# COMPARATIVE STUDY ON TWITTER SENTIMENT ANALYSIS

**MANOJ PERAVALI[1], K. SHASHIKANTH[2], M. NAVADEEP REDDY[3], M. SAI ROHIT[4], AND DR. SUBHRANGINEE DAS[5]**

[1] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India*

*Compiled April 23, 2024*

**Sentiment Analysis also known as Opinion Mining refers to the use of natural language processing, and text analysis to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to reviews on survey responses, online social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. In this paper, we aim to perform a Sentiment Analysis of product-based reviews. Data used in this project are customer reviews collected from "twitter.com". We expect to do review-level categorization of review data with promising outcomes.**

http://dx.doi.org/10.1364/ao.XX.XXXXXX

## 1. INTRODUCTION

Sentiment, defined as a thought, attitude, or judgment prompted by feeling, holds significant value in business intelligence, aiding companies in enhancing their services and products. Various online platforms, including social media, forums, and micro-blogs, serve as avenues for individuals to express their sentiments, generating vast amounts of data available for analysis. However, sentiment analysis encounters challenges, primarily stemming from the subjective nature of opinions shared online and the absence of a reliable ground truth for validation. For instance, a movie review's complexity lies in deciphering nuanced expressions, such as describing a film as "quite humdrum" while acknowledging positive aspects. Sentiment analysis entails extracting evaluative terms, determining polarity and opinion strength, and categorizing reviews into sentiment classes based on the sentiments expressed. This multistep process aims to provide insights into the overall sentiment orientation of the opinions contained within textual data.

## 2. LITERATURE REVIEW

| Title | Author | Description |
|---|---|---|
| Sentiment Analysis Using Twitter Data | Qi, Y.Shabrina, Year = 2023 | Explored sentiment analysis methods, comparing lexicon-based and machine-learning-based approaches. Highlighted the importance of analyzing social media content for understanding emotions and perceptions |
| Comprehensive Review of Twitter Sentiment Analysis | Wang, Y., Guo, J., Yuan, C., Li, B. Year:2022 | Categorized recent articles based on different TSA methods. Discussed lexicon-based and machine learning approaches for extracting opinions and sentiments from Twitter data. |
| Sentiment Analysis of Twitter Data for Financial Market Prediction | Ostrovskiy, Stanislav et al. Year:2018 | This paper explores the use of sentiment analysis on Twitter data for financial market prediction. The authors propose methods to extract relevant financial tweets and analyze their sentiment. They investigate the correlation between public sentiment on Twitter and stock market movements, demonstrating the potential of sentiment analysis for financial applications. |
| Sentiment Analysis on Social Media | Fredrico Ner, Carlo Alparindi, Monstreerat Cuadros Year:2012 | This paper describes a Sentiment Analysis study performed on over than 1000 Facebook posts about newscasts, comparing the sentiment for Rai - the Italian public broadcasting service-towards the emerging and more dynamic private company La7 |
| Sentiment Analysis in Social Media and Its Application: Systematic Literature Review | Zulfadzli Drus. Haliyana Khalid Year:2019 | They conducted a systematic literature review that provides information on studies on sentiment analysis in social media. It shows what is the method used in analyzing sentiment in social media, most common type of social media site to extract Information and the application of sentiment analysis in social media. |
| Sentiment Analysis on Social Media Data Using Intelligent Techniques | Kassinda Francisco Martins Panguila, Dr. Chandra J Year:2019 | Sentiment analysis using intelligent techniques approach was proposed to deal with social media data. According to their experiment results Neural Networks methods such as Multi-layer Perceptron (MLP) and Convolutional Neural Networks (CNN) performed better than other classifiers in general. |

**Table 1.** Literature Review on few papers.

## 3. ARCHITECTURAL COMPONENTS

### A. LOGISTIC REGRESSION

Logistic regression is used for understanding the meaning of words in sentences because it's good at predicting things. It's like a helpful friend who can tell if something is likely to happen or not. People use logistic regression because it's straightforward and works well with lots of data. It's especially handy for figuring out probabilities, like whether a sentence is positive or negative. Logistic regression helps make sense of words by guessing the chances of different meanings, making it easier to understand what's being said in a bunch of writing. In terms of odds, logistic regression can also be stated as follows:

The formula for odds(y=1 | x) is P(y=1 | x) $\frac{}{1-P(y=1|x)=e^z}$.

By taking both sides' natural logarithms (log), we obtain:

The expression

$$\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \ln(e^z)]$$

The formula is

$$\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = z]$$

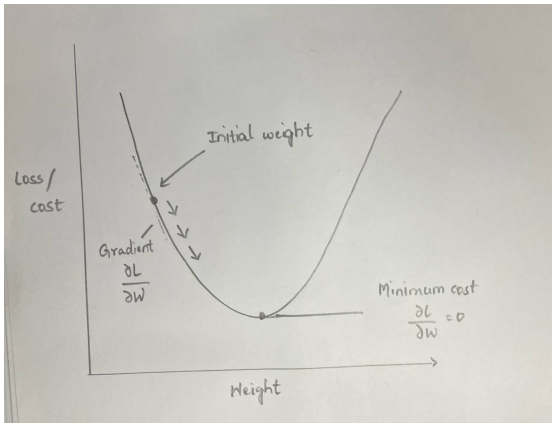
where $z$ is the previously defined linear combination of the input features and their coefficients. Since logistic regression coefficients indicate the change in the log-odds of the positive class for a one-unit change in the corresponding input feature, they are frequently interpreted using this formulation.

### B. NAIVES BAYES

Naive Bayes is widely used in semantic analysis because of its ease of use, sharpness, and efficiency while processing textual input. Understanding the meaning included in textual information, such as emails, documents, and posts on social media, is the main objective of semantic analysis. Because Naive Bayes classifiers can effectively classify text based on the likelihood that a given document belongs to a specific class or category, they are especially well-suited for this purpose. Naive Bayes has several advantages for semantic analysis, one of which is the simplicity of handling massive amounts of textual input. Naive Bayes classifiers are also appropriate for analyzing unstructured text data, which frequently contains noise and irrelevant information because they can handle irrelevant features gracefully and are robust to noisy data.

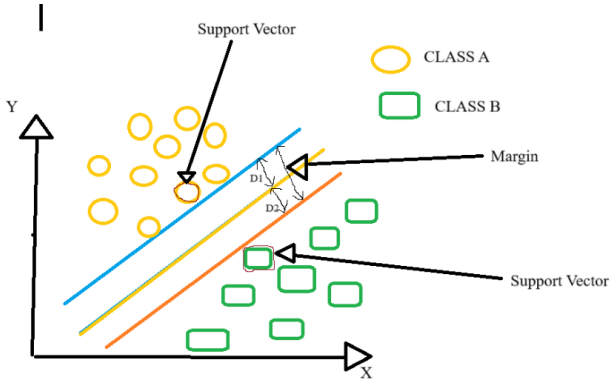
### C. STOCHASTIC GRADIENT DESCENT

A stochastic Gradient Descent (SGD) classifier is used for understanding the meaning of words in sentences because it's fast and efficient. It's like a speedy worker who quickly learns from examples to figure out what words mean. People like using SGD because it can handle large amounts of text data without taking too much time. It's especially good at learning from lots of examples and making decisions based on them. SGD classifier helps make sense of words by quickly adjusting its guesses to match what's being said in a bunch of writing, making it easier to understand the meaning behind the words.




### D. SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) are used for understanding what words mean in sentences because they're good at finding patterns in data. They're like detectives looking for clues to figure out the meaning behind text. People like using SVM because they're reliable and can handle big amounts of text easily. They're especially good at separating different types of text, like figuring out if something is positive or negative. SVM helps make sense of words by finding the best way to group them together, helping us understand the meaning in a bunch of writing.




## 4. DATASET AND EVALUATION

This substantially introduces the evaluation styles of open-source datasets and generated rulings in this field. Data, computational power, and algorithms are the three major rudiments of the current development of artificial intelligence. They complement and enhance each other. It is said to be a good dataset that can make the algorithm or model more effective. The image description task is analogous to machine restatement, and its evaluation system extends from machine restatement to form its unique evaluation criteria.

### A. Data Collections

The Sentiment140 Dataset. it is a big collection of tweets—those short messages people share on Twitter. We gathered these tweets from a dataset Seniment140 and focused on specific keywords related to emotions like sadness or distress. The dataset contains 3085 positive tweets (like little rays of sunshine) and 2310 negative tweets (conveying frustration or other not-so-happy feelings). Each tweet got a special label: “0” for positive and “1” for negative. Researchers and data scientists use this information to teach computers to read emotions based on words. It’s like decoding feelings in text form! We found this valuable dataset on Kaggle, a platform where data enthusiasts share their discoveries.

### B. Data Preprocessing

Data preprocessing cleans and organizes data for better computer understanding. Steps in Preprocessing: Remove Retweets: Think of retweets as echoes—sometimes we hear the same thing twice. We want unique tweets, so we remove these echoes. Lowercase Everything: Imagine if “Happy” and “Happy” were two different things. To avoid confusion, we convert everything to lowercase. Get Rid of Useless Words: Some words (like “and,” “or,” etc.) don’t add much meaning. We remove them out. Goodbye Usernames and URLs: Usernames and web links don’t help us understand feelings. We replace them with generic tags or just remove them. Stemming: This is like trimming words to their roots. For example, “running” becomes “run.” It helps reduce word variations. No Special Characters or Digits: Symbols and numbers don’t express emotions. We kick them out. Create a Word Dictionary: We make a list of important words and remove the rest. Understand Slang and Abbreviations: If someone says “LOL,” we know it means laughter. We expand these shortcuts. Fix Spelling Mistakes: Imagine if “hppy” meant “happy.” We correct such typos. Tagging Words: We label each word as a noun, verb, etc. It’s like sorting ingredients into categories.

### C. Feature Extraction

Some of the machine learning algorithms were created using statistical methods. Natural Language Processing was utilized to extract features from the Twitter data that will be used in Machine Learning. Extraction of the feature was achieved by removing the punctuation mark, word tokenizing, removing the words to avoid, marking sections, and building frequency distribution. All features extracted were randomized to remove all biases. Feature extraction has then assigned vector values to the preprocessed data and these have been used as a training dataset and testing dataset. A set of words has been given by the Philippine Psychology Association to determine if the posted sentiments on Twitter are posing for possible mental health crisis. Previous posts have been crawled also to further check if the current post shows a mental health crisis

.

## 5. RESULTS

The graphical representation of performance scores of the four machine learning models. The dataset is split into a training set of 80 percent tweets and a testing set of 20 percent tweets to analyze the precision, recall, f1 score, and accuracy of different classification models.
Based on the results, Logistic Regression performs better than the other models in the detection of potential mental health crisis tweets. shows the precision, recall, f1 score, and accuracy of the four machine learning techniques. Naïve Bayes Classifier has the highest precision compared to the other models.
Support Vector Classifier has the the lowest accuracy of 69 percent only whereas Stochastic Gradient Descent, Naive Bayes, and Logistic Regression obtained an accuracy of 75.5 percent, 75 percent, and 77 percent respectively.
Also, Logistic Regression has the highest recall and f1 score among the four algorithms. The performance scores indicate that the sentiment analysis model performed well in classifying positive and negative tweets.
The best-fitted machine learning algorithm was logistic regression since it has the highest accuracy, recall, and f1 score compared to the other model.

### A. Comparison of Drawbacks Used:

| Machine Learning Models | Precession | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 76.79 | 79.7 | 78.22 | 77.8 |
| Naive Bayes | 77.06 | 72.84 | 74.89 | 75 |
| Stochastic Gradient Descent | 72.68 | 82 | 77.06 | 75.5 |
| Support Vector Machine | 72.13 | 75.86 | 73.95 | 69 |

**Table 2.** Comparison of machine learning models with Twitter dataset.

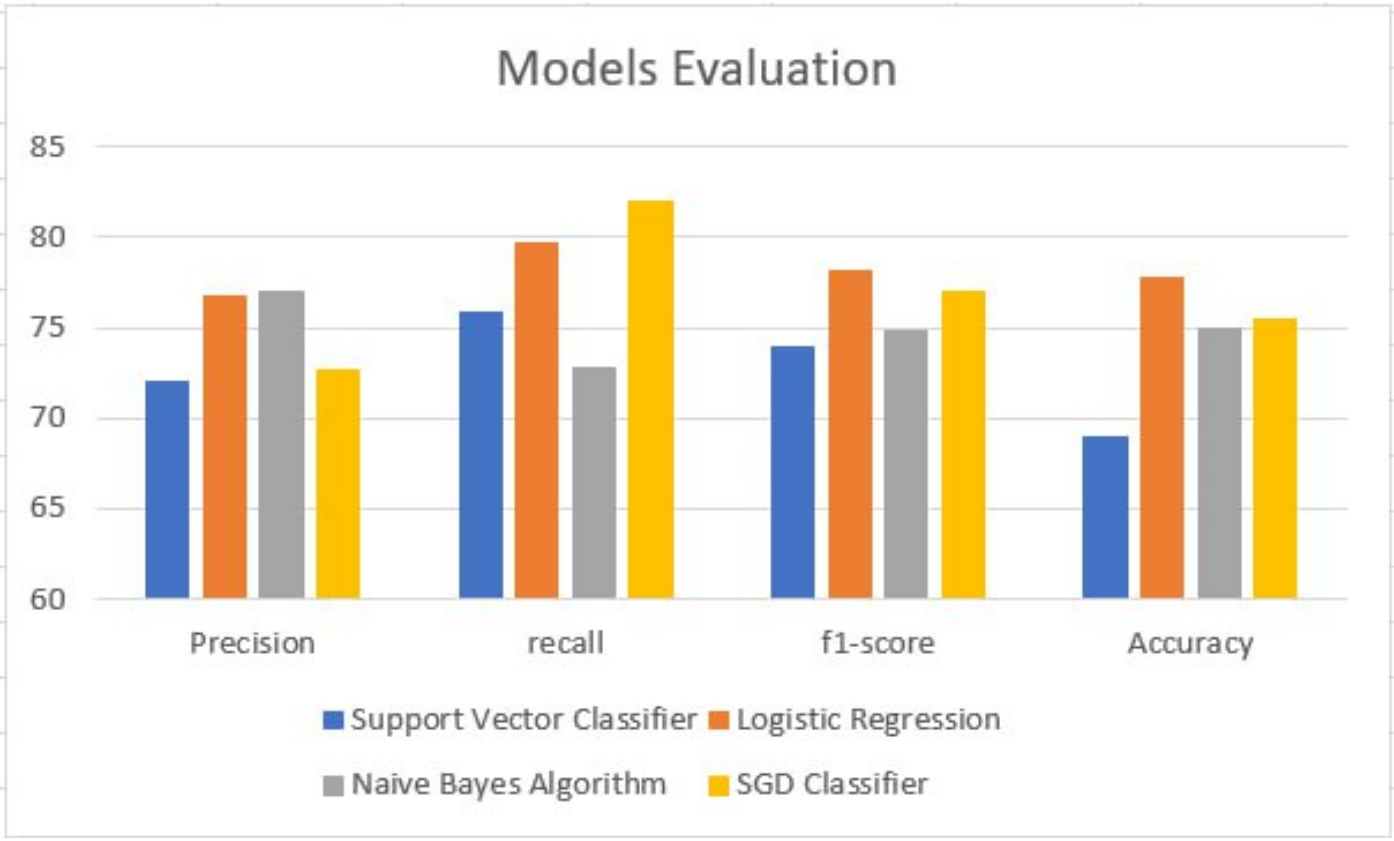



**Fig. 1.** The evaluation of Machine Learning models

## 6. CONCLUSION

The data pre-processing handles in the presented Twitter sentiment analysis include converting all text to lowercase and eliminating stopwords, special characters, recurring characters, and URLs. After tokenizing the text, lemmatization and stemming are applied. After choosing a subset of the data for analysis, the dataset is further examined by counting the number of unique values in the target variable, dividing the tweets into positive and negative ones, and then merging them. Next, a list of English stopwords is defined and the text is changed to lowercase. After transforming the data with the TF-IDF vectorizer, three distinct models are trained using the dataset. The ROC-AUC curve, classification report, and confusion matrix are some of the evaluation metrics that were employed.

While several metrics like BLEU, METEOR, ROUGE-L, and SPICE provide valuable insights, CIDEr (Consensus-based Image Description Evaluation) is often considered a good metric for image captioning on datasets like Flickr8k. CIDEr takes into account the consensus among multiple reference captions, providing a measure of overall quality and diversity. It is particularly suitable for assessing the ability of a model to generate diverse, contextually relevant captions that align with human judgments. In conclusion, a CNN-LSTM model with a Visual Attention Mechanism, evaluated using the CIDEr metric, is a strong combination for image caption generation on the Flickr8k dataset. It's essential to experiment, fine-tune hyperparameters, and perform thorough validation to optimize the performance of your chosen model on your specific task.

## 7. FUTURE SCOPE

Technological developments in natural language processing (NLP) and machine learning promise great things for Twitter sentiment analysis going forward. Sentiment analysis models are positioned to advance in accuracy and nuance as NLP techniques continue to develop, enabling them to comprehend the minute details and context that are present in tweets. It is anticipated that the use of multimodal analysis—which includes text, photos, emojis, and videos—will enhance sentiment analysis's depth and provide a more thorough comprehension of user sentiment. Moreover, domain-specific models designed for the financial, political, and medical domains will improve the accuracy and application of sentiment analysis. Businesses and organizations will be able to use real-time analysis capabilities to rapidly derive insights from Twitter data for market estimation, crisis management, and brand monitoring.

## REFERENCES


1. P. J. M. Loresco, I. C. Valenzuela and E. P. Dadios, "Color space analysis using KNN for lettuce crop stages identification in smart farm setup," TENCON 2018-2018 IEEE Region 10 Conference , pp. 2040-2044, 2018
2. A. . PappuRajan and S. P. Victor, "Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data," International Journal of Computer Applications, vol. 96, no. 6, pp. 33-37, 2014.
3. Carenini, G., Ng, R. and Zwart, E. Extracting Knowledge from Evaluative Text. Proceedings of the Third International Conference on Knowledge Capture (K-CAP'05), 2005.
4. W. Gordon, "Understanding OAuth: What Happens When You Log Into a Site with Google, Twitter, or Facebook," 2020.
5. Bordes A, Glorot X, Weston J, Bengio Y (2014) A semantic matching energy function for learning with multi-relational data. Mach Learn 94(2):233–259.
6. A. . PappuRajan and S. P. Victor, "Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data," International Journal of Computer Applications, vol. 96, no. 6, pp. 33-37, 2014..
7. P. e. a. Peduzzi, "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis," Journal of Clinical Epidemiology, vol. 49, p. 1373–1379, 1996.
8. H. Uysal, A Genetic Programming Approach to Classification Problems, GRIN Verlag, 2016.
9. A. U. Aquino, M. E. M. Fernandez, A. P. Guzman, A. A. Matias, I. C. Valenzuela and E. P. Dadios, " An Artificial Neural Network (ANN) Model for the Cell Density Measurement of Spirulina (A. platensis)," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp. 1-5, 2018.
10. P. Xu, F. Davoine and T. Denoeux, "Evidential Logistic Regression for Binary SVM Classifier Calibration," In International Conference on Belief Functions, pp. 49 57, 2014.
11. S. M. Kamruzzaman and C. M. Rahman, "Text Categorization using Association Rule and Naive Bayes Classifier," arXiv: Information Retrieval, vol. , no. , p. , 2010.
12. D. e. a. Mittal, "An Effective Hybridized Classifier for Breast Cancer Diagnosis 2015," IEEE International Conference on Advanced Intelligent Mechatronics (AIM), 2015.
13. F. Kabir, S. A. Siddique, M. R. A. Kotwal and M. N. Huda, "Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier," 2015 International Conference on Cognitive Computing and Information Processing (CCIP), pp. 1-4, 2015.
14. Nasukawa, Tetsuya, and Jeonghee Yi. "Sentiment analysis: Capturing favorability using natural language processing." In Proceedings of the 2nd international conference on Knowledge capture, ACM, pp. 70-77, 2003.
15. P. . Domingos, "A few useful things to know about machine learning," Communications of The ACM, vol. 55, no. 10, pp. 78-87, 2012.
16. A. Aquino, Ma. Veronica Bautista, C. Diaz, I. Valenzuela and E. Dadios, "A Vision-Based Closed Spirulina (A. Platensis) Cultivation System with Growth Monitoring using Artificial Neural Network," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp. 1-5, 20186.
17. I. Valenzuela, R. Baldovino, A. Bandala and E. Dadios, "Pre-Harvest Factors Optimization Using Genetic Algorithm for Lettuce," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), pp. 1-4, 2018
18. H. Saif, Y. He, M. Fernandez and H. Alani, "Semantic Patterns for Sentiment Analysis of Twitter," The Semantic Web – ISWC 2014, Springer International Publishing, p. 324–340, 2014
19. G. M. Fung, O. L. Mangasarian and J. W. Shavlik, "Knowledge-Based Support Vector Machine Classifiers," Advances in neural information processing systems, pp. 537-544, 2002.
20. G. P.-S. C. M. W.J. Frawley, "Knowledge discovery in databases: An overview," Knowledge Discovery in Databases, pp. 1-27, 1991.