INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 07 | JULY - 2023 SJIF ATING: 8.176 ISSN: 2582-3930

Opinion Mining with Hypertuned Random Forest

Dhruvin Hasmukh Barot
(crce.9116.ecs@gmail.com)
(Electronics and Computer Science Engineering college of engineering)
Fr. Conceicao Rodrigues College of Engineering, Mumbai, India.

Arpita Ashok Khot
(khotarpita9@gmail.com)
(Electronics and Computer Science Engineering college of engineering)
Fr. Conceicao Rodrigues College of Engineering, Mumbai, India.

Praveen Raju Bhandari
(bhandaripraveen3105@gmail.com)
(Electronics and Computer Science Engineering college of engineering)
Fr. Conceicao Rodrigues College of Engineering, Mumbai, India.

Abstract *- E-tailing has seen a lot of development, while brick-and-mortar had a significant downfall. A few reasons for that was that lack of traveling, comparisons between best prices, offers, sales etc. A beneficial feature in online shopping is the reviews about the product. The purchasers give their product experience which helps other consumers to select what is best suited for themselves. But the purchasers are in huge amounts and thus their reviews are also in greater quantity, therefore it becomes very difficult for the consumer to find the Positive, Neutral and Negative reviews. Here Machine Learning comes into picture, by using various ML models one can segregate all types of reviews and classify them into different sections so that the consumer can directly get a comparative analysis about his/her product. This is known as Opinion Mining also commonly known as sentimental analysis. Applications are that its uses cut across sectors, aiding market research, customer satisfaction enhancement, and decision-making. Businesses and organizations may gain a competitive edge and better serve the requirements of their target audience by comprehending and utilizing the power of public opinion. The Fundamental idea of this paper is to discuss the sentimental evaluation and categorize them into their respective polarities based on the given sentence.*

Keywords: *Sentimental analysis, Positive, Neutral, Negative, Natural Language Processing, Random Forest, SVC, Logistic Regression, Gaussian Naive Bayes, KNN, Grid search view, Random Search, Cross Validation, Lemmatization,Tokenization, Vectorization .*

INTRODUCTION

Sentiment evaluation, also known as opinion mining, is a captivating area of natural language processing (NLP) that entails analyzing text information to determine the emotional tone or sentiment expressed by the author. With the explosion of social media and purchaser reviews, sentiment evaluation has become increasingly relevant for companies, governments, and researchers alike. Understanding how people experience products, services, or topics that can provide insights for decision-making, logo control, and escalate engagement. The fundamental ideas of the algorithm are sentiment evaluation and classifying the polarity of a given text in a sentence whether or not the expressed opinion in a sentence is effective, negative, or neutral.

I. DATASET & DATA PREPROCESSING

A. Brief about the Dataset

Three datasets were used:

1. Amazon reviews - This dataset contains a huge number of product reviews posted by Amazon consumers. The reviews span a wide range of topics and contain information about the reviewers, review ratings, timestamps, and the review texts themselves.

2. Twitter and Reddit sentiment analysis - This dataset is made up of Twitter tweets and their related sentiment classifications. The clean_comment feature provides preprocessed versions of tweets, whereas the category feature gives sentiment categories (positive, negative, or neutral) to each tweet. The dataset is appropriate for analyzing sentiment trends and constructing sentiment analysis algorithms for social media data.

3. Medicine Review Data Set with side effects - This dataset focuses on medicine reviews and side effects. It includes features such as interaction identifiers (interaction.id), article URLs (article_url), review content (content), timestamps (time), relevance indicators (relevant), sentiment labels (sentiment), gender information, and specific side effects experienced by users (dizziness, convulsions, heart palpitations, shortness of breath, headaches, etc.). This dataset is useful for analyzing the relationship between medications, user experiences, and side effects. It can support research in pharmacovigilance, drug safety analysis, and personalized medicine.

B. Preprocessing Techniques

Analyzing and understanding the features in the dataset and deciding on which features are relevant and properly pre-processing these features. The features in our dataset were the textual review of the medicine along with details on whether the user experienced any side effects and whether they are overall satisfied with the medicine or not.

Pre-processing of the feature with techniques such as:

• Lemmatization - means grouping the inflected forms of a word so they can be analyzed as a single item.

• Removing Stopwords - These words are so common they are ignored by typical tokenizers eg. a, the, then, etc. Removing Stop Words are those objects in a sentence that are not necessary for any sector of text mining.

• Tokenization - This means splitting the input data into a sequence of meaningful parts and removing relevant text with the help of regular expression. It is the process of separating a sequence of strings into individuals such as words, keywords, phrases, symbols, and other elements known as tokens. Tokens can be individual words, phrases, or even whole sentences

• Vectorization - The process of converting text input into a numerical form that machine learning algorithms can understand. It entails converting written documents or words into numerical vectors, which computing systems can then process and analyze.

Pre-process the data by cleaning, normalizing, and transforming it into a suitable format for analysis. This includes tasks such as selection appropriate features, text tokenization, stop word removal, vectorization, and feature engineering. Perform exploratory data analysis (EDA) to gain insights into the characteristics of your data. Analyzing the distribution of sentiment labels, exploring patterns and trends in the data, and identifying any potential challenges or biases that may impact the analysis. Next, to choose appropriate machine learning or NLP techniques for sentiment analysis, we choose classification models such as Logistic Regression, K Nearest Neighbour, Gaussian Naive Bayes, Support Vector Classifier, and Random Forest classifier. Evaluate the performance of the trained model(s) on the test set and compare the performance of these models and techniques to identify the most effective model for sentiment analysis. The accuracy of these data is compared using a confusion matrix as all the reviews were narrowed down to Positive, Negative, and Neutral. A confusion matrix for each model analyzed the accuracy of these models.

II. METHODOLOGY

Supervised learning is achieved by analyzing input-output pairs of examples (instances) of an unknown function that maps inputs to outputs. The learned model will approximate the function based on a few assumptions, which gives rise to multiple supervised learning algorithms each suited for a specific set of assumptions. To learn a model, supervised learning algorithms are fed a large number of training examples containing the input data and their corresponding labeled outputs (ground truth) to estimate the parameters of the learned model[11]. Our dataset rated the sentiment feature of the reviews on a scale of 1 to 5. Where reviews with sentiment value 1 and 2 were classified as Positive, 3 was considered Neutral and 4 and 5 was concluded to be a Negative review.

- *Logistic Regression*

Is a supervised learning algorithm that can be used in several problems including text classification. It is a regression model which generalizes the logistic regression to classification problems where the output can take more than two possible values[12].

- *K-nearest neighbor*

KNN is a distance-based classifier. Generally, KNN calculates the distance of one test data with all existing data trains using Euclidean distance[14]. The process after calculating the distance for each data train is voting. Voting aims to determine the class or label of a data test. Voting is done by taking as much as K-nearest distance and counting how many of each class is contained. If the results of voting are more positive classes, then the data tested is a positive class and vice versa. The disadvantage of KNN is one needs to determine the right K for all data tests and data trains that are not overfitted(only good for training data) [13].

- *Gaussian Naive Bayes*

The Naïve Bayes method is a classification method for text mining used in sentiment analysis. This approach is theoretically good in terms of data consistency and calculation classification[15]. It is called "naive'' because it assumes that the input is independent of each other. The Naive Bayes classifier gives us an excellent result when one uses it for text data analysis. Such as Natural Language Processing. Naive Bayes algorithm gives us a probability analyzing the data set we have given. Naïve Bayes classifier is used as a probabilistic classifier[16].

- *Support Vector Classifier*

SVC is a variant of SVM. Support Vector Machine is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems[11]. In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well[4].

- *Random forest*

Random forest, which was formally proposed in 2001 by Leo Breiman and Adèle Cutler, is part of the automatic learning techniques. This algorithm combines the concepts of random subspaces and "bagging". The decision tree forest algorithm trains on multiple decision trees driven on slightly different subsets of data[4] .

Applied Model	Accuracy
Gaussian NB	31.477
KneignbourNeighor	51.177
Logistic Regression	60.814
SVC	62.384
Random Forest	63.169

Table 1. Accuracy of applied algorithms.

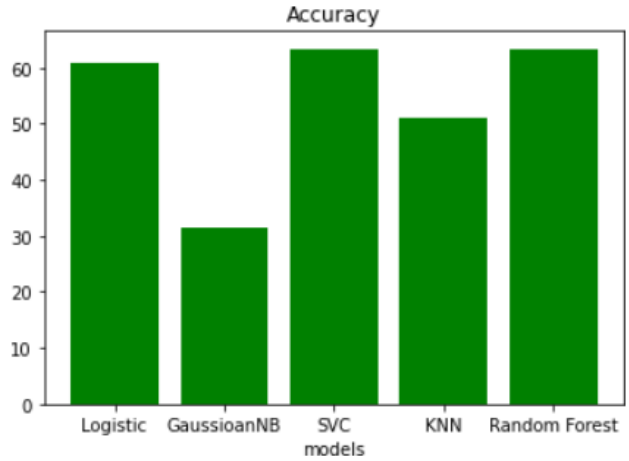

Fig.1. Comparison graph illustrating the algorithm's accuracy.

Random Forest achieved the highest accuracy. Therefore, for hypertuning we decided to choose Random Forest.

III. HYPERTUNING

When creating ML models, the selection of proper parameters is essential for the best results. Different hyper-parameter sets for each ML method were tried to make the models more accurate. Also, picking the right hyperparameters is one of the most important parts to improve the model's accuracy[17]. The hyperparameter combinations in the search space, ultimately select the best-performing hyperparameter combination. Grid search is a decision-theoretic approach that involves exhaustively searching for a fixed domain of hyperparameter values. Random search is another decision-theoretic method that randomly selects hyper-parameter combinations in the search space, given limited execution time and resources. In GS(Grid Search) and RS(Random Search), each hyperparameter configuration is treated independently[18]. Random Forest Classifier ought to achieve the highest level of accuracy to improve the accuracy the model was hyper-tuned on a combination of the parameters for this model. Hyperparameter tuning is an important step that involves optimizing the parameters of a model to achieve better performance. A total of 648 combinations were fitted and cross-validated over 3 folds.

IV. RESULTS

Confusion matrix is used to describe how well a classification system performs. The output of a classification algorithm is visualized and summarized in a confusion matrix. The calculation for the True Positive, True Negative, False Positive, and False Negative values in a 3x3 matrix is given in Table No. [3] Diagonal values in blue color from the top left to the bottom right show the ‘’true positives’’ of negative, neutral, and positive sentiment classes, respectively[7] .

The classification metrics considered for the sentiment analysis are Accuracy, Precision, Recall, and F-Measure and these parameters are evaluated based on the calculated positivity and negativity of reviews by the proposed hybrid approach[4].

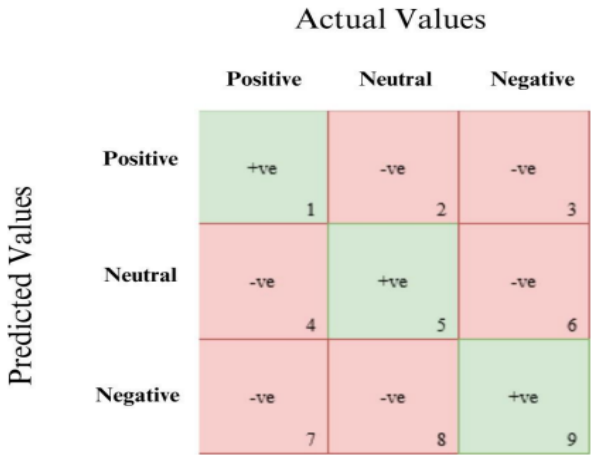

Fig 2. 3x3 Confusion Matrix

Positive	Neutral	Negative
TP = $Cell_1$	TP = $Cell_5$	TP = $Cell_9$
FP = $Cell_2 + Cell_3$	FP = $Cell_4 + Cell_6$	FP = $Cell_7 + Cell_8$
TN = $Cell_5 + Cell_6 + Cell_8 + Cell_9$	TN = $Cell_1 + Cell_3 + Cell_7 + Cell_9$	TN = $Cell_1 + Cell_2 + Cell_4 + Cell_5$
FN = $Cell_4 + Cell_7$	FN = $Cell_2 + Cell_8$	FN = $Cell_3 + Cell_6$

Table 3. Calculation for 3x3 Matrix

Accuracy is a common measure for the classification performance and it’s proportional of correctly classified instances to the total number of instances, whereas the error rate uses incorrectly classified rather than correctly

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Recall is the percentage of correct items that are selected. recall of 1 means that all the positive examples were found

$$Recall = \frac{TP}{TP+FN}$$

Precision is the report between the number of the true positive and the sum of the true positives and the false positive. A value of 1 expresses the fact that all the positive classified examples were real

$$Precision = \frac{TP}{TP+FP}$$

Computational analysis by these parameters from the Confusion Matrix for Random Forest given in Fig. 3 is shown in Table 4. These values proved a strong base for Random Forest providing the highest accuracy and furthermore improved via hyper tuning.

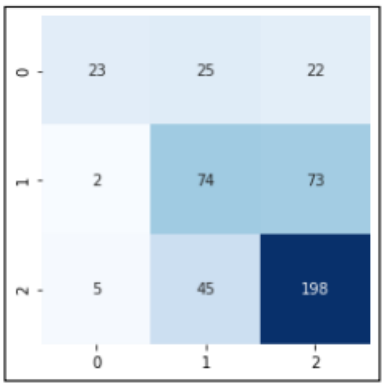

Fig. 3. Confusion matrix for Random Forest

	Precision	Recall	Accuracy
Positive	0.32	0.76	0.88
Neutral	0.49	0.51	0.68
Negative	0.79	0.67	0.67

Table 4. Precision, Recall and Accuracy for RF.

The hyper tuning is performed by GrideSearchView the 'cv' parameter which is called 'cross validation'. Different models are developed utilizing distinct training and non-overlapping test sets in cross-validation. The performance on test sets is then combined to provide better outcomes. These are results that display which combination of tested values for parameters like max_depth, max_features, and n_estimator [19] for Random Forest gives the best score alone.

max_depth	max_features	n_estimator	Accuracy
500	Sqrt	400	0.604
70	Log2	600	0.500
10	Log2	100	0.499
70	Log2	100	0.506
None	Auto	400	0.690
500	auto	800	0.603

Table 2. Hypertunning Different Parameters Accuracy

Upon hyper tuning Grid Search View the best parameters - criterion=entropy, max_depth=None, max_features=auto, n_estimators=400; resulted in an escalated accuracy score giving the best score of- 69.32%. with a computation time of 1.9mins.

V. FUTURE WORK

After studying various approaches in Sentimental Analysis and different levels of emotions one can explore several courses of action in this respected field. The shortcoming of sentimental analysis is the way of speaking(tone) this usually differs from person to person verbally and interpreting it will be very difficult, another aspect is about sarcasm usually irony-sarcastic types of people use sarcasm in even casual conversations and this can be misleading to find out the true context behind the person that if the sentence is positive or negative by the training models, the other issue is about emoticon(emojis) this is a problem with the social media contents Instagram, Facebook, etc. that are text-based which usually includes emojis. As emojis are mostly considered in the special characters category by the analyzer and are removed while refining a sentence. Therefore all the points that are discussed above are major issues that result in depletion in the accuracy of the training model.

This Limitation can be overcome by different techniques which reduces the issue and helps in enhancing the accuracy of the model. One of the techniques that is implemented above is Hyper tuning. This algorithm selects the most optimal value for the hyperparameter in the dataset of the model, based on rate, strength, max-depth, min-depth, and number of layers all these factors are taken into consideration. According to the dataset used above, Random Search with Hyper tuning is the most effective technique to get the best and most accurate results. One can also explore other techniques that include Grid based search, Bayesian Optimization, and automated libraries as well.

Thus, one can develop a Hybrid Model by taking more than one classifier to get the best results.

REFERENCES

[1] Discrete Opinion Tree Induction for Aspect-based Sentiment Analysis Chenhua Chen1,2 , Zhiyang Teng1,2 , Zhongqing Wang3 and Yue Zhang1,2 1School of Engineering, Westlake University, China Institute of Advanced Technology, Westlake Institute (2021)

[2] K. L. Tan, C. P. Lee, K. M. Lim and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," in IEEE Access, vol. 10, pp. 103694-103704, 2022, doi: 10.1109/ACCESS.2022.3210182.(2022)

[3] Bhatt, A., Patel, A., Chheda, H., & Gawande, K. (2015). Amazon review classification and sentiment analysis. International Journal of Computer Science and Information Technologies, 6(6), 5107-5110.

[4] Al Amrani, Yassine, Mohamed Lazaar, and Kamal Eddine El Kadiri. "Random forest and support vector machine based hybrid approach to sentiment analysis." Procedia Computer Science 127 (2019)

[5] M. T. H. K. Tusar and M. T. Islam, "A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), Khulna, Bangladesh, 2021

[6] Hasib, Khan Md. "Sentiment analysis on Bangladesh airlines review data using machine learning." PhD diss., Brac University, 2022.

[7] Almuayqil, S.N.; Humayun, M.; Jhanjhi, N.Z.; Almufareh, M.F.; Javed, D. Framework for Improved Sentiment Analysis via Random Minority Oversampling for User Tweet Review Classification. Electronics 2022, 11, 3058.

[8] Oghu, Emughedi, Emeka Ogbuju, Taiwo Abiodun, and Francisca Oladipo. "A Review of Sentiment Analysis

Approaches for Quality Assurance in Teaching and Learning." (2023).

[9] Tan, Leonard, Ooi Kiang Tan, Chun Chau Sze, and Wilson Wen Bin Goh. "Emotional Variance Analysis: A new sentiment analysis feature set for Artificial Intelligence and Machine Learning applications." Plos one 18, no. 1 (2023): e0274299.

[10] Rosenberg, Emelie, Carlota Tarazona, Fermín Mallor, Hamidreza Eivazi, David Pastor-Escuredo, Francesco Fuso-Nerini, and Ricardo Vinuesa. "Sentiment analysis on Twitter data towards climate action." (2023).

[11] Saif, Waddah S., et al. "Machine learning techniques for optical performance monitoring and modulation format identification: A survey." *IEEE Communications Surveys & Tutorials* 22.4 (2020): 2839-2882.

[12] W. P. Ramadhan, S. T. M. T. Astri Novianty and S. T. M. T. Casi Setianingsih, "Sentiment analysis using multinomial logistic regression," *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*, Yogyakarta, Indonesia, 2017, pp. 46-49, doi: 10.1109/ICCEREC.2017.8226700.

[13] P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose, CA, USA, 2001, pp. 647-648, doi: 10.1109/ICDM.2001.989592.

[14] Daeli, N. O. F., & Adiwijaya, A. (2020). Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. *Journal of Data Science and Its Applications*, *3*(1), 1-7.

[15] Pristiyono, et al. "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm." *IOP Conference Series: Materials Science and Engineering*. Vol. 1088. No. 1. IOP Publishing, 2021.

[16] Rahat, Abdul Mohaimin, Abdul Kahir, and Abu Kaisar Mohammad Masum. "Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset." *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*. IEEE, 2019.

[17] Dawei Yang, Ping Xu, Athar Zaman, Thamer Alomayri, Moustafa Houda, Abdulaziz Alaskar, Muhammad Faisal Javed,Compressive strength prediction of concrete blended with carbon nanotubes using gene expression programming and random forest: hyper-tuning and optimization,Journal of Materials Research and Technology,Volume 24,2023.

[18] Yang, Li, and Abdallah Shami. "On hyperparameter optimization of machine learning algorithms: Theory and practice." *Neurocomputing* 415 (2020): 295-316.

[19] https://arxiv.org/abs/1309.0238