Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.4, May : 2024

SENTIMENTAL ANALYSIS ON TWITTER DATA PREDICTION USING NLP

Dr. Moturi Sireesha Associate professor Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet, Andhra Pradesh sireeshamoturi@gmail.com
Sanapathi Akshaya kumar Student Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet, Andhra Pradesh akshayakumarsanapathi@gmail.com
Adhi Gangadhar Rao Student Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet, Andhra Pradesh gangadharadhi@gmail.com
Rolla Narasimha Rao Student Computer Science and Engineering Narasaraopeta Engineering College (Autonomous) Narasaraopet, Andhra Pradesh rollanarasimharao@gmail.com

Abstract-
Sentimental analytics is a phenomenon that displays many people's thoughts, feelings, and opinions. Sentiment analysis facilitates improved client communication and output development for writers. It is the computational process of locating and obtaining different user perspectives or alternatives inside that particular situation. Sentiment analysis involves four main stages: gathering data, cleaning it up using preprocessing techniques, analyzing it, and interpreting the findings using machine learning algorithms. It also goes by the name "opinion meaning," which denotes the underlying motional tone of a text. Sentiment analysis is often implemented using machine learning methods like Naive Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM). These algorithms analyze the text data and classify it into different sentiment categories based on patterns and features extracted from the text.

In addition to SVM, LR, and NB, other algorithms such as Random Forest, K-Nearest Neighbors (KNN), and Decision Trees can also be utilized for sentiment analysis, providing alternative approaches to analyzing and interpreting text data. Evaluation of algorithm performance is typically conducted in terms of metrics such as accuracy, F1-score, precision, and recall often using tweets sourced from social media platforms [1, 2, 3].

Keywords- Sentiment Analysis, Emotional Tone, Opinion Mining, Machine Learning Algorithms, K- Nearest Neighbors (KNN), Support Vector Machine (SVM), Precision, Logistic Regression (LR), Recall, F1-Score, Naive Bayes (NB), Random Forest, and Accuracy.

I.INTRODUCTION

As a fundamental component of artificial intelligence, machine learning works in tandem with human intelligence by obeying commands to interact, alter, and evaluate data. These algorithms usually use input data to anticipate output values, categorized into supervised, unsupervised, reinforcement, and semi-supervised learning, primarily focusing on tasks like classification and regression. Examples include Naive Bayes, logistic regression and support vector machines.

Deep learning, emphasizes complex neural network structures and algorithms, facilitating faster processing and commonly applied in tasks such as image recognition, speech recognition and natural language processing (NLP). Noteworthy, algorithms in deep learning includes Random Forest, Long Short-Term Memory (LSTM) and Naive Bayes.

Understanding consumer sentiment in the modern world is crucial for many industries, including customer service management, online communications, and feedback systems [3]. This is especially true for customer sentiment as it is represented on social networking sites like Twitter, LinkedIn, and Facebook and in newspapers.

Twitter analysis focuses on evaluating audience responses to content, optimizing data for future utilization. This analysis encompasses various metrics such as tweet impressions, profile visits, mentions, and follower counts. Sentiment analysis and emotion recognition play vital roles in

extracting user opinions, aiding authors in effective communication and output development. The process involves data collection, cleaning through preprocessing techniques, analysis, and result interpretation using machine learning algorithms.

Twitter data analytics boasts a vast user base of approximately 237 million individuals across age groups globally. A proposed 85% accuracy rate is achievable in sentiment analysis of user reviews through mathematical calculations [4].

Statistical analyses reveal that countries like the USA and Japan exhibit high engagement with tweets, expressing opinions and emotions towards various topics. Notably, newspaper reporting and journalistic tweets command higher attention percentages, indicating greater time spent compared to other social media content.

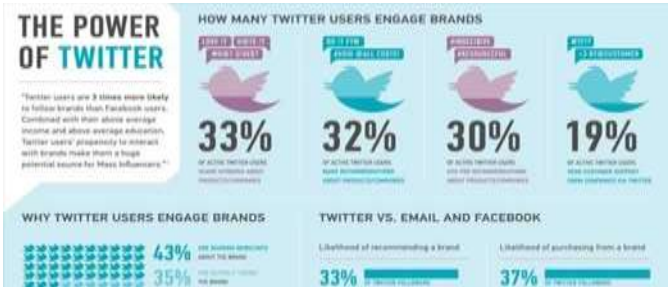
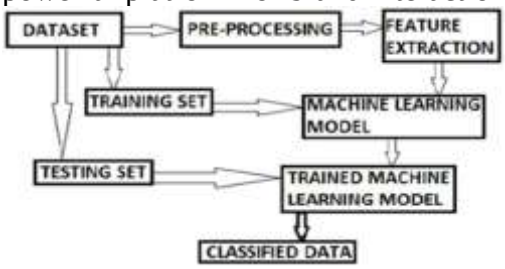


Fig. 1. Prevalency of Twitter

Fig. 1 shows how Twitter users actively engage with brands, with approximately 33% following brands, 32% talking about them, and 30% retweeting brand content. Additionally, 19% seek customer service via Twitter each month. Compared to email and Facebook users, Twitter users are more likely to recommend brands 43% and make purchases 37%. Overall, Twitter's engaged user base makes it a powerful platform for brand interactions.



II. LITERATURE SURVEY

Numerous recent studies have delved into sentiment analysis on the data from the twitter, employing a plethora of the methodologies and approaches to gain insights into users' emotions and opinions. For instance, researchers have explored the effectiveness of kernel models supplemented by emoticon and acronym dictionaries, as well as Unigram and tree kernel models, demonstrating the superiority of unigrams plus senti-features over kernels in sentiment prediction. Additionally, attempts have been made to classify tweets into positive, neutral, and negative categories using the Naive Bayes algorithm, albeit with limited success due to noisy training data, resulting in only 40% accuracy for the neutral class.

In the realm of deep machine learning, sentiment analysis on Twitter data has exhibited higher accuracy through the utilization of maximum entropy with Bigram and stop words in slang feature combinations. With Twitter producing millions of tweets and massive volumes of data, researchers have employed classification and clustering techniques to discern relationships between tweets with subjectivity and emotive ratings based on polarity principles. Furthermore, multiclassification techniques, including random forest, have achieved considerable accuracy, reaching 77.4% accuracy, 77.4% recall, and 76.6% precision.

Hybrid models, combining rule-based and lexicon-based algorithms, have shown promise, achieving a harmonic mean of 83.3% across six datasets. The challenge of sentiment analysis on Twitter data is exacerbated by the sheer scale of the population and data volume, prompting researchers to explore solutions using open-source frameworks like Apache Spark, with the expectation that classification and regression models will enhance efficiency and output quality.
Cloud computing technology has emerged as a valuable tool for sentiment analysis in various sectors, including business and banking, with examples such as SAAS, email, and social networks [4]. Despite advancements, the need for new artificial methods to improve accuracy and results remains paramount. Moreover, sentiment analysis has been applied to social media discussions on stock market movements, with an average accuracy of 56% using topic modeling techniques.

Additionally, sentiment analysis using Hadoop has been explored, with researchers claiming 80.85% accuracy employing classification techniques. Logistic regression classifiers have also shown promise, achieving accuracies ranging from 77% to 82% depending on data size. However, challenges persist, particularly in sentiment classification across different language datasets, where under-sampling methods have proven inadequate.
In conclusion, while existing sentiment analysis works have made strides in understanding customer emotions and opinions on Twitter, there is a consensus among researchers that further advancements, particularly in artificial methods, are necessary to enhance accuracy and results. These studies collectively underscore the importance of sentiment analysis in deciphering user sentiments and opinions in the digital age.

III PROPOSED SYSTEM

3.1 Data Analysis:

we utilize the dataset (Twitter dataset) from the Kaggle website. This dataset contains 31961 row and 4 columns using these data we predict the sentimental analysis [5].

3.2 Data Preprocessing:

3.3

we can preprocess the data by using some NLP preprocessing Techniques like:

3.3.1. Text Cleaning:

a. Text data often comes with noise and irrelevant information, such as HTML tags, URLs, special characters, and punctuation, which can distract the model from understanding the content.
b. Removing HTML tags, URLs, and special characters: Useful for cleaning web data.
c. Lowercasing: Converts all characters in the text to lowercase to ensure uniformity and reduce the vocabulary size.
d. Removing stop words: Stop words are common words like "is", "and", "the", etc., that are often removed since they carry little meaningful information for analysis.
e. Removing punctuation: Punctuation marks can be removed as they usually don’t add much value to the meaning of a sentence for many NLP tasks.

3.2.2 Tokenization:

This involves breaking down text into smaller units, such as words or phrases. It’s the foundational step in turning text into data that can be analyzed.
a. Word Tokenization: Splits text into words
b. Sentence Tokenization: Segments text into sentences.

```
0    [when, father, dysfunctional, selfish, drags, ...
1    [thanks, #lyft, credit, cause, they, offer, wh...
2                              [bihday, your, majesty]
3                    [#model, love, take, with, time]
4                  [factsguide, society, #motivation]
Name: clean_tweet, dtype: object
```

Fig. 3. word Tokenization

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.4, May : 2024

3.2.2 Normalization:

Normalization aims to convert all variations of a word into a canonical form, so they can be analyzed as a single item.

a. Stemming: A crude process of chopping off word endings to return a word to its base or root form.

b. Lemmatization: More sophisticated than stemming, it involves using vocabulary and morphological analysis to return words to their base or dictionary form (lemma).

3.2.2 Part-of-Speech Tagging:

Assigns parts of speech to each word (like nouns, verbs, adjectives, etc.), based on its definition and context. It's useful for disambiguation and for subsequent processing steps.

3.4 Data Visualization:

Data visualization transforms complex data sets into intuitive graphical representations, making it easier to identify patterns, trends, and outliers at a glance. It serves as a vital tool in decision-making processes by presenting data in a clear, accessible manner across various industries. Through charts, graphs, and maps, data visualization enables both experts and non-experts to uncover insights from vast amounts of information efficiently.

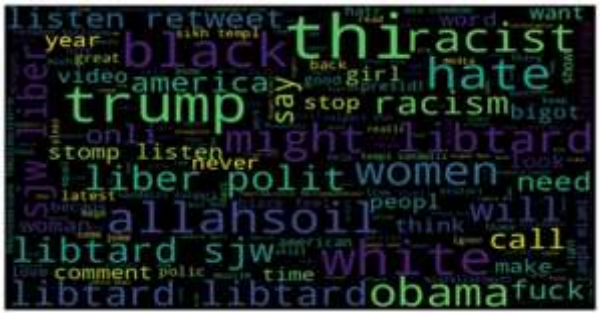


Fig. 4. Word cloud for the positive tweet Fig. 5. Word cloud for the Negative tweet

The word cloud of the dataset's positive and negative tweets is displayed in the visualization in Fig. 4 and Fig 5.

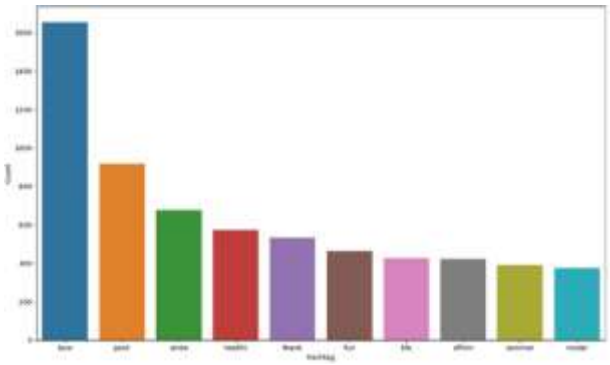


Fig. 6. Popular Hashtag and count

The Fig. 6 visualization specifies the popular hash tags and its count in the dataset

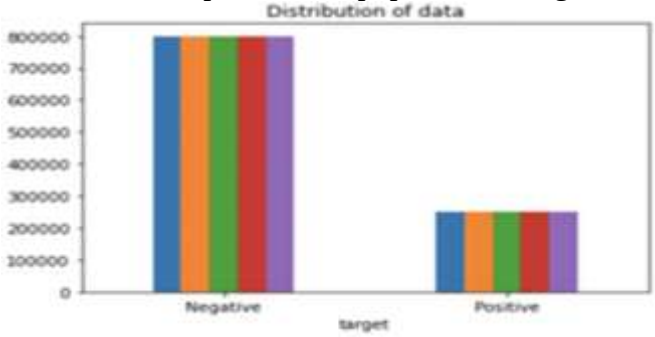


Fig. 7. Target of negative positive tweets

The parameters used in the dataset of positive or negative tweets are specified in the Fig. 7 visualization, which is displayed in the bar graph below

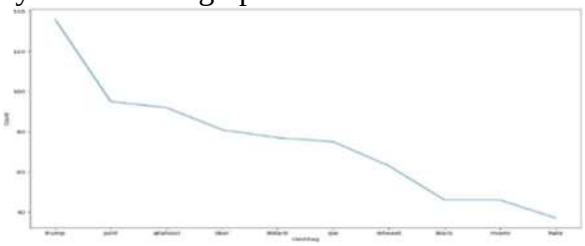


Fig. 8. Data visualization of popular hashtags

The Fig. 8 visualization specifies the popular hashtags and it count numbers in the dataset.

3.1 Model Evaluation:

Using the already prepared data, an appropriate machine learning algorithms are chosen and taught. In order to reduce the discrepancy between the model's anticipated output and the actual output found in the training set, the parameters are optimized. The model is tested on a different validation dataset to gauge its performance after training. The evaluation metrics used depend on the type of problem and the performance criteria. Common evaluation metrics include accuracy, F1 score, precision, Area Under the Curve (AUC) and recall. Based on the results, the model can be further refined by adjusting any of its parameters or selecting a different algorithm. The model may be used in a production setting to make predictions or judgments based on new data, once it has been trained and assessed.

3.2 Accuracy

Accuracy is a common metric used to evaluate the performance of a machine learning algorithm. It calculates the percentage of each instance in the test dataset that is properly categorized.

The accuracy rates for various algorithms on a given dataset are as follows: Logistic Regression achieved 94% accuracy, Naïve Bayes also achieved 94% accuracy, Support Vector Machine achieved 94% accuracy, Random Forest achieved 95% accuracy, KNN achieved 93% accuracy, and Decision Tree achieved 94% accuracy.

ALGORITHM	**ACCURACY**
Logistic Regression	94%
Naïve Bayes	94%
Support Vector Machine	94%
Random Forest	95%
KNN	93%
Decision Tree	94%

Table -1 Accuracy of different algorithms

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.4, May : 2024

Example Code

```
from sklearn.ensemble import RandomForestClassifier

# Initialize Random Forest Classifier
random_forest = RandomForestClassifier()

# Train Random Forest Classifier
random_forest.fit(x_train, y_train)

# Predictions
random_forest_pred = random_forest.predict(x_test)

# Evaluation metrics for Random Forest
random_forest_metrics = [accuracy_score(y_test, random_forest_pred),
                         precision_score(y_test, random_forest_pred),
                         recall_score(y_test, random_forest_pred),
                         f1_score(y_test, random_forest_pred)]

# Update results dictionary
results['Random Forest'] = random_forest_metrics

# Create DataFrame
results_df = pd.DataFrame(results)
print("Performance Metrics:")
print(results_df)
```

#Output:0.95

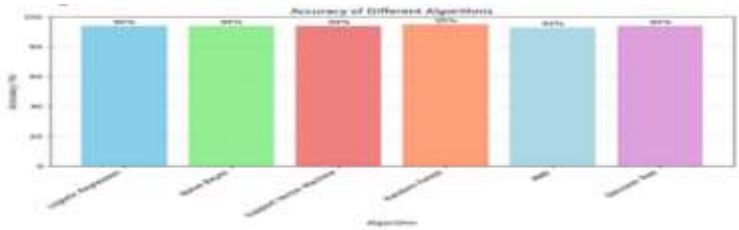


Fig. 9. Accuracy of different algorithms

The bar plot visualization, as shown in Fig.9, effectively represents various machine learning algorithms performance based on their accuracy percentages. Every bar in the figure represents a different algorithm, and the height of the bar represents the accuracy the algorithm was able to accomplish. The use of distinct colors for each bar enhances the visual appeal and aids in distinguishing between the different algorithms. Additionally, the inclusion of annotations at the top of each bar displaying the accuracy percentage provides clear and concise information to the viewer.

By comparing the heights of the bars, viewers can easily identify which algorithm performs the best in terms of accuracy. In this particular plot, the Random Forest algorithm stands out with the highest accuracy rate of 95%, closely followed by Logistic Regression, Naïve Bayes, and Decision Tree, all achieving an accuracy of 94%. On the other hand, KNN lags slightly behind with an accuracy rate of 93%.

This visual representation allows for a quick and intuitive comparison of algorithm performance, enabling users to choose the most appropriate algorithm with knowledge for the given task. The ability to visually assess the relative performance of different algorithms based on their accuracy rates can be particularly valuable in fields such as machine learning and data analysis, where algorithm selection plays a crucial role in achieving desired outcomes.

The bar plot provides a concise and informative visualization of algorithm performance, highlighting the strengths and weaknesses of each algorithm in terms of accuracy. This type of visualization can serve as a valuable tool for researchers, practitioners, and decision-makers seeking to optimize algorithm selection for their specific needs.

II CONCLUSION AND FUTURE SCOPE

Sentiment analysis, or sentimental analytics, is a valuable approach for understanding the emotions, opinions, and sentiments conveyed in text data. Utilizing ML algorithms such as SVM, Logistic

Regression (LR), NB, as well as Random Forest, K-Nearest Neighbors (KNN), and Decision Tree. sentiment analysis can effectively categorize text into various sentiment categories.
Through the process of data collection, cleaning, analysis, and result interpretation, sentiment analysis aids authors and businesses in better understanding their customers and developing products or services that resonate with them. The performance of sentiment analysis algorithms is often assessed using metrics like accuracy, precision, recall, and F1-score, which offer insights into the models' effectiveness. Among these algorithms, Random Forest achieved the highest accuracy of 95%, outperforming SVM, LR, NB, KNN, and Decision Trees, making it the most effective algorithm for sentiment analysis in this study.
The field of sentimental analysis presents several exciting avenues for future exploration and advancement. Continued research into machine learning algorithms and techniques may lead to the development of more sophisticated models for sentiment analysis, enhancing accuracy and efficiency. Exploration of deep learning techniques, such as

Integration of text data with other modalities, such as images, videos, and audio, could offer a more comprehensive understanding of sentiment, opening up new possibilities for sentiment analysis in multimedia content.

Contextual data, such user demographics and social context, might provide a more nuanced understanding of sentiment, which could improve the accuracy of sentiment analysis algorithms.

Development of real-time sentiment analysis systems capable of analyzing streaming data could enable businesses to respond promptly to changing customer sentiments and trends.

By exploring these areas of future research, sentiment analysis can continue to evolve, providing valuable insights into customer opinions and emotions and helping businesses make informed decisions to enhance customer satisfaction.
convolutional neural networks (CNNs) and recurrent neural networks (RNNs), could further improve sentiment analysis by capturing intricate patterns in text data.

III REFERENCES

[1] Priyavrat Chauhan, Nonita Sharma and Geeta Sikka, "The emergence of social media data and sentiment analysis in election prediction", Journal of Ambient Intelligence and Humanized Computing, February 2021.
[2] S. M. Kayes, M. S. Islam, P. A. Watters, A. Ng and H Kayesh, "Automated measurement of attitudes towards social distancing using social media: A COVID-19 case study", Tech. Rep., Oct. 2020.
[3] C. K. Pastor, "Sentiment analysis on synchronous online delivery of instruction due to extreme community quarantine in the Philippines caused by Covid-19 pandemic", Asian J. Multidisciplinary Stud., vol. 3, no. 1, pp. 1-6, Mar. 2020.
[4] Dubey AD, Twitter sentiment analysis during COVID- 19 outbreak, April 2020, [online] Available:
[5] G. Barkur, Vibha and G. B. Kamath, "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India", Asian J. Psychiatry, vol. 51, no. 102089, Jun. 2020.
[6] Singh N. K, Tomar D. S and Arun Kumar S, "Sentiment analysis: a review and comparative analysis over social media", Springer Science and Business Media LLC 2020, 2020.
[7] K. Chakraborty, S. Bhattacharyya and R. Bag, "Asurvey of sentiment analysis from social media data", IEEE Trans. Comput. Soc. Syst, vol. 7, no. 2, pp. 450-464, Jan 2020.
[8] Arwa A. Al Shamsi, Reem Bayari and Said Salloum, "Sentiment Analysis in English Texts", Advances in Science Technology and Engineering Systems Journal, January 2021.
[9] Ashraf Elnagar, Sane Yagi, Ali Bou Nassif, Ismail Shahin and Said A. Salloum, Sentiment Analysis in Dialectal Arabic: A Systematic Review, AISC, vol. 1339, 2021.
[10] https://www.kaggle.com/code/muhammadimran11223 3/eda-twitter-sentiment-analysis-using-nn