Exploring Emotions Towards COVID-19 Vaccines Using NLP and Statistical Analysis

Muhammet Batuhan Doğan mbdn19156@gmail.com
Alper Mert alperm76@gmail.com

[1] Hacettepe University, Computer Engineering Department, Software Engineering Research Group, Ankara, Turkey

Abstract
Context: Sentiment Analysis and Classification
Purpose: Analysing Public Sentiment Change About COVID-19 Vaccines During Time From Tweets
Method: VADER for Sentiment Analysis, Logistic Regression for Prediction
Results: Analyzes show that public opinion has changed positively about vaccines. Logistic Regression models have very good ability on predictions.

***Keywords*:** NLP, Sentiment Analysis, Classification, COVID-19, Vaccines

1. INTRODUCTION

Context The COVID-19 pandemic has significantly impacted global health and socio-economic systems, bringing vaccine development and public acceptance to the forefront. As vaccines became available, public perception and sentiment around them varied widely, influenced by factors such as misinformation, trust in pharmaceutical companies, and governmental communication strategies. Social media platforms like Twitter have become pivotal in shaping and reflecting these public opinions. Analyzing these discussions provides a unique opportunity to understand public sentiment and trends on a large scale, offering insights for improving health communication strategies.

Problem The core problem addressed by this project is understanding public sentiment and perception toward different COVID-19 vaccines, based on discussions on Twitter. Traditional methods like surveys are costly, time-intensive, and limited in scope, making them unsuitable for capturing large-scale, real-time public sentiment. Social media platforms, with their vast amount of user-generated data, offer an alternative means to analyze public opinion. However, extracting meaningful insights from such data poses challenges, including noise, varying contexts, and the need for sophisticated natural language processing (NLP) techniques. This project seeks to fill this gap by systematically analyzing sentiment, trends, and public reactions to different vaccine brands over time.

Solution This project uses NLP techniques and statistical methods to analyze Twitter discussions about COVID-19 vaccines. It aims to classify public sentiment into categories such as negative, non-negative, and possibly neutral, focusing on tweets about different vaccine brands like Pfizer, Moderna, AstraZeneca, and others. By employing sentiment analysis (using models like VADER, or logistic regression), topic modeling, and time-series analysis, the project will uncover sentiment dynamics and trends in response to significant events. Insights derived from this study will help health officials, vaccine producers, and policymakers understand public perception better, enabling more effective communication strategies and fostering trust in vaccination campaigns. These findings may also provide a framework for addressing similar health crises in the future.

Structure of the paper:

Abstract:
A concise summary of the paper, outlining the context, problem, methodology, and key findings. This section provides a snapshot of the study to help readers quickly grasp the essence of the work.

Introduction:
Introduces the study by providing context, defining the problem, presenting the research question, and outlining the goals and objectives. It also briefly highlights the significance and potential impact of the work.

Background and Related Work:
Reviews relevant literature and previous studies on sentiment analysis, topic modeling, and time-series analysis in the context of public health and vaccine sentiment. This section positions the paper within the existing body of knowledge, highlighting the research gap this work addresses.

Methodology:
Details the dataset used (e.g., "COVID-19 All Vaccines Tweets" from Kaggle) and describes the preprocessing steps, sentiment analysis models (e.g., BERT), topic modeling approaches (e.g., LDA), and time-series analysis techniques. This section also explains the evaluation metrics and criteria for comparing model performances.

Experiments:
Presents the experimental setup and results, including comparisons of sentiment analysis models, insights from topic modeling, and observations from sentiment fluctuations over time. This section emphasizes key findings, supported by visualizations like graphs and tables.

Conclusion:
Summarizes the study's contributions and key insights, discusses implications for public health communication, and outlines potential applications. It also acknowledges the study's limitations and suggests directions for future research.

References:
Lists all cited works in a consistent format, ensuring proper attribution for literature, datasets, and tools used.

Appendix:
Includes supplementary materials such as detailed tables, additional visualizations, or extended explanations of methods and results that support the main text but are too detailed for inclusion in the primary sections.

The remainder of this article is organized as follows. Section 2 provides an overview of related studies and highlights the need for this research with respect to related studies. Section 3 explains the methodology employed while carrying out this study... Section 4 provides the results in correspondence with the research questions. Section 5, we provide overall conclusions and plans for future work.

2. BACKGROUND AND RELATED WORK

In Table 1 we summarize these studies with year, title, objective …

Table 1: Summary of the related work

Year [ref], Venue	Title	Objective	Datasets	Findings w.r.t. ....
2023,Elsevier's Healthcare Analytics	A natural language processing approach for analyzing COVID-19 vaccination response in multi-language and geo-localized tweets	Analyze global social media discussions about COVID-19 vaccines to understand public perceptions and attitudes toward different vaccine brands. By examining multilingual and geo-localized tweets using Natural Language Processing (NLP) techniques, the research evaluates emotional trends, discussion intensity, and the influence of language. The study aims to provide data-driven insights to better address public concerns and improve vaccine acceptance.	Source: Tweets from Twitter collected using the Twitter StreamAPI (Tweepy). Timeframe: April 15 to September 15, 2022. Tweet Volume: ~9.5 million total, filtered to ~8.3 million after preprocessing.	Pfizer was the most debated vaccine, with significant concerns around side effects on children, pregnant women, and heart-related issues. Language influenced tweet content significantly, with some terms causing unique biases (e.g., "moderna" in Spanish). Temporal patterns showed a higher spread of negative news compared to positive news.
2023,MDPI's Electronics	TSM-CV: Twitter Sentiment Analysis for COVID-19 Vaccines Using Deep Learning	Develop a deep learning model (TSM-CV) to analyze public emotions and opinions about COVID-19 vaccines. Using historical and real-time data collected from Twitter, the research aims to understand and classify emotional trends related to vaccines. The proposed model integrates methods like FastText, VADER, and RMDL to evaluate vaccine hesitancy and public attitudes toward vaccination. This analysis seeks to contribute to combating misinformation and improving vaccine	Source: Historic Data: "All COVID-19 Vaccines Tweets" dataset from Kaggle (125,906 tweets). Real-Time Data: Tweets collected using the snscrape tool from January 2020 to June 2021. Volume: 4,554,258 tweets.	The model achieved 94.81% accuracy and an F1 score of 97.50%, outperforming traditional ML methods like SVM, KNN, and Naïve Bayes. The AUC-ROC value of 92.59% indicated high recognition power. Results emphasized the effectiveness of combining FastText with RMDL for sentiment analysis.

		acceptance within society.		
2024,Springer Nature's Cognitive Computation	Emotion Analysis of COVID-19 Vaccines Based on a Fuzzy Convolutional Neural Network	Analyze emotional tendencies toward COVID-19 vaccines, identify shifts in public attitudes about vaccination, and uncover the reasons behind these changes.	Source: Public datasets: NLPCC2013, NLPCC2014, simplified Weibo, and comment datasets. Private dataset: Microblogs collected via Python crawlers from January to September 2021. Volume: 31,335 Chinese microblogs after preprocessing.	Negative sentiments outweighed positive ones throughout most months, peaking in April. Word clouds and LDA identified "allergies" and "outbreak" as common themes tied to negative sentiments. Emotional trends were linked to vaccination rates and public events, highlighting the importance of policy-driven communication strategies.
2023, Expert Systems With Applications	Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset	Investigate public hesitancy toward COVID-19 vaccines by analyzing Twitter data using text mining, sentiment analysis, and machine learning. It aims to identify the sentiment trends over time, understand the key themes in vaccine-related discussions, and evaluate the effectiveness of different sentiment analysis tools and machine learning models in classifying public opinion. This research seeks to provide insights into the changing attitudes toward COVID-19 vaccines and the factors influencing vaccine hesitancy.	Source: Twitter API Timeframe: September 26, 2021 – November 7, 2021 Size: 42,796 tweets	Key findings indicate that public sentiment about COVID-19 vaccines has become more positive over time. The best-performing model combined TextBlob sentiment scores with TF-IDF vectorization and LinearSVC, achieving an accuracy of 96.75%. This suggests that attitudes toward COVID-19 vaccines are improving and highlights the efficacy of combining sentiment analysis with machine learning for social media data analysis.
2021, Vaccines	COVID-19 Vaccine and Social Media in	Explore public emotions and	Source: Brandwatch platform, tweets in	Key findings include a decrease in

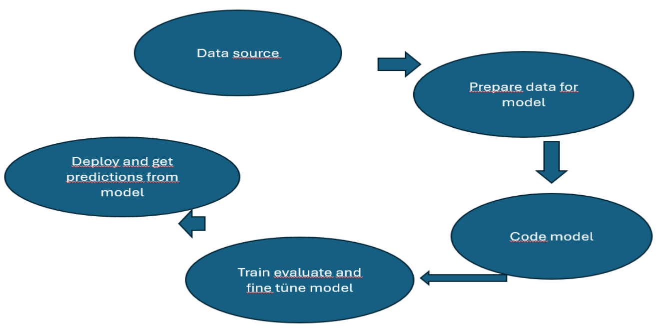
	the U.S.: Exploring Emotions and Discussions on Twitter	discussions regarding COVID-19 vaccines in the U.S. using Twitter data. It aims to identify the sentiment trends over time, discover key topics in vaccine-related conversations, and compare the focus of negative and non-negative tweets.	English from the United States Timeframe: November 1, 2020 – February 28, 2021 Size: 185,953 tweets	negative sentiments and an increase in non-negative sentiments about COVID-19 vaccines over time. Negative tweets focused more on vaccine hesitancy and political issues, while non-negative tweets highlighted vaccination stories, effectiveness, and management. The results demonstrate the potential of using social media data for public health insights, particularly in tracking sentiment trends and identifying key public concerns.

Both studies show the potential of using social media data for public health research, demonstrating its ability to track sentiment trends and identify key themes in vaccine discussions. These systems may be beneficial to understand public opinion about micro or macro health problems and governments/healthcare organizations may determine their strategies according to those opinions.

There are a lot of fake accounts on social media sites. These accounts may take the majority by posting spam tweets. This can lead normal people to impose wrong ideas. It is hard to detect fake accounts, and these systems' results can be affected by fake accounts.

Some of the models used may have high computational costs in real-time applications, especially the multi-language models may be very expensive.

3. METHODOLOGY



1. Business Requirements

The objective of this project is to analyze public sentiment, opinions, and trust levels regarding COVID-19 vaccines using Twitter data. The scope includes assessing general sentiment toward different vaccine brands such as Biontech and Sinovac, analyzing changes in public opinions over time, and conducting location-based analysis to identify regional differences in sentiment. The research focuses on answering questions about the general sentiment toward each vaccine brand, regional differences in sentiment, and changes in public trust levels over time.

2. Data Requirements

The analysis will utilize a dataset from Kaggle containing 228,208 tweets. This dataset includes features such as user information, tweet content, hashtags, and engagement metrics. These features provide comprehensive details for conducting sentiment analysis, exploring temporal trends, and performing regional comparisons.

3. Data Preparation

The preparation phase involves understanding the dataset, exploring its characteristics, and preparing it for analysis. Exploratory data analysis will help uncover trends and anomalies, while visualization techniques will provide insights into the distribution and patterns within the data. Preprocessing steps include cleaning the dataset by removing irrelevant information, handling missing values, and preparing the text for analysis through tokenization and the removal of stop words.

The preparation phase involves understanding the dataset, exploring its characteristics, and preparing it for analysis. Exploratory data analysis will help uncover trends and anomalies, while visualization techniques will provide insights into the distribution and patterns within the data. Preprocessing steps include cleaning the dataset by removing irrelevant information, handling missing values, and preparing the text for analysis through tokenization and the removal of stop words.

We transformed the tweets lowercase to get rid of meanings of uppercase. Then we transformed the date features into datetime format to do time serial analysis. We filled the empty data in "hashtag", "user_location" features. Then we dropped the features "user_name", "source", "user_description", "user_friends" features since we thought those are irrelevant.

As nlp preprocessing techniques, we removed stop words from tweets. We used predefined stop word list to get rid of all such words. We also preprocessed the tweets to eliminate url, emoji.

4. Model Development

We used VADER to do sentiment analysis. VADER is effective on sentiment classification. We calculated "positive", "negative" and "neutral" possibilities for each tweets by analysing the words in tweets. Then we calculated compound scores for each tweet. We classified sentiments as "Positive", "Negative" and "Neutral" by using compound scores. We decreased the threshold from 0.05 to 0.03 to decrease the count of Neutral tweets.

After made the visualizations, we implemented Logistic Regression model. We chose logistic regression for this project due to the following reasons:

a. **Simplicity and Interpretability:** Logistic regression is straightforward to implement and provides interpretable results. The coefficients of the model indicate the contribution of each feature (e.g., TF-IDF scores of words) to the classification task, helping us understand the importance of specific terms in predicting sentiment.

b. **Efficiency for High-Dimensional Data:** In text classification tasks, where the input data is transformed into high-dimensional TF-IDF vectors, logistic regression performs well without requiring extensive computational resources. Its ability to handle sparse data effectively aligns with our preprocessed TF-IDF features.

c. **Good Baseline Model:** Logistic regression is often used as a baseline for classification tasks. Its performance can set a benchmark to evaluate more complex models like support vector machines, random forests, or deep learning models, if needed.

d. **Compatibility with TF-IDF Features:** Logistic regression is inherently linear, making it a good match for the numerical TF-IDF vectors, where each dimension represents the importance of a word or phrase. This makes the model particularly suitable for text classification tasks, such as sentiment analysis.

e. **Probabilistic Outputs**: Logistic regression outputs probabilities for each class, allowing us to interpret not only the predicted class but also the confidence of the prediction. This is especially useful in tasks where understanding uncertainty is critical

We used two logistic regression models: One is default model(parameters are not changed, default), the other is optimized model. After doing GridSearch, we found that C=10(C: hyperparameter controls the strength of the regularization)[6] and penalty=l2 (The L2 penalty in logistic regression, also known as L2 regularization or Ridge regularization, is a technique used to prevent overfitting by adding a penalty term to the loss function that is proportional to the sum of the squares of the model's coefficients.)[7] are best parameters.

5. Training, Evaluation, and Fine-tuning

We splitted our data as training(60%), validation(20%) and test(20%). We trained our default Logistic Regression model with using preprocessed texts to make the model learn the relationship between TF-IDF vectors and classified sentiments.

We evaluated the model’s performance with validation set to ensure the generalization and avoid overfitting. We used predefined key evaluation metrics such as accuracy, precision, recall to evaluate our models.

After we finished the evaluation of default model, we did grid search operation to find optimal parameters for Logistic Regression in our task. We then defined the C parameter as 10 and penalty parameters as l2. Then we did the same process to test our parameter changed model.

6. Deployment and Predictions

The deployment phase of this project focuses on applying the trained sentiment analysis model to real-world textual data, enabling automated sentiment classification for incoming content. This section outlines the deployment considerations, the process for predicting sentiment, and the practical implications of our results.

The deployed model classifies text into three sentiment categories: **Positive**, **Negative**, and **Neutral**. Predictions are made in the following steps:

1. **Preprocessing**: Input text is preprocessed to remove noise, lowercased, and transformed.
2. **Sentiment Analysis:** Compound Score calculations for input text
3. **Prediction**: The Logistic Regression model predicts sentiment probabilities, assigning the highest probability class to the input text.
4. **Output**: Results include the predicted sentiment label and confidence scores for each class, offering interpretable outcomes.

4. EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets

We used the "vaccination_all_tweets.csv"[8] dataset from kaggle. This dataset contains 228208 rows and 16 columns. This dataset includes features such as user information, tweet content, hashtags, and engagement metrics. These features provide comprehensive details for conducting sentiment analysis, exploring temporal trends, and performing regional comparisons. The most important feature of our data is "text" obviously. Date feature is also important to do time serial analysis.

4.2 Experiment Results

We calculated a classification report for both cases of our models. Here are the reports:

Default Model's Validation Classification Report:

Validation Set Classification Report:

	precision	recall	f1-score	support
Negative	0.91	0.76	0.83	6726
Neutral	0.92	0.98	0.95	22023
Positive	0.95	0.93	0.94	16892
accuracy			0.93	45641
macro avg	0.93	0.89	0.91	45641
weighted avg	0.93	0.93	0.93	45641

Default Model's Test Classification Report:

Test Set Classification Report:

	precision	recall	f1-score	support
Negative	0.91	0.77	0.83	6725
Neutral	0.92	0.98	0.95	22024
Positive	0.95	0.93	0.94	16893
accuracy			0.93	45642
macro avg	0.93	0.89	0.91	45642
weighted avg	0.93	0.93	0.93	45642

Parameter Changed Model's Validation Classification Report:

Validation Set Classification Report:

	precision	recall	f1-score	support
Negative	0.89	0.80	0.85	6726
Neutral	0.94	0.98	0.96	22023
Positive	0.95	0.94	0.94	16892
accuracy			0.94	45641
macro avg	0.93	0.91	0.92	45641
weighted avg	0.94	0.94	0.94	45641

Parameter Changed Model's Test Classification Report:

Test Set Classification Report:

	precision	recall	f1-score	support
Negative	0.89	0.81	0.85	6725
Neutral	0.94	0.98	0.96	22024
Positive	0.95	0.93	0.94	16893
accuracy			0.94	45642
macro avg	0.93	0.91	0.92	45642
weighted avg	0.94	0.94	0.94	45642

5. CONCLUSION

Both model's performances are very good, maybe too good. There may be overfitting risk, or the tweets may be easy classifiable according to sentiments. Since dataset is huge and we cannot control the tweets one by one, it might be useful to try it with a dataset that includes tweets where it is certain that the emotions are difficult to understand.

We wanted to use more complex sentiment analyser than VADER. We tried to implement BART and wanted to compare their analysis, but we had some issues(most probably because of versions), so we used only VADER to do sentiment analysis. Using more complex sentiment analysis may change the results.

Funding

There is no funding in this project.

Conflict of interest

The authors declare that they have no conflict of interest.

Data Availability

https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets

REFERENCES

[1]: https://pmc.ncbi.nlm.nih.gov/articles/PMC10088351/
[2]: https://www.mdpi.com/2079-9292/12/15/3372
[3]: https://link.springer.com/article/10.1007/s12559-022-10068-6
[4]: https://www.sciencedirect.com/science/article/pii/S0957417422017407
[5] :https://pmc.ncbi.nlm.nih.gov/articles/PMC8540945/
[6]:
https://medium.com/@rithpansanga/logistic-regression-and-regularization-avoiding-overfitting-and-improving-generalization-e9afdcddd09d#:~:text=The%20%E2%80%9CC%E2%80%9D%20hyperparameter%20controls%20the,and%20a%20more%20complex%20model.
[7]:
https://www.geeksforgeeks.org/what-is-l2-penalty-in-logistic-regression/#:~:text=in%20Logistic%20Regression%3F-,Answer%20%3A%20The%20L2%20penalty%20in%20logistic%20regression%2C%20also%20known%20as,2%20min%20read
[8]: *https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets*
Github
Link:https://github.com/AlperMRT/Data-Intensive-Applications-Project-COVID-19-Vaccines-Sentiment-Analysis