Twitter Sentiment Analysis using Machine Learning Algorithms: A Comparative Analysis

Rupam Singh[1], Narayan Kulshrestha[1], Aparajita Sinha[1], Monika Agarwal[1], Bishal Sinha[2]

Dayananda Sagar University, India[1]
National Institute of Technology, Mizoram[2]

Corresponding author: Aparajita Sinha, Email: aparajitasinha-aiml@dsu.edu.in

Data posted by people, or the users of a particular social network, has increased dramatically due to the changing behavior of various social networking sites like Instagram, Twitter, Snapchat, etc. Innumerable millions and billions of bytes of audio, video, and text are uploaded daily. This is because millions of people use a particular website. These folks are interested in sharing their thoughts and opinions on any topic they choose. People also want to know if most people will see an incident favorably, unfavorably, or neutrally. In this paper, the data is classified into Positive, Negative, or Neutral opinions, and it presents a detailed survey of Sentiment analysis of Twitter data using various Machine learning algorithms like Naïve Bayes, Support Vector Machine (SVM), Logistic regression, and decision tree. Additionally, the accuracy and F1 scores of the aforementioned algorithms are examined on two distinct Twitter datasets, and a comparison is made between the algorithms respective accuracies in the two datasets.

Keywords: Twitter, Machine learning, Twitter Sentiment Analysis, Naïve Bayes, SVM, Decision tree.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 135–144. DOI: https://doi.org/10.56155/978-81-955020-7-3-12

1 Introduction

The world is evolving very fast in the technology field. New and improved work is being done continuously in the technology field. People use various social media platforms to express themselves and their thoughts about anything in day-to-day life, be it the products they use or the services offered by various companies. The users express these thoughts and reviews without any bias, totally based on their experience. The topics people post about could be a product from an organization such as a laptop, a phone, a car, or something else. These expressions are used as data by the giants for understanding consumer behavior and the quality and thoughts on their products and services. Orit could be a famous personality or any other thing. Sentiment analysis can be used as a complement to other systems such as recommendation systems, information extraction and question-answering systems.

In order to study the methods and procedures used for text classification, this research paper surveyed multiple studies that looked at Twitter's data categorization and analysis for better understanding. An insight into sentiment analysis is given in this paper. This paper's main objective is to provide a concise introduction to sentiment analysis, as well as a look at the different problems that researchers have run into, as well as some current solutions. We have covered a variety of analysis-related steps in the subsequent section of this study. In Section II we discussed the literature survey. Section III describes the various data characteristics of Twitter, followed by Section IV which gives us a detailed overview of the methodology used. The further discussion contains Section V which is the Results and analysis part discussing the results obtained in the analysis. Section VI concludes all the desired results and analysis.

2 Literature Survey

The research community is actively evaluating the significant impact of Twitter applications on various companies today, with a particular focus on the consistent analysis of Twitter sentiment. One key challenge in this analysis lies in the intricate structure of the retrieved data and the diverse nature of speech.

In a study by Masoud et al. [1], data from two distinct datasets with different characteristics underwent analysis using four classification algorithms and ensemble techniques to enhance reliability. Surprisingly, the tests revealed that the use of a single algorithm slightly outperformed ensemble techniques. Additionally, the analysis concluded that employing 50% of the data as training data yielded results comparable to using 70% of the data for training.

Another investigation conducted by K. Arun et al. [2] centered on the analysis of Twitter data related to demonetization. Utilizing the R programming language, the study presented graphical plots with word clouds based on the tweet analysis. The plotted results led to the conclusion that a considerable majority of individuals expressed acceptance of demonetization compared to those who rejected it.

In the realm of Twitter data studies, Aliza Sarlan et al. [3] took a straightforward approach by extracting tweets in JSON format and determining tweet polarity using the Python Lexicon Dictionary. On the contrary, Mandava Geeta, Bhargavav, and Duvvada [4] adopted a more sophisticated strategy, leveraging learning approaches to enhance accuracy. Focusing on cryptocurrency data, they applied SVM (Support Vector Machine) and Naïve Bayes algorithms, revealing that the Naïve Bayes classifier outperformed SVM in accuracy.

In a distinct investigation, Agarwal et al. [5] utilized the unigram model as a baseline, comparing it with experimental models based on features and kernel trees. The results highlighted the superior performance of the kernel tree-oriented model over both unigram and feature-oriented models, while the feature-oriented model exhibited a slight edge over the unigram model.

An unconventional approach was taken by Akshi Kumar and Teeja Sebastian [6], who combined a corpus-based approach with a lexicon-based one, a rarity in the predominantly machine learning-focused research landscape.

Kaur et al. [9] delved into public opinion on geographical flood data collected from Twitter, employing the Naïve Bayes algorithm to achieve a 67% accuracy rate. Kaur emphasized the importance of gathering diverse measures from the public to enhance situational management.

A paper [10] focused on analyzing feedback encompassing interjections, emotional expressions, and opinions extracted from various posts, retweets, and updates. The research aimed to ascertain the utility of sentiment analysis for both customers and online businesses, exploring the demand and impact of this analytical approach.

Researchers Patodkar and Imran R. Shaikh, along with Vaibhavi N [7], aimed to discern the emotional responses of viewers to a random television program. Collecting remarks from a selection of diverse TV broadcasts, these comments served as data for training and testing a Naive Bayes classifier model. The output, presented as a pie chart, revealed a prevalence of negative tweets over positive ones in terms of polarity.

In 2010, Tumasjan et al.'s study [8] delved into the potential of Twitter data in predicting elections. Analyzing political sentiment expressed within Twitter's 140-character limit, the research explored the correlation between Twitter activity and election outcomes. This investigation shed light on the use of social media as a predictive tool for political events.

The paper authored by Devika, Sunitha, and Ganesh in 2016 [17] concentrated on sentiment analysis, a vital facet of natural language processing that involves extracting and understanding opinions expressed in text. Providing a comparative analysis of various sentiment analysis approaches, the study offered insights into their strengths, weaknesses, and application domains.

Shahare et al. [11] harnessed social media and news data for sentiment analysis, utilizing Naïve Bayes and Levenshtein methods to categorize sentiments from diverse sources. This approach not only enhanced lead-to-term accuracy but also demonstrated robust performance in real-time news content on social media. Notably, the Levenshtein formula emerged as a swift and efficient means of processing a substantial amount of information with high accuracy levels.

3 Data Characteristics

These days, there are many social networking sites available, like Facebook, Instagram, Twitter, and others. The primary focus of this research is Twitter, a micro blogging platform. These days, there are far too many social networking sites to list; nevertheless, for this paper, we will just be discussing Twitter. Twitter's unique writing format has made it extremely popular these days.

Below are a few of the qualities of tweets:

- Tweet length: The tweets are short messages or reviews written in140 words.
- Tweet availability: With over 1.2 billion tweets produced every day, Twitter has grown to be one of the most popular social networking platforms in use today.
- Topics discussed: A variety of individuals tweet on a range of subjects, from politics to just about anything that comes to mind.

- Writing form or methods: Tweets have a very hazy, unpolished style with plenty of typos, frequent utilization of slang, and frequent usage of smileys. Thus, preprocessing of these tweets is necessary for this.

- Real-time: Tweets are updated often due to their brief nature.

- Emoticons: These are written representations of a user's facial emotions. These emoticons are created by combining punctuation with other characters.

- Mentions: The "@" character is used to emphasize a person and mention them in a message.

- Hashtags: When tweeting about a topic, the '#' character is utilized to highlight that topic.

4 Methodology

The methodology process consists of a series of steps which are displayed in the form of flowchart in Figure 1.
The four stages of the method for sentiment analysis used in this research paper are as follows-

1. Data collection
2. Data Preprocessing
3. Model Selection,
4. Model Evaluation

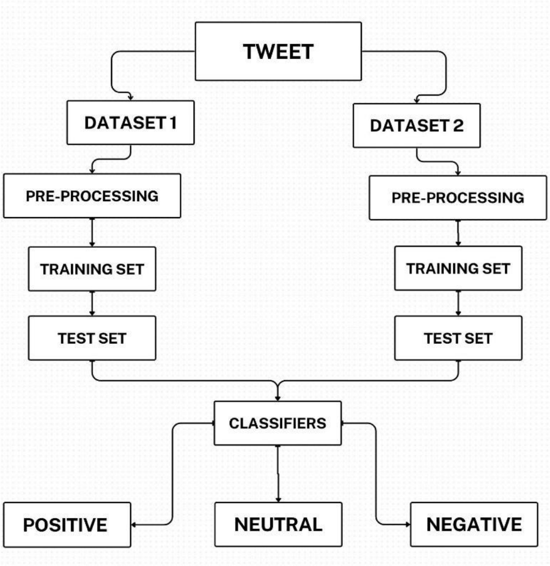

Figure1. Flowchart of Methodology of Twitter Sentiment Analysis

4.1 Data Collection

The initial step in the analysis of data is the collection of data. We have employed the Python programming language as a tool in our experiments. We obtained two different datasets from an online re-source named Kaggle, one of the largest online data science communities in this world. The first dataset [12] consists of around 1.62 lakh tweets labeled as positive, negative, and neutral. The second dataset [13] consists of around 69 thousand tweets labeled as positive, negative, and neutral. Datasets are represented in below Table 1.

Table 1. Data Distribution Table

	Dataset 1	Dataset 2
Number of Tweets	1,62,980	69,497
Positive Tweets	23,365	20,831
Negative Tweets	19,802	18,318
Neutral Tweets	12,685	22,342

4.2 Data Preprocessing

Pre-processing of data is the process of transforming raw data into a form that can be used by a classifier to increase its accuracy. Several procedures are used for data preparation, including:

1. Utilising the dataset to create training and testing sets. Eighty percent of the training dataset and twenty percent of the testing dataset make up the two datasets used in the study.

2. Removing Twitter handles, URLs, hashtags, and special characters like emojis, and then the text is converted into lowercase for better generalization.

3. Diffuser Eliminate stop words like "is," "the," "at," and so forth.

4. The tokenization process divides the text into a set of tokens. Tokenization is the act of dividing a text or sentence into discrete units.

5. The Porter Stemming approach is used to first reduce the words to their root form. Stemming is a rule-based method of removing suffixes; usually, it eliminates the "s," "es," "ing," and so on from a given sentence phrase

4.3 Model Selection

The pre-processed data must then be given to a classification model for additional processing. These models were created using 4 different categorization methods.

4.3.1 Support Vector Machine

The described classification approach operates on a non-probabilistic basis and demands an extensive amount of training data. Utilizing a hyperplane with dimensions of (d-1), the classification of particles is achieved. The Support Vector Machine (SVM) is employed to pinpoint the hyperplane with the most considerable margin. Central to the SVM concept are decision planes that establish limits for decision-making. Different class memberships within an object group are discerned by a chosen plane [14].

Functioning as a machine-learning technique, SVM effectively categorizes data into multiple classes, with distinctions here being positive, neutral, and negative. The primary objective of the SVM technique is to determine the optimal line, or decision boundary, segregating data points into various classifications. In high-dimensional feature spaces, this boundary is represented as a hyperplane. The term "mapping" or "transformation" is employed because a mathematical function, known as a kernel, is utilized to map or reconstruct the original elements. Post-transformation, the mapped items retain linear distinctness. This process is particularly crucial as it enables the use of curves to delineate boundaries, facilitating the effective handling and classification of intricate arrangements while avoiding complexity.

4.3.2 Logistic Regression

Logistic regression, a supervised machine learning technique, is widely applied to classification issues where the goal is to predict whether a given instance will fall into a specific class or not [15]. This kind of statistical method looks at the relationship between a set of binary dependent variables and a collection of independent variables. It works well as a tool for decision-making.

4.3.3 Decision Tree

We utilize the Decision Tree classifier from the sci-kit-learn sklearn. tree module to build our model. Decision Tree [16] is a supervised learning technique that may be used for both regression and classification problems, albeit it performs best for classification-related tasks. It is a tree-structured classifier where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result.

The Decision Node and Leaf Node are the two nodes of a decision tree.

4.3.4 Naïve Bayes

It is primarily utilized if the training dataset is smaller and is a probabilistic classifier. It belongs to the family of sample probabilistic classifiers in machine learning that depends on the Bayes theorem. The Bayes rule calculates the conditional probability that an event A occurs given evidence B using the formula-:

$$P(A/B) = P(A)\, P(B/A) / P(B) \qquad (1)$$

The Naive Bayes Classifier systematically considers each feature within the feature vector independently, as each feature in the vector is presumed to be equally independent of the others. Specifically designed for classification tasks, such as text classification, the Naive Bayes classifier stands out as a supervised machine learning algorithm. Furthermore, it falls within the generative learning algorithm family, aiming to replicate the input distribution of a specified class or category.

For practical implementation, the Python NLTK package serves as a valuable resource for training and classifying using the Naive Bayes Machine Learning approach. In the context of Naive Bayes classification, the MultinomialNB module from the sci-kit-learn sklearn.naive_bayes package is employed, providing a robust tool for accurate and efficient classification tasks.

4.3.5 Model Evaluation

In this section of our inquiry, we import all the necessary Python packages, including Seaborn, SKlearn, NLTK, Numpy, Pandas, and others, in order to analyze the text reviews. Furthermore, we calculated the precision and recall values by evaluating the model with TF-IDF.

TF-IDF: The occurrence rate of a word in a specific document is commonly known as its frequency. To assess a term's importance within a given document, the term "Index Document Frequency," abbreviated as IDF, is utilized. IDF gauges the significance of each word in a text relative to the entire corpus, aiding in the identification of key phrases conveying neutral, positive, or negative sentiments.

In the context of sentiment analysis, TF-IDF proves valuable as it effectively manages vast amounts of text data. It excels in recognizing words and phrases within a text, assigning greater weight to distinctive expressions, thereby enhancing its ability to discern and analyze sentiments.

$$\text{TF=} \quad \frac{Numberoftimesaword^{F}x^{F}appearsinadocument}{Numberofwordspresentinadocument} \tag{2}$$

$$\text{IDF=log} \ \{ \ \frac{Numberofdocumentspresentinacorpus}{Numberof\ documentswhere\ word\ \ Xappeared} \ \} \tag{3}$$

$$\text{TF-IDF=}TF*IDF$$

Precision: precision demonstrates how frequently the classifier's anticipated outcome is accurate when it signals true. The recipe for accuracy is:

$$Precision = \frac{True\ Positive}{True\ Positive+Fal\quad Negative} \tag{4}$$

Recall: It displays the classifier's genuine positive rate. The formula for the recall is-:

$$\text{Recall=} \frac{True\ Positive}{True\ Positive+Fa\quad Negative} \tag{5}$$

5 Results and Analysis

In this section, we start analyzing the prediction analysis through different algorithms we have used in our analysis. The accuracies of both the datasets and the respective F1 scores are shown below in Table 2 and Table 3.

Table 2. Accuracy table for Dataset 1 and Dataset 2

Classification Algorithm	Dataset 1	Dataset 2
Naïve Bayes	70%	83%
SVM	80.02%	83%
Logistic Regression	80%	83%
Random Forest	77%	96%

Table 3. F1 Score table for Dataset 1 and Dataset 2

Classification Algorithm	Dataset 1 (F1 Score)	Dataset 2 (F1 Score)
Naïve Bayes	75%	82%
SVM	84%	83%
Logistic Regression	82%	82%
Random Forest	81%	95%

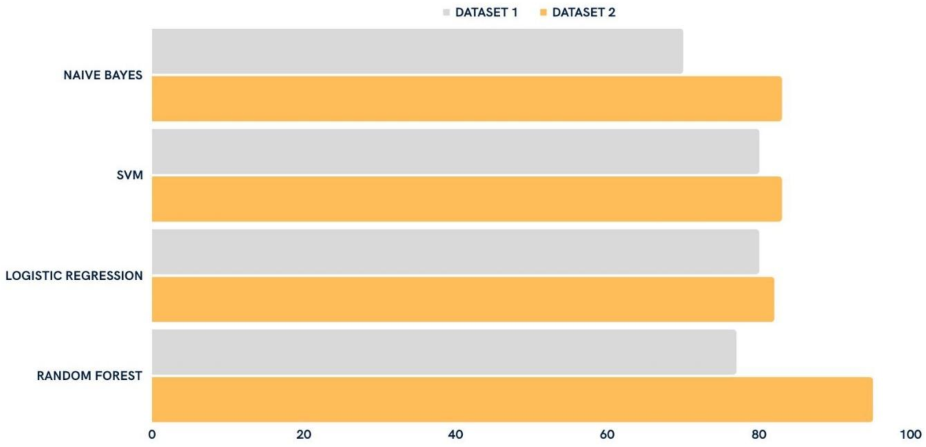

Figure 2. Accuracy score graph for dataset 1 and dataset 2

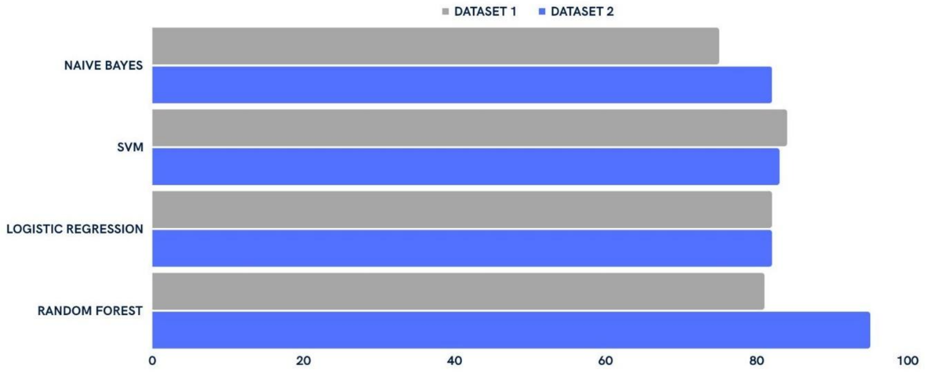

Figure 3. F1 Score graph for dataset 1 and dataset 2

5.1 Naive Bayes and SVM

From Table 2 we can analyze that SVM showed better performance with the large dataset (dataset 1) in comparison with Naïve Bayes, and the difference is so apparent. While in the small dataset (dataset 2) both SVM and Naïve Bayes have almost similar accuracy. The maximum accuracy in the larger dataset (dataset 1) for Naïve Bayes is 70 % while it reached to 83% in the smaller dataset (dataset 2), while, the maximum accuracy for SVM in dataset 1 is 80.02% which increased to 83% in dataset 2.

In larger dataset SVM algorithm performed better with an accuracy of 80.02% whereas in smaller dataset both the algorithms performed similar having an ac curacy of 83%.

From Table 3 we can analyze that the F1 score for the Naïve Bayes algorithm for the large dataset (dataset 1) is 75% which reached 82% in the smaller dataset (dataset 2) while the F1 score of the SVM algorithm is almost similar in both datasets, unlike its accuracy.

5.2 Logistic Regression and Random Forest

From Table 2 we can analyze that Logistic regression performed better with the larger dataset (dataset 1) in comparison with the Random forest classifier, While in the smaller dataset (dataset 2) Random forest classifier performed better than logistic regression. The accuracy for Logistic regression with the larger dataset(dataset 1) is 80% while it reached 83% in the smaller dataset (dataset 2) while the accuracy for the Random forest classifier with the larger dataset (dataset 1)comes out to be 77% which reached to 96% in the smaller dataset (dataset 2).

In the larger dataset Logistic algorithm performed better with an accuracy of 80% whereas in the smaller dataset Random forest algorithm performed better with 96% accuracy.

From our analysis, we got an F1 score(Table 3) for the Logistic regression algorithm for both dataset 1 and dataset 2 to be the same. While it differed for the Random Forest algorithm for both datasets. For dataset 1 it came out to be 81% while it increased drastically to 95% in dataset 2.

Figure 2 depicts the graph of the accuracy of Dataset 1 and Dataset 2. From the graph, it is clearly evident that SVM performed better in comparison with all other classifiers used in dataset 1, while the Random forest classifier performed better in comparison with all other classifiers used in dataset 2.
Figure 3 represents the graph of the F1 score for dataset 1 and dataset 2 respectively.

6 Conclusion and Future Work

This paper gives an outline of recent studies on the classification and analysis of sentiment. This study uses a variety of classification algorithms, including decision trees, Naive Bayes, SVM, and logistic regression. When compared to other algorithms, the experimental research leads us to the conclusion that the random forest approach provides good accuracy of 96%. The experimental research demonstrates that the support vector machine provides the best accuracy of 80.02% for the huge dataset (dataset 1). The experimental research shows that the random forest classification method provides the best accuracy of 96% for the tiny dataset (dataset 2).

In conclusion, the Random Forest classifier gave a better accuracy of 96% than other classifiers, and the performance was better with the small dataset(dataset 2).

Due to the abundance of data sources and practical applications, sentiment analysis is a subject that is growing. It is being subjected to an increasing number of demands. It was initially sufficient to determine if the document was primarily positive or negative.

More challenging, polarity at the level of a particular component of an item or service subsequently became required. Subsequent research endeavors will focus on contrasting sentiment analysis methodologies utilizing Deep Learning algorithms.

References

[1] AminiMotlagh M, Shahhoseini H, Fatehi N. A reliable sentiment analysis for classification of tweets in social networks. Soc Netw Anal Min. 2023;13(1) 7. doi:10.1007/s13278-022-00998-2. PMID: 36532862; PMCID: PMC9742011.

[2] "Arun k, Sinagesh a & Ramesh m, twitter sentiment analysis on demonetization tweets in India using r language", international journal of computer engineering in research trends,vol.4, no.6, (2017), pp.252-258"

[3] "Aliza sarlan, chayanit nadam, shuib basri ," twitter sentiment analysis 2014 in- ternational conference on information technology and multimedia (ICIMU), no- vember 18 – 20, 2014, putrajaya, malaysia 978-1-4799-5423-0/14/$31.00 ©2014 ieee 212"

[4] "Mandava geetha bhargava , duvvada rajeswara rao ," sentiment analysis on social media data using R" , international journal of engineering & technology, 7 (2.31) (2018) 80-84"

[5] "Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. "Sentimen- tanalysis of twitter data." InProceedings of the ACL 2011,Workshop on Lan- guages in Social Media, pp. 30–38, 2011.

[6] "IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012 ISSN (Online): 1694-0814"

[7] "International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 6, Issue "7 July 2017 "

[8] "Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I."Predictingelectionswith twitter: What 140 characters reveal about political sentiment.." In Proceedings of theFourth International AAAI Conference onWeblogs and Social Media (2010), pp. 178–185, 2010".

[9] Kaur, H. J. (2015). Sentiment Analysis from Social Media in Crisis Situations. IEEE, (pp. 251-256)

[10] Mittal, S., Goel, A., & Jain, R. (2016). Sentiment analysis of E-commerce and social networking sites. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2300-2305.

[11] Shahare, F. F. (2017). Sentiment Analysis for the News Data Based on the social Media. IEEE, pp. 1365-1370.

[12] https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset.

[13] https://www.kaggle.com/datasets/jp797498e/twitter-entitysentimentanalysis?re-source=download&select =twitter_validation.csv

[14] https://www.techtarget.com/whatis/definition/support-vector-machine- SVM

[15] https://www.geeksforgeeks.org/understanding- logistic-regression/

[16] https://www.javatpoint.com/machine-learning-decision-tree-classifi- cation-algorithm

[17] Devika, M.D. & C, Sunitha & Ganesh, Amal. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science. 87. 44-49. 10.1016/j.procs.2016.05.124.