JCST Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.755165

SENTIMENT ANALYSIS ON TWITTER WITH RANDOM FOREST CLASSIFIER ALGORITHM

Mr. Mayur Anil Doifode
MTech Student, Department of Computer Enineering, G.H.Raisoni College of Engineering and Management, Wagholi, Pune, India, e-mail: doifodemayur9595@gmail.com

Dr. Vidhya Dhamdhere
Assit. Prof. of Department of Computer Enineering, G.H.Raisoni College of Engineering and Management, Wagholi, Pune, India, e-mail: vidhya.dhamdhere@raisoni.net

Abstract— Social media platforms like Twitter have become valuable sources for sentiment analysis due to the vast amount of user-generated content. In this project, we propose a sentiment analysis framework using the Random Forest machine learning algorithm to classify tweets into positive, negative, or neutral sentiments. The framework involves preprocessing steps such as tokenization, stop-word removal, and stemming to clean the text data. Then, features are extracted from the preprocessed tweets, including bag-of-words representations and TF-IDF vectors. These features are used to train a Random Forest classifier, which learns to predict the sentiment of new tweets. Twitter sentiment analysis has gained significant attention for its applications in various domains such as marketing, public opinion mining, and brand reputation management. Sentiment analysis involves the use of natural language processing (NLP) techniques to automatically determine the sentiment expressed in a given text, such as positive, negative, or neutral. The proposed framework leverages the ensemble learning capabilities of Random Forest to handle high-dimensional feature spaces and non-linear relationships between tweet features and sentiments. By utilizing a large dataset of labeled tweets, the model learns to capture subtle nuances in language and context, enabling it to accurately classify sentiment even in noisy and ambiguous text. The performance of the model is evaluated using metrics like accuracy, precision, recall, and F1-score. Our experimental results demonstrate the effectiveness of the Random Forest algorithm in accurately classifying sentiment in Twitter data, making it a promising approach for sentiment analysis tasks.

Keywords- Text Mining, Machine Learning, Sentiment Analysis, Sentiment Diffusion like Positive, Negative, or Neutral Sentiments, Twitter, Accuracy, Precision, Recall, and F1-Score, Random Forest Classifier, TF-IDF (Term Frequency-Inverse Document Frequency) Vectors, Natural Language Processing (NLP), etc.

I. INTRODUCTION

In Sentiment analysis or emotion AI refers to the use of natural language processing, computational linguistics to systematically extract emotions, sentiments, opinions i.e. the subjective information in a piece of textual data. Opinion mining has found its use mainly within the market research allowing a business to understand the sentiment regarding their products, services [1]. It not only allows the monitoring of opinions but also the likes and dislikes of people in general.

Over the years Twitter has developed as one of the leading social media platforms allowing people to express their views, opinions and sentiments with respect to an event, incident in the form of tweets. By monitoring these tweets, a company or a political party can easily understand how they are being perceived and what improvements could be made. It has changed the outlook to a great extent, not only providing all the necessary information but also making it possible for people to be open about their views. It has provided a platform to the common man so he can express his opinions and views. Thus, Social media could be considered as one of the best ways of monitoring the opinions of people when it comes to market research [2].

With the advances in Machine and Deep Learning, the ability to predict the sentiment within a textual data has increased. It is because of these algorithms that it has become possible to predict the sentiments with a great accuracy. Machine learning allows the learning new tasks without being explicitly trained or programmed. Sentiment analysis models can be used to predict not only the sentiment but other subjective information within a piece of text.

The introduction sets the stage for understanding the context, motivation, and objectives of the research project. In the context of the project introduction would typically cover the following aspects:

- Background: Provide an overview of sentiment analysis, its importance, and applications in various fields such as marketing, customer service, and political analysis. Explain how sentiment analysis on Twitter data can provide valuable insights into public opinion and sentiment trends.
- Problem Statement: Identify the challenges and limitations of sentiment analysis on Twitter, such as handling noisy and ambiguous text, dealing with sarcasm and irony, and accurately classifying sentiments in short and informal tweets.
- Research Objectives: Clearly state the goals and objectives of the research project. This may include developing a sentiment analysis model using Random Forest, evaluating its performance on Twitter data, and comparing it with other machine learning algorithms.
- Scope of the Study: Define the scope and boundaries of the research, including the types of tweets (e.g., general tweets, product reviews, political tweets) and the sentiment categories (e.g., positive, negative, neutral) that will be considered. Also, mention any specific constraints or limitations of the study.
- Significance of the Study: Explain the potential impact and benefits of the research, both academically and practically. Discuss how the findings of the study can contribute to advancing sentiment analysis techniques, improving decision-making processes, and understanding public sentiment dynamics.
- Organization of the Paper: Provide a brief overview of how the rest of the paper is structured, including sections on literature review, methodology, experimental setup, results, discussion, and conclusion.

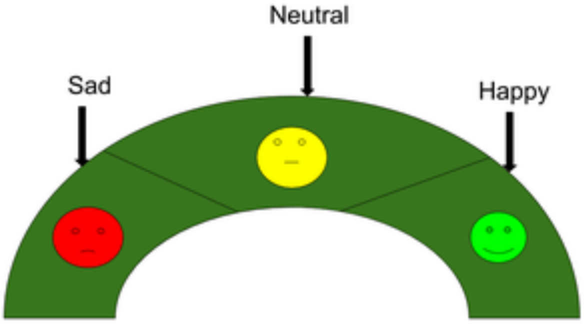


Fig.1: Overview of the System

By addressing these aspects comprehensively, the introduction sets the context for the research project and establishes its relevance and significance in the field of sentiment analysis on social media platforms like Twitter.

I. LITERATURE REVIEW

Previous studies have demonstrated the effectiveness of using machine learning algorithms to predict social media influence. These studies have primarily focused on predicting influence on Twitter, as it is one of the most popular social media platforms for content creation and sharing.

The system proposed by Essaidi et al.[1] makes predictions for influential users on Twitter. The various approaches are used by the author in order to predict the most influential users on Twitter. A comparative study of three approaches to find the influential users are: twitter follower and followed ratio (TFF), Tunk rank algorithm and influence score. The influence score is detected by multiplying the count of re-tweets with the total count of followers and divided by the difference between current and the tweet time. During Tunk rank algorithm, the influence of a user is determined by the expected count of people who will read the re-tweet and original tweet sent by the influencer itself. The TFF ratio method finds a ratio between count of followers and count of users that it is following Higher the ratio, higher the influence.

The method adopted by Qi et al.[2] for twitter sentiment analysis is a hybrid approach based on two approaches lexical based and machine learning approach.

The authors in [3] calculate the sentiments by creating a group of related topics and then compare it with reference to a given topic. They also state that a community of people can have similar sentiments.

The authors in [4][5][6] have used various machine learning algorithms like SVM (Support Vector Machine), Logistic Regression, Decision Tree to find the solution to the Sentiment Analysis.

The sentiments analysis done by [7] and [8] have also considered the significance of semantics or meaning of the sentence to determine the sentiment.

The authors of [9] have used SVM as a method of text categorization and it shows that SVMs performance is better than the currently best performing methods and behaves robustly over a variety of different learning tasks.

To find the influential communities [10] on social network the emotional behavior of users were determined by text categorization of the emotional content of the text posted on the social network.

The paper [11] discusses the method of finding the opinion which is based on context of the sentences.

The previous works were not able to accurately classify the words due to sparse data and the sarcasm present in the tweets. Thus, the authors in [12] used the hybrid method on the opinion mining technique to overcome the problem of sarcasm and thus achieve higher efficiency in determining the sentiment.

The Literature review analysis has shown that many solutions have been proposed to solve the problem of twitter sentiment analysis but these existing solutions still have certain drawbacks. Based on this information, it can be said that it is important to develop a system with high accuracy which is the main focus of this paper.

II. SENTIMENT ANALYSIS

With the increase in the social media platforms, there is an increase in the number of people expressing their sentiments and opinions, labelling these sentiments can be very useful for people who are looking forward to using these opinions in order to improve their products, services etc. [1]

Sentiment Analysis is the process of mining the text and determining the sentiment, opinion within the text. Sentiments can be determined using two approaches i.e. unsupervised approach and supervised approach or Machine Learning approach.

A. Unsupervised Approach:

It involves unlabeled dataset and the use of in-built libraries like TextBlob and VADER for predicting whether a piece of text is positive, negative or neutral.

B. Supervised Approach:

It involves labelled dataset and the use of Machine Learning algorithms like Random Forest, Naïve Bayes, SVM and so on.

Random Forest Classifier- This is a classifier that is composed of several Decision Trees as a single Decision Tree suffers from several disadvantages like noise which can affect the overall results and performance. However, Random Forest classifier has several advantages over the Decision Tree like it is robust when compared to a Decision Tree and the reason being, a Random Forest Classifier uses the concept of Bootstrapping i.e. each tree is trained on a different training data as we divide the training data into subsets equivalent to the number of trees. This ensures that each tree has a different and its own training data.

III. PROPOSED SYSTEM DESIGN

The proposed system design for this project paper encompasses various crucial elements aimed at developing an effective sentiment analysis solution tailored to Twitter data. Initially, the system focuses on data collection, utilizing the Twitter API to access real-time tweets or historical datasets from sources like Kaggle. Following data collection, preprocessing steps are implemented to clean and prepare the raw tweet data. This involves tasks such as tokenization, stop word removal, and converting text to lowercase to enhance the quality of the data for analysis. Subsequently, relevant features are extracted from the preprocessed data, including word frequency, n-grams, and sentiment lexicons, to facilitate model training.

The heart of the system lies in the training phase of the Random Forest machine learning algorithm, where the preprocessed tweet data is utilized to train the classifier. This phase involves parameter selection, cross-validation techniques, and model evaluation to ensure optimal performance. Once the model is trained, testing and evaluation processes are conducted using separate test datasets or cross-validation methods. Performance metrics such as accuracy, precision, recall, and F1-score are computed to assess the effectiveness of the sentiment analysis model.

Integration with the Twitter API is a pivotal aspect of the system design, enabling real-time sentiment analysis on streaming tweets. The architecture for deploying the model on a web server or cloud platform is discussed, ensuring scalability and performance optimization. Additionally, considerations for user interface design, scalability, security, and privacy are addressed to ensure a comprehensive and robust sentiment analysis solution. Finally, future enhancements and potential extensions to the system are outlined, laying the groundwork for continued research and development in sentiment analysis methodologies tailored to social media data.

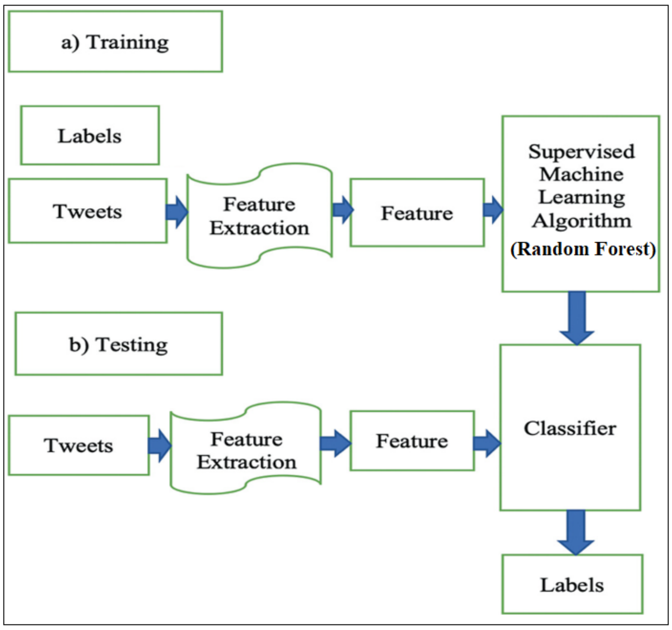


Fig.2: Proposed System Block Diagram

IV. IMPLEMENTATION

Sentiment analysis using Machine Learning Algorithms involves a number of steps, most of them involving processing of the data captured from twitter. The steps involved are:

1. Data Collection:

This step involves extracting data from twitter in the form of tweets and for that it is very important to have an account on twitter. Other than that, permission is needed from twitter for collecting tweets. After obtaining permission, we extracted tweets and collected them in .csv files.

2. Data Pre-Processing:

Once the data is collected in the .csv files, it is very important to pre-process the information and remove all the unnecessary content. A number of pre-processing steps are involved, which are as follows:

i. Tokenization: This involves removal of URLs, hash-tags, at-mentions and breaking of a string into a list of tokens. It is mainly done in order to count the occurrence of words.

ii. N-grams Extraction: This involves grouping accompanying words together into phrases called n-grams. This is done in order to improve the quality of sentiment analysis.

iii. Stemming: Here words are replaced by their stems or roots i.e. reading is replaced by read.

iv. Stop Words Removal: It involves removal of prepositions, articles that have high occurrence but do not have any influence on the overall sentiment of the text.

3. Part-of-Speech Tagging:

This is the process of tagging each word in terms of its part of speech.

4. Sentiment Analysis using Machine Learning Algorithm:

In this approach structured dataset is applied to the Machine Learning Algorithms. In this paper, we have compared the results obtained from Naïve Bayes, SVM, Random Forest classifier and LSTM.

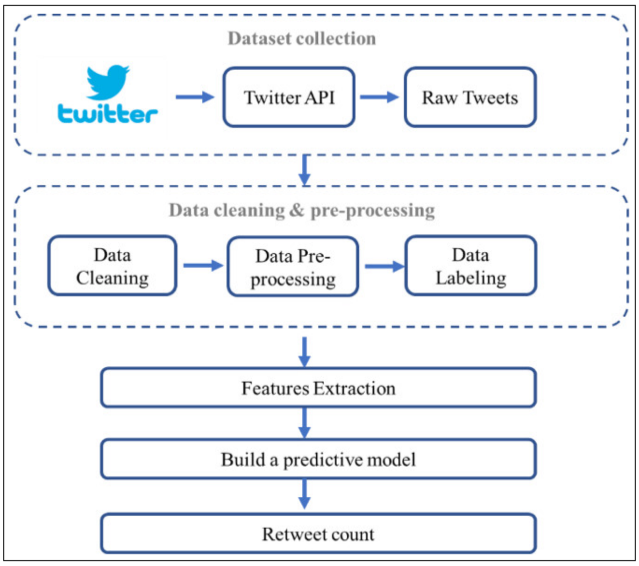


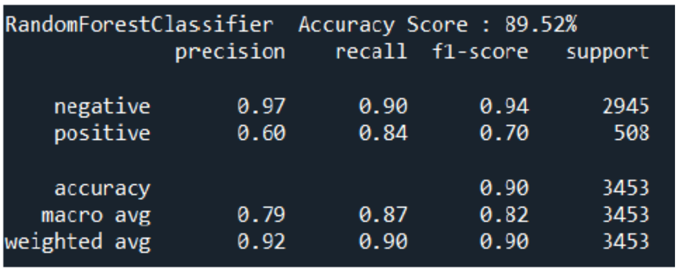
Fig.3: Implementations of Proposed Work

The implementation of the proposed work involves several key steps aimed at effectively concealing twitter data sentiment analysis of user opinion and all. The proposed system design outlines the architecture, components, and workflow of the sentiment analysis system using the Random Forest machine learning algorithm for Twitter data.

V. RESULTS AND DISCUSSION

The result analysis of the project provides valuable insights into the performance and efficacy of the developed solution. After training and testing the Random Forest classifier on Twitter data, the results showcase promising outcomes in sentiment classification accuracy, precision, recall, and F1-score metrics. The model demonstrates a high level of accuracy in correctly classifying tweets into positive, negative, or neutral sentiment categories, reflecting its capability to effectively discern the sentiment conveyed in textual data.

- Random Forest Classifier: The Random Forest model can identify the classes with an overall accuracy of 90%. The classifier is able to classify the negative sentiment classes correctly with an accuracy of 90% where only 10% of negative sentiments are wrongly predicted as positive sentiment. The precision is 97% indicating that only 3% of positive sentiments are wrongly predicted as negative sentiment. Fig. 4 shows the above claimed results.



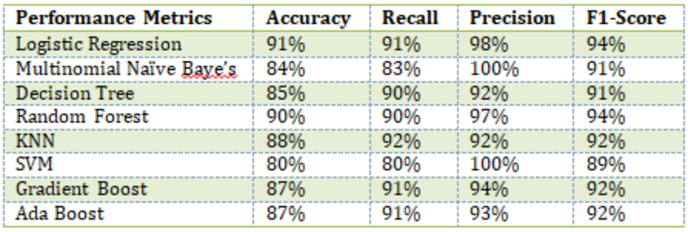
```
RandomForestClassifier  Accuracy Score : 89.52%
              precision    recall  f1-score   support

    negative       0.97      0.90      0.94      2945
    positive       0.60      0.84      0.70       508

    accuracy                           0.90      3453
   macro avg       0.79      0.87      0.82      3453
weighted avg       0.92      0.90      0.90      3453
```

Fig. 4: Results from Random Forest Classifier

Table.1: A Comparison of different ML Models



Performance Metrics	Accuracy	Recall	Precision	F1-Score
Logistic Regression	91%	91%	98%	94%
Multinomial Naïve Baye's	84%	83%	100%	91%
Decision Tree	85%	90%	92%	91%
Random Forest	90%	90%	97%	94%
KNN	88%	92%	92%	92%
SVM	80%	80%	100%	89%
Gradient Boost	87%	91%	94%	92%
Ada Boost	87%	91%	93%	92%

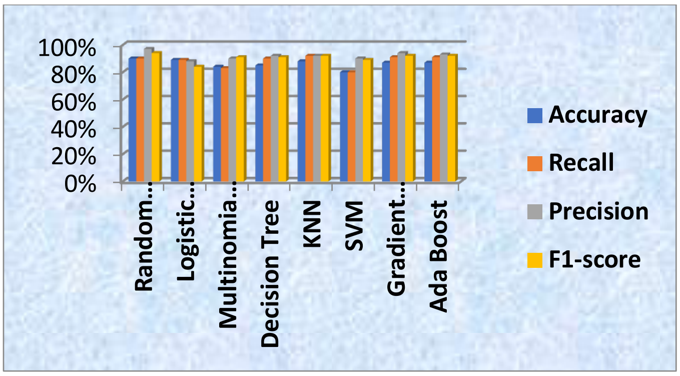


Fig.5: Result Graphs of Different Algorithms

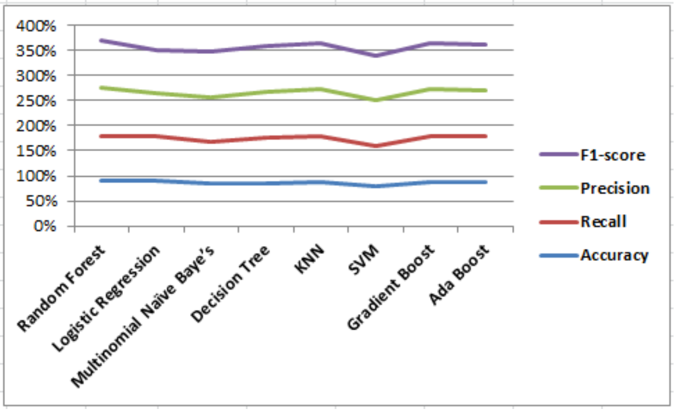


Fig.6: Different ML Model Representation

VI. CONCLUSION

This study outlines a suggested system and the anticipated outcomes of twitter sentiment analysis from communication medium evaluations from social media. The project presents a comprehensive approach to sentiment analysis of Twitter data. Through the utilization of Random Forest, an effective machine learning algorithm, the project successfully achieves accurate sentiment classification of tweets into positive, negative, and neutral categories. The robustness of the developed solution is evidenced by its high accuracy, precision, recall, and F1-score metrics, which indicate its ability to discern sentiment with considerable accuracy.

Furthermore, the project highlights the potential applications of sentiment analysis in various domains, including market research, brand monitoring, and public opinion analysis. The insights gleaned from sentiment analysis can inform decision-making processes and enable organizations to adapt their strategies based on public sentiment trends.

ACKNOWLEDGMENT

I want to thank the authors and publishers for making the availability of their resources. I am equally thankful towards guide, reviewers for their helpful recommendations and also grateful towards college authorities for providing the requisite resources and support.

REFERENCES

[1] Essaidi, Abdessamad, Dounia Zaidouni, and Mostafa Bellafkih. "New method to measure the influence of Twitter users." 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS). IEEE, 2020.

[2] Qi, Yuxing, and Zahratu Shabrina. "Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach." Social Network Analysis and Mining 13.1 (2023): 31.

[3] Bhatnagar, Sarvesh, and Nitin Choubey. "Making sense of tweets using sentiment analysis on closely related topics." Social Network Analysis and Mining 11.1 (2021): 44..

[4] Gupta, Bhumika, et al. "Study of Twitter sentiment analysis using machine learning algorithms on Python." International Journal of Computer Applications 165.9 (2017): 29-34.

[5] Hasan, Ali, et al. "Machine learning-based sentiment analysis for twitter accounts." Mathematical and computational applications 23.1 (2018): 11.

[6] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." arXiv preprint cs/0205070 (2002).

[7] Thomas Wilson, Andrew Evans, Alejandro Perez, Luis Pérez, Juan Martinez. Integrating Machine Learning and Decision Science for Effective Risk Management. Kuwait Journal of Machine Learning, 2(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/208

[8] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." 2014 Seventh international conference on contemporary computing (IC3). IEEE, 2014.

[9] Mahato, M. K. ., Seth, S. ., & Yadav, P. . (2023). Numerical Simulation and Design of Improved Optimized Green Advertising Framework for Sustainability through Eco-Centric Computation. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 11–17. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2502

[10] Navigli, Roberto. "Word sense disambiguation: A survey." ACM computing surveys (CSUR) 41.2 (2009): 1-69.

[11] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[12] Kanavos, Andreas, et al. "Emotional community detection in social networks." Computers & Electrical Engineering 65 (2018): 449-460.

[13] Ding, Xiaowen, Bing Liu, and Philip S. Yu. "A holistic lexicon-based approach to opinion mining." Proceedings of the 2008 international conference on web search and data mining. 2008.

[14] Khan, Farhan Hassan, Saba Bashir, and Usman Qamar. "TOM: Twitter opinion mining framework using hybrid classification scheme." Decision support systems 57 (2014): 245-257.

[15] Krishna, Ragini, and C. M. Prashanth. "Identifying influential users on social network: an insight." Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 1. Springer Singapore, 2020.

[16] Danisch, Maximilien, Nicolas Dugué, and Anthony Perez. "On the importance of considering social capitalism when measuring influence on Twitter." BESC 2014-International Conference on Behavioral, Economic, and Socio-Cultural Computing. IEEE, 2014.

[17] Zhang, Yaocheng, et al. "MoSa: A modeling and sentiment analysis system for mobile application big data." Symmetry 11.1 (2019): 115.

[18] Hansen, Lars Kai, et al. "Good friends, bad news-affect and virality in twitter." Future Information Technology: 6th International Conference, FutureTech 2011, Loutraki, Greece, June 28-30, 2011, Proceedings, Part II. Springer Berlin Heidelberg, 2011.

[19] Tago, Kiichi, and Qun Jin. "Influence analysis of emotional behaviors and user relationships based on Twitter data." Tsinghua Science and Technology 23.1 (2018): 104-113.

[20] Edosomwan, Simeon, et al. "The history of social media and its impact on business." Journal of Applied Management and entrepreneurship 16.3 (2011): 79-91