Sentiment Analysis Accuracy of Machine Learning Models against Deep Learning for Vaccination Data and its Data Analysis

***Abstract*—As we know in this time, we are encountering different kinds of diseases and pandemic and to protect ourselves and others, public needs to be vaccinated. But despite different vaccination encouragement and drive by government and medical team, people are hesitant to it due to some reasons which can cause pathogens and disease-causing elements to spread and increase in human lives loss. So, government and agencies need to quickly get the public sentiment for vaccination so that appropriate actions may be taken to increase vaccination among the public. For this we are building a sentiment analysis model using both Machine Learning model as well as Deep Learning and selecting the one that performs best. Afterwards, the model is used to further analysis on unseen data on this vaccine dataset.**

Keywords—Sentiment Analysis, Machine Learning, Deep Learning, Vaccination Data Analysis

I. INTRODUCTION

In the modern times, we have seen and come across various kinds of diseases that are either life threatening or become so if appropriate measures are not taken. They may increase in the population if not stopped with appropriate measures which can lead to large loss of lives. One such method is vaccination, a medical treatment to inject liquid into the patients to fight the diseases and make human immune to them and check its spreading. Yet, in this modern times, people don't trust the vaccination that can be traced to number of reasons which can lead to unintended consequences and red alert for healthcare system.

For this we need to get the public opinion what they think of the vaccination based on the sentiments of their thoughts. This can be done by checking their tweets in Twitter and know the public perception towards vaccination. This needs a tool that can automatically get the sentiment of the tweet text using Sentiment Analysis tools. Although there are available tools but their accuracy on vaccination data is very low and hence we need to train our own model to detect Sentiment with greater accuracy. For this we will train the traditional Machine Learning models like SVM, Random forest etc. on the labelled vaccination data along with the Deep Learning models like BERT. We will compare them against each other to get the best mode out of the two and use the Sentiment Analysis model to analyse the unseen unlabelled vaccination data to understand what are the reasons and factors that make people hesitant towards the vaccination or supportive to it.

Both the Machine Learning (ML) and Deep Learning (DL) algorithms have their own advantages and disadvantages like simplicity, accuracy, fast inference and lastly accuracy. ML algorithms can give much better accuracy with simple architecture but prone to underfitting or performance reaching saturation level while DL algorithms can better capture complexity and hence good for complicated datasets but prone to overfitting .We need to see how the ML algorithms prediction accuracy varies against the DL algorithms for Sentiment Analysis. Secondly, we will use the trained model to identify the causes and factors that make people perceive vaccination in different way either positively, negatively or neutral. Hence, my research question is " *How does Machine Learning models performance vary against Deep Learning models for vaccination text data in sentiment analysis and how it can be used for vaccination data analysis to identify causes and factors for their sentiment on vaccination*?"

Therefore we have two objectives: 1) To check which model performs best in terms of accuracy on labelled dataset and 2) how to use model for vaccination data analyses on unlabelled dataset to identify patterns and reasons for variation in sentiments.

Before working in this project a research was conducted for papers on this topic. Paper [1] uses TextBlob for twitter data while the paper [2] uses conventional Machine Learning algorithms to create model. Paper [3] uses DistillRoBERTa by fine tuning to create model.

II. DATASET

For this task we will need a labelled dataset for training a Sentiment Analyzer model and an unlablled dataset to analyse the vaccination data using the model. The dataset should contain the perception of people regarding the vaccination along with the sentiment polarity.

For this reason, we will be using vaccination dataset with text from the Twitter regarding vaccination. It has been prepared by an organization "Zindi" which provides us the data with above properties. For the supervised data, it contains the tweet ID, tweet text and sentiment polarity in a single file. Next another file is provided with unlabeled dataset which we will use to do analyses for further discoveries. It just contains the tweet ID and its text.

The vaccination dataset contains tweets regarding vaccination and different diseases like measles, flu, viral disease etc. There are three polarities in the dataset as "positive", "negative" and "neutral". Before training and evaluating the model, the labelled dataset is partitioned in two parts: "train" and "Val" to compare all the algorithms and select the best one.

III. SENTIMENT ANALYSIS

In Sentiment Analysis, we predict the sentiment of the text in one of the multiple categories. It can be simply positive and negative or multiple levels of 5 or 10 based on the degree of the polarity of the sentiment in the dataset. We first used the conventional standard libraries to test the sentiment of the validation data using TextBlob and NLTK, two popular NLP libraries. We then ran the code from these two and got sentiment scores but with unsatisfactory results. The TextBlob and NLTK gave 40% and 36.9% accuracy.

At this point it became clear that we need to train models of our own to get well performing models on vaccine unseen data. We started to train the ML algorithms after pre-processing and bringing them to trainable form. At the beginning we used different pre-processing techniques to process the data. We replaced the emojis to text, used regular expressions to clean the text data from noisy and unnecessary characters and finally tokenize them. Thereafter we removed stopwords tokens and lemmatized the token using Porter Stemmer from NLTK library. This was done to both train and val data to be trained thereafter. Finally the train and val data were vectorized using TF-ID vectorizer. This prepared data will be used for multi-class classification for sentiment polarity.

Different models were trained to achieve good accuracy. First simpler model were used like Naïve Bayes and KNN classifier to predict sentiment. Although, model is simple still Naïve Bayes achieved 69.6% accuracy compared to TextBlob with 40%. To increase the accuracy, we increased the slight complexity in the model using SVM models both with linear and RBF kernel that gave us 73.15% and 72.9% accuracy, a gain of 3.55% from Naïve Bayes. To further attempt to increase accuracy, we trained the Decision Tree and Random forest algorithm with greater complexity that gave us accuracy of 66.9% and 71.7% which resulted in decrease in accuracy that implies that model has overfitted. Hence in ML algorithms, SVM with linear kernel performed best against other models. Below table gives accuracy between 0 and 1.

Next, we need to train the Deep Learning architectures to predict the sentiment polarity using transformer networks, although heavy in size. We trained two models, BERT and RoBERTa. Both of them are complex networks and may be prone to overfitting. So we trained to check the results. BERT gave us the 76.3% and RoBERTa gave **78.5%** accuracy which is 5.34% gain in accuracy.

	Model	Accuracy
0	Naive Bayes	0.6960
1	SVM (Linear)	0.7315
2	SVM (RBP)	0.7290
3	Decision Tree	0.6695
4	Random Forest	0.7175
5	KNN	0.6680

Table 3.1 ML Models and their accuracy

Once all the models are trained, we compute the bar graph and compare the models against each other. From the graph of figure 3.1, it is observable that Machine Learning models are less accurate than the deep learning algorithms. Among the deep learning algorithms, RoBERTa model gives the best accuracy so far with 78.5%. Hence for our first objective, we trained the Sentiment Analysis model and Deep learning models beat the Machine Learning models in terms of accuracy or performance. We will be using RoBERTa model for

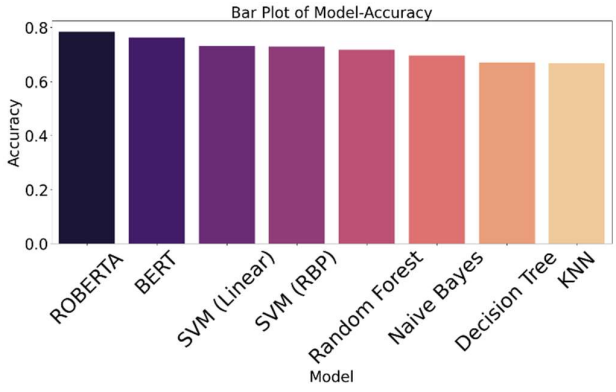


Figure 3.1 DL and ML models accuracy graph comparison

our data analysis on unlabelled text of vaccine data.

IV. DATA ANALYSIS

Now that our Sentiment Analysis model is ready, we can use it to analyse the unlabelled data to make the discoveries about people sentiment for vaccination. First we run the model on the data and get the distribution of polarities. From the figure 4.1, it is clearly visible that positive tweets and neutral tweets together account for 53+41=94% of the tweets which is a positive remark. It shows that people are aware of the seriousness of vaccination and if not, they don't see vaccination as a repulsive threat. The tiny 6% of negative sentiment gives relief that vaccination can be easily done to 94% of people and hence greater coverage. Yet it is not small enough as with regard to population when encountered with communicable viral or bacterial diseases or when people are large in population.

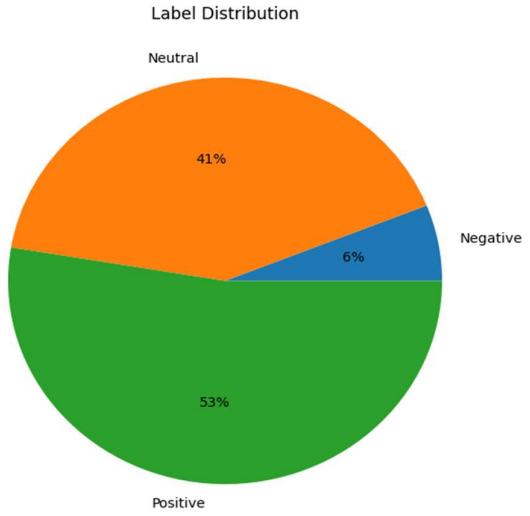
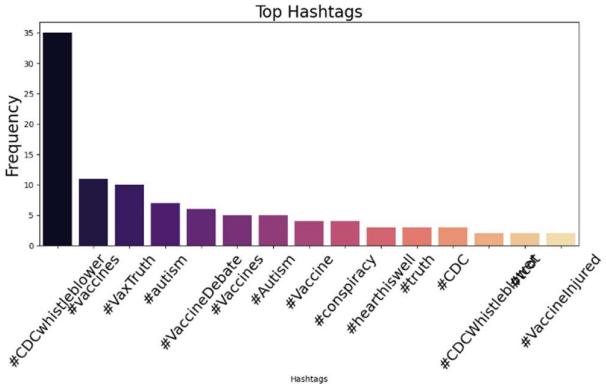


Figure 4.1 Label distribution on unlabelled data

We now focus on negative tweets because they are the section of people of concern. We split the data based on predicted sentiment and take out with negative sentiment. We search for the hashtags and we find the



results in figure 4.2.

Figure 4.2 Hashtags distribution in Negative Sentiment tweets

The above bar graph shows the hashtags of #CDCwhistleblower, #VaxTruth, #autism and so on gives the signal that the person is not having positive attitude to vaccination. This can create trouble because either the person if not vaccinated can cause death to arrive early or contribute to spread the disease to other person at a rapid rate that would be disastrous. Using the keyword of "CDCWhistleblower" in tweets of negative sentiment and analysing their tweets reveals certain things. They are as follows: People believe in rumour that vaccine cause child autism although scientifically its incorrect. Some people think Afro-American children are more prone to autism with no evidence. Some people have illusionary thinking that vaccine causes insomnia but medical science says isn't but by vaccine anxiety. People have gone this extreme of making allegation that vaccine makes child low-IQ. Lastly, many think vaccination as a business and commercialization means to earn dollars and money.

Finally we take top 15 words in the text body and we get this set : ['vaccine', 'url', 'user user', 'autism', 'url user', 'user', 'CDCwhistleblower', 'amp', 'U', 'CDC', 'disease', 'kid', 'Health', 'Measles', 'cause']. From the list, we can see the top order and no of times "user" and "url" occurs. It shows the relevance of online materials as well as people with negative attitude to vaccination can cause more people to cause negative sentiment to vaccination. For this the government and agencies must create proper website, online resources like in social media to spread awareness and curb the spread of fake and invalid news regarding vaccination.

V. FUTURE WORK

For future work, we can look for more advance models with greater accuracy but running in real time to make sentiment analysis. For analysis, currently pure text have been taken, but can be extended to image with texts or video subtitles to get large and broader area for analysis.

VI. CONCLUSION

From the training and analysis part, it is clear that deep learning model perform superior against the Machine Learning models with greater margin. Secondly, in terms of analysis, it seems that the spread of rumours and unverified news as well as mere anticipations without medical team investigation and misleading content from websites and users can spread the negative sentiment towards the vaccination. People must be made aware that vaccination is not for business making but to save the lives from deadly disease through various media. The usage of sentiment analysis to segregate data and doing deeper analysis gives us the better discoveries.

VII. REFERENCES

1) *Sentiment analysis of Covid 19 Vaccines using Twitter Data*

2) *Sentiment Analysis of COVID-19 Vaccine Tweets Using Machine Learning.*

3) *Fine-tuned Sentiment Analysis of COVID-19 Vaccine-Related Social Media Data: Comparative Study*