UNIFIED APPROACH TO DISCOVER SENTIMENT ANALYSIS OF COVID-19 TWITTER DATA UTILIZING MACHINE LEARNING CLASSIFIERS

Abstract

Sentiment analysis has been emerging factor from Covid-19 wave. Finding out polarity of data cloud is not enough. Human emotions always give better idea about behavioral characteristics. Machine learning classifiers and its result surely gives impactful idea about specific condition. This chapter will give comparative and unified approach among machine learning classifiers.

Keywords: ML, Sentiment Analysis, NLP, Classifiers

Authors

Sudeep Kisan Hase
Research Scholar
Department of Computer Science Enginering Oriental University, Indore, India.

Dr. Rashmi Soni
Professor,
Dayananda Sagar Academy of Technology & Management, Bangalore
Research Supervisor, Oriental University, Indore, India.

I. INTRODUCTION

COVID-19 vaccines have brought much relief and newfound optimism to so many after a long time of sickness, devastation, grief, and hopelessness. Every day, news stories and Twitter spheres are filled with discussions about vaccination progress, accessibility, efficacy, and side effects. In spite of this, as online users, our visibility is very limited to the echo chambers that we create within ourselves. Hence, this chapter was motivated by a desire to increase my understanding of the global pandemic through Twitter data[2].

Since its first discovery in the Chinese town of Wuhan in Dec. 2019, the highly epidemic corona-virus affliction (COVID-19) has been transmitted to 212 countries and territories, influencing tens of millions of people. The disease was identified for the first time in a student travelling from Wuhan on the last day of January in 2020 in India, a country with a population of over 1.3 billion people[6].

COVID-19 disease and vaccines have been the subject of a lot of tweets, making it nearly impossible for a human to read through it all. Thus, the urge to better understand the global epidemic using Twitter data was the driving force behind this initiative.There have been so many tweets about 19 vaccines that it would take a human being a very long time to read them all.Natural language processing (NLP) allows us to acquire insight into an enormously complicated and broad conversation by examining narrative aspects, doing sentiment analyses, and visualizing word clouds[1][3].

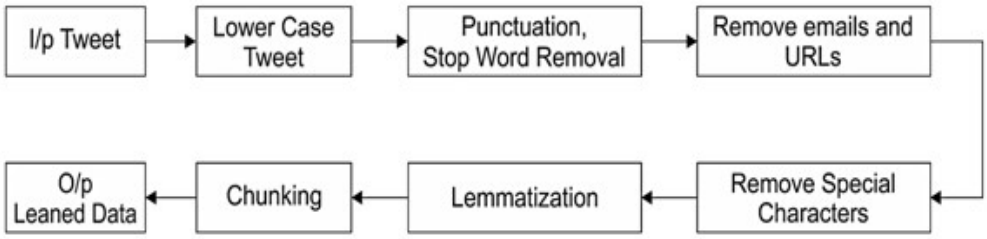


Figure 1: Sentiment Analysis Preprocessing Steps

The data which is fetched from twitter API has been preprocessed with some steps[4]. Lower case word will required for processing, so all the words are converted into lower case words. Comma, Apostrophe, Hyphen, Ellipsis, Full Stop, Exclamation Mark, Questions Marks, Colon, Bracket, Splash, Quotation mark etc. punctuation and stop words will be removed. The URL links and email IDs should be removed[5][7][8]. Word should be reduced to its base form. Then Noun phrases from sentence are extracted. At the end we will get output in the forms of words.

Text is scanned, processed, analyzed, and interpreted by a natural language processing system for textual data. The technology first preliminary processes the text via a number of phases to create a format that is more organized. The term "preprocessing stream" refers to a procedure where the outcome from one stage functions as the input for the subsequent one[11][12].

II. LITERATURE SURVEY

Table 1: Different Survey papers and accuracies of classifier

Classifiers used	Accuracy	Data Set Used	References
ML-CNN, ML-BERT, ML-CNNBERT	Accuracy of ML-CNNBERTis around 80%	Commodity Review	Ref 1 (2020)
ML-Text Based Classifier, ML-Senti-Reversal Prediction, Senti-diff	Senti-diff accuracy is around 80%	China Text Mining Service Provider	Ref 2 (2018)
ML-LR, ML-SVM, ML-NB	Accuracy of ML-LRis around 86%	Manual and Kaggle Dataset	Ref 3 (2020)
ML-LR, ML-NB	Accuracy of ML-LR is around 80%	Rutweet corporation twitter text	Ref 4 (2020)
ML-LR, ML-SVM, ML-RF	Accuracy of ML-LR under Word2Vec model is 80%	Myanmar FB dataset	Ref 5 (2020)
8 Various classifiers and LSTM-gate CNN	Accuracy of LSTM-gate CNN 73%	Global AI Chinese Dataset	Ref 6 (2020)
Occurrences, weightage and emotions	Different classifiers	5 different countries	Ref 7 (2020)
ML-Voted Classifier, ML-LSTM, ML-CNN_LSTM,	ML-LSTM gives 97% accuracy	Bitcoin, IMDB, GOP Debate live dataset	Ref 8 (2020)
ML-NB, ML-SVM, LM-RF	Accuracy of LM-RFis Higher	Amazon product Review	Ref 9 (2015)
ML-BERT, ML- (RMSE)	Accuracy of RMSE is around 93%	Humor data analysis	Ref 10 (2019)
ML- (CNN)	GPU Parallelism	Manual Twitter Data	Ref 11 (2017)
ML-XgBoost	Accuracy of XgBoost is around 97%	Covid Senti Dataset	Ref 12 (2021)
ML-(LSTM), and ML-(CNN)	Accuracy of LSTM is around 84%	US airline, IMDB and GOP debate dataset	Ref 13 (2020)
ML-SVM,ML-(ULMFit SVM)	Accuracy of ULMFit SVM is around 99%		Ref 14 (2022)

In Ref.1 typical CNN and BERT classifiers are compared with BERT-CNN model. Mobile text dataset of JD mall used for this experiment. This model gives more than 80% accuracy. Deep earning method explores features from commodity reviews. Ref. 2 introduced dataset through data mining service provider in china. They have examined tweets and retweets and added special algorithm to find out relationship between textual messages. Different methods and sentiment diffusion collectively work better upto 78%. PR-AUC percentage for invented senti-diff method works well on proposed algorithms Ref. 3 used

manual and kaggle dataset having 4 attributes and 1 million instances extracted from twitter. ML-LR, ML-SVM, ML-NB classifiers tested. The ML-LR classifier gives highest accuracy among these 3 classifiers. Ref4. In this work Russian language data extracted from Rutweetcorporation. Around 1 Lakh data has been processed. ML-LR, ML-NB classifiers applied on N-grams where ML-LR gives 80% accuracy. Ref. 5 Dataset taken from Myanmar FB pages. The Myanmar special font then converted into English language. ML-LR, ML-SVM, ML-RF these algorithms are applied for training and testing. Word2vec and tfidf model used for the experiment. ML-LR give 80% accuracy in this experiment.Ref. 6 Global AI Chinese dataset used for the experiment. Some aspect terms and 4 emotions studied. Among 8 classifiers LSTM-gate CNN Classifiers gives 73% highest accuracy. This model is the combination of 2 classifiers.Ref. 7 From year 2007-2010 various ecommerce dataset has been extracted and studied different classifiers. Data from 5 countries evaluated. Occurrences and weightage calculated on different themes. Ref. 8 Kaggle and Twitter website data used. Bitcoin, IMDB, GOP Debate live dataset used for experiment. Among different classifiers ML-LSTM gives highest accuracy of 97%. Ref 9. Work is done on Amazon product review dataset. ML-NB, ML-SVM and LM-RF classifiers are used for experiment. Based on ROC curve LM-RF gives 98% coverage.

III. WORKING

1. Preprocessing

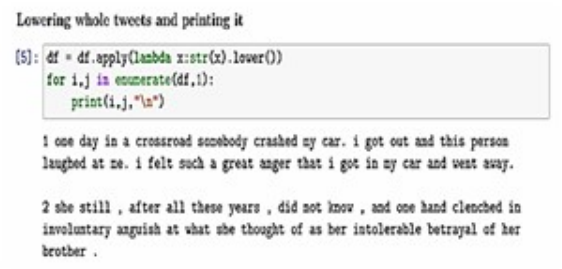


Figure 2 : Step1 -Lowercase String

Figure 3 : Step2-Counting words

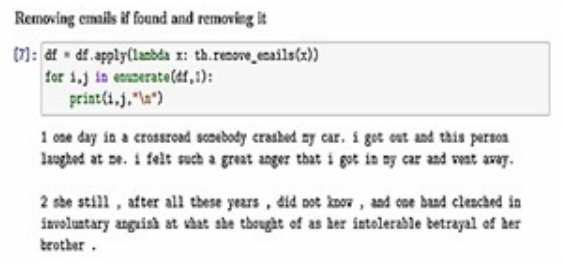


Figure 4 : Step3–Deletion of Email IDs

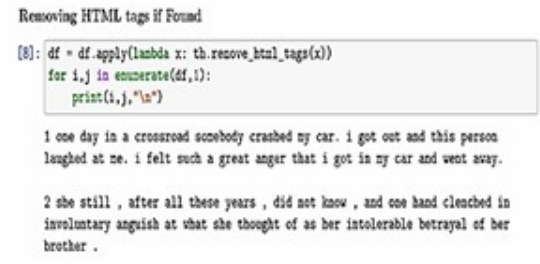
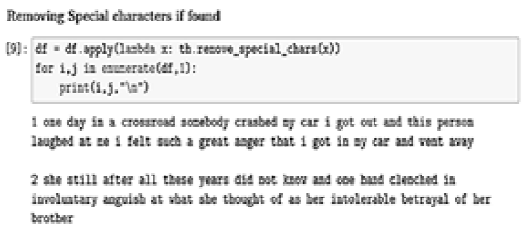


Figure 5 : Step4- Delete Hypertext ML Tags



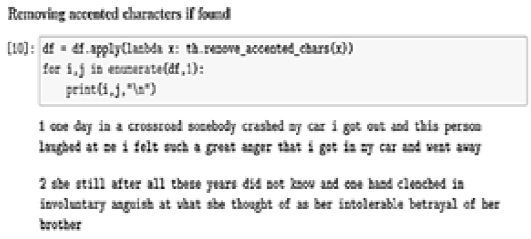


Figure 6: Step5–Deleting Special Characters

Figure 7: Step6- Deleting Western Lang Accented Words

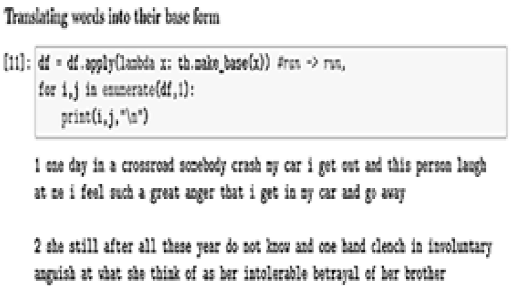


Figure 8: Step7–Returning String to Base Form

Figure 9: Step8- Deleting Stopwords

These preprocessing steps have cleaned twitter data so that we can perform training and testing phases on it. If these steps are not follow then we will get false result in classifier accuracy

2. **Experimental Workflow:** The first and foremost process is carrying out data from the Twitter. Tweepy API will give interface to extract data from twitter login. Kaggle provides service of machine learning dataset, which is community based model.

   Registering with this service, we will be able to get experimental and worked dataset[14].

   Natural Language Toolkit, Scipy and other preprocessing packages are available. We can remove unstructured data from the dataset. Word count and token created for data set with the help of tokenizer.

   As you can see in the fig.10 and 11, for training and testing 5 different classifiers are taken for the experiments. The output of this training and testing classified in 6 different human emotions. After probability distribution accuracy of these classifiers drawn[9][13].

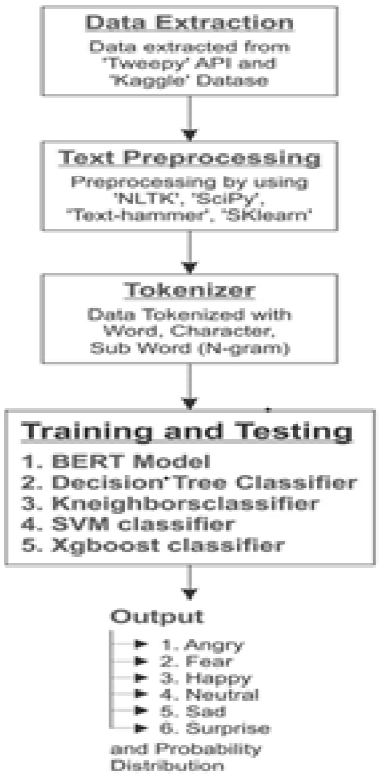


Figure 10: Experimental Workflow for Covid-19 Twitter Data Sentiment Analysi

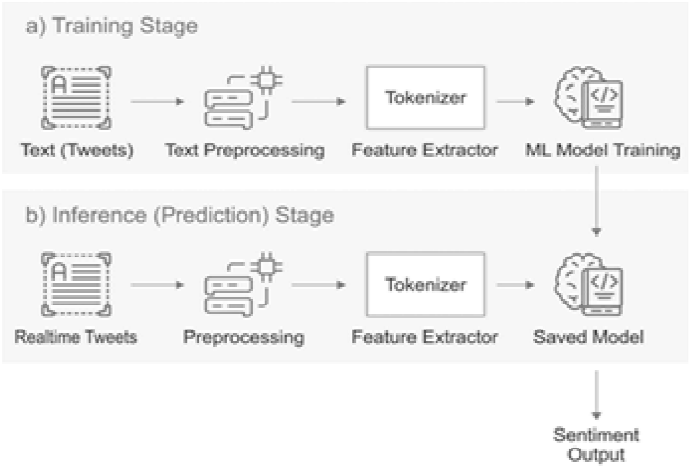


Figure 11: Training and Prediction Phase of Covid-19 Twitter Data Sentiment Analysis

IV. RESULTS

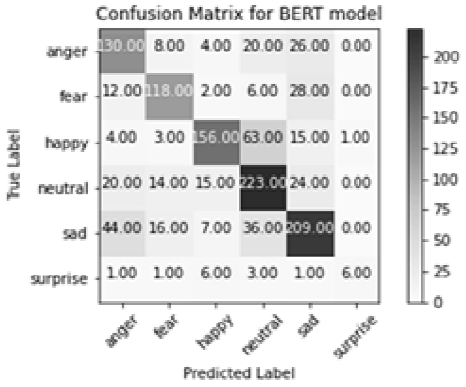


Figure 12: BERT model Confusion matrix

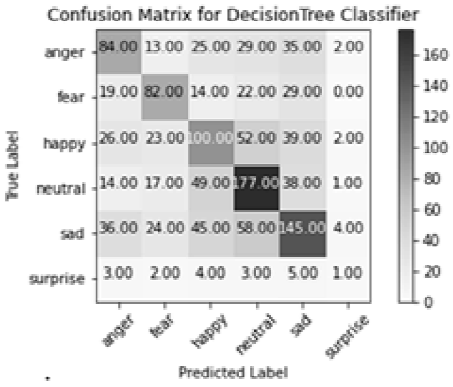


Figure 13: DT Confusion Matrix

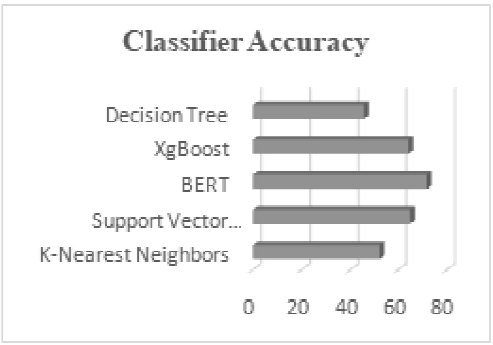


Figure 14: Five Classifier Accuracy

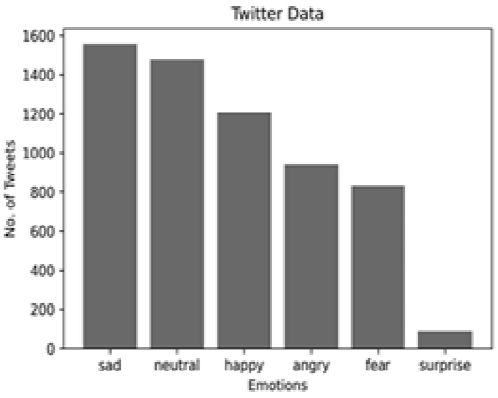


Figure 15: Human Emotion based on

As we can see in fig. 12 and 13 we have carried out confusion matrix for 5 classifiers. After Calculating precision, Recall, F-1 score for 5 classifiers, we came to know that BERT classifier gives best result in terms of accuracy (Fig. 14). Result also shows people are sadder in Covid-19.

V. CONCLUSION

After Covid-19, it is very important to know human emotions based on twitter data. This work provides best result on five classifiers and on Kaggle dataset. In future, more classifiers will be experimented on same data.

REFERENCES

[1] J Dong, F He, Y Guo, H Zhang, "A Commodity Review Sentiment Analysis Based on BERT-CNN Model", 5th International Conference on Computer and Communication System, IEEE Xplore 2020

[2] Lei, Niu, Yu," SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis", IEEE, journal of latex class files, vol. 14, no. 8, august 2018, 1041-4347 (c) 2018

[3] Purnima A, K. Priya," A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques", 6th International Conference on Advanced Computing & Communication Systems (ICACCS), 978-1-7281-5197-7/20/$31.00 ©2020 IEEE

[4] Egor, Evgeniy, Agbozo," Assessing the Impact of Text Preprocessing in Sentiment Analysis of Short Social Network Messages in the Russian Language", International Conference on Data Analytics for Business and Industry: Way towards a Sustainable Economy (ICDABI), (ICDABI) | 978-1-7281-9675-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICDABI51230.2020.9325654

[5] Mar Su, Win Pa," Analysis of Word Vector Representation Techniques with Machine-Learning Classifiers for Sentiment Analysis of Public Facebook Page's Comments in Myanmar Text", 2020 IEEE Conference on Computer Applications(ICCA), DOI: 10.1109/ICCA49400.2020.9022842, IEEE Xplore: 05 March 2020

[6] Cao, Gao,"LSTM-Gate CNN Network for Aspect Sentiment Analysis", th International Conference on Information Science, Computer Technology and Transportation (ISCTT), 978-1-7281-8575-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISCTT51595.2020.00084

[7] Casas, Faz," The synergic relationship between e-commerce and Sentiment Analysis: A content analysis of published articles between 2007 and 2020", 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) | 978-1-7281-9675-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICDABI51230.2020.9325689

[8] Chandra, Jana," Sentiment Analysis using Machine Learning and Deep learning", 7th International Conference on Computing For Sustainable Global Development (INDIACom), 978-93-80544-38-0 /20/$31.00 2020 IEEE

[9] Fang, Zan,"Sentiment analysis using product review data", Journal of Big Data, a SpringerOpen Journal, DOI 10.1186/s40537-015-0015-2, 2015

[10] Mao, Liu, "A BERT-based Approach for Automatic Humor Detection and Scoring," Proceedings of t heIberian Languages Evaluation Forum (IberLEF 2019).

[11] Campos, Sastre, Francesc and Maurici and Bellver, Nieto, Xavier and Torres, Jordi, "Distributed training strategies for a computer vision deep learning algorithm on a distributed GPU cluster", journal Procedia Computer Science, volume 108, pages 315324, year 2017, publisher Elsevier.

[12] Jalil, Abbasi, Javed AR, Khan M, Abul Hasanat MH, Malik KM and Saudagar AKJ (2022), "COVID-19 Related Sentiment Analysis Using State-of-the-Art Machine Learning and Deep Learning Techniques." Front. Public Health 9:812735. doi: 10.3389/fpubh.2021.812735https://xgboost.readthedocs.io/en/stable/

[13] Kariya, "Twitter Sentiment Analysis", 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020

[14] AlBadani, Dong J., "A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM." Appl. Syst. Innov. 2022, 5, 13. https://doi.org/10.3390/asi5010013

[15] Hase Sudeep Kisan, Hase Anand Kisan, Aher Priyanka Suresh, "Collective intelligence & sentiment analysis of twitter data by using StandfordNLP libraries with software as a service (SaaS)", 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2016.7919697, Electronic ISSN: 2473-943X

