International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Twitter Sentiment Analysis Using Machine Learning

Alamelu Mangai P[1] , Dr. Nancy Jasmine Goldena[2]

[1]*Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Affiliated to Manonmaniam Sundaranar University, Tirunelveli - 627007,mangaialamelu199@gmail.com*

[2]*Associate Professor, Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Affiliated to Manonmaniam Sundaranar University, Tirunelveli - 627007, nancy_lordwin@rediffmail.com, ORCID: 0000-0002-8626-2604*

ABSTRACT:

Twitter sentiment analysis uses machine learning and natural language processing (NLP) to determine the sentiment of tweets. It can classify tweets as positive, negative, or neutral based on their content. This project aims to develop a Machine Learning model to analyse and classify the sentiment of tweets-specifically categorizing them as positive, negative, or neutral. Utilizing Natural Language Processing (NLP) techniques, the project gathers tweets via the Twitter API, pre-process the text and transform it into numerical format using techniques like TF-IDF or word embeddings. Twitter is the most popular micro-blogging site that allows users to express their views and opinions in 280 characters. As companies and political leaders take to the online social media platform to establish and develop their brand, one cannot ignore the amount to data being generated on Twitter. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiment expressed in the tweets. The proposed system aim to extract and analysis tweets and classify them as positive or negative with the help of machine learning techniques and algorithms, and finally subject to performance evaluation techniques. A comparison of logistic regression and Naive Bayes execution is considered to determine which algorithm works better for given dataset in terms of recall, accuracy and score.

Keywords: Sentiment Analysis, Machine Learning, Natural Language Processing, Social Media, Text Pre-processing.

1. Introduction:

Twitter sentiment analysis uses machine learning and natural language processing (NLP) to determine the sentiment of tweets. It can classify tweets as positive, negative, or neutral based on their content [1]. Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining [2]. We aim to assess the efficacy of ML system in conducting sentiment analysis on Twitter using NLP methodologies. Genism is an open-source python package for NLP. The effectiveness of these algorithms will be evaluated based on several criteria, including F1 score, accuracy, recall, and precision. This information can be used to understand the opinions of people on social media, each day, millions of tweets posted, reflecting a wide range of sentiment on various topics, from politics to entertainment [3]. This analysis aims to identify the most effective algorithm for the dataset, providing insights into public sentiment on demonetization as expressed on Twitter [4]. Analysing these sentiment can provide valuable insights for businesses, researchers, and policymakers. Sentiment analysis to twitter data, we can understand public sentiment regarding specific events, brands, or social issues.

Twitter sentiment analysis examines the overall feeling or emotion expressed in tweets. It employs machine learning and natural language processing techniques to automatically categorize tweets as good, negative, or neutral depending on their content. You can do it manually by analysing each tweet and evaluating whether it is positive or negative [5]. But it is a time-consuming process. Twitter is a popular micro-blogging service in which users post status messages, called “tweets”, with no more than 140 characters. The millions of statues appear on social networking every day. In most cases, its users generated content approximately 200 million user’s post400 million tweets per day [6]. Tweets can express opinions on different topics, which can help to detract marketing campaigns so as to share consumers’ opinions concerning brands and products, outbreaks of bullying, events that generate insecurity, polarity prediction in political and sports discussions, and acceptance or rejection of politicians, all in an electronic word-of mouth way. Secondly, Twitter users post messages on a variety of topics, unlike blogs, news, and other sites, which area tailored to specific topics. Big challenges can be and negative ones. This different from other sentiment analysis domains, which tend to be predominantly positive or negative; and tweets are very short and often show limited sentiment cues. The project will helpful to the companies, political parties as well as to the common peoples. It will be helpful political party for reviewing about the program that they are going to do or the program that they have performed. Similarly, companies also can get reviews about their new product on newly released hardware or software.

2. Literature Review:

Afzaal et al. (2023)[8] have recommended a novel approach of aspect-based sentiment Classification, which recognized the features in a precise manner and attained the best classification accuracy. Moreover, the scheme was developed as a mobile application, which assisted the tourists in identifying the best hotel in the town, and the proposed model was analysed using the real-world data sets. The results have shown that the presented model was effective in both recognition as well as classification.

Feizollah et al. (2023)[9] have concentrated on tweets related to two halal products such as halal

Cosmetics and halal tourism. By utilizing Twitter search function, Twitter information was extracted, and a new model was employed for data filtering. Later, with the help of deep learning models, a test was performed for computing and evaluating the tweets. Moreover, for enhancing the accuracy and building prediction methods, RNN, CNN, and LSTM were employed. From the outcomes, it was seemed that the combination of LSTM and CNN attained the best accuracy.

Mukhtar et al. (2023)[5] have performed the sentiment analysis to the Urdu blogs attained from

several domain with Supervised Machine learning and Lexicon-based models. In Lexicon-based models, a well-performing Urdu sentiment analyser and an Urdu Sentiment Lexicons were employed, whereas, in Supervised Machine learning algorithm, DT, KNN, and SVM were employed. The data were combined from the two sources for performing the best sentiment analysis. Based on the tests conducted, the outcomes were shown that the Lexicon-based model was superior to the supervised machine learning algorithm.

Kumar et al. (2023)[10] have presented a hybrid deep learning approach named Convent-SVMBoVW that dealt with the real-time data for predicting the fine-grained sentiment. In order to measure the hybrid polarity, an aggregation model was developed. Moreover, SVM was used for training the BoVW to forecast the sentiment of visual content. Finally, it was concluded that the suggested Convent-SVMBoVW was outperformed by the conventional models.

Abdi et al. (2021)[10] have proffered a machine learning technique for summarizing the opinions of the users mentioned in reviews. The suggested method merged multiple kinds of features into a unique feature set for modelling accurate classification model. Therefore, a performance investigation was done for four best feature selection models for attaining the best performance and seven classifiers for choosing the relevant feature set and recognized an effective machine learning algorithm. The suggested method was implemented in various datasets. The outcomes have demonstrated that the combination of IG as the feature selection approach and SVM-based classification approach enhanced the performance.

3. Methodology:

3.1Data Collection:

- **Twitter API:** Utilize the Twitter API to collect tweets based on specific Keywords, hashtags, or user accounts.
- **Dataset:** Gather a diverse dataset to ensure the model can generalize well. This will include tweets with varying sentiment.

3.2 Data Pre-processing:

- **Cleaning**: Remove unnecessary elements such as URLs, mention, hashtags, and special characters.
- **Tokenization:** Split text into individual words or tokens.
- **Normalization:** Convert text to lowercase and remove stop words.
- **Lemmatization/steaming:** Reduce words to their base or root form.

3.3 Model Training and Evaluation:

- **Split Dataset:** Divide the dataset into training and testing sets.
- **Training:** Train the models using the training dataset.
- **Evaluation Metrics:** Assess model performance using accuracy, precision, recall, F1 score, and confusion matrix.

3.4 Challenges:

- Handling sarcasm and slang in tweets.
- Dealing with imbalanced datasets where one sentiment class is more prevalent than others.

3.5 Text vectorization:

Use techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word Embedding (Word2Vec, Glove) to convert text data into numerical format for model training.

>>> From sklearn.model_selection import train_test_split X_train,

X_test, y_train, y_test = train_test_split (df. Text, df.label_num, test_size=0.2, # 20% samples will go to test dataset random_state=15, stratify=df.label_num)

Where X_train and X_test are the sample tweets and y_train, y_test are the numerical classes assigned to label neutral, sexist, and racist classes. The ML models will be trained and tested with this data.

3.6 Visualization and Reporting:

- Generates visual reports and dashboards to showcase sentiment trends over time.
- Offers sentiment distribution pie charts, line graphs for sentiment over time, and word clouds for frequently used terms

4. Result and Analysis

4.1 Data Overview

Display the first few rows of the dataset to give a snapshot of the data.

 will output the string "Data Overview:" to the console.

 will display the first 5 rows (by default) of the Data Frame df. This is useful for a quick preview of the data you're working with.

4.2 Text Pre-processing Summary

Text pre-processing is the first step in any NLP pipeline and is designed to convert raw textual data into a format that is easier for algorithms to understand and process. The goal is to remove noise, standardize the data, and extract meaningful patterns or features. Without pre-processing, the text might contain inconsistencies, irrelevant information, or redundant elements, which could negatively affect the performance of downstream models.

4.3Word Cloud Visualization

Generate and display a word cloud to visualize the most common words in the dataset.

- **plt.figure(figsize=(10, 5))**: Sets the size of the figure to 10 inches wide and 5 inches tall.
- **plt.imshow(wordcloud, interpolation='bilinear')**: Displays the word cloud image with bilinear interpolation, which smooths the image.
- **plt.axis('off')**: Hides the axis labels and ticks around the image for a cleaner look.
- **plt.title("Word Cloud of Text Data")**: Adds a title "Word Cloud of Text Data" at the top of the plot.
- **plt.show()**: Displays the plot with the word cloud and the title on the screen.

This code is typically used to visualize a word cloud generated from text data.Text pre-processing completed: tokenization, stemming, stop word removal.

Accuracy Score: 0.85

The sentimental words are categorized into positive and negative words, visually represented using word cloud. In this we mainly focus on sentimental words confusion matrix, hashtag counts and ROC Curve used to determine Accuracy using Naïve bayes classifier.

4.4 Model Evaluation

Accuracy: As far as the accuracy of the model is concerned, Logistic Regression performs better than SVM, which in turn performs better than Bernoulli Naive Bayes.

F1-score: The F1 Scores for class 0 and class 1 are:

(a) For class 0: Bernoulli Naive Bayes(accuracy = 0.90) < SVM (accuracy =0.91) < Logistic Regression (accuracy = 0.92)

(b) For class 1: Bernoulli Naive Bayes (accuracy = 0.66) < SVM (accuracy = 0.68) < Logistic Regression (accuracy = 0.69)

In our problem statement, Logistic Regression follows the principle of Occam's razor, which defines that for a particular problem statement, if the data has no assumption, then the simplest model works the best. Since our dataset does not have any assumptions and Logistic Regression is a simple model.

The models is summarized in the table 4.2.1

Table4.2.1Average accuracies of different models

S. no	**Classifier**	**Accuracy**
1	DAN2	86.06%
2	SVM	85.0%
3	Logistic Regression	74.84%
4	Naïve Bayes	66.24%
5	Random Forest Classifier	87.5%
6	Neural Network	89.93%
7	Maximum Entropy	90.0%
8	k-NN	96.0%

Figure 4.2.1 Average accuracies comparison of Machine Learning models

Average accuracies of different models" refers to the mean performance score (accuracy) of various machine learning models over multiple runs or across different datasets. It helps in evaluating how consistently a model performs compared to others. By computing the average accuracy, we get a more reliable measure of a model's performance, minimizing the impact of any random variations that might occur during training or testing.

The phrase "SVM 85.0%" refers to the performance of a **Support Vector Machine (SVM)** model, a popular type of supervised machine learning algorithm, in terms of its accuracy or effectiveness in a particular task.

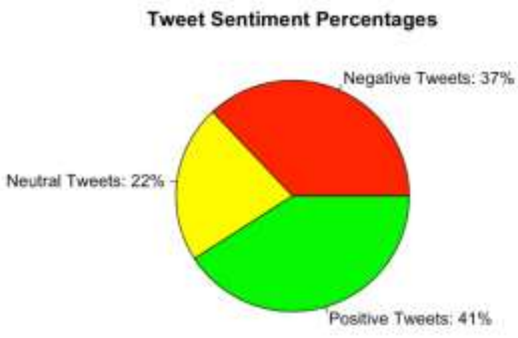

Figure1.Tweet sentiment percentages

The phrase "neural tweets 37%" likely refers to a metric or result related to a machine learning or natural language processing (NLP) model's performance, specifically regarding tweets. In this context, "neural" refers to the use of neural networks—advanced machine learning models designed to process and analyse data, especially textual data like tweets.

The phrase "negative tweets 22%" likely refers to a metric indicating that 22% of a set of tweets have been classified or identified as expressing negative sentiment.

The phrase "positive tweets 41%" likely refers to a sentiment analysis result where 41% of a set of tweets are classified as expressing positive sentiment. In the context of natural language processing

5. Conclusion

Twitter sentiment analysis falls under text and opinion mining. It involves analysing the sentiments expressed in tweets, training machine learning models on this data, and evaluating their accuracy for future applications. The process includes steps such as data collection, text pre-processing, sentiment detection, sentiment classification, and model training and testing. Implements a comprehensive workflow for analysing textual data using various Python libraries. It begins with data pre-processing steps, including tokenization and stemming, to prepare the text for further analysis. The program transforms text into a numerical format suitable for machine learning. Subsequently, it splits the dataset into training and testing sets, allowing for robust evaluation of model performance. A Logistic Regression model is employed for sentiment classification. . Visualizations, including word clouds and confusion matrix displays, enrich the analysis, allowing for intuitive understanding of the data and results. Logistic Regression achieved an accuracy of 74.84%.These results highlight the limitations of these traditional methods in effectively capturing sentiment nuances in tweets.

6. FUTURE ENHANCEMENT

Social media platforms are rich sources of real-time sentiment data. Future trends in sentiment analysis we can likely focus on improving the analysis of social media content, which often contains slang, emoji, and informal language. Additionally, analysing the sentiment of trending topics and viral content will be essential for understanding public opinion and managing brand reputation. Personalization is a growing trend in various fields, and sentiment analysis is no exception. Future sentiment analysis tools may offer personalized sentiment scoring, considering an individual's unique preferences and emotional responses. As sentiment analysis becomes more powerful and pervasive, ethical concerns come to the forefront. Ensuring the responsible and ethical use of sentiment analysis tools will be a significant focus in the future. Sentiment analysis is evolving rapidly, driven by advances in NLP, machine learning, and data availability. As we move forward, we can expect sentiment analysis to become more accurate, nuanced, and integrated into various applications and industries.

7. References

1. P. Domingo and M. Pazzani, “On the optimality of the simple bayesian classifier under zero-one loss,” Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997.
2. R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.
3. A. Balahur, J. M. Hermida, and A. Montoyo, “Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model,” Affective Computing, IEEE Transactions on, vol. 3, no. 1,pp. 88– 101, 2012.
4. Dubey, Pushkar. "Twitter Sentiment Analysis Model." International Journal for Research in Applied Science and Engineering Technology 9, no. VII (July 10, 2021): 296–302. http://dx.doi.org/10.22214/ijraset.2021.36309
5. Gupta, Itisha, and Nisheeth Joshi. "A Review on Negation Role in Twitter Sentiment Analysis." International Journal of Healthcare Information Systems and Informatics 16, no. 4 (October 2021): 1–19.