WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Sentiment Analysis Using Machine Learning on Twitter Data

Thanuku Askhith [a], A Sravya [b], Kurakula Jashwanth [c], Utla srinivas [d], M.Deenababu [e*]

[a,b,c,d] Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

[e] Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT-

In today's interconnected world, social media platforms have become dominant arenas for the expression and dissemination of public and private opinions across a diverse range of subjects. Among these platforms, Twitter stands out as a particularly popular and dynamic space, offering organizations an invaluable opportunity to gain rapid and insightful access to customer perspectives, a critical factor for achieving market success. To capitalize on this potential, this paper outlines the design and implementation of a sentiment analysis program specifically tailored for extracting and analyzing vast quantities of Twitter data. Utilizing the robust capabilities of Python, alongside essential modules such as NumPy for numerical operations, Pandas for data manipulation, and TextBlob for natural language processing, the program aims to computationally measure and classify customer sentiments expressed in tweets. The core objective is to categorize these sentiments into positive and negative polarities, providing a clear and quantifiable understanding of customer feedback. The results of this analysis are then presented in both a visually intuitive pie chart and a detailed tabular format, enabling organizations to readily grasp the overall distribution of customer sentiment and make informed decisions based on these insights. This approach provides a streamlined and efficient method for organizations to monitor and understand customer opinions, crucial for navigating the complexities of the modern marketplace.

Keywords: Analysis, Python, Sentiments, TextBlob, Natural Language Processing, NumPy

1. Introduction

Sentiment analysis has become a vital tool for businesses to understand public perception by computationally categorizing textual emotions. Applying this to Twitter data, with its massive user base and real-time information flow, provides a unique and immediate window into public opinion, making it crucial for marketing and customer service. Twitter's open platform allows direct interaction with a vast audience, but also presents risks due to the rapid spread of negative content[1]. Social listening, particularly on Twitter, is essential for businesses to monitor conversations, understand their audience, track brand mentions, and identify trends. This proactive approach enables quick responses to potential crises and fosters stronger customer relationships[2]. While quantitative metrics offer a basic understanding of Twitter activity, sentiment analysis delves into the qualitative aspects, revealing the emotional tone behind mentions. This insight is crucial for businesses to accurately gauge public perception, tailor marketing strategies, improve customer service, and address issues before they escalate[3]. This project aims to provide a comprehensive guide to Twitter sentiment analysis, covering data collection tools, program implementation, and detailed explanations of relevant libraries and tools. It will also address working with Twitter's API, text preprocessing techniques, and different sentiment analysis approaches, including lexicon-based, machine learning-based, and deep learning-based methods[4]. Quantitative metrics like mention counts and retweets provide a surface-level understanding of Twitter activity, but they fail to capture the nuances of public sentiment. Sentiment analysis, on the other hand, delves into the qualitative aspects, revealing the emotional tone behind mentions[5]. This deeper understanding is crucial for businesses to make informed decisions about marketing, customer service, and product development. By understanding the underlying opinions and feelings of customers, companies can tailor their strategies and address potential issues before they escalate[6]. Finally, the project emphasizes data visualization for presenting sentiment analysis results and evaluating model accuracy using metrics like precision, recall, and F1-score. By providing practical guidance, it aims to equip users with the tools and knowledge to effectively monitor and analyze public opinion on Twitter, enabling businesses to maintain a positive brand image and stay competitive[7].

2. Related Work

Research in Twitter sentiment analysis frequently explores the comparative effectiveness of machine learning algorithms, particularly Random Forest, Logistic Regression, and Naive Bayes. Studies consistently emphasize the critical role of data preprocessing and feature engineering, which significantly influence model performance. Common preprocessing steps include cleaning tweets by removing irrelevant elements, tokenizing text, removing stop words, and applying stemming or lemmatization[8]. Feature extraction techniques, such as TF-IDF and n-grams, are also crucial for transforming textual data into a format suitable for machine learning models. Furthermore, the selection of appropriate datasets and addressing class imbalance are recurring themes in this domain. Researchers often utilize publicly available datasets, but they must also contend with the inherent challenges of Twitter's informal

and often ambiguous language[9]. Findings from these studies reveal that while Naive Bayes can achieve surprisingly high accuracy, as evidenced by claims of up to 94% in some instances, Logistic Regression and Random Forest also demonstrate strong performance. Logistic Regression is particularly effective in high-dimensional spaces, which is typical of text data, and provides interpretable results[10]. Random Forest, as an ensemble method, excels at capturing complex relationships and is robust to noisy data. The context in which sentiment analysis is applied, such as political analysis, brand monitoring, or market research, also plays a significant role in determining the most suitable algorithm and feature set. The nuances of Twitter data, including slang, emojis, and sarcasm, necessitate specialized techniques and a contextual understanding for accurate sentiment classification[11]. Ultimately, the related work underscores the importance of a comprehensive approach that combines effective data preprocessing, careful feature engineering, and appropriate algorithm selection. While Naive Bayes can achieve impressive accuracy when combined with strong feature engineering, other algorithms like Logistic Regression and Random Forest offer distinct advantages in terms of interpretability and robustness. The application context and the unique characteristics of Twitter data further influence the choice of methodology. Researchers continue to refine these techniques to improve the accuracy and reliability of Twitter sentiment analysis, which has become an increasingly valuable tool for understanding public opinion and trends[12].

3. Proposed Methodology

The process we followed to reach our results, we present and describe the stages in figure 1. This system aims to provide a robust methodology for early prediction and classification of tweets such as "positive, negative and neutral". The system is built on a foundation of ML models, like Random Forest, Logistic regression and Naive Bayes Classifier. Each model evaluated using key metrics like Accuracy, Precision, Recall, F1-Score The feature selection process employs Feature Elimination using Recursive method to prioritize the most influential variables, ensuring the models remain both efficient and interpretable.

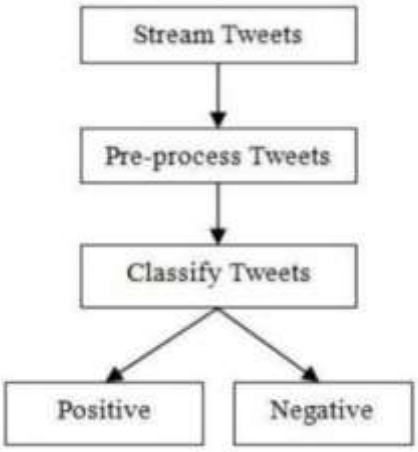

Figure 1: Demonstration of Proposed System

3.1 Feature Extraction

Chi-Square test identifies the top 4 features highly correlated with the target variable then fed into the classification models. Features are reduced to 4 key features after feature extraction and each one is denoted by

$X = [[f1, f2, \ldots\ldots, fn], Y \in \{0,1\}]$----------(1)

Where X and Y represents feature vector and target variable and Y=0, poor performance and y=1, successful performance.

$X' = [f1', f2', f3', f4']$----------(2)

Chi-Square, $\chi 2 = \sum_{k=1}^{K} (O_k - E_k)^2 / E^k$----------(3)

Where, O_k and E_k are observed frequency and expected frequency.

The extracted features help the models to differentiate between students likely to perform well and those at risk of dropping out.

3.2 Logistic Regression

Logistic regression is used for binary classification and uses sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

$z = w \cdot X + b$----------**(4)**

Kernel Function**:** For non-linearly separable data, a kernel function p(y,w) to map data into higher dimensions:

$$P(Y = 1 \mid x, \mathbf{w}) = \frac{1}{1 + e^{-(\mathbf{w} \cdot \Phi(\mathbf{x}) + b)}} \text{----------(5)}$$

3.3 Random Forest

In training stage several decision trees are constructed then merge their results to get better accuracy and then reducing over-fitting. In the implementation, parameters such as n_estimators=1 and max_depth=0.9 restrict the model's ensemble capability, causing it to behave similarly to a single tree. Prediction Aggregation for classification

$y^{\wedge} = Mode\{T1(x), T2(x), \ldots, TK(x)\}$----------(6)

Where $T_k(x)$ is the prediction from the k-th tree.

3.4 Naive Bayes Classifier

The probability of given set of inputs for all possible values of the class variable *y* and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \text{----------(7)}$$

3.5 Evaluation Metrics

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table. The performance metrics used for classification and regression

Metric	**Formula**
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$2 * \frac{\mathrm{R} * \mathrm{P}}{\mathrm{R} + \mathrm{P}}$

4. RESULT AND DISCUSSION

In sentiment analysis of tweets, Random Forest often yields the highest accuracy due to its robust handling of complex data and prevention of overfitting, though it's computationally intensive. Logistic Regression provides a strong balance of accuracy and efficiency, particularly with linear relationships, while Naive Bayes, known for its speed and effectiveness with high-dimensional text data, can be surprisingly competitive despite its simplifying independence assumption; the ultimate perform ance of each model heavily depends on dataset quality, preprocessing, and feature extraction, evaluated through metrics like accuracy, precision, recall, and F1-score.

	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. &
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!!! RT @UrEnemDBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? Y
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than
5	3	1	2	0	1	!!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me ,claim you so faithful and down for somethin
6	3	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch . I got too much shit going
7	3	0	3	0	1	!!!!“@swiftlequenster: cause I'm tired of you big bitches coming for us skinny girls!!”
8	3	0	3	0	1	" & you might not get ya bitch back & that's that "
9	3	1	2	0	1	"
10	3	0	3	0	1	" Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh
11	3	0	3	0	1	" Murda Gang bitch its Gang Land "
12	3	0	2	1	1	" So hoes that smoke are hoers ? " yea ... go on lil
13	3	0	3	0	1	" bad bitches is the only thing that I like "
14	3	1	2	0	1	" bitch get up off me "
15	3	0	3	0	1	" bitch nigga miss me with it "
16	3	0	3	0	1	" bitch plz whatever "
17	3	1	2	0	1	" bitch who do you love "
18	3	0	3	0	1	" bitches get cut off everyday bi "
19	3	0	3	0	1	" black bottle & a bad bitch "
20	3	0	3	0	1	" broke bitch cant tell me nothing "
21	3	0	3	0	1	" cancel that bitch like Nino "
22	3	0	3	0	1	" cant you see these hoes wont change "

Figure 1: Sample Dataset

	precision	recall	f1-score
Random Forest Baseline	0.854	0.9285	0.8897
Logistic Regression Baseline	0.9064	0.8921	0.8992
Naive Bayes Baseline	0.8756	0.9212	0.8978

Figure.2. Evaluation Results

Algorithm	Precision	Recall	F1-score
Random Forest	0.854	0.928	0.889
Logistic regression	0.906	0.892	0.899
Naive Bayes	0.875	0.921	0.897

The performance metrics—precision, recall, and F1-score—for three machine learning algorithms: Random Forest, Logistic Regression, and Naive Bayes, in a baseline classification task. Logistic Regression demonstrates the highest precision (0.9064), indicating a low rate of false positives, while Random Forest and Naive Bayes exhibit higher recall (0.9285 and 0.9212, respectively), suggesting better identification of positive instances. The F1-scores, which balance precision and recall, are consistently high across all models, ranging from 0.8897 to 0.8992, indicating strong overall performance.

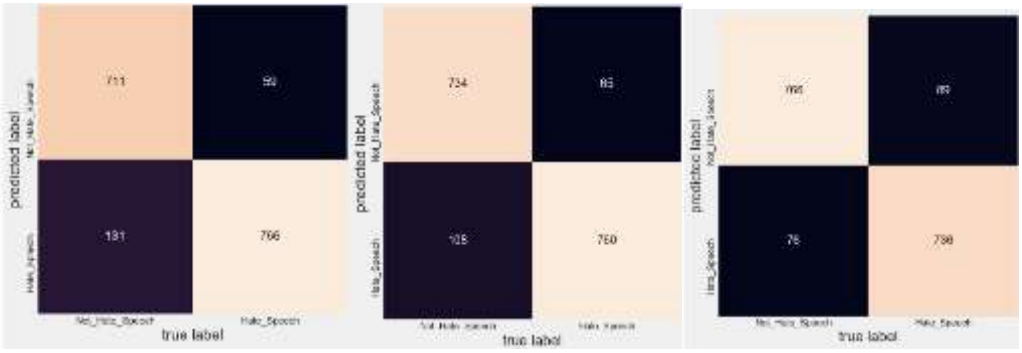

Figure .4. Random Forest, Naive Bayes and Logistic Regression confusion matrix

The performance of different machine learning algorithms. The Random Forest has high accuracy (88%), Naive Bayes (84%), Logistic regression (83.5%).

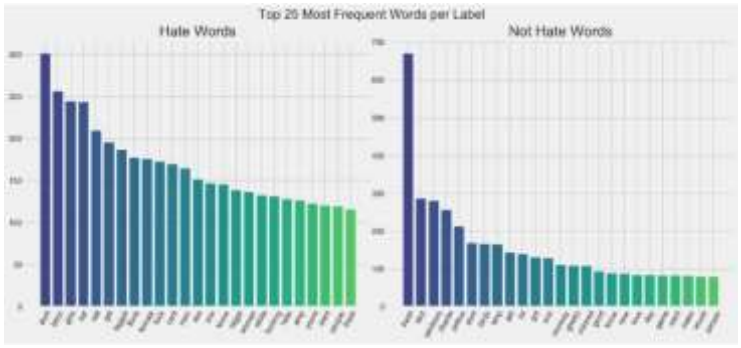

Figure 5: The Classification of Tweets

Table 3. Comparative Summary of Models

Algorithm	**Accuracy (%)**	**Key Characteristics**
Logistic Regression	83.5%	Maximum likelihood function, interpretable, and non-linear relationships.
Naive Bayes	84%	Simplicity and efficiency; effective high dimensional data.
Random Forest	88%	Robustness to overfitting, Feature Importance, Versatility.

The results validate the efficiency of feature selection in improving model performance, and the Decision Tree algorithm emerged as the most reliable model for accurate predictions in this application. Finally, the results are analyzed to uncover meaningful insights to understand how different features impact student performance, identifying potential areas of improvement in the model, and visualizing the results for better interpretability.

5. CONCLUSION

In this research, we established a data-driven approach for sentiment analysis for tweets using machine learning algorithms. Twitter sentiment analysis, a specialized field within text and opinion mining, has achieved notable advancements, with models now demonstrating efficiency rates between 85% and 90%. This process encompasses several stages, from data collection and preprocessing to sentiment detection and model training. Its applications span various industries, enabling businesses to monitor brand perception, analyze customer feedback, and conduct market research. Similarly, academic performance prediction employs machine learning to forecast student success, highlighting the crucial role of prior academic performance and engagement metrics. Models like Random Forest, Naive Bayes classifier have proven effective in capturing complex relationships, and interpretability is enhanced through techniques. Future advancements in both areas will likely involve AI-driven techniques, including multimodal data analysis, transfer learning, and adaptive learning systems. Despite its progress, Twitter sentiment analysis remains an evolving field with potential for further refinement. Future research should focus on enhancing unigram models by incorporating negation effects, exploring the impact of bigrams and trigrams with larger datasets, and integrating Part-of-Speech information. Addressing class imbalance through dataset expansion and conducting context-specific sentiment analysis are also essential. Furthermore, modelling human confidence in sentiment labels could improve the reliability of analyses. In academic performance prediction, future efforts should emphasize leveraging multimodal data, employing transfer learning, and developing AI-powered adaptive learning systems to personalize educational experiences. Both fields illustrate the increasing importance of AI and machine learning in extracting valuable insights from data, facilitating more accurate predictions and informed decision-making.

REFERENCES

[1]. Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. Discovery Science, Lecture Notes in Computer Science, 2010, Volume 6332/2010, 1-15

[2]. Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.

[3]. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining.

[4]. Bermingham, A., & Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results.

[5]. Joulin, A., Grave, E., Mikolov, T., Ruder, S., & Dupret, G. (2017). Bag of Tricks for Efficient Text Classification.

[6]. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis.

[7]. Sarker, A., & González, H. (2017). Comparitive study of sentiment analysis of twitter data.

[8]. Qiu, L., & Zhang, Z. (2016). Analyzing and predicting social media sentiment dynamics.

[9]. Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys.

[10]. Zhang, Z., & Liu, Y. (2018). A survey of sentiment analysis on Twitter.