Analyzing Public Perceptions and User Sentiments on Tweets: A Machine Learning Approach

Khushi Sarkar[1]
School of Computer Science and Engineering
Lovely Professional University
Punjab,INDIA
sarkar.khushi122002@gmail.com[1]

Parshotam Pal[2]
School of Computer Science and Engineering
Lovely Professional University
Punjab,INDIA
parshotampal90@gmail.com[2]

Shaurya Verma[3],
School of Computer Science and Engineering
Lovely Professional University
Punjab,INDIA
shauryaverma713@gmail.com[3]

Ayush Singh[4]
School of Computer Science and Engineering
Lovely Professional University
Punjab,INDIA
ayushraja.arsb@gmail.com[4]

Samyak Dahat[5]
School of Computer Science and Engineering
Lovely Professional University
Punjab,INDIA
dahatsamyak07@gmail.com[5]

Sarvesh Kumar[6]
School of Computer Science and Engineering
Lovely Professional University
Punjab, INDIA
sarveshk7499@gmail.com[6]

Shakshi Kumari[7]
School of Computer Science and Engineering
Lovely Professional University
Punjab,INDIA
pandeyshakshi000564@gmail.com [7]

***Abstract*—**

The blast of social media has driven to a tremendous sum of user-generated information, especially on stages like Twitter where individuals express their sees and feelings openly. Opinion mining develops as a capable instrument to extricate profitable bits of knowledge from this unstructured information. Twitter serves as a rich resource for comprehending user sentiments and public viewpoints across diverse domains such as politics, current events, consumer behaviors, and brand perception. With the exponential growth in tweet volume, manual analysis of this vast dataset proves impractical. This paper digs into opinion investigation strategies particularly custom fitted for Twitter. Therefore, we are going to analyze different opinions or emotions of users and for this we will use a sentiment analysis approach. Sentiment analysis is an approach to analyzing data and capturing the emotions it embodies. Twitter Sentiment Analysis is the application of sentiment analysis to data from Twitter (Tweets) to extract the sentiments sent by users. The paper moreover examines common assessment measurements utilized to survey the adequacy of these procedures. By giving a comparative investigation, this overview points to prepare analysts and specialists with a comprehensive understanding of estimation investigation on Tweets. This information can be saddled for different applications, such as gaging open conclusion on current occasions, observing brand notoriety, or illuminating promoting techniques.

Keywords
Dataset**,** Sentiment analysis**,** Random Forest, Challenges and Future, Stopford's, NLP etc.

I. INTRODUCTION

The online buzz around reviews creates a win-win situation. Consumers get valuable information before they buy, and businesses gain the feedback they need to improve their offerings. It's a constantly evolving conversation that benefits everyone involved. Twitter has become one of the most significant weblog sites with over one billion people posting over millions of tweets daily.

Twitter's enormous exposure has attracted people to exhibit their views on any concerning issue, brands, any misconducts, or other noteworthy field of study. Thus, Twitter's data is commonly used for analyzing sentiment of different people's thought process.

Twitter allows users to contribute their thoughts and perspectives regardless of the size limitation, the text can be an emoji to a long paragraph or story, slang, abbreviations etc. Additionally, people channel their sentiments by using texts full of sarcasm, vagueness, and bluntness.

Therefore, it is appropriate to say that Twitter language is unstructured.

With the aim of extricating different sentiments from tweets, we use sentiment analysis [1,2,3,6]. The result from this analysis can be efficacious.

in many fields such as investigating and observing public perception or views on various brands or products. In recent decades, research in this field has grown steadily. The reason is that there is a large amount of data tweet data which is required to analyze the sentiment and difficult to process. However, the advent of machine learning methodologies,

particularly natural language processing (NLP), enables automated processing and sentiment analysis of tweets.
Over the past decades, multiple research studies have been consistently increased. The actual cause is the small format of tweets which is difficult to process.
Thus, we aim to re-evaluate some research in this field and perform sentiment analysis [1,2,3,6] on Twitter data [6] using Python to categorize the tweets and estimate favorable outcomes. The main objective of this paper is executing multiple machine learning algorithms, methodologies and making comparisons according to the accuracy of these models.

II. LITERATURE REVIEW

Analyzing different opinions, emotions which comes under the field of Natural language process (NLP) [8] that is centered on retrieving tweets and then examining the points of view, their thinking and attitude which are textually exhibited. Twitter data [6] are commonly used because it has potential in different fields such as finance, branding etc. This literature review focuses on auditing existing research papers, their methods, approaches, exceptions, and enhancements over the years in this domain.

III. SENTIMENT ANALYSIS

Public perception or opinion mining [5] can be esteemed as a process that extracts emotions, attitude, opinions from a particular set of instructions or we can say that a text, emoji, sentence, or a paragraph.
In this process, we will decide which are the traits which will fall in the positive, negative, or neutral category. For that we must make a class of entities which are to be performed. Here, the classification will occur in three class tweets that are (positive, negative, and neutral).
The paper explains how machine learning approaches effectively overcome the constraints presented by Twitter data [6], including its informality, contextuality, and shortness. It also contains a thorough examination of these challenges. The article explores sentiment analysis [1,2,3,6] and describes important steps in the process, such as feature extraction, data preparation, model selection, training, and evaluation techniques. Many machine learning algorithms such as Recurrent, Support Vector Machines, and Naive Bayes. The applicability, advantages, and disadvantages of transformer-based models, convolutional neural networks, and neural networks are discussed along with their strengths and weaknesses in sentiment analysis [1,2,3,6].
The procedure includes dataset collection from Kaggle or any other website, preprocessing, extraction, training, sentiment exploration, testing and machine learning methods.

A. Twitter sentiment analysis

The main goal of this analysis on tweets is to properly classify the sentiments precisely. Numerous methodologies have been used in the past,
Thus, we are here to propose the best fit trained model which will give fruitful result, and which will have best accuracy or efficiency. [2]

Different difficulties in twitter dataset [2]
1)Compact tweets
2)Slangs, abbreviations, emojis, repeated words
3)Text variation. URL, HTML, Special Character (@, #, $)

IV. METHODOLOGY

The first step for any analysis, the basic step, is to gather different twitter dataset in which we will perform various approaches. We have taken the twitter data [6] from Kaggle. The dataset which we download from the websites are raw data which means it contains stop words [8], special characters, repeating words, spaces etc., which are not required in the process.
Therefore, pre-processing takes place, it is a process in which all the unwanted records that are not essential and cannot make any changes to the data for the further process get removed. It makes the data more machine friendly than its raw form.

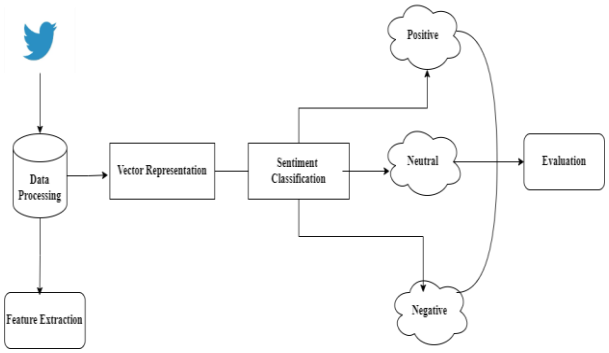

Fig. 1. General approach for sentiment analysis

A. Data mining

Data mining refers to the collection or gathering of data of choice. We can get the twitter data [6] from Twitter API, or from any other online websites (for example KAGGLE). After collection of data, the next step is very critical role as the accuracy of the model will depend on the correct division of dataset into training and testing. The training will be the fundamental key for the efficiency of models.

B. Twitter dataset pre-processing

Pre-processing is the crucial step in determining the efficiency and accuracy of any algorithm. The raw data needs some processing where cleaning of data is required for the best outcomes. This step helps the data to be more machine friendly to reduce the chances of polysemy. Additionally, feature engineering may be used to extract relevant details, for example word frequency counts or TF-IDF [6] scores. pre-processing steps includes following steps:[2] [3]

- Removing of same tweets.
- Upper case to lower case: As python is a case sensitive language, so it might take same words again which will affect the analysis.

- URL removal: Usernames and URLs are not crucial, and they are not required in the future processing. So, URLs and usernames are removed.
- Special character (hashtag handling) and number removal: these are also not needed in the processing as they do not have any usage in sentiments. Therefore, removing them can help connect two words which were not considered distinct.
- Deletion of stop words [8]: it filters out common words like (or, and, is) because these words are not effective.
- Stemming: it is the most important step in pre-processing as it converts the word into root word which makes it easy to identify and reduces complexity in understanding.
- deletion of punctuation mark
- expanding the abbreviations
- checking the spelling
- Word Clouds: Highlighting frequent words within each sentiment category.

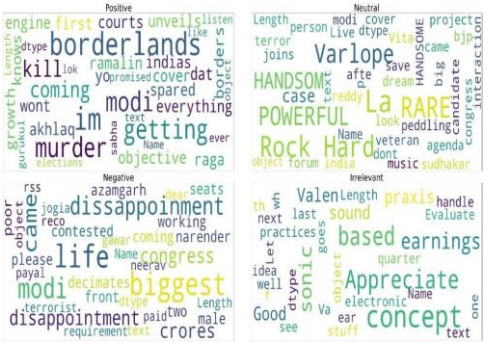

Fig. 2. General approach for sentiment analysis

V. FEATURE SELECTION

The next step after preprocessing is the addition of features that are going to be used in training the random forest [2, 7] model. These features are crucial as they play a vital role in capturing the important information. These features consist of word embedding, numerical representations and unordered list of texts**.** [3]

VI. RANDOM FOREST APPROACH

Random forest [2, 7] is used for classification of classes including public perception like detection of sentiment by the help of tweets. It is a member of the collaborative family, which is known for the integrating various models to make prediction more precisely and accurately than the models which work alone. So, it is a simple yet resilient model to perform analysis.

A. Building the forest

For this approach, we must prepare collection of decision trees. These decision trees are made by using the subset of the training data and randomly distributed subset of features [10]. The reason being the features are distributed randomly is that it helps in creating diverse number of trees that will give more accurate predictions together. During the composition of these trees, it recursively breaks the data based on features which results in maximizing the purity of subsets.

B. Voting Mechanism

After all the decision trees are processed, each tree is going to predict individually. Here, each tree is going to predict the feeling (positive, negative, or neutral) of a tweet as per its features. The final prediction is actuated through voting technique which is going to choose the majority sentiment among all trees.

C. Hyperparameter Tuning

It is a mechanism like grid search and used to find the ideal combination of hyperparameter [9]. This model has a few hyperparameters that can be fine- tuned to make its performance better. This process includes the number of trees, maximum depth, and number of features.

VII. ANALYSIS CONTENT

- Aims to divide the tweets as positive, negative, and neutral.

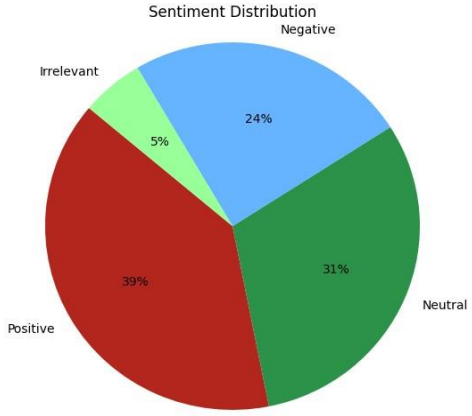

Fig. 3. Pie chart showing different types of tweets.

- Preprocessing begins by removing noise, irrelevant information. (also, URLs, html tags, special characters).
- After preprocessing, the data is split into training and testing using the train_test_split function from the sklearn.model_selection module in a ratio of 80% and 20%.

- A TfidfVectorizer is used to convert the text data into numerical features which represent the importance of each word in the dataset.
- Then, the model is instantiated and trained on the pre-processed data with the use of a pipeline which contains the TF-IDF [6] vectorizer and the classifier.
- The model's performance is evaluated using the accuracy score on the testing set, and the trained model is saved using pickle for future use.

VIII. FLASK WEB APPLICATION COMPONENT

- The Flask web application provides a user-friendly interface for interacting with the sentiment analysis [1,2,3,6] model.
- It begins by defining routes for different functionalities, such as home page rendering, file upload, sentiment prediction, and progress update.
- Users can upload CSV files containing tweet data for model training using the /generate_pkl route. The uploaded file is processed in a separate thread to avoid blocking the main application.
- Progress updates during the model training process are provided in real-time through the /update progress route, allowing users to monitor the status of the analysis.
- Users can input individual tweets through a form on the web interface for sentiment prediction using the /predict_tweets route.
- The sentiment analysis [1,2,3,6] model is loaded based on the selected model name, and the input tweet is pre-processed before being passed to the model for prediction.
- The predicted sentiment (positive/negative) and the prediction time are returned to the user via JSON response, allowing for real-time feedback.

Overall, the code seamlessly integrates sentiment analysis [1, 2, 3, 6] functionality with a user-friendly web interface, enabling users to analyze sentiments in tweets conveniently. The sentiment analysis [1,2,3,6] component preprocesses tweet data, trains a Random Forest Classifier model, and evaluates its performance. The Flask web application component facilitates user interaction by providing features for uploading files, inputting individual tweets, and monitoring the progress of sentiment analysis [1, 2, 3, 6] tasks in real-time.

As, we have observed that in previous research papers the accuracy in the random forest model using TF-IDF [6] is less, that is 87.5% [2] and in other it is 66% [1] which is quite low, but we have achieved more accuracy in the same model that is approximately 91.7% which is quite impressive. The model performed very effectively and hence it can be used to get a higher accuracy score.

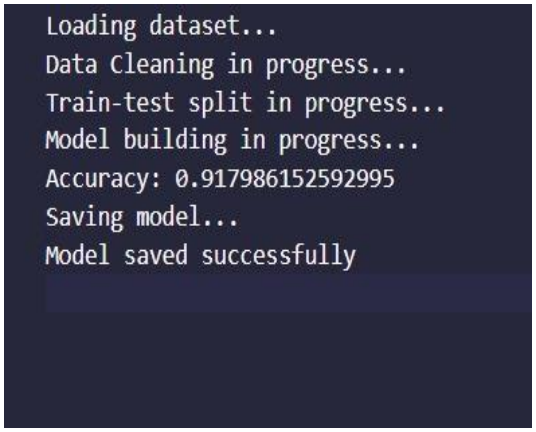

Fig. 4. Accuracy score

Research Source	Random Forest Model (TF-IDF) Accuracy	Observations
Previous Paper 1	**87.5%**	Reasonable accuracy
Previous Paper 2	**66%**	Low accuracy
Our Research	**91.7%**	Impressive improvement

Fig. 5. comparison between previous paper's accuracy and this research paper.

X. CONCLUSION

In conclusion, this project successfully demonstrates the implementation of Analyzing public perception on Twitter data [6] using Python. By ascendancy of machine learning techniques and a Random Forest Classifier model, tweets are explicitly sorted as positive, negative, and neutral feelings. With a splendid accuracy of 91.7%, the model exhibits its effectiveness in discerning public opinion on various topics discussed on Twitter.

The integration of an adaptable web interface using Flask enhances attainability, allowing users to conveniently input tweets for sentiment analysis [1, 2, 3, 6] and receive

predictions in real-time. Progress updates during model training further contribute to transparency and user engagement.

Overall, as the result shows that the project proves to have high accuracy and adaptable environment and can be used in future predictions and research in this sentiment field, which will be very beneficial for offering insights into the current fashion or trend and into the points of view of society.

REFERENCES

[1] Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning- based approach by Yuxing Qi1 and Zahratu Shabrina.

[2] Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python by Bhumika Gupta, PhD Assistant Professor, C.S.E.D G.B.P.E.C, Pauri, Uttarakhand,
India Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani B.Tech,
C.S.E.D G.B.P.E.C Uttarakhand, India [2]

[3] Sentiment Analysis of Twitter Data: A Survey of Techniques Vishal A. Kharde Department of Computer Engg, Pune Institute of Computer Technology, Pune University of Pune (India) S.S. Sonawane Department of Computer Engg, Pune Institute of Computer Technology, Pune University of Pune (India).

[4] Sebastiani, F. Machine learning in automated text categorization. ACM Comput. Surv. 2002, 34, 1–47.

[5] Anjaria, M.; Guddeti, R.M.R. Influence factor-based opinion mining of twitter data using supervised learning. In Proceedings of the 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, 6–10 January
2014.

[6] Ramos, J. Using TF-IDF to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, Banff, AB, Canada, 27 February–1 March 2003; pp. 133–142.

[7] Segnini, A.; Motchoffo, J.J.T. Random Forests and Text Mining. Available online:
http://www.academia.
Edu/11059601/Random_Forest_and_Text_Mining

[8] Mishra, A., Ranjan, A., Kumar, A., Sharma, V., & Biswas, P. (2023). Sentiment Analysis Using NLP. International Journal of Research, 10(11), 105–112.
https://doi.org/10.5281/zenodo.10211177

[9] Elgeldawi, E.; Sayed, A.; Galal, A.R.; Zaki, A.M. Hyperparameter Tuning for
Machine Learning Algorithms Used for Arabic Sentiment Analysis. *Informatics* **2021**, *8*, 79. https://doi.org/10.3390/informatics8040079

[10] A. S. Zharmagambetov and A. A. Pak, "Sentiment analysis of a document using deep learning approach and decision trees," *2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)*, Almaty, Kazakhstan, 2015, pp. 1-4, doi: 10.1109/ICECCO.2015.7416902.