Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 8, 2025

ANALYZING SENTIMENTS ON X USING MACHINE LEARNING

Umar Ashraf[*1], Ms. Ayesha Akram[2], Saba Ashraf[3], Armeen Pervez[4], M Ehsan Alvi[5]

[*1,4,5]*Constituent College Toba Tek Singh University of Agriculture, Faisalabad, Pakistan*
[2]*Lecturer, of Computer Science. Constituent College Toba Tek Singh University of Agriculture, Faisalabad, Pakistan*
[3]*Department of Computer Science, The Islamia University of Bahawalpur, Pakistan*

[*1]umarashraf60@gmail.com [2]ayeshaakram443@gmail.com [3]ashrafsaba429@gmail.com
[4]shoaibarmeen22@gmail.com [5]mehsanalvi435@gmail.com

DOI: https://doi.org/10.5281/zenodo.17034665

Keywords
Sentiment Analysis, Ensemble Learning, Machine Learning, Social Media Analytics, X (Twitter), Text Classification

Article History
Received: 30 May 2025
Accepted: 05 August, 2025
Published: 30 August, 2025

Corresponding Author: *
Umar Ashraf

Abstract
Nowadays social networking platform has become a quick source of information and a part of daily life routine. Among all the emerging technologies X formerly known as twitter is the most popular platform to share information whether it is from daily life, industry, economy, politics or related to any other field people are tweeting every minute. There is a huge opportunity for data analysts to predict the trends by analyzing the data. This huge number of tweets attracts the attention of data scientists for analyzing the sentiments. By utilizing the sentiment analysis technique, they can classify the data into different categories like positive, negative, irrelevant or neutral and get important information from this. This study will show us how we can use social media for analyzing sentiments to get useful information. We will categorize the reviews of different users into different category using different algorithm. Categorizing the reviews help us to give information about different topics that how people feels about specific things and what are their opinions related to product through analysis we can improve strategies, advertisement, and known people interests. X formally twitter categories sentiments datasets sourced from Kaggle, which comprises user reviews about games, sentiments will be labeled as positive, negative, neutral and irrelevant. Python will employ model classification utilizing Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and Ensemble Learning algorithms. Additionally, Python will be utilized for sentiment categorization into positive, negative, neutral, and irrelevant categories. The success rates of the classification algorithms will be compared to find out their respective performance levels. Experimental results revealed that ensemble learning exhibited the highest accuracy performance, achieving a remarkable accuracy of 91% compared to other methodologies.

1. INTRODUCTION

Social media has transformed into a powerful medium for communication, opinion sharing, and real-time information exchange. Platforms such as X (formerly Twitter) have become hubs for public discourse, where users post millions of short, informal messages daily. These posts

often convey sentiments, opinions, and attitudes about a variety of topics, from consumer products to political events. Understanding such sentiments in real time can provide actionable insights for businesses, policymakers, and researchers.

Sentiment analysis, or opinion mining, applies natural language processing (NLP) and machine learning (ML) techniques to determine the polarity of text whether positive, negative, neutral, or irrelevant. However, the nature of social media content poses unique challenges: informal language, abbreviations, slang, emojis, sarcasm, and rapidly evolving topics make accurate classification difficult.

To address these challenges, machine learning algorithms have been widely used for sentiment classification. Yet, individual models often struggle to generalize across diverse and noisy datasets. Ensemble learning combining predictions from multiple algorithms has emerged as a robust approach to improve accuracy and stability, particularly in complex classification tasks.

This research focuses on developing an ensemble learning framework for sentiment analysis on X, specifically in the gaming domain. The approach integrates four machine learning classifiers Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN) and aggregates their predictions to improve sentiment classification performance. A labeled dataset from Kaggle containing game-related tweets was used to evaluate the models.

1.1 Problem Statement

While many sentiment analysis studies have applied individual ML algorithms, few have systematically compared their performance against ensemble methods in the context of X data. The dynamic, noisy, and high-volume nature of tweets makes it challenging to process, classify, and interpret sentiments accurately. This study addresses these challenges by implementing an ensemble learning model to improve classification accuracy, robustness, and scalability.

1.2 Research Objectives

The objectives of this research are:

1. To implement multiple machine learning algorithms (LR, DT, RF, KNN) for sentiment classification on X data.
2. To develop an ensemble learning model that integrates the strengths of individual classifiers.
3. To compare the performance of ensemble learning against individual models using metrics such as accuracy, precision, recall, and F1-score.
4. To assess the applicability of the proposed approach for real-time sentiment monitoring in noisy social media environments.

2. LITERATURE REVIEW

Sentiment analysis on social media has evolved significantly over the past decade, transitioning from rule-based methods to advanced machine learning and deep learning techniques. X (formerly Twitter), with its vast, dynamic, and informal content, presents both an opportunity and a challenge for sentiment classification tasks.

2.1 Traditional Machine Learning Approaches

Early sentiment analysis models relied on supervised machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Decision Trees [1], [2]. These models, when paired with preprocessing techniques such as stemming, lemmatization, and TF-IDF feature extraction, provided reasonable accuracy but were limited in handling sarcasm, abbreviations, and domain-specific slang [3]. Logistic Regression and K-Nearest Neighbors were also used effectively for text classification, although they often suffered from overfitting on small datasets [4].

2.2 Ensemble Learning in Sentiment Analysis

Ensemble learning emerged as a robust alternative, combining multiple classifiers to reduce variance and bias [5]. Bagging techniques like Random Forest aggregate

multiple decision trees to improve stability [6], while boosting methods such as AdaBoost focus on misclassified instances to enhance model accuracy [7]. Voting and stacking ensembles have also been applied to Twitter sentiment analysis, achieving superior results compared to single classifiers [8]. Recent works show that hybrid ensembles combining traditional ML and deep learning models outperform individual approaches, particularly in noisy datasets [9], [10].

2.3 Deep Learning and Transformer Models

The advent of deep learning brought Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks into sentiment analysis, enabling models to capture contextual and sequential dependencies [11]. However, transformer-based models like BERT and its Twitter-specific variant BERTweet [12] have set new performance benchmarks. While these models excel at contextual understanding, their computational requirements and complexity often make them less suitable for lightweight or real-time sentiment monitoring. Hybrid approaches combining deep embeddings with lightweight classifiers in ensemble frameworks have been proposed to address this limitation [13].

2.4 Domain-Specific Applications

Several studies have focused on domain-specific sentiment analysis, such as political discourse [14], disaster response [15], and product reviews [16]. In the gaming domain, sentiment analysis has been used to assess player feedback, community engagement, and market trends [17]. Given the prevalence of slang, sarcasm, and domain-specific terms in gaming tweets, ensemble learning offers a promising approach to improve classification accuracy.

2.5 Research Gap

While ensemble learning has been explored in sentiment analysis, few studies have compared multiple traditional ML models and ensembles in the specific context of noisy, domain-focused X datasets. This research addresses that gap by implementing and evaluating an ensemble classifier against four baseline ML models on a gaming-related Twitter dataset.

3. METHODOLOGY

This section outlines the approach used for sentiment classification on X (Twitter) data, including dataset description, preprocessing steps, machine learning algorithms, ensemble strategy, and evaluation metrics. The methodology follows a structured pipeline to ensure reproducibility and accuracy.

3.1 Dataset Description

The dataset used in this research was sourced from Kaggle and contains game-related tweets labeled into four sentiment categories: **Positive**, **Negative**, **Neutral**, and **Irrelevant**. Two CSV files were used:

- **Training Set**: 74,681 tweets
- **Testing Set**: 999 tweets

Each record contains:

1. **ID** – Unique identifier for each tweet.
2. **Game** – Name/title of the game referenced.
3. **Sentiment** – Labeled sentiment category.
4. **Text** – Actual tweet content.

The dataset provides a balanced representation across most sentiment categories, although minor class imbalance was observed, with **Negative** tweets slightly more frequent than others.

3.2 Preprocessing

Given the noisy and informal nature of Twitter data, extensive preprocessing was performed to improve classification accuracy:

1. **Lowercasing** – All text converted to lowercase.
2. **Removal of URLs & Mentions** – Eliminated hyperlinks and @user mentions.
3. **Punctuation & Special Character Removal** – Stripped symbols, hashtags (while retaining keywords), and unnecessary characters.
4. **Tokenization** – Split text into words/tokens.

5. **Stopword Removal** – Removed non-informative words (e.g., "the", "and").
6. **Lemmatization** – Reduced words to base form (e.g., "playing" → "play").
7. **Vectorization** – Converted processed text into numerical format using **TF-IDF** representation.

3.3 Machine Learning Algorithms

Four baseline machine learning classifiers were implemented:

1. **Logistic Regression (LR)** – Effective for binary and multi-class classification with interpretable coefficients.
2. **Decision Tree (DT)** – Rule-based classifier capable of handling non-linear relationships.
3. **Random Forest (RF)** – Bagging-based ensemble of decision trees for improved stability.
4. **K-Nearest Neighbors (KNN)** – Instance-based learner using majority voting among nearest neighbors.

3.4 Ensemble Learning Approach

To leverage the strengths of individual models, a **Voting Classifier** ensemble was implemented, combining LR, DT, RF, and KNN:

- Ensemble learning reduces bias, improves generalization, and enhances performance on noisy datasets.

3.5 Evaluation Metrics

Model performance was evaluated using:

- **Accuracy** – Proportion of correctly classified tweets.
- **Precision** – Ratio of true positives to total predicted positives.
- **Recall** – Ratio of true positives to actual positives.
- **F1-Score** – Harmonic mean of precision and recall.

These metrics were calculated for each model and compared to identify the most effective approach.

4. DATA ANALYSIS AND IMPLEMENTATION

This section describes the experimental setup, tools, and step-by-step process used for training and evaluating the sentiment classification models.

4.1 Experimental Setup

The experiments were conducted on a personal computer with the following specifications:

- **Processor:** Intel Core i7 (8th Gen) @ 2.2 GHz
- **RAM:** 16 GB
- **Operating System:** Windows 10 (64-bit)
- **Programming Language:** Python 3.9
- **Libraries Used:** Pandas, NumPy, Scikit-learn, NLTK, Matplotlib, Seaborn

4.2 Data Exploration

Initial exploratory data analysis (EDA) revealed:

- Tweets ranged from **5 to 280 characters**.
- Vocabulary contained domain-specific gaming terms such as *lag*, *loot*, and *multiplayer*.
- Sentiment distribution: Negative tweets were slightly more common, followed by Positive, Neutral, and Irrelevant.
- Noise in the dataset included hashtags, emojis, URLs, and misspellings.

4.3 Implementation Steps

Step 1 – Preprocessing: All tweets were cleaned, tokenized, and transformed into TF-IDF vectors as described in Section 3.2.

Step 2 – Model Training: Each baseline classifier (LR, DT, RF, KNN) was trained on the TF-IDF matrix of the training set. Hyperparameters were tuned using grid search where applicable:

- LR: Regularization parameter C tuned in the range [0.1, 10]
- DT: Maximum depth tuned between 10–50
- RF: Number of trees varied between 100–500
- KNN: Optimal *k* determined between 3–15

Step 3 – Testing: The testing dataset of 999 tweets was used to evaluate model performance using Accuracy, Precision, Recall, and F1-Score.

4.4 Data Visualization

Visualization played a crucial role in understanding feature distributions and model performance:

- **Word Clouds** – Generated for positive and negative tweets to identify frequent terms.
- **Confusion Matrices** – Plotted for each classifier to highlight classification patterns and misclassifications.
- **Performance Comparison Chart** – Compared accuracy scores of all models side-by-side.

5. RESULTS AND DISCUSSION

This section presents the experimental results of the four baseline machine learning classifiers and the proposed ensemble learning model, evaluated on the testing dataset.

5.1 Performance of Individual Models

Table 1 summarizes the performance metrics Accuracy, Precision, Recall, and F1-Score of each classifier.

Table 1: Performance of Baseline Classifiers

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	88.2	88.0	87.9	88.0
Decision Tree	83.5	83.2	83.1	83.1
Random Forest	89.0	88.9	88.8	88.8
K-Nearest Neighbors	85.4	85.0	85.1	85.0

The Random Forest achieved the highest accuracy among the individual models at **89.0%**, followed closely by Logistic Regression at **88.2%**. Decision Tree and KNN showed comparatively lower performance, primarily due to overfitting and sensitivity to noisy text data.

5.2 Ensemble Model Performance

The ensemble voting classifier, integrating predictions from LR, DT, RF, and KNN, outperformed all individual models with an accuracy of **91.0%**.

Table 2: Performance of Ensemble Classifier

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Ensemble (Voting)	**91.0**	90.9	90.8	90.8

The improvement in accuracy and balanced precision-recall scores demonstrate the ensemble model's robustness in handling noisy, short-text data from X.

5.3 Confusion Matrix Analysis

The ensemble model showed fewer misclassifications in Positive and Negative classes compared to individual classifiers. However, the Neutral and Irrelevant classes occasionally overlapped, likely due to ambiguous wording in tweets.

5.4 Comparative Analysis with Previous Studies

Compared to related works [5], [8], [13], which reported accuracies in the range of 85–89% for Twitter sentiment classification using individual models, the proposed ensemble approach achieves a noticeable performance boost. This validates the effectiveness of integrating diverse classifiers to mitigate weaknesses inherent in individual models.

5.5 Discussion

The results suggest that ensemble learning provides better generalization for sentiment classification in

gaming-related X data. By combining linear, tree-based, and instance-based classifiers, the model benefits from complementary decision boundaries and reduced overfitting.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This research presented a sentiment analysis framework for X (Twitter) data using ensemble learning, with a focus on gaming-related tweets. Four baseline machine learning models Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors were implemented and compared against an ensemble voting classifier.

The experimental results demonstrate that the ensemble model achieved an accuracy of **91%**, outperforming all individual classifiers. This improvement underscores the effectiveness of combining diverse classifiers to capture different aspects of noisy, short-text social media data.

The study contributes to the field of sentiment analysis by:

1. Demonstrating the robustness of ensemble learning in noisy, domain-specific datasets.
2. Providing a comparative analysis of traditional ML algorithms for social media sentiment classification.
3. Offering a practical approach that can be extended to other domains requiring real-time opinion monitoring.

6.2 Future Work

While the ensemble model showed strong performance, several areas for future research remain:

1. **Deep Learning Integration** – Incorporating transformer-based models such as BERT or BERTweet in an ensemble framework to capture richer semantic context.
2. **Sarcasm Detection** – Enhancing preprocessing to identify and correctly classify sarcastic tweets, which remain challenging for current models.
3. **Real-Time Deployment** – Implementing the model in a streaming environment to analyze live tweets for immediate sentiment monitoring.
4. **Multi-Language Support** – Expanding the model to handle multilingual tweets to widen applicability in global contexts.
5. **Domain Adaptation** – Testing the framework on other industries such as politics, healthcare, or finance to evaluate generalizability.

By addressing these areas, future work can further improve sentiment analysis accuracy and adaptability, making such models more effective for large-scale, real-time applications.

REFERENCES

[1] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford University, 2009.

[2] B. Liu, Sentiment Analysis and Opinion Mining, 1st ed. San Rafael, CA: Morgan & Claypool Publishers, 2012.

[3] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, 2010, pp. 1320–1326.

[4] C. K. H. Lee, Y. C. Cheung, and F. L. Chung, "A comparative study of machine learning algorithms for text classification," in Proceedings of the International Conference on Neural Information Processing (ICONIP), 2005, pp. 126–131.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.

[8] H. Hasan, N. Ahmad, and M. A. Azmi, "Twitter sentiment analysis using ensemble learning approach," Journal of Theoretical and Applied Information Technology, vol. 96, no. 14, pp. 4563–4572, 2018.

[9] M. S. Akhtar, A. Kumar, and A. Ekbal, "A hybrid deep learning model for sentiment analysis,"

Pattern Recognition Letters, vol. 125, pp. 129–135, 2019.
[10] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," Concurrency and Computation: Practice and Experience, vol. 33, no. 23, pp. 1–15, 2021.
[11] Y. Kim, "Convolutional neural networks for sentence classification," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1746–1751.
[12] D. Nguyen, R. Vu, and M. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
[13] S. Z. Qamar, A. Khan, and S. R. Ahmad, "Hybrid deep learning model for aspect-based sentiment analysis," IEEE Access, vol. 8, pp. 181350–181359, 2020.
[14] K. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM), 2010, pp. 178–185.
[15] K. Rudra et al., "Summarizing situational tweets in crisis scenarios," Proceedings of the 27th ACM Conference on Hypertext and Social Media, 2016, pp. 137–147.
[16] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.
[17] H. Mall, A. K. Singh, and S. R. Ahmad, "Opinion mining of social media data for the gaming industry," International Journal of Computer Applications, vol. 179, no. 27, pp. 1–5, 2018.