importance [Jung, 2018].
The cell of an LSTM network contains an explicit memory state, which is regulated by the input, forget, and output gates. These gates control changes to the memory state such as what and how much information is retained, which helps to avoid vanishing gradient problems [Koehn, 2017]. An LSTM model can be built using Keras, with layers added to define its forward pass. An optimizer, batch size, and number of epochs is specified. These parameters respectively work to reduce loses, determine how many data samples are incorporated in an iteration, and how many times the model adjusts its weights.

3.5 Text Summarization

Summaries can be developed in the form of keywords most relevant to the text, sentence extraction, and generating new sentences to summarize the text. Keywords provide limited insight, while summary generation using machine learning models is an inherently difficult task, and therefore sentence extraction will be explored as an alternative for providing more accessible insights. Models of sentence centrality, such as those based around a goal centroid or finding central sentences based on cosine distance, are used in combination with word weights to determine important sentences based on their position or score. Word weights can be obtained using tf-idf, log-likelihood ratio, or external vectors such as GloVe [Jurafsky and Martin, 2014].

4 Methodology

4.1 Data Acquisition and Cleaning

The dataset has been obtained with the use of Python's snscrape library, which requires Python version 3.8. Snscrape is suitable for scraping various social network services (SNS), hence the naming. For this analysis, it was used to scrape Tweets from Twitter throughout December 2020 and May 2021, for the first day of each month. We chose this period as the rollout for mass vaccination has started in early December 2020, and we then covered every month uniformly until May 2021, to be able to show any progression that occurred in this timeframe. Furthermore, different keywords were used for the text query in order to obtain a dataset with a larger variety, as listed here: 'covid vaccine', 'moderna', 'pfizer', 'astrazeneca', and 'sputnik v'. The search query collected 500 Tweets per keyword per month, so the initial data being input contains 15,000 rows overall.

As the data was received in a raw state from Twitter, many cleaning steps were required. We first filtered the Tweets down to only those in English, which reduced the dataset down to 8,126 rows, and included some additional columns. For example, a CountryCode column was added, whose value was extracted from the initial column called place. However, as this information was missing in about 98% of cases, we decided not to include location-related analysis in this study.
Further steps were taken to clean the Tweets themselves. Cleaning the text puts it into a standardized and simplified format which allows for the learning models to be better able to utilize it. The following steps were taken:

i **Remove usernames** in the format of @user, which are used throughout twitter to refer to a specific account. They are unique and do not benefit our analysis.

ii **Remove webpage links** which our learning models will not be able to interpret.

iii **Remove "\n" and "\xa0"** HTML code incorporated throughout the Tweets.

iv **Remove non-alphanumeric characters** such as punctuation, the hashtag symbol, and emojis.

v **Lowercase**. Text was made to be lowercase all over since ML models are case-sensitive.

vi **Remove repeating letters** beyond two repeating letters. This helps to catch some misspellings although the applications are limited since two repeating letters must be allowed since it occurs in correctly spelled words, e.g. "smoooooth" will be corrected to "smooth", but "waaaay" only to "waay".

vii **Expand contractions and standardize Covid-19 name** to simply "covid", to standardize the text.

Although emojis often convey sentiment and could be beneficial to sentiment analysis, they have been removed and the focus will be on the sentiment of the language used. The nltk stopwords library was not used for the sentiment analysis portion, since emitting some of the words included could alter the meaning of the text. However, it was later applied for the improvement of the sentence extraction and text generation models. Stemming and lemmatization could also influence how sentences are interpreted for sentiment analysis. However, lemmatization was applied to the text to optimize the word cloud results. A few custom stopwords that do not carry sentiment in our case were also removed (such as "covid").

4.2 Sentiment Analysis

Data Labelling

To label the tweets, we use the Python library TextBlob. TextBlob assigns a sentiment score in the range of [-1,1] to each tweet, where -1 stands for negative and 1 for positive sentiment. In figure 1, the average sentiment scores are displayed with respect to time and vaccination brand. As the average sentiment scores are in a range of [0.05, 0.14], the tweets tend to be more positive than negative on average.

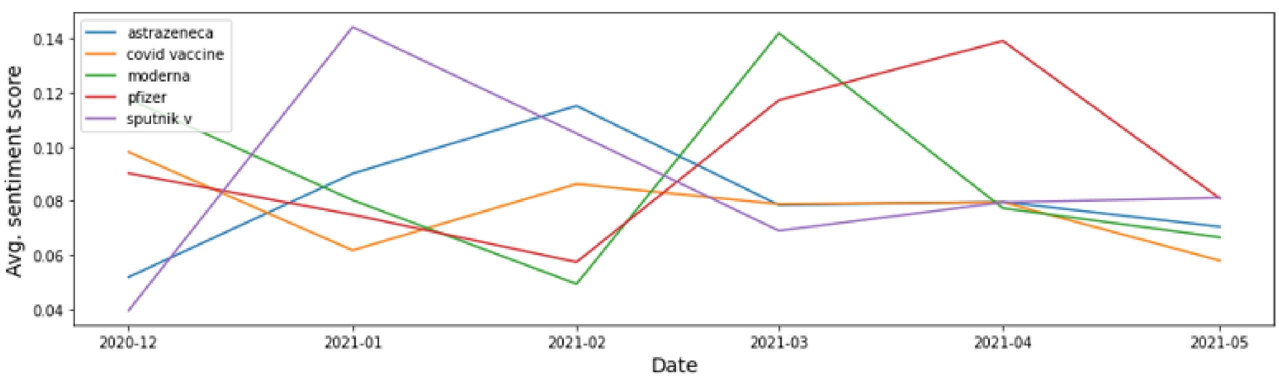

Figure 1: Variance of sentiment with respect to time and vaccination brand

As a next step, the sentiment scores [-1,1] obtained from TextBlob are grouped into three sentiments: Scores are classified as either negative [-1,-0.1), neutral [-0.1,0.1] or positive sentiment (0.1,1]. To understand which common topics arise in the tweets and how these topics are perceived, word clouds are plotted. Figure 2 displays the most frequent negative, neutral and positive words used in the tweets.

Figure 2: Frequent words used to express negative, neutral and positive sentiment in tweets

In the tweets classified as negative sentiment, words such as ‘sick’, ‘cold’ and ‘ill’ were frequently used, referring to the side effects of the vaccination. Other common negative words are ‘government’, ‘fail’, ‘behind’, and ‘shipments’, which indicate that the people in the tweets often complain about the government’s vaccination plan. On the other hand, people are grateful about the (government’s) vaccination plan indicated by words such as ‘available’, ‘effective’ and ‘free’ in the positive word cloud. Also, ‘first’, ‘happy’ and ‘great’ occur frequently in tweets classified as positive. In the neutral word cloud, rather informative and objective words such as each vaccine’s name, ‘commence’ and ‘approve’ are prevalent.

Feature extraction using Count Vectorization After labeling the tweets as positive, neutral, and negative sentiment, we used vectorization to convert texts into numerical feature vectors. This strategy of tokenizing, counting and normalizing is called ‘Bag of Words’ or ‘Bag of n-grams’ representation. Texts are described by word occurrences while the position of the words in the texts and sentences is ignored. A popular feature extraction technique is the Count Vectorization which produces a sparse representation of word counts. In the sparse model, each dimension corresponds to a word in the vocabulary. Count Vectorization returns integers while other sparse models, such as tf-idf returns floats. We decided to use Count instead of tf-idf Vectorization, as models such as the Multinomial Naïve Bayes require integer instead of fractional feature counts. [Jurafsky and Martin, 2014]

Perform training-test split The data is then split into 80% as training and 20% as test set. Splitting before applying models ensures that the models are only fit to the training but not to the test set, as otherwise this would be data leakage.

Tackle class imbalance using SMOTE When looking at the number of tweets per sentiment class, it is clear that 50% of the tweets are assigned to the positive sentiment, while only 35% and 15% are assigned to neutral and negative sentiment, respectively. To counteract this class-imbalance, the Synthethic Minority Oversampling Technique (SMOTE) is introduced. As the name suggests, it oversamples minority classes by generating synthetic examples to improve their recognition. SMOTE is only applied on the training set, as original data must be used for testing the model [Chawla et al., 2002].

Train model using SMOTE training set Based on the feature set, three machine learning classifiers, namely Multinomial Naïve Bayes, Support Vector Machine and Multinomial Logistic Regression are applied. They are all used to predict sentiment categories acquired using the TextBlob library for each tweet. As there are three sentiment labels present in the dataset, multinomial instead of binomial models are used as they can have more than two outcomes. GridSearch was used in order to find the optimal parameter configuration setting. Of the options presented, the ones found best suited for the three classification problems are presented in figure 3. For the Naïve Bayes model, the alpha parameter has been tuned to determine the required level of additive Laplace smoothing. The best kernel, C, and gamma was chosen for our SVC, and the L2 regularization strength C and solver parameter used for the optimization algorithm for our Logistic Regression model.

Classifier	**Parameter**
Naïve Bayes	alpha = 0.25
Support Vector Machine	C = 1, kernel = linear, gamma = 0.01, random_state = 42
Logistic Regression	C = 10, solver = saga, random_state = 42

Figure 3: Hyperparameters tuned and chosen by GridSearch

Test model and evaluate its performance using Classification metrics Based on the trained models, the test set is used for testing and predicting. The evaluation of the different models is conducted with the help of various classification metrics, such as Confusion Matrix, Accuracy, Precision, Recall, and F1-measure. The confusion matrix visualizes the relationship between the true and predicted labels, giving the ratio for the true and false positives and negatives that occurred during the testing. The accuracy score gives the percentage of all tweets that are labeled correctly. Precision represents the percentage of selected items that are correct, while Recall displays the percentage of correct items that are selected. A measure that incorporates aspects of both precision and recall is the F-measure. We used the F1-measure here, i.e. when Precision and Recall are equally balanced [Jurafsky and Martin, 2014].

4.3 Text Summarization

Long Short-Term Memory The LSTM model utilizes a cleaned, tokenized, and flattened dataset consisting of the content of the Tweets obtained. Stop words are removed from the text for this portion of the analysis so they are not given a disproportionate amount of weight due to their frequency. The Keras library Tokenizer is fit to the text to generate an index of the words in the text and their frequencies.

The text is then transformed into a sequence of integers, and sub-sequences are created and split between X and y for the training of the model. A pretrained GloVe word embedding consisting of 400,000 words is used to generate an embedding matrix of word weights, and then a model is built. Word embeddings allow for the mapping of words and their similarity based on their context in the text [Koehn, 2017]. The learning model consists of an Embedding, LSTM, and Dense layer. The Embedding layer is used with neural networks based on text data and makes them more efficient using low-dimensional vector representations. Since the output was represented with one-hot encoding, the categorical cross-entropy loss function is selected. The Adam optimizer, a fairly standard and efficient option, is also used. The accuracy metric is used to evaluate the model's performance. A function is defined to simplify the generation of text once the model is built, and allows for a seed text and input for the number of words desired.

Summary Extraction Sentence extraction uses word vectors based on weights, which can be provided by pre-trained GloVe word embeddings, to compare sentences based on the cosine similarity taken for each sentence pair. Although there are other ways of obtaining word weights, with some more customizable to specific cases, the GloVe vector is quite comprehensive and should provide a sufficient base. The large matrix of cosine similarities can then be converted into a graph depicting sentence overlap, for the PageRank algorithm to provide sentence rankings based on importance in the text. The highest-ranking sentences are extracted to serve as the summary of the text. Stop words were also removed here so the model can focus on more important words in the text and use them for ranking. The final summary is converted to a string to be more readable, however, it would still benefit from further formatting and cleaning before reaching a final form. The sentiment is also taken, to assist with the interpretation of the output.

5 Results

What is the overall sentiment relating to Covid-19 vaccinations compared to that for certain vaccines?

In reviewing tweets relating to the Covid-19 vaccines in figure 1, it was found that the sentiment is skewed toward positive with just under half of the tweets collected falling into this category. The sentiment for “covid vaccine” was approximately an average of that of the other vaccines, and more stable over the months covered despite a slight dip as mass vaccination picked up in mid-December 2020. Tweets about the Pfizer vaccine dropped slightly in sentiment leading up to February 2021. Although the cause is difficult to pinpoint, vaccine shortages, poorly handled vaccine roll-outs, and the fear of the vaccine not protecting against variants of the virus might have played a role. After February 2021, the Pfizer and Moderna vaccines started to gain more favor. In the case of the Pfizer vaccine, possibly because it has the highest efficiency rate out of all the Covid-19 vaccines. AstraZeneca was viewed relatively positively until a drop starting in February which corresponds with news of side effects emerging and countries choosing to ban the vaccine. As far as the Sputnik V is concerned, its perception was increasing sharply until January 2021, after then it has turned to a slow but steady decrease. The Russian authorities applied for the registration of the vaccine in the EU in January 2021, this can explain the decrease with people starting to talk more about it in general on social media. Overall, it becomes obvious in figure 1 that the sentiments regarding the vaccine brands considerably vary over time.

Which model is best suited for classifying the sentiment of tweets?

The confusion matrices in Figure 4 show the performance of the three classification models. Based on the confusion matrices, the logistic regression gives the best results, having 41.70% true positives, 26.01% true neutrals, and 12.12% true negatives each. The SVM also shows decent results but produces a disproportionate amount of false negatives. Both the SVM and the Logistic regression faultily predict a positive sentiment for around 5% of the tweets, when they have, in fact, neutral or negative sentiment.

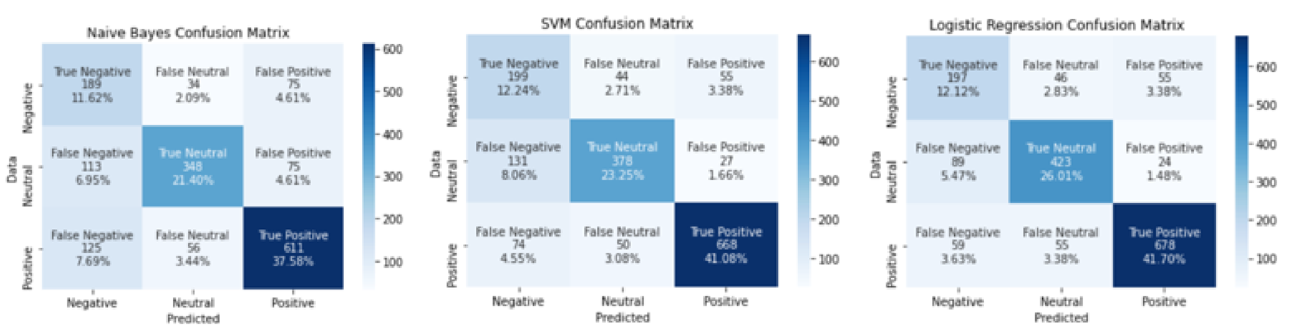

Figure 4: Confusion matrices of the three models

These observations are confirmed by the classification metrics of the three models in figure 5.

	Accuracy	Precision			Recall			F1 Score		
		neg	neu	pos	neg	neu	pos	neg	neu	pos
Naïve Bayes	0.71	0.44	0.79	0.80	0.63	0.65	0.77	0.52	0.71	0.79
Support Vector Machine	0.77	0.49	0.80	0.89	0.67	0.71	0.84	0.57	0.75	0.87
Logistic Regression	0.80	0.57	0.81	0.90	0.66	0.79	0.86	0.61	0.80	0.88

Figure 5: Classification metrics of the three models

Comparing the accuracy scores, the logistic regression model achieved the highest accuracy of 0.80. SVM and Naive Bayes achieved a lower accuracy of 0.77 and 0.71, respectively. The logistic regression model also achieved the highest Precision, Recall and F1-score for 'pos' values. For example, the Precision score for 'pos' shows that out of the tweets labelled as positive, 90% are truly positive. The Recall for 'pos' as positive sentiment class equals 0.86, which means that the 'pos' positive sentiment class has identified 86% of existing positive cases. The performance of the SVM follows close behind with a Precision and Recall score for 'pos' of 89% and 84%, respectively.

Is a model able to produce reliable statements/summaries related to tweets regarding Covid-19? The LSTM model developed for text generation provided an unclear result. When an eight word output was requested, it returned "covid vaccine rollout unacceptably slow tortoise virtually covid vaccine". In its current state, the model has a tendency for repeating words. Although the mention of the vaccine rollout being slow is informative, this unfortunately does not serve the desired purpose of providing a reliable overview. With further optimization and increased training time the model could of course be improved, but summarization based on text extraction provides a simplified alternative that still results in significant insight. The extracted sentences discuss different vaccines, comparing one and two dose vaccines, and criticizing delays in receiving second doses. Health concerns relating to the vaccine, as well as the importance of following authority guidelines is also touched upon. Although this is undoubtably limited given the over 8,000 Tweets within the data set, it provides a brief overview of topics discussed which can likely be considered reliable since sentences are ranked and extracted based on having the greatest similarity to other sentences in the text. The sentiment of the summary is taken simply to understand how it relates to the sentiment analysis performed previously. A polarity of 0.13 and subjectivity of 0.38 is returned. In reading the summary, a judgement can be made as far as to the accuracy of these scores, which can then shed light on the overall sentiment analysis performed.

6 Discussion

The main goal of this study was to identify the changing attitude of Twitter users towards Covid-19 vaccination. The overall sentiment of the tweets was slightly positive on average; however, a certain level of fluctuation was detected, especially in the case of individual vaccines. As we have seen, these fluctuations can be explained by policy decisions (of vaccine launches and suspensions) and news coverage (of side effects) that were made just before such sentiment shifts occurred in the tweets. Towards spring 2021, a drop can be observed in the sentiment scores for every vaccine. This could mean that the general scepticism towards vaccination is increasing, with the spread of misinformation gaining space on Twitter. The phenomenon can be counteracted with educational posts explaining the working of the vaccines in an easily understandable manner on social media, and to draw people's attention not to believe the news they receive from probably unreliable sources.

Investigating the suitability of three supervised models for classifying the sentiment of tweets, we found that the Logistic Regression and SVM show a better performance than the Naïve Bayes. The Naïve Bayes assumes conditional independence of the feature probabilities which may not hold, impacting the accuracy of the model. Also, the Naïve Bayes is a generative classifier, while the Logistic Regression and SVM are discriminative. Discriminative models are often better for sentiment classification because they classify directly rather than solving a general problem as an intermediate step, as generative models do. Based on the good results obtained from discriminative models, we recommend doing further research in this direction, for example, trying out other discriminative models, such as a Random Forest and XGBoost for sentiment classification. Neural Networks could be suitable when classifying large datasets.

There were some limitations we have faced during the analysis. Among them, a prevalent problem was related to the labeling accuracy of the TextBlob library. For example, it categorized tweets like 'I tested positive' with a positive sentiment, which is the exact opposite of what we would have expected regarding the problem we have raised. A solution for this could be the labeling of tweets by hand. Nevertheless, that task can be quite time-consuming, which is why we have decided to work with the TextBlob categories in the end. The ambiguity of the content of tweets can pose another limitation, as the negative and positive sentiments neutralize each other, resulting in an entirely neutral tone. This kind of information loss can be crucial during the interpretation of the results. Furthermore, often there is no clear sentiment of tweets, or they are only noticeable because of an exclamation point or hashtag, or an attached image.

At full potential, an LSTM model would allow for unique summaries generated based on the text provided. However, implementing a sufficient LSTM model with limited time and computational resources does not provide the desired result and therefore the alternative route of text extraction for summarization should be considered. The summary provided through text extraction aligns fairly well with the sentiment and topics covered in earlier portions of the analysis and allows for an interested party to read a portion of the dataset without arbitrarily selecting Tweets. This can be further adjusted to filter for sentences addressing sub-topics of interest, like the previous look at specific versions of the Covid-19 vaccine. Overall, we would recommend further research into improving the LSTM solution for more sophisticated insights which consider the full text in determining a unique output but believe the summary extraction method achieves the desired result and has the benefit of being an easily adaptable solution.

7 Conclusion

As a social media platform used by millions of people, Twitter takes on a serious responsibility when it comes to deceptive information spreading. It must have the appropriate means at hand to protect people from untrue rumors for the social good, and for this task, a sentiment analysis like this one can be of great help. Through the monitoring of sentiment trends related to Covid-19 vaccines, extreme behavior can be detected and handled accordingly. Moreover, our analysis can also be useful for governing bodies that are primarily invested in the well-being of the public. Medical institutions and doctors can be aware of the general emotional state of the community and provide help for those who need it.

In analyzing the feasibility of sentiment classification, we concluded that the two discriminative models Logistic Regression and SVM achieved a good degree of accuracy. We believe this provides a foundation for further research into discriminative models for sentiment classification purposes. Rather than using a single feature extraction method, an ensemble approach could lead to an improvement. Using such advanced practices, as well as hand-labeled data, can provide a base for exploring the cultural effects of the Covid-19 pandemic and the vaccination effort.

References

[Avery, 2017] Avery, E. J. (2017). Public information officers' social media monitoring during the zika virus crisis, a global health threat surrounded by public uncertainty. *Public Relations Review*, 43(3):468–476.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique.

[Feehan and Apostolopoulos, 2021] Feehan, J. and Apostolopoulos, V. (2021). Is covid-19 the worst pandemic? *Maturitas*.

[Fung et al., 2016] Fung, I. C.-H., Duke, C. H., Finch, K. C., Snook, K. R., Tseng, P.-L., Hernandez, A. C., Gambhir, M., Fu, K.-W., and Tse, Z. T. H. (2016). Ebola virus disease and social media: a systematic review. *American journal of infection control*, 44(12):1660–1671.

[Géron, 2019] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

[Hussain et al., 2021] Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., Ali, A., and Sheikh, A. (2021). Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study. *Journal of Medical Internet Research*, 23.

[Jung, 2018] Jung, A. (2018). Machine learning: Basic principles. *arXiv preprint arXiv:1805.05052*.

[Jurafsky and Martin, 2014] Jurafsky, D. and Martin, J. H. (2014). Speech and language processing. vol. 3. *US: Prentice Hall*.

[Koehn, 2017] Koehn, P. (2017). Neural machine translation. *arXiv preprint arXiv:1709.07809*.

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Praveen et al., 2021] Praveen, S., Ittamalla, R., and Deepak, G. (2021). Analyzing the attitude of indian citizens towards covid-19 vaccine – a text analytics study. *Diabetes Metabolic Syndrome: Clinical Research Reviews*, 15.

[Sharma et al., 2017] Sharma, M., Yadav, K., Yadav, N., and Ferdinand, K. C. (2017). Zika virus pandemic—analysis of facebook as a social media health information platform. *American journal of infection control*, 45(3):301–302.

[Thelwall et al., 2021] Thelwall, M., Kousha, K., and Thelwall, S. (2021). Covid-19 vaccine hesitancy on english-language twitter. *El profesional de la información*.

[Zhou et al., 2019] Zhou, F., Zhang, Q., Sornette, D., and Jiang, L. (2019). Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. *Applied Soft Computing Journal*, 84:105747.