Public Opinion on Vaccines during COVID19 pandemic through Twitter (1802 words)

Himani Anil Deshpande

Data Science, Indiana University Bloomington

May 1, 2022

1 Introduction

The highly contagious SARS-CoV-2 virus responsible for COVID-19 outbreak had created an enormous disruption in not only the way of living but also public health systems. The Covid-19 outbreak was declared a global pandemic by World Health Organization on 11th March, 2020.[1]

The CDC declared using masks, social distancing and frequent sanitation as the initial preventive measure, while stating that vaccine will be important in breaking the infectious transmission of SARS-COV-2 virus.[2]

A vaccine is an instrument in controlling diseases and preventing death. It helps in developing adaptive immunity by administering very small and safe amounts of attenuated or killed viruses and bacteria to our body.[3]

Even with accelerated development of COVID-19 vaccines, the vaccination drive saw a slow growth in prevention due to different causes like vaccine hesitancy, misinformation and conspiracy theories, mistrust in vaccine efficacy, health-care systems, political ideology and anti-vaccination movements.[4]

Research Question

In this project I have used Topic modeling methods like LDA and a hybrid approach inspired by paper[5] of sentiment lexical analysis technique like EMPATH and Machine Learning models like SVM and Random Forrest to analyze the different sentiments that the population displays [4,6] with below objectives:

What is the public opinion of population in USA about Vaccinations developed during the COVID19 pandemic?

What are the major topics or sentiments of population about these vaccinations?

2 Methods

Using Twitter Developer API v2 to gather at 4430 random USA originated English tweets. I have gathered data from September 2020 to May 2021 with below collected hashtags and keywords.

Keywords:

#CovidVaccine, vaccination, vaccinations, vaccine, vaccines, immunization, vaccinate, vaccinated, #COVID-19, coronavirus, CORONA, covid-19 Vaccines, SARS-CoV-2, vaccination Refusal, #Pandemic, vaccination drive, immunity, immune, rollout, side effect, immunise, immunize, immunisation, astrazeneca, novavax, #pfizerbiontech. [7, 8, 9, 10, 11]

The 9 months are divided into 3 Timelines; September 2020 to November 2020 are considered as **Before** the Vaccination drive started. December 2020 to February 2021 is considered **During** vaccination drive and March 2021 to May 2021 is **After** the vaccination Drive.

I have collected, parsed and stored data about the Tweet Id, Author ID(user ID), Tweet content, country, place name, public metrics like reply, likes, quotes and retweet counts, source and timestamp of tweet using Tweepy package.

For better results [10], I have cleaned the data to remove stopwords, URLs, usernames or user mentions and special character like emoticons and punctuation's and converting text into lowercase. I have removed the special character '#' in the tweet but not removed the keywords as they are part of a sentence and provides information.[9]

2.1 Analysis

I have analyzed the data using different Models mentioned below.

Sentiments	Definition	Seed Words
Hesitation	Skeptic attitude and reluctance towards being vaccinated due to multiple negative factors affecting an individual's opinions.	anxious, nervous, fear, consequences, uncertain, reluctant, hesitation, suspicion, harm
Sorrow	Dissatisfaction and disapproval towards the different phases of COVID-19 vaccine production and distribution	sad, hopeless, worst, dissatisfaction, disappointment, setback
Faith	Signifies strong belief and confidence in vaccines along with optimistic behavior towards the success of vaccines.	faith, optimism, vaccines, work, assurance, trust, belief, grateful
Contentment	Signifies a state of happiness, appreciation, and acceptance of the COVID-19 vaccines.	satisfy, glad, proud, gratitude, great, joy
Anticipation	State of urgent demand and necessity of vaccines.	anticipate, urgently, expectancy, priority, quick, await
Rage	Anger or aggression is associated with conflict arising from a particular situation.	angry, annoyance, hate, mad, pathetic
Misinformation	Propagation of false information such as misinterpreted agendas and conceiving vaccines as conspiracy or scam.	propaganda, conspiracy,fraud, fake, poison, deception, plot, trick, plan, frame-up, false
Vaccine Rollout	Availability and distribution of vaccines through campaigns and mass vaccination drives.	vaccinate, distribution, supply, mass, dose, vaccine rollout, vaccine campaig, vaccination drive
Inequities	Socioeconomic disparities are based on societal norms such as caste, race, religion	socioeconomic, deprive, racial, injustice, racism, underrepresented
Health Effects	Mentions of health-related adverse events caused by or affected by vaccines, including diseases, symptoms, and pre-existing conditions	nausea,swelling,muscle pain, vomitting,chills,malaise,high temperature, irritation, cough, nasal congestion, dizziness,headache, fatigue, inflammation, fever, bodyache,drowsiness

Figure 1: 10 Categories of Sentiments

2.2 EMPATH

EMPATH[12] is a freely available tool with 200 built in categories and allows the flexibility of creating new categories by using deep learning neural embedding to discover new words and phrases which are similar to the seed words we provide while creation.

By adding a few seed words I have created 11 new categories based on paper[11]. EMPATH provides a normalized value in the range of 0 to 1 for each category as per the tweet content, I will be using the sentiment with highest value to label the tweet. This label will be used for training the Machine learning models later.

Figure 1 shows the seed words and definition of categories.

2.3 Latent Dirichlet Allocation (LDA)

The Latent Dirichlet Allocation (LDA)[14] is a generative probabilistic model which efficiently produces the latent topics available in documents and is publicly available in gensim package. It is a unsupervised clustering model which assumes that a document comprised of multiple Topics and each Topic is made up of different words.[15]

Apart from the Data Processing steps mentioned previously, I had to lemmatize the Twitter Text, which is to get the dictionary based-root word of the words in the cleaned text. Using PyDaViz package I have created an HTML file with interactive visualization to check overlapping of the 5 different topics.

2.4 Machine Learning

I have used 2 supervised machine Learning models which use the results of EMPATH as Labels for sentiments classification of Tweets. I have split the 80% of data into training set and remaining into test set and evaluated them based on their scores like F1, precision, recall and accuracy. The model input is a vectorized by TF-IDF(Term Frequency-Inverse Document Frequency).

2.4.1 Support Vector Machines (SVM)

SVM is a supervised classification model[16], which uses the technique of maximum separation of hyperplanes and is available in sklearn[17] library. It is widely used in classification, regression and sentiment analysis problems due to its high accuracy and performance even with large dimensional dataset while avoiding over-fitting the data. [18]

Study [19] shows that SVM performs best when trained on a mix of strictly and laxly labeled data with a more fine-grained labeling yielding the best result.

By using GridSearchCV, which finds the best model after tuning the hyperparameters like gamma and kernel, I got a model with an accuracy and recall of 81.15%, precision of 81.03% and F1 Score of 79.26%.

2.4.2 Random Forrest (RF)

Random Forrest is a combination of a number of decision trees to create an ensemble model with high accuracy and ability for parallel learning due to bagging of sub-trees.[20,21] The core idea of RF is that these decision tree models are uncorrelated, but together as an ensemble their predictions and accuracy outperforms each of these individual tree's predictions or accuracy.

After tuning the estimators and depth,the RF model I had created gave me an accuracy and recall of 85.78% with precision of 85.72% and F1 Score of 85.01%.

3 Results

3.1 Distribution of Sentiments classified using EMPATH

Figure 2 shows a Bar chart with the distribution of the 10 categories of sentiments that the 4430 tweets were identified with using EMPATH. 65.40% of my data is classified as Vaccine Rollout. 6.75% belongs to contentment and 6.16% to Anticipation. Health effects contribute to 0.81% and 4.81% belongs to Misinformation.

3.2 Yearly Distribution of Sentiments classified using EMPATH

Figure 3 shows a Bar chart with the distribution of the 10 categories of sentiments in the 2 different Years of 2020 and 2021. We can see that Vaccine Rollout is dominating in both the years and has increased in the year of 2021. We see Misinformation at its highest in 2020,

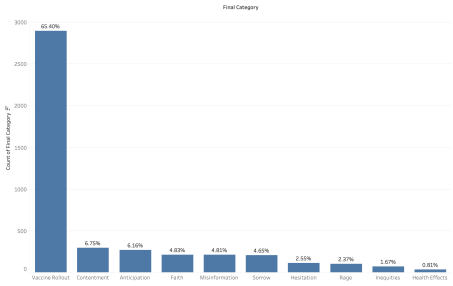

Figure 2: Distribution of Sentiments in captured Data

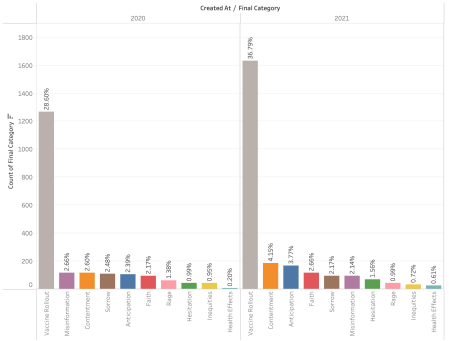

Figure 3: Yearly Distribution of Sentiments in captured Data

but has decreased in 2021. While Contentment, Faith and Hesitation has significantly increased in 2021 compared to 2020; Sorrow, Rage and Inequities has decreased in 2021.

3.3 WordCloud of Tweet Content

I have created a wordcloud of cleaned tweet content with word count above 50. From Figure 4 we can see that Covid19, first,effect, virus, received,infection, death, health, sarscov2, trump, public, mask, appointment, staff,life, getting, antibody, worker, response, protect, state, community, everyone and school have been repeatedly used.

3.4 Top 10 Cities in USA with Twitter posts on Covid Vaccines

Figure 5 shows a Bar chart with top 10 Cities in USA. Manhattan captured 19.19% of the dataset. LA generated 18.29% of tweets in my Timeline. It follows Manhattan in the During and Before discussions on Twitter. 16.61% of my data originates from Washington, where 6.73% of which are leading the After discussion followed by LA and Manhattan.

Figure 4: Wordcloud of Tweet content with occurrence above 50

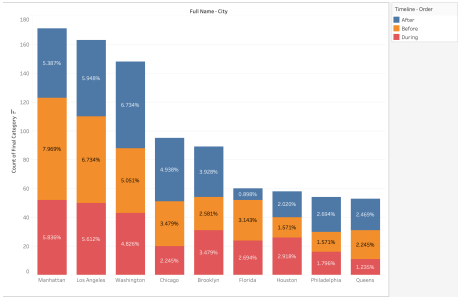

Figure 5: Top 10 Cities with high Twitter posts

3.5 Trend of Mean Engagement and Retweets by EMPATH Sentiments and Time

Figure 6 shows a Bar chart with the trend of EMAPATH sentiments over Time periods using Mean engagement, which is average of retweets and likes. Similarly, Figure 7 shows the Trend of sentiments using Retweet counts.

Contentment is the only category which shows a different trend among both the Figures. According to Retweet counts in Figure 6, the feeling of Contentment is at peak During the drive. But according to Figure 5 and Mean Engagement, the feeling is at peak Before the drive and it is steadily decreasing.

Misinformation and Faith is also at peak Before the drive in both the Figures and is at its lowest During the vaccination drive. Hesitation is at at peak After the drive. Health effects, Inequities and anticipation are almost at a plateau with barely any change. Sorrow is at its peak During the drive. Conversation about Vaccine Rollout is at its highest Before the drive and lowest during the drive.

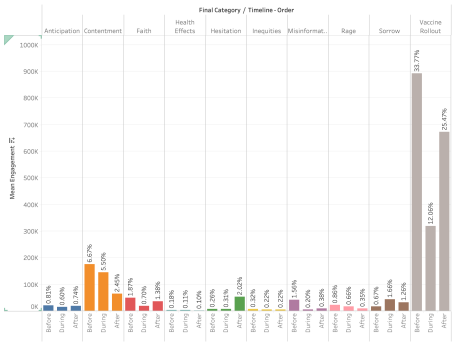

Figure 6: Trend of Tweets by Mean Engagements

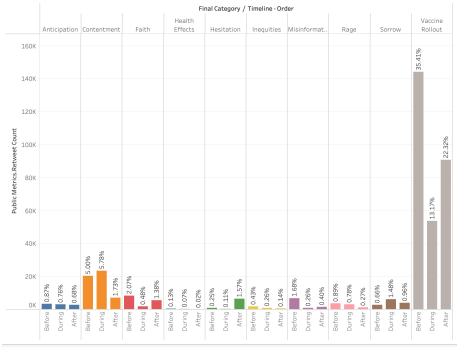

Figure 7: Trend of Retweets by Sentiments

3.6 SVM and Random Forrest Model Comparison

Papers [20, 22] have used Accuracy, Precision, Recall and F1 scores for evaluating the Machine Learning models. From Figure 8 we can see that Random Forrest model is outperforming SVM in all these criteria.

3.7 LDA Vizualization by PyDavis Package

The interactive LDA visualization was created using PyDavis package, which shows the 5 Topics that I have created.

Figure 9 shows the the top 30 most frequently used words in the tweet contents. It also shows that Topic 1 and 5 are overlapping which means they are almost similar and shows that we have 4 unique Topics in out dataset. Topic 1 contains 23.3% of the whole dictionary corpus.

From Figure 10, the biggest circle for Topic 1 shows that it comprises of highest percentage of tweets. It shows that covid19 which occurs

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Machine(SVM)	81.15	81.03	81.15	79.26
Random Forrest Model	85.78	85.72	85.78	85.01

Figure 8: Machine Learning Model Comparison

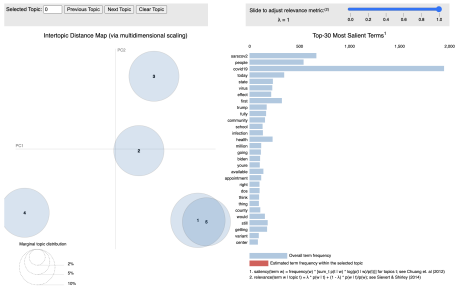

Figure 9: LDA Topics

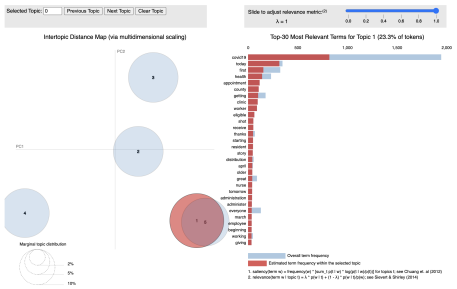

Figure 10: LDA Topics 1 and 5

2000 times in the dataset and it occurs almost 800 times in Topic 1 itself.

4 Conclusions

Apart from the Vaccine Rollout category which contributes to most of my dataset, We can clearly see that the top emotions that public shows are the feeling of **Contentment, Anticipation** and **Faith** as seen from Figure 2, 5 and 6.

From Figure 9 and 10 which shows the LDA Topics, I can conclude that there are 4 diverse topics in the whole dataset.

While comparing the two Machine Learning models using the different scores like Precision, Recall and F1 as well as accuracy, Random Forrest model is the winner and shows better performance in classification than SVM model. Random Forrest has a F1 score of 85.01% and accuracy of 85.78% compared to SVM which shows F1 score of 79.26% and accuracy of 81.15%

5 Limitations and Future Works

The dataset is focused on US population with English language tweets which limits the research as people all over the world are talking about the vaccines which are developed. Also, the vaccination drive in USA started earlier than most of the countries, it would have sparked a

debate in other countries so it would have been beneficial for my research if I could expand into different languages and countries.

Sentiment analysis using EMPATH is limited as there are multiple seed words which define multiple categories. This overlapping of words and therefore the sentiments/categories make it difficult to analyze the sentiments properly. Built-in categories like 'science' and 'health' have similar seed words.

The code book I have used [11] code be improved with other different sentiments and improved seed words. It would have been more convenient to remove a few seed words which EMPATH learned from neural embedding.

Human intervention is required to verify and correct the different sentiments and Topics in which a tweet was classified using EMPATH and LDA respectively.

In Future I would have liked to use the normalized numerical values generated by EMPATH for each tweet to predict the overlapping sentimentality of a text using regression algorithms. I would like to build different deep learning models instead of using already existing Machine Learning models. I believe that these models would be a better fit for natural language and sentiment analysis.

References

1. Resta, E., Mula, S., Baldner, C., Di Santo, D., Agostini, M., Bélanger, J. J., Gützkow, B., Kreienkamp, J., Abakoumkin, G., Khaiyom, J. H. A., Ahmedi, V., Akkas, H., Almenara, C. A., Atta, M., Bagci, S. C., Basel, S., Kida, E. B., Bernardo, A. B. I., Buttrick, N. R., ... Leander, N. P. (2022). 'We are all in the same boat': How societal discontent affects intention to help during the COVID-19 pandemic. Journal of Community Applied Social Psychology, 32(2), 332– 347. https://doi-org.proxyiub.uits.iu.edu/10.1002/casp.2572

2. Centers for Disease Control and Prevention. (2022, Feb 25). How to Protect Yourself Others. CDC. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html

3. Dai, X., Xiong, Y., Li, N., Jian, C. (2019). Vaccine Types. In (Ed.), Vaccines - the History and Future. IntechOpen. https://doi.org/10.5772/intechopen.84626

4. Alamoodi, A.H., Zaidan, B.B., Al-Masawa, M., Taresh, S. M., Noman, S., Ahmaro, Y.Y.I., Garfan, S., Chen, J., Ahmed, M.A., Zaidan, A.A., Albahri, O.S., Aickelin, U., Thamir, N.N., Fadhil, J.A., Salahaldin, A. (2021, December) Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy, Computers in Biology and Medicine, Volume 139, 2021, 104957, ISSN 0010-4825. URL

5. Shofiya, C.; Abidi, S.(2021, June 3) Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data. Int. J. Environ. Res. Public Health 2021, 18, 5993. https://doi.org/10.3390/ijerph18115993

6. Baldha, T., Mungalpara, M., Goradia, P. and Bharti, S., (2021) "Covid-19 Vaccine Tweets Sentiment Analysis and Topic Modelling for Public Opinion Mining," International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021, pp. 1-6, doi: 10.1109/AIMV53313.2021.9671000.

7. Lyu J, Han E, Luli G (2021, June 6) COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. J Med Internet Res 2021;23(6):e24435 URL DOI: 10.2196/24435

8. Ansari, Md Tarique Khan, Naseem. (2021). Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. Electronic Journal of General Medicine. 18. em329. 10.29333/ejgm/11316. URL

9. Jang H, Rempel E, Roe I, Adu P, Carenini G, Janjua N (2022, March 29) Tracking Public Attitudes Toward COVID-19 Vaccination on Tweets in Canada: Using Aspect-Based

Sentiment Analysis. J Med Internet Res 2022;24(3):e35016 URL DOI: 10.2196/35016
10. Hussain A, Tahir A, Hussain Z, Sheikh Z, Gogate M, Dashtipour K, Ali A, Sheikh A (2021, April 5) 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 J Med Internet Res 2021;23(4):e26627 URL DOI: 10.2196/26627
11. Chopra, H., Vashishtha, A., Pal, R., Ashima, Tyagi, A., Sethi, T. (2021). Mining Trends of COVID-19 Vaccine Beliefs on Twitter with Lexical Embeddings. ArXiv, abs/2104.01131.
13. Fast, E., Chen, B., Bernstein, M.S. (2016). Empath: Understanding Topic Signals in Large-Scale Text. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.
14. Blei, D. M., Ng, A. Y. Jordan, M. I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res., 3, 993–1022. doi: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993
15. Sutrave, K., Rajesh, G., Jun, L., (2021) "Understanding the Public Sentiment and Discourse on COVID-19 Vaccine". Dakota State University Beadle Scholar. AMCIS 2021 Proceedings. 24.
16. Lilleberg J., Zhu Y. and Zhang Y., (2015) "Support vector machines and Word2vec for text classification with semantic features", IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC), pp. 136-140, 2015.
17. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.
18. Jakkula, V.R. (2011). Tutorial on Support Vector Machine (SVM).
19. Kunneman, F., Lambooij, M., Wong, A. (2020) Monitoring stance towards vaccination in twitter messages. BMC Medical Informatics and Decision Making. https://doi.org/10.1186/s12911-020-1046-y
20. Reshi, A. A., Rustam, F., Aljedaani, W., Shafi, S., Alhossan, A., Alrabiah, Z., Ahmad, A., Alsuwailem, H., Almangour, T. A., Alshammari, M. A., Lee, E., Ashraf, I. (2022). COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset. Healthcare (Basel, Switzerland), 10(3), 411. https://doi.org/10.3390/healthcare10030411
21. Madani, Y., Erritali, M., Bouikhalene, B., (2021, June) Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. Results in Physics, Volume 25, 2021, 104266, ISSN 2211-3797, https://doi.org/10.1016/j.rinp.2021.104266. URL
22. Gerts, D., Shelley, C. D., Parikh, N., Pitts, T., Watson Ross, C., Fairchild, G., Vaquera Chavez, N. Y., Daughton, A. R. (2021). "Thought I'd Share First" and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study. JMIR public health and surveillance, 7(4), e26527. https://doi.org/10.2196/26527