Intelligent Sentiment Analysis of Pharmaceutical Reviews Using TF-IDF and Logistic Regression

MANDA LAKSHMI NAGA SAI SRAVANTHI,
PG scholar, Department of MCA, DNR college, Bhimavaram, Andhra Pradesh.
A. Durga Devi
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: ***The various diseases attacking the human body, such as the coronavirus and so on; nowadays, due to the increase in infections, there are no systems and medical experts so that patients can take medicines at their own risk. Still, they cause severe damage to the patient's body and cause death. To solve that problem, the author introduces the drug recommendation system based on machine learning and sentiment in this paper. They can take the name of the disease from the patient, recommend the drug for the given condition, and provide the SENTIMENT based on the experience of earlier users. If the rating is high for the predicted disease, then patients recommend and trust the drug. The TF-IDF (Term frequency-inverse document frequency) algorithm is used to extract features. We use different machine learning algorithms to determine accuracies, such as the SGD classifier, Multilayer perceptron classifier, Nave Bayes, Ridge classifier, Linear SVC, Logistic Regression, WORVEC, and BAG of WORDS. These features extracted will be applied to the different machine learning algorithms. We use the TF-IDF feature extraction algorithm among all the algorithms because it performs best. The UCI machine learning website used a DRUG REVIEW dataset to implement the project.***

Keywords: Machine learning classification, Feature Extraction, Sentiment analysis, Drug recommendation system.

I. Introduction

The drug recommendation system offers machine learning by using feature engineering and sentiment analysis on patient reviews on a specific condition. In the paper, we develop a strategy for medicine recommendation that uses patient reviews using different vectorization processes. That is manual feature analysis, TF-IDF, Word2Vec, and Bow to predict the sentiment that can help us with the recommendation of the top drug by various algorithms of classification for a disease.

We see a shortage of doctors in rural areas with an exponentially increasing number of coronavirus cases. Compared to urban areas and rural areas, there is a shortage of medical experts in rural areas. It takes approximately 6-12 years to acquire the necessary qualifications. The number of doctors cannot be rapidly increased in a short period. In this challenging time, the telemedicine framework ought to emerge as soon as possible. The drug recommendation system uses machine learning to analyze sentiment in drug reviews.

Because of the prescription mistake, In the USA and China, over 200 thousand individuals and 100 thousand individuals are affected annually, respectively. Since experts only have a limited amount of expertise, they often make mistakes while prescribing medicine—more than 40% of the time [2][3]. Choosing a high level of medication is essential for the patient. A specialist needs to know a wide range of information, such as patients, antibacterial medications, and microscopic organisms.[6]

A new clinical study with tests, drugs, and other components is launched every day. As a result, choosing medication or treatment for a patient based on past clinical history and indications proves to be ever more difficult for doctors. Acquiring worldwide items has become an integral and imperative factor of the web with the exponential development of the web business industry.

To analyze the reviews, individuals worldwide become adjusted. To buy a thing, visit the website before settling on a choice. While the majority of prior research focused on evaluating expectations and proposals for the E-Commerce

industry, the area of healthcare or therapeutic medicine has only sometimes been covered. The number of people searching online for a diagnosis because they are concerned about their health has increased.

In 2013, the centre survey demonstrated in Pew American research [5] that roughly 60% of grownups searched online for health-related subjects. On the web for diagnosing health conditions, around 35% of users looked. A medication recommendation system is necessary to aid patients and doctors in expanding their knowledge about medications for particular medical conditions. A typical system known as a recommender framework makes an item recommendation to the user based on their benefit and need. These frames employ the customer's surveys to suggest a piece of advice and break down their sentiment for their exact needs.

Machine learning is offered in the drug recommendation system by using feature engineering and sentiment analysis on patient reviews on a specific condition—tools for extracting and distinguishing emotional data from a language, like attitudes and opinions. Sentiment analysis is a progression of strategies, tools, and methods. [7] In contrast, the "feathering engineering" process involves adding new features to existing ones to enhance model performance.

This research project is divided into five sections: For this research, the introduction section presents a brief explanation of the need; on this topic, the related works section provides a quick overview of prior studies; research methodologies include the methodology section, and the conclusion section summarizes the findings. The applied model results are evaluated using a variety of metrics in the result phase, and the discussion segment lists the framework's limitations before moving on to the conclusion.

II. Literature survey

To recommender frameworks applying deep learning and machine learning strategies, there has been an exertion in AI advancement with a sharp increment. The restaurant, e-commerce, and travel industry have very regular recommender frameworks. In the production of the drug, the medication reviews contain clinical terminologies like names of infections, reactions, and synthetic characters that are used; they are much more challenging to analyse. Using sentiment analysis, as a result, there are sadly few studies in the field of drug proposal framework [8]. A semantic-powered online framework, Galen OWL, is described in a survey to help specialists discover details [9]. For a patient, the paper depicts a framework that suggests drugs, such as drug interactions, sensitivities, and patient infection.

Using international standards like UNII and ICD-10, clinical data and terminology were correctly combined with Galen OWL clinical information. . For a patient to locate the best treatment prescription, examined a large-scale Leilei sun on [10] treatment records. The plan was to estimate the similarities between treatment data using a powerful semantic clustering technique. The author also developed a framework to evaluate the suitability of the recommended course of treatment. This framework can provide optimal treatment plans according to new patients' demographics and medical difficulties. For testing from numerous clinics, electrical medical records of patients are gathered.

The outcome demonstrates how this paradigm raises the cure rate. In that study, multilingual sentiment analysis was carried out using Recurrent Neural Network and Naive Bayes [11]. (RNN). Multilingual tweets were translated into English using the Google Translator API. The outcomes show that Naive Bayes with 95.34% RNN outperformed as compared to 77.21%. The study [12] is founded on the idea that the patient's capacity should determine whether medication is advised. A reliable remedy should be recommended if the patient's immunity is low.

A risk technique level classification was suggested to determine the immunity of patients. For instance, more significant than the 60 risk variables, such as alcoholism, hypertension, and other characteristics, has been adopted and evaluated on how the patient can protect from the infection. A web-based prototype was constructed using a decision support system to assist physicians in selecting first-line medications. Xiaohong Jiang et al. [13], based on treatment data, investigated three various algorithms the support vector machine (SVM), the back propagation neural network and the decision tree algorithm,

Medication errors can be challenging to underestimate and understand despite contributing significantly to patient morbidity and mortality. On medication errors for practising physicians that focus that article provides a review.1) Legal and consequences and disclosure, 2) Avoidance strategies, 3) risk factors, 4) incidence, and 5) definitions and terminology. In the medication use process, at any point, a medication error is any error that occurs. The Institute of Medicine has calculated that pharmaceutical errors result in 1 of 854 inpatient deaths and 1 of 131 outpatient fatalities. Medication errors may be influenced by healthcare professional factors, patients and medication.

Following drug errors, doctors may have adverse effects such as Loss of patient trust, medical board reprimands, legal lawsuits and criminal accusations. Methods for preventing drug errors have been tried with various degrees of effectiveness. To avoid the mistakes in future communication of efforts and apology, in personal revelation, patients want quick disclosure when an error is identified to the patient to provide safe care—learning more about the medication can enhance health care.

Medication mistake is a significant cause of mortality and morbidity, but it is an underestimated and complex concept. A pharmaceutical error is any error that occurs during the medication administration process. Medical board discipline, criminal charges, civil litigation, and Loss of patient trust are all expected outcomes for physicians who make drug errors. Methods for preventing drug errors have been tried with various degrees of effectiveness. When an error is identified, most patients want measures to avoid future mistakes, an apology, in-person disclosure, and immediate disclosure. **[2]**

1. Poor prescribing is likely the most common source of preventable pharmaceutical errors in hospitals and many more incidents involving newly graduated junior doctors. Prescribing is a complex skill that requires a thorough knowledge of drugs, an awareness of the capacity to weigh, preferably, clinical pharmacology principles, and experience of benefit and risk. Unsurprisingly, mistakes occur.
2. Being a prescriber is likely more difficult now than it has ever been. In the last 20 years, medical education has evolved dramatically, reflecting concerns about an overcrowded curriculum and a lack of emphasis on social sciences. In the United Kingdom, these developments have resulted in less clinical pharmacology and practical prescription as required components of undergraduate training and assessment. There is a rising concern among students that medical school education does not adequately prepare them for the pressures of becoming prescribers. Other countries have expressed similar worries. There is circumstantial evidence that these changes are related to pharmaceutical errors discovered in practice, but there is no irrefutable evidence.
3. According to a systems analysis of errors, knowledge and training are essential elements in error causation, and concentrated instruction enhances prescription performance. We feel that there is already enough data to thoroughly examine how students are trained to become prescribers and how these abilities are developed in the postgraduate years. We offer a set of guiding concepts upon which training could be based.

Better medical education has evolved dramatically frequently in recent years. However, it is regrettable that particular therapeutics courses supporting safe, successful treatment and clinical pharmacology have been lost. Some believe that once students are exposed to the clinical setting, this learning area will 'take care of itself.' This has been proven untrue. We believe that kids must be thoughtfully prepared and actively guided to get the most from this type of learning. For safeguarding patients from prescription errors, Prescriber education and training are only of the strategies. Help from other colleagues will be essential, as will the growth of electronic prescribing with decision support. Still, we believe that prescribers' expertise and intuition will be their most important defence against illogical and hazardous medication usage. **[3]**

In the United States, due to infectious diseases, Lower respiratory tract infections are the

leading cause of death worldwide and the leading cause of death. Recent breakthroughs in the discipline include microbial detection technologies, antibiotic agents and the discovery of new diseases. Despite substantial research, few medical disorders are as contentious as treatment.

The current guidelines are the IDSA's amended recommendations. These guidelines are designed to reflect updated facts, provide more thorough suggestions in certain areas, and illustrate an opinion evolution compared to previous approaches. These treatment guidelines apply to immunocompetent persons with community-acquired pneumonia (CAP). Recommendations are ranked alphabetically based on their Roman numeral and strength based on the quality of supporting evidence. This is the usual practice for IDSA quality standards [5]. A set of guidelines cannot address the plethora of variables that influence antibiotic selection, location of care, and diagnostic evaluation. As a result, these principles should not be used to replace sound clinical judgement**. [6]**

To determine what medical ailment, 35% of U.S. adults especially say they have gone online to try them or someone else may. These results stand based on a national poll conducted by Pew American Life Project and the Centre's Internet research. In this research on the internet, we refer to individuals who sought answers as online diagnoses in all sections. If the information when asked they discovered online led them to believe in the need to pay attention to a medical professional, 46% of internet diagnoses claim it did. 38% of online diagnoses said it was something they could handle at home, while 11% said it was both or something in between. When we asked respondents by the medical expert about the accuracy of their original diagnosis, they stated that 41% of internet diagnoses were confirmed.

Another 2% claim a medical professional partially validated it. 35% admit they did not seek professional advice from a clinician. 18% report consulting a medical expert who disagreed or had a different view of the condition. With a clinician, 1% of people report that their talk was inconclusive. Women are more likely than males to look up a probable diagnosis online. It is critical to understand what these findings and what they do not mean. To answer their questions about their health at home and make personal decisions, People have traditionally attempted about whether or not to visit a professional. Many people have now added to the personal health toolbox on the internet, allowing them and their loved ones to grasp better what is wrong with them. It assesses the scope of the action but not its outcome. **[7]**

EBM is the intentional, transparent, and sensible use of current best evidence in making decisions about individual patients' care. Extracting information from large amounts of text is both beneficial and challenging. The collection is studied using a generative probabilistic aspect mining technique. In aspect-based opinion mining, the frequency-based approach extracts high-frequency noun phrases, while the relation-based method detects aspects based on the aspect-sentiment relationship in reviews. However, in drug reviews, the authors do not expressly state the elements, and the depiction of side effects and people's experiences varies.

The co-occurrence of words in reviews is used to identify aspects using Topic Modelling. A finer-grained aspect level opinion mining is employed in this research Topic Modelling Based on Probabilistic Approach. Using the model to discover elements related to distinct data segmentation, such as different age groups or other attributes, is intriguing. Working with aspect interpretation is also fascinating because a collection of keywords now represents aspects. Performance and understanding will improve if a few phrases can be automatically retrieved or constructed to summarise the keywords.

Opinion mining is a popular study topic these days. Many firms are now using modern web technologies to improve their offerings. With this rate of advancement, the analytics techniques employed for these web bodies must also advance. Thus, commercial and data analytics strategies are applied in the Online Reviewing section of web technology. Several methods for extracting, assessing, and interpreting vast data corpora are available as study topics. More fine-grained analytics approaches are required with the daily increase in data volumes. Its widespread use in review-related websites, as a subcomponent technology, and in business intelligence makes it a beautiful topic of study. **[8]**

III. Proposed method:

This paper used a dataset from the Drug Review Dataset (Drug.com) taken from the UCI ML Repository. That dataset consists of 6 attributes, Condition of a patient, suggested proper count, the number of Individuals who found the review helpful, Determining a patient's contentment, 10-star Patient Rating overall, Review of a patient, Date of review entry, and Name of drug used. A total of 215063 instances are shown in fig.1. It contains four stages, recommendation, evaluation, classification and data preparation.

A. Data cleaning and visualizations: Standard data preparation approaches were used in this study, including checking for deleting duplicate rows, null values, and text from rows and removing superfluous values. To avoid duplication, ensure that a unique id is indeed impressive.

B. Feature extraction : For sentiment analysis to build classifiers, a proper data set-up is required after text pre-processing. The text should be converted to numbers, so machine learning algorithms can process it. Textbooks cannot be processed directly by these algorithms, specifically numerical vectors. In this study, the feature extraction bag of words TF-IDF, Word2Vec and (Bow) are known as the simple methods from text data. To manually extract features using some feature engineering approaches from the review column, another model called manual part was created in addition to TF-IDF, Word2Vec, and Bow.

i) Bow:

A technique called "bag of words" is used processing in natural language to count how many times each token appears in a text or review.

ii) TF-IDF:

The TF-IDF weighing approach, in which words are supplied with the weight but not count, is very common. The idea was to give terms that frequently appear in the dataset a low priority, implies that TF-IDF determine relevance rather than recurrence. The Term frequency (TF) is a measure of how often it is that a term will be found in a document.

iii) Word2Vec:

In the various natural languages preparing tasks by using the TF-IDF and TF is popular vectorization, methods. They disregard the syntactic and semantic likenesses between the words. For instance, although being almost similar, the phrases lovely and pleasant are classified as two distinct terms in both TF and TF-IDF vectorization algorithms.

iv) Manual Features:

To improve the models accuracy is feature engineering a common idea that helps. We employed fifteen features, including full count, label encoding of the condition column using the label encoder function from the Scikit package, and using the Date Time function in pandas the development of day, month, and year features from the date column.

C. Train Test Split: We created four datasets using manual features, Word2Vec, TF-IDF, and Bow. For testing and training purposes, we split the dataset into 25% and 75% resp. We selected an equal random state when breaking the data to ensure that the train-test split of all four produced datasets used the same set of random values.

D. Smote:

After the Train Test split, to avoid the class imbalance issue, only the training data was subjected to the synthetic minority over-sampling technique (Smote). Smote is a method of oversampling that created new data from old data. By linearly combining a minority instance, a was randomly chosen, and its k nearest neighbours, instance b, in the feature space, Smote creates new minority class data.

E. Classifiers:

The distinct machine learning classification algorithms predict the sentiment to build a classifier. Ridge,

perceptron, Logistic regression, linear support vector classifier, multinomial naïve Bayes, classifier experimented with TF-IDF, gradient descent, stochastic gradient descent. Ridge, perceptron, Logistic regression, linear support vector classifier, multinomial naïve Bayes, classifier experimented with TF-IDF, gradient descent, stochastic gradient descent experimented with the Bow, the model TF-IDF they have highly sparse matrices, thus using classifiers based on trees would take a lot of time. Used Word2Vec and manual features model with random forest, cat boost classifier, decision tree, and LGBM.

F. Metrics

Five metrics, including AUC score, f1score (F1), accuracy (Acc.), precision (Prec), AUC score, and recall (Rec), were used to assess the predicted sentiment

Tp = true positive.

Tn = True negative

Fp = False positive,

Fn = False negative, f1score, accuracy, Precision, recall shown in equations given below,

$$Precision = \frac{Tp}{Tp+Fp}$$

$$Recall = \frac{Tp}{Tp+Fn}$$

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn}$$

$$F1score = 2.\frac{Precision.Recall}{precision+Recall}$$

G. Drug Recommender system:

The four best findings are selected, giving the combined prediction after evaluating the metrics. For the medicine to produce an overall score, a properly normalised count was multiplied by the combined data for a given condition. Higher the score better the medication. Examining the functional count distribution was the impetus for standardising the helpful count.

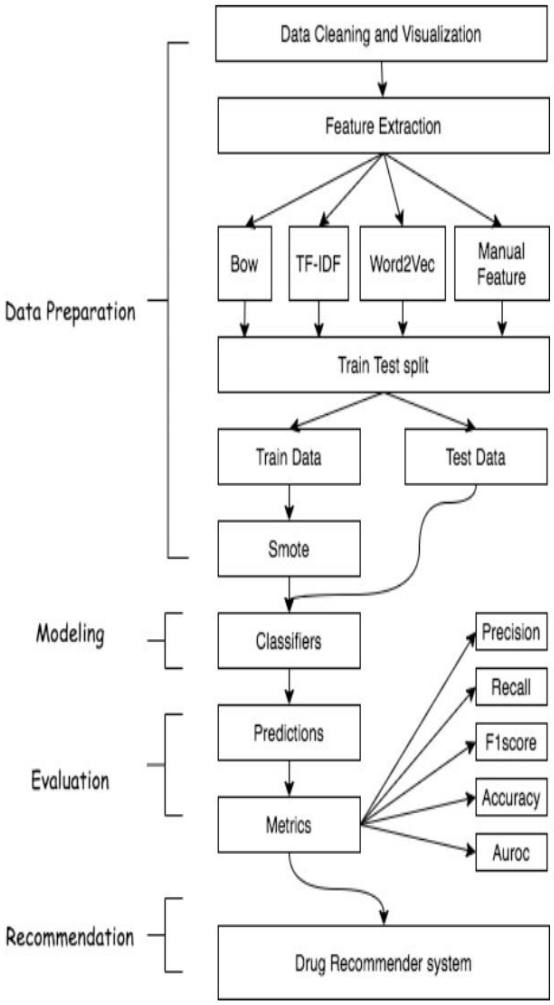

Fig.1 Flowchart of the proposed model.

Result

We are using machine learning in the sentiment analysis of the drug reviews in the Drug Recommendation System.

Various diseases attack the human body, such as the coronavirus. Nowadays, due to the increase in infections, there are no systems and medical experts, so at their risk, patients can take medicines which cause severe damage to the patient's bodies and cause death. To solve that problem, the author introduces the drug recommendation system based on machine learning and sentiment in this paper.

They can take the name of the disease from the patient, recommend the drug for the given condition, and provide the SENTIMENT based on the experience of earlier users. The patient recommends and trusts the drug if the rating is high for the predicted disease. The TF-IDF (Term frequency-inverse document frequency) algorithm is used to extract features.

We designed a paper to implement the modules;

1. Upload the drug Review dataset: To the application, we will upload the dataset using that module.
2. Pre-process and Read dataset: We will review all drug names using that module from a feature array and rating from the dataset.
3. TF-IDF Extraction feature: It finds the average frequency for each word and then replaces that word with a vector and the frequency value. If the word is not found in the sentence, then put 0. To RATINGS and machine learning, all reviews will consider input features. And as a class label, the drug name will be considered.
4. Train the machine learning Algorithm: We will input TF-IDF using that module to all machine learning and then train a model. This model will apply to test data to determine the prediction algorithm's accuracy.
5. Comparison Graph: We will plot a graph of the accuracy of using that module for each algorithm.
6. Recommend the drug from the test data: We upload the name of the disease using that module then machine learning predicts the rating and character of the drug.

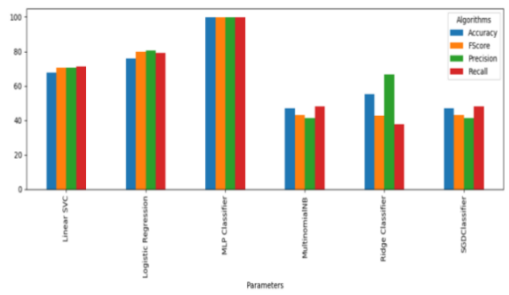

Fig.2.Comaprion Graph

We use a 6 different machine learning algorithm for the determination of accuracy such as SGD classifier, Multilayer perceptron classifier, Naïve Bayes, Ridge classifier, Linear SVC, Logistic Regression. From that algorithm we get more accuracy in MLP classifier, so we recommend predicting the MLP classifier.

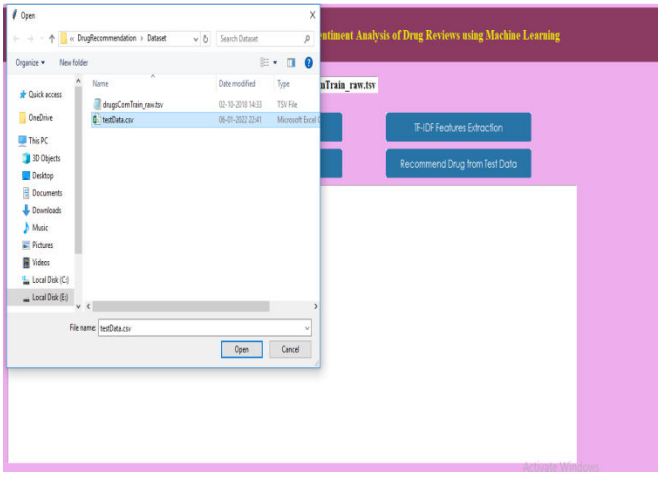

Fig.3.Select and upload test data

Select and upload the testData.csv file and load the test data

Fig.4. Recommended drug and rating

We get a name of the each disease, recommendation drug and the rating according to given disease name.

Conclusion

Now, Reviews have become an essential part of our daily life. As if we go to some

restaurant, purchase some online, or go shopping, we check how many reviews are there to make the correct decision. In this paper, we build the recommendation system of sentiment analysis of drug reviews using various types of machine learning. That classifiers are linear SVC, logistic regression, ridge classifier, Perceptron, stochastic gradient descent, multinomial naïve Bayes, applied on the bow, TF-IDF and the classifiers such as cat boost, LGBT, random forest, decision tree applied on the manual feature method. We determine them using the five various metrics, AUC score, accuracy, F1 score, recall, and precision. On IF-IDF, the linear SVC outer performs all the models with an accuracy of 93%.

References

[1] Telemedicine, https://www.mohfw.gov.in/pdf/Telemedicine.pdf

[2] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. Mayo Clin Proc. 2014 Aug;89(8):1116-25.

[3] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. Practical Journal of Clinical Medicine, 4.

[4] Drug Review Dataset, https://archive.ics.uci.edu/ml/datasets/Drug%2BReview%2BDataset%2B%2528Drugs.com%2529#

[5] Fox, Susannah, and Maeve Duggan. ”Health online 2013. 2013.” URL: http://pewinternet.org/Reports/2013/Health-online.aspx

[6] Bartlett JG, Dowell SF, Mandell LA, File TM Jr, Musher DM, Fine MJ. Practice guidelines for the management of community-acquired pneumonia in adults. Infectious Diseases Society of America. Clin Infect Dis. 2000 Aug;31(2):347-82. doi: 10.1086/313954. Epub 2000 Sep 7. PMID: 10987697; PMCID: PMC7109923.

[7] Fox, Susannah & Duggan, Maeve. (2012). Health Online 2013. Pew Research Internet Project Report.

[8] T. N. Tekade and M. Emmanuel, ”Probabilistic aspect mining approach for interpretation and evaluation of drug reviews,” 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1471-1476, doi: 10.1109/SCOPES.2016.7955684.

[9] Doulaverakis, C., Nikolaidis, G., Kleontas, A. et al. Galen OWL: Ontology-based drug recommendations discovery. J Biomed Semant 3, 14 (2012). https://doi.org/10.1186/2041-1480-3-14

[10] Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and Yanming Xie. 2016. Data-driven Automatic Treatment Regimen Development and Recommendation. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). Association for Computing Machinery, New York, NY, USA, 1865–1874. DOI:https://doi.org/10.1145/2939672.2939866

[11] V. Goel, A. K. Gupta and N. Kumar, ”Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing,” 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 208-212, doi: 10.1109/CSNT.2018.8820254.