RECOMMENDATION SYSTEM USING GEO-LOCATION BASED ON SENTIMENT ANALYSIS OF COVID-19 TWEETS

Prof. Dhanashri Patil , Mugdha Joshi, Ritul Yadav, Shivani Kulthe

Professor, Student, Student, Student

Department of Computer Engineering,

Sinhgad College of Engineering, Pune, India.

Abstract – In the year 2020, we all faced a big challenge of covid-19. The WHO announced it as pandemic and we were all forced to shut down and stay indoors. While the frontline workers like doctors, scientists, army officials, police were fighting to stop the spread of this deadly virus. Today, after almost a year we are being threatened by a second wave of covid-19 which is believed to be much more dangerous than the previous virus. In this difficult situation where we are forced to stay at our home, work from home, it has affected our mental health in many ways. Some people find it amusing that we get to stay home but others find it hard, lonely and tiresome. For the mental well-being of people we have proposed this project wherein we analyse the mental state of people in an area. Based on this analysis we give recommendation to the user as to what measures to take in order to stay mentally and emotionally healthy.

Index Terms - sentiment analysis; COVID-19; tweets; machine learning; geo-location; recommendation; dataset.

I. INTRODUCTION

In this 21st century we take to social media to express our feelings, post what we do in our day to-day lives and interact with other people. Sentiment analysis of these posts help us know what a group of people are going through and their mind set. We have used sentiment analysis to find out the current mental state of people in this pandemic. From this analysis we give user personalised recommendations based on their area as to how to maintain your mental health. We have gathered data of tweets posted on Twitter. We got this dataset from Kaggle website.

II. RELATED STUDIES

Sentiment analysis is the most common text classification tool, it analyses the incoming text and tells whether the sentiment is positive, negative or neutral. With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Various scientists and researchers have been attracted to this field since analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics, also, it allows us to gain an overview of the wider public opinion behind certain topics. [12]

In Sentiment Identification in Covid-19 Specific Tweets by Manoj Sethi, Sarthak Pandey, Prashant Trar, Prateek Soni [1], it uses TF-IDF for tweets which is useful. It compares various ML algorithms by their accuracy on both bi-class and multi-class classification and also does cross-dataset evaluations.

Saad, Shihab Elbagir, Jing Yang performed study includes balancing and scoring of features [4]. The ML algorithms like SVR, Decision tree, Random Forest were used for the classification of the tweets, are inspired by ordinal regression. It was observed that the results were far more better using the ML algorithm, Decision tree, with an accuracy of 91.81%.

[2] Anuja P Jain, Padma Dandannavar use a hybrid model and use machine learning and lexicon based approach to perform classification. NLP algorithms are used for data preprocessin. The results obtained concluded that the decision tree model gives an accuracy of 100%.

The research by Jim Samual, G.G.Ali, Md. Mokhlesur Rahman, Ek. Esawi [7] in COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification, includes using Logistic regression and decision tree for the classification of tweets. According to their results, the logistic regression method provides a reasonable accuracy of 74% with shorter tweets and Naïve Bayes method gave strong classification accuracy of 91% for short tweets while both the methods showed relatively weaker performance for longer tweets.

In the paper by Zhao Jianqiang, Gui Xiaolin [11] a new word embeddings method that combines with n-grams features and word sentiment polarity score features to form a sentiment feature set of tweets has been included. On comparing this model with the baseline model the results show that their model performs better based on the accuracy and F1 measure for sentiment classification of tweets.

III. METHODOLOGY

This section explains the methodology of how we have implemented the project. It's done in four steps. The first being, gathering and collecting of data/dataset. This data has been collected from kaggle website as discussed earlier. Further, it is pre-processed to reduce complexity and remove noise which in turn increases the accuracy of the data to continue onto the next process. In the second step, relevant features are extracted in order to construct the model for classification further. In the third step, LSTM algorithm is used to classify tweets into positive, negative and neutral. Finally, in the fourth step we have used this analysis to give appropriate recommendations to the user.

Data Collection and Pre-processing:

i.Data Collection: The dataset collected from Kaggle website is used as the primary data source. The dataset comprises of 10,000 entries and includes columns titled username, location of the tweet, time at which the tweet has been posted, and the original tweet.

ii. Dataset Pre-processing: This is an essential step in order to build an efficient machine learning model, as the results depend on how well the data has been pre- processed which in turn impacts on the performance of the system. In general, the raw tweets contain many inconsistencies that include many misspellings, missing values and invalid data. This "dirty" data is cleaned to produce accurate and valid results. Therefore, the data is pre-processed before actually extracting features from them. This is done in the following steps, refer Fig. 1.:

Fig. 1. Steps followed during data pre-processing

IV. MACHINE LEARNING TECHNIQUES

Different machine learning models can be used to classify tweets according to their sentiments. The technique used in this study is as follows:

1. Long Short -Term Memory:

Long Short-Term Memory (LSTM) networks are a modified version of recurrent neural networks. These networks make it easier to remember past data in memory. The vanishing gradient problem, which is a drawback of RNN networks, is overcome by the LSTM networks. It uses long-term memory as opposed to the other algorithms that use short-term memory. These networks use internal gates to decide which data is relevant and discard the unnecessary data. It gives higher accuracy than other algorithms like SVM, Logistic Regression, etc.

Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points, but also entire sequences of data.

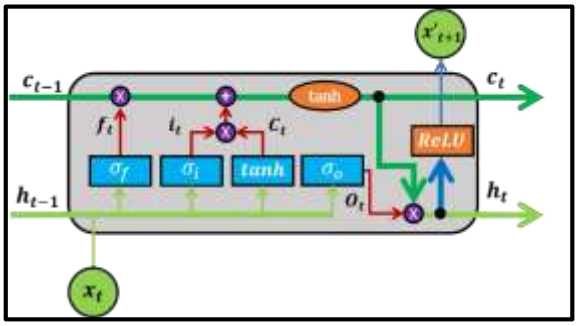

Fig. 2. LSTM architecture

According to the Fig. 2., a LSTM network consists of different memory blocks called cells that are responsible for remembering information and manipulations in this information is done by using three gates namely: Forget gate, Input gate and Output gate.

- *Forget gate:*

It is responsible for removing the information that is no longer required by LSTM to understand things.

- *Input Gate:*

This gate is responsible for the input activities into the cell state, and ensures the information that is added is important and not redundant.

- *Output Gate:*

This gate is responsible for all the output activities and the flow of call activations into the rest of the network.

Our model also considers the bidirectional property of LSTM network in order to read and process longer tweets easily without delays. The neurons of a regular LSTM are split into two, one for processing in forward state and the other for processing in backward state.

Further, the LSTM layers have an activation function, that transforms the weighted input from the node to the output node, basically, it is responsible to decide which values to keep and which ones to discard. In this paper, we have used ReLu which is a linear function that outputs the input directly if it is necessary, otherwise, it will output 0 to discard a piece of information.

V. GEO-LOCATION AND RECOMMENDATION SYSTEM

1. Geo-Location:

We have used Google API for geo-location mapping. Based on the output of LSTM network, we plotted dots on the global map. According to the polarity of sentiment, the colour of the dot is decided i.e. blue depicting neutral sentiments, green depicting positive sentiments and red depicting negative sentiments.

For geo-location mapping, we created a new training and testing dataset which contains latitudes and longitudes of every city. By comparing the 'city' in the tweets dataset with 'city' in the above mentioned dataset, we have plotted dots accordingly.

2. Recommendation system:

After clear analysis of tweets into positive, negative and neutral sentiments, we plotted this on world map as shown in the fig. Personalised recommendations are given to user after the user gives access to his location. The location is fetched and spotted on map using google API. The area where this location is pointed is compared with already existing information in our dataset regarding the polarity of tweets (above analysis using LSTM). Appropriate recommendation is given according to severity of negative sentiments to help user have a positive perspective.

VI. RESULTS AND DISCUSSION

The main aim of this study is to build a machine learning model that can analyse sentiments within COVID-19 specific tweets into positive, negative and neutral sentiments. Further, this system provides personalised and generalised recommendations. The experiments were done using Scikit-learn [11, 12], which is
an open-source assembly in python comprising many software packages.
Following figures Fig. 3. And Fig. 4. show classification of tweets according to the polarity of sentiments.

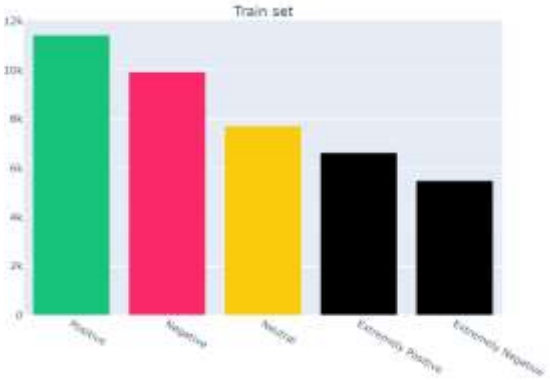

Fig. 3.

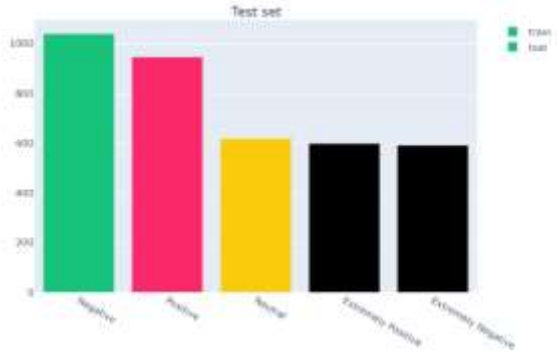

Fig. 4.

The following figure Fig. 5. shown is a confusion matrix of LSTM algorithm that we have used for classification of tweets into positive, negative and neutral sentiments. The table below shows overall accuracy of the system is 83%.

Fig. 5. Confusion matrix of LSTM

Table 1.Accuracy table

	Precision	Recall	F1-score	support
Negative	0.83	0.83	0.83	1633
Neutral	0.90	0.72	0.80	619
Positive	0.80	0.87	0.84	1546
Accuracy			0.83	3798
Macro avg	0.85	0.81	0.82	3798
Weighted avg.	0.83	0.83	0.83	3798

The figure Fig. 6. shows recommendation given to the user according to polarity of sentiments and location of user.

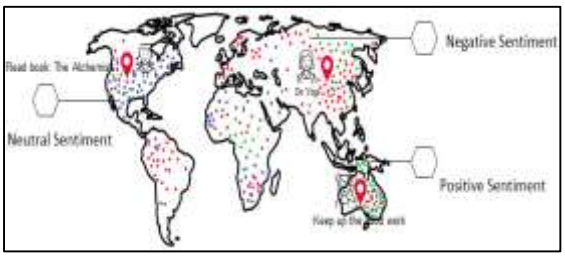

Fig. 6. Global map showing polarity of tweets in different locations

VII. CONCLUSION

The study in this project was done with the goal of creating a model which can effectively predict the sentiment expressed by people on social media platforms amidst this COVID-19 pandemic. From the experiments performed in this study, it is concluded that LSTM has performed extremely well and it was more robust and consistent throughout all the experiments. LSTM gave us correct analysis of tweets with accuracy of 83%. We have successfully plotted the analysis on the world map to understand the distribution better. Recommendations were given to user based on location and polarity of tweets. Hence recommendation using geo-location was successfully implemented.

VIII. FUTURE WORK

For the future work, this system can be improved by processing and working on bigger datasets and real-time data. Further, it can also be extended by processing different media options like images or audio/video by using image processing models.

IX. ACKNOWLEDGMENT

We would like to thank our mentor and guide Mrs. Dhanashri Patil, Faculty of Computer Department, Sinhgad college of Engineering, for her expertise, insights and guidance for this research. We would also like to express our immense gratitude towards Department of Computer Science And Engineering, Savitribai Phule Pune University for providing us the opportunity to explore and pursue research in the field of Computer Science.

X. REFERENCES

[1] Manoj Sethi, Sarthak Pandey, Prashant Trar, Prateek Soni, Sentiment Identification in COVID-19 Specific Tweets, ICESC 2020, IEEE Explore, e Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4, August 2020.

[2] Jain, Anuja P., and Padma Dandannavar. "Application of machine learning techniques to sentiment analysis." In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 628-632. IEEE, 2016.

[3] Saif, Hassan, Miriam Fernandez, Yulan He, and Harith Alani. "Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset , the STS-Gold." (2013).

[4] Saad, Shihab Elbagir, and Jing Yang. "Twitter Sentiment Analysis Based on Ordinal Regression." IEEE Access 7 (2019): 163677-163685.

[5] Naimul Hossain; Md. Rafiuzzaman Bhuiyan; Zerin Nasrin Tumpa; Syed Akhter Hossain "Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM"(2020)

[6] Saeed Mian Qaisar "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory"(2020)

[7] Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi and Yana Samuel, COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification, information, MDPI, 2020.

[8] World Wide Web

[9] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort , Vincent Michel, Bert rand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit -learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[10] Scikit -Learn: Machine Learning in Python. Accessed: Feb. 10, 2020. [Online]. Available: ht tp://scikit -learn.org/stable/.

[11] Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for twit ter sent iment analysis." IEEE Access 6 (2018): 23253-23260.

[12] https://towardsdatascience.com/senti ent-analysis-concept-analysis-and applications-6c94d6f58c17