International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

TWITTER SENTIMENT ANALYSIS

Jayant Mishra
MCA (Data Science & Artificial Intelligence), Babu Banarasi Das University, Lucknow, Uttar Pradesh, India- jrnarayan@gmail.com

Abhishek Kumar
MCA (Data Science & Artificial Intelligence), Babu Banarasi Das University, Lucknow, Uttar Pradesh, India- abhishek221197@gmail.com

Dr. Prabhash Chandra Pathak
Head of Department, School of Computer Application, Babu Banarasi Das University, Lucknow, Uttar Pradesh, India- Hod.ca@bbdu.ac.in

Abstract
Sentiment analysis is a popular research field that aims to extract opinions, attitudes, and emotions from social media platforms like Twitter. Traditionally, sentiment analysis has focused on analyzing textual data. Twitter, being a well-known microblogging platform where users post updates in the form of tweets, is often used as a data source for sentiment analysis studies.

In this particular paper, the authors utilize a publicly available labeled dataset from Kaggle. They also propose a set of preprocessing steps to enhance the manageability of tweets for natural language processing techniques. Since the dataset consists of pairs of tweets and their corresponding sentiment labels, supervised machine learning is employed. The authors propose sentiment analysis models based on naive Bayes, logistic regression, and support vector machine algorithms with the objective of achieving more effective sentiment analysis. In Twitter sentiment analysis, tweets are typically categorized into positive or negative sentiment. Machine learning classifiers can be utilized for this purpose. These classifiers can be valuable for various entities such as businesses, political parties, and analysts, as they can evaluate the sentiments expressed towards them. By appropriately training the machine learning models with labeled data, tweets can be accurately classified without the need for a predefined word database. Consequently, machine learning techniques offer a superior and faster approach to performing sentiment analysis.

Keywords: Twitter; sentiment; Web data; text mining; SVM; Bayesian algorithm; hybrid; ensembles

1. Introduction

The widespread use of the internet has allowed people worldwide to express their opinions on various platforms such as blogs, online forums, and product review sites. In today's world, individuals heavily rely on user-generated content to make informed decisions. For instance, before purchasing a product, many
people search for reviews and comments to gather information [1]. However, it is impractical for a person to manually go through every single review available as it would be time-consuming. To address this issue, automation can be employed, and Machine Learning (ML) plays a crucial role in this regard. Sentiment Analysis (SA), a branch of ML, helps systems understand the sentiment conveyed in a statement.

The research indicates that ML methods have outperformed knowledge- and dictionary-based approaches in determining the polarity, which refers to the semantic orientation ranging from 0 to 1 or positive to negative [2]. The paper proposes a system

that extracts data from Twitter to perform sentiment analysis. The extracted data is stored in a data frame, and several cleaning and preprocessing steps are applied to ensure accurate information is utilized for training the ML model. This model is then used to predict labels for unknown, cleaned, and preprocessed data samples.

Twitter is chosen as a data source due to its vast amount of available data. To achieve accurate outcomes in sentiment analysis, the paper employs supervised machine learning methods such as multinomial naive Bayes, linear support vector classifiers, and logistic regression classifiers. Twitter allows users to compose tweets in any form without following strict rules, leading to the usage of abbreviations, spelling mistakes, exaggerated reviews, and emoticons [3]. While these formats may pose challenges for analysis, techniques like feature extraction and mapping emoticons to their actual meanings can be employed to investigate the tweets.

This paper primarily focuses on product reviews as a source of data for evaluating products, mainly from vendors, manufacturers, entrepreneurs, and similar domains. The messages in these reviews can range from general opinions to individual thoughts [4].

Additionally, movie and item reviews, as well as discussions on religious and political issues, are easily accessible and serve as valuable sources of sentiment.

2. Related Work

Sentiment analysis involves carefully examining how emotions and opinions expressed in natural language can be associated with one's feelings and attitudes towards an event. Twitter's profile information is chosen as the primary data source because it provides subjective data [5]. Recent developments demonstrate the significant success of sentiment analysis, surpassing simple positive versus negative classification and expanding to encompass the entire spectrum of behaviors and emotions across different networks and topics. Numerous research studies have been conducted in the field of sentiment analysis using various techniques.

Pang and Lee (2002) proposed a framework that determined the sentiment as positive or negative based on the ratio of positive words to total words in a text [6]. In 2008, the authors developed an approach that selected tweet outcomes based on specific terms within the tweets. Go et al. [7] conducted another study on sentiment analysis of Twitter, treating it as a two-class classification problem to categorize tweets as positive or negative.

M. Trupthi, S. Pabboju, and G. Narasimha proposed a system that utilized Hadoop and Twitter's streaming API to extract data. The extracted tweets were loaded into Hadoop and preprocessed using map-reduce functions. They employed a uni-word naive Bayes classification approach [8]. Another paper [9] focused on the application of sentiment analysis in business contexts. It demonstrated the text analysis process for evaluating customer sentiment towards a specific brand and highlighted the hidden information that can be utilized for decision-making after performing the analysis.

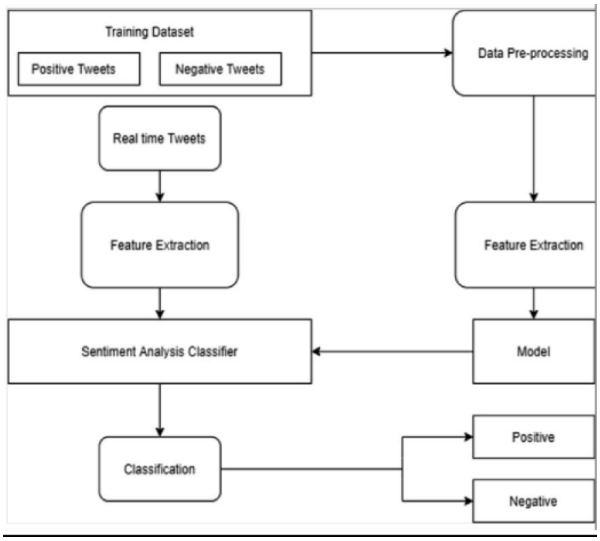
In another paper [10], sentiment analysis was conducted in four phases. Real-time tweets were collected up to a given limit, followed by tokenization as part of the preprocessing step. The tweets were then compared with a predefined bag of words, and finally, they were classified as positive or negative. The proposed system in this paper is domain-specific and includes a user interactive GUI where users can enter keywords related to commercial products. This specificity distinguishes the system from existing ones and reduces processing time by focusing only on tweets relevant to specific products based on the entered keywords. Additionally, the paper aims to

compare various machine learning algorithms and select the one that achieves the highest accuracy in producing results.

3. Proposed System

The system aims to conduct sentiment analysis on tweets gathered from the Twitter dataset. Several algorithms have been employed and tested on the dataset, and the most suitable algorithm has been selected. The process of sentiment analysis is outlined in Figure 1.

Once the dataset is cleaned and divided into



training and testing datasets, it undergoes preprocessing using the mentioned techniques. Feature extraction is then performed to reduce the dataset's dimensionality. A model is created, which is fed into a classifier to classify the tweets as positive or negative. Real-time tweets are also tested using this classifier.The system is specifically designed to perform sentiment analysis solely on tweets related to products in the market. It does not engage in sentiment analysis across all domains. Users are provided with an interactive GUI where they can enter keywords or sentences associated with a particular product. All tweets related to that product are displayed, along with the number of positive and negative statements made by others. This information helps users revise their production and work strategies, leading to improvements in their businesses. Outline of proposed system:

Below are steps involved in handling large incoming data:

1. Data cleaning:

i. Use of various data tools that can help in cleaning the dataset.

ii. Use of several AI tools that help in identifying duplicates in large corpora of data and eliminate it.For correcting the corrupted data, the source of errors should be tracked and monitored constantly

iii. Validate once, when the existing data is cleaned.

2. Data pre-processing:

i. Assessing data quality.

ii. Identification of inconsistent values to know what the data type of the features should be.

iii. Aggregate the features to give better performance

Data Cleaning

The information gathered was not in the proper format. Information cleaning is the process of ensuring that information is correct, predictable, and useable. Typically, datasets must be cleansed because they contain a large amount of noisy or undesired data known as outliers. The presence of such outliers may result in incorrect results. Data cleaning guarantees that such data is removed and improved, resulting in a far more dependable and stable dataset. There are several methods for cleansing data:Monitors errors. The entry point or source of errors should be tracked and monitored constantly. This will help in correcting the corrupted data.

- Process standardization. The point of entry should be standardized. By standardizing the data process, the risk of duplication reduces.

- Accuracy validation. Data should be validated once the existing database is

cleaned. Studying and using various data tools that can help in cleaning the datasets is very important.

identification of duplicate data is a very mandatory process. Several AI tools help in identifying duplicates in large corpora of data.

Data Pre-processing

The subsequent stage after data cleaning involves data pre- processing, which is a crucial step in machine learning. It entails transforming or encoding the data into a format that can be easily understood by the machine learning algorithms. In simpler terms, it involves preparing the dataset so that its features can be effectively interpreted by the algorithms. Features refer to measurable properties of the observed entity, such as height, age, or gender in the case of a person.

When extracting tweets from the Twitter stream, the collected data is typically unstructured. To prepare this unstructured data for classification, it needs to undergo pre-processing, which involves tokenization and cleaning. Initially, any HTML content present in the tweets is removed by identifying URL structures. Subsequently, the cleaning process involves eliminating non-letter characters or images using regular expressions in Python
In this pre-processing step, efforts are made to filter out slang words and correct misspellings before extracting the features. This helps ensure the quality of the data before further analysis. The following steps can be followed for data pre-processing.

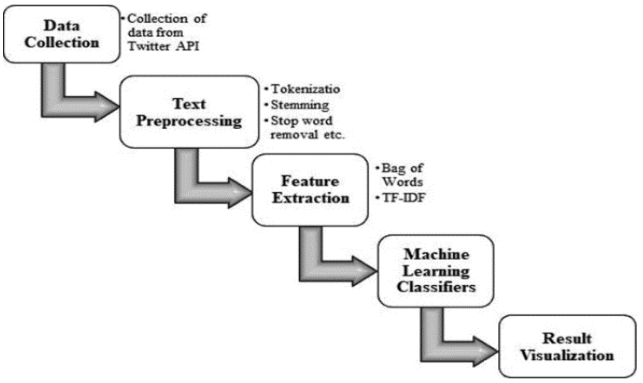
- **Data quality assessment**. Since the data is collected from multiple sources, it will be unrealistic to consider it to be perfect. Assessing the data quality must be the first step while pre-processing it.

- **Inconsistent values**. Data can be inconsistent at times. Like the "address" field can contain a phone number. Hence, the assessment should be done properly like to know what the data type of the features should be.

- **Feature aggregation**. As the name says, features are aggregated to give better performance. The behavior of aggregated features is much better when compared to individual data entities.

- **Feature sampling**. It is a way of selecting a subset of the original (first) dataset. The central matter of sampling is that the subset should have nearly the same properties as the original dataset.

Coming to the proposed system, the pre-



processing done will be as follows:

- Converting tweets to lowercases.
- Supplant at least two dots with spaces.
- Replace extra spaces with a single one.
- Remove spaces and quotes at the end of tweets. The dataset is used to extract two types of features: unigrams and bigrams. For the retrieved features, a frequency distribution is generated. The top N unigrams and bigrams are then picked to conduct the analysis. Tweets also include unique elements such as URLs, user names, emoticons, and so on. Tweets include retweets as well. These characteristics are not required for sentiment analysis. As a result, these features are replaced with common keywords or markers such as "URL," "USER_MENTION," and "EMO," as appropriate. Again, stop words must be removed and lemmatization is required.

Stop words. Stop words will be words that don't have any criticalness in search inquiries.

For example, “I like to write.” After removing stop words becomes, “like write.” “I” and “to” are termed as stop words.

Stemming. It is an element procedure of delivering morphological variations of a base word. The words like “chocolatey,” “chocolates” are converted to their root word “chocolate.”

Lemmatization. Lemmatization decreases the inflected words appropriately guaranteeing that the root word has a place with the language.

4. Results

The Sentiment140 dataset was utilized for training the model. It is a balanced dataset of 1.6 million tweets, of which 8 lakh tweets are favorable and the remaining 8 lakh tweets are negative.

The train_test_split method is used to split the data, with a test size of 0.20. The model is trained using 12 lakh tweets, and it is tested using the remaining 4 lakh tweets.

4.1. Logistic Regression

Table 1 gives the classification report of the logistic regression model. The accuracy of the model is 82.47. It also shows the precision and recall of the model. Precision is the positive predictive value, and recall is the sensitivity of the model [14].

Table 1

Classification report of logistic regression

	Precision	Recall	F1 Score	Support
0 (Negative label)	0.78090	0.85148	0.81466	160,156
1(Positive label)	0.83637	0.76064	0.79671	159,844
Accuracy			0.80610	320,000
Macro avg	0.80864	0.80606	0.80569	320,000
Weighted avg	0.80861	0.80910	0.80569	320,000

4.2. Multinomial Naive Bayes

Table 2 gives the classification report of multinomial naive Bayes model. The accuracy of the model is 80.61.

Table 2

Classification report of multinomial naive Bayes

	Precision	Recall	F1 Score	Support
0 (Negative label)	0.85970	0.90315	0.88889	160,011
1 (Positive label)	0.78447	0.70515	0.74270	79,989
Accuracy			0.83716	240,000
Macro avg	0.82288	0.88415	0.81179	240,000
Weighted avg	0.83462	0.83716	0.83483	240,000

4.3. Linear Support Vector Machine

Table 3 shows the classification report of the linear SVC model. The accuracy of the model is 83.71.

Table 3

Classification report of linear support vector machine

0 (N				
1 (Positive label)	0.83236	0.81280	0.82246	159,844
Accuracy			0.82472	320,000
Macro avg	0.82490	0.82471	0.82469	320,000
Weighted avg	0.82490	0.82472	0.82470	320,000

ROC curves are suitable when there is a balance between the observations in each class. Since the dataset used for training and testing in this case is balanced, ROC curves are used to measure the performance of the model. AUROC (Area Under the ROC Curve) is considered a superior measure of classifier performance compared to accuracy. This is because accuracy is biased by the size of the test or evaluation data, whereas AUROC takes into account various aspects of performance and provides a single comprehensive

measure. Figure 5 presents a comparison of the three algorithmsemployed for sentiment analysis. Among them, linear SVC achieves the highest accuracy of 83.71%. However, its AUROC is lower compared to logistic regression, which has an accuracy of 82.47%. Therefore, logistic regression is chosen for the purpose of classification, considering its AUROC and overall performance

5. Conclusion and Future Work

The objective of this research paper is to classify a large volume of Twitter data into two sentiment categories: positive and negative. By utilizing sentiment features instead of traditional text classification methods, a higher accuracy rate is achieved. These sentiment features can be valuable for different entities such as businesses, organizations, and entrepreneurs to assess their products and gain a better understanding of public opinions. In future work, the scope will be expanded beyond the English language to include regional languages. Additionally, the analysis will incorporate more complex emotions like sarcasm, and a hybrid classifier will be developed to enhance accuracy and performance.

References

1. Neethu MS, Rajasree R (2013) Sentiment analysis in twitter using machine learning techniques. In: 2013 Fourth international conference on computing, communications and networking technologies (ICCCNT), Tiruchengode, pp 1–5

2. Kumar A, Sebastian TM, Sentiment analysis on twitter. Department of Computer Engineering, Delhi Technological University Delhi,India

3. Joshi R, Tekchandani R (2016) Comparative analysis of Twitter data using supervised classifiers.In:2016 International conference on inventive computation technologies (ICICT). ISBN: 978-1-5090-1285-5

4. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R, Sentiment analysis of twitter. Passonneau Department of Computer Science Columbia University New York, NY 10027 USA

5. Hasan A, Moin S, Karim A, Shamshirband S, Machine learning-based sentiment analysis for twitter accounts. Department of Computer Science, Air University, Multan Campus, Multan 60000

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

7. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1, no 12

8. Trupthi M, Pabboju S, Narasimha G (2017) Sentiment analysis on twitter using streaming API. In: 2017 IEEE 7th international advance computing conference (IACC), Hyderabad, pp 915–919

9. Halibas AS, Shaffi AS, Mohamed MAKV (2018) Application of text classification and clustering of twitter data for business analytics. In: 2018 Majan international conference (MIC), ISBN: 978-1-5386-3761-6

10. Neethu MS, Rajasree R (2013) Sentiment analysis intwitter using machine learning techniques. In:2013 Fourth international conference on computing, communications and networking technologies (ICCCNT),

Tiruchengode, pp 1– 5

11. Shamantha RB, Shetty SM, Rai P (2019) Sentiment analysis using machine learning classifiers: evaluation of performance. In: 2019 IEEE 4th international conference on computer and communication systems (ICCCS), Singapore, pp 21–25

12. Tyagi P, Tripathi RC (2019) A review towards the sentiment analysis techniques for the analysis of twitter data. In: Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE)

13. Yadav N, Kudale O, Gupta S, Rao A, Shitole A (2020) Twitter sentiment analysis using machine learning for product evaluation. In: 5th International conference on inventive computation technologies (ICICT-2020)

14. Shitole A, Devare M (2018) Optimization of person prediction using sensor data analysis of IoT enabled physical location monitoring. J Adv Res Dyn Control Syst 10(9):2800–2812. ISSN: 1943-02