Tikrit Journal of Pure Science Vol. 30 (4) 2025

DOI: https://doi.org/10.25130/tjps.v30i4.1916

IRAQI Academic Scientific Journals

Tikrit Journal of Pure Science

ISSN: 1813 – 1662 (Print) --- E-ISSN: 2415 – 1726 (Online)

Journal Homepage: https://tjpsj.org/

A Social Media Sentiment Analysis Using Machine Learning Approaches

Noor Salah Irzooqi Al-Agele[1], Didem KIVANÇ TÜRELİ[2]

[1]*Department of Advanced Electronics and Communication Technology, University of Istanbul Okan-Turkey*

[2]*Department of Electrical and Electronics Engineering, University of Washington-United States of America.*

Received: 23 Feb. 2025 **Received in revised forum:** 25 May. 2025 **Accepted:** 29 May. 2025

Final Proof Reading: 14 Aug. 2025 **Available online:** 25 Aug. 2025

ABSTRACT

Social media platforms like Twitter provide major means for individuals to express their opinions on various topics; therefore, a need for complex tools to distinguish between negative and positive attitudes in textual content. With consideration for the most suitable models for precisely classifying sentiments within social media data, this study aims to evaluate the efficacy of machine learning algorithms in analyzing sentiment text that people post or comment on Twitter, thus bridging the research gap in the analysis of sentiment in an understudied domain. a set of machine learning algorithms was applied along with feature extraction methods, including TF-IDF and Natural Language Processing (NLP). With an accuracy of 93%, the Random Forest (RF) model proved to be the most effective among other models, Because of its exceptional capacity and generating accurate and dependable results on textual data, the Random Forest (RF) model proves in the study to be the most optimal choice for sentiment analysis textual.

Keywords: Sentiment Analysis, Machine Learning (ML), Random Forest (RF), Natural Language Processing (NLP), Feature Extraction, Fake news.

Name: **Noor Salah Irzooqi Al-Agele** **E-mail:** **noor.salah9457@gmail.com**

Tikrit Journal of Pure Science Vol. 30 (4) 2025
DOI: https://doi.org/10.25130/tjps.v30i4.1916

تحليل المشاعر في وسائل التواصل الاجتماعي باستخدام أساليب التعلم الآلي

نور صلاح ارزوقي العجيلي[1]، ديدم كيفانتش تورلي[2]

[1] قسم تكنولوجيا الالكترونيات والاتصالات المتقدمة، جامعة إسطنبول أوكان – تركيا

[2] قسم الهندسة الكهربائية والإلكترونية، جامعة واشنطن ، الولايات المتحدة الأمريكية

الملخص

تُوفر منصات التواصل الاجتماعي، مثل تويتر، وسيلةً رئيسيةً للأفراد للتعبير عن آرائهم حول مواضيع مُختلفة؛ ولذلك، تبرز الحاجة إلى أدوات مُعقدة للتمييز بين المواقف السلبية والإيجابية في المحتوى النصي .مع الأخذ في الاعتبار أنسب النماذج لتصنيف المشاعر بدقة ضمن بيانات وسائل التواصل الاجتماعي، تهدف هذه الدراسة إلى تقييم فعالية خوارزميات التعلم الآلي في تحليل نصوص المشاعر التي ينشرها الناس أو يُعلقون عليها على تويتر، وبالتالي سد الفجوة البحثية في تحليل المشاعر في مجالٍ لم تُدرس بشكلٍ كافٍ .طُبقت مجموعة من خوارزميات التعلم الآلي إلى جانب أساليب استخراج السمات، بما في ذلك TF-IDF ومعالجة اللغة الطبيعية (NLP). أثبت نموذج الغابة العشوائية (RF) أنه الأكثر فعاليةً بين النماذج الأخرى بدقةٍ تبلغ 93%. ونظرًا لقدرته الاستثنائية وتوليده نتائج دقيقة وموثوقة على البيانات النصية، فقد أثبت نموذج الغابة العشوائية (RF) في الدراسة أنه الخيار الأمثل لتحليل المشاعر النصية .

الكلمات المفتاحية: تحليل المشاعر ، التعلم الآلي(ML) ، الغابة العشوائية(RF) ، معالجة اللغة الطبيعية(NLP) ، استخراج السمات، الأخبار الكاذبة.

INTRODUCTION

In recent times, the rapid evolution of information technology and the widespread usage of social media platforms such as Twitter, Instagram, WhatsApp, and Facebook have become integral parts of people's daily lives [(1)]. As of 2024, Twitter (now known as X) continues to be a major social media platform globally, with approximately 556 million active users worldwide [(2)]. Regarding daily activity, Twitter users generate around 500 million tweets each day, making it one of the most active social networks in terms of content creation [(2)]. These platforms serve as central hubs for individuals to express their thoughts and opinions on various subjects, including products and services. In such a potentially limitless universe of textual materials, there is an urgent need for tools that would allow for the classification of positive and negative attitudes within the context of texts. Feelings are interdependent with people's relations, affect their views on occurrences, subjects, practices and so on. As a result, sentiment analysis is used in a range of sectors for the purpose of monitoring opinion shifts and improving business insights, optimizing recommendation systems, understanding customer feedback, and optimizing message filtering [(2)]. Sentiment Analysis (SA) is one of the most prominent and impactful applications of Natural Language Processing (NLP), widely used in analyzing user-generated content (UGC) on social media platforms, like tweets and comments, where the nature and context of the text differ considerably[(3)]. This data is highly valuable in revealing individuals responses to products, political opinions, racism, and behavioral intentions through the analysis of comments, feedback, and posts[(4)]. Nevertheless, simple keyword-based analysis is often insufficient to fully understand underlying emotions, requiring more sophisticated and intelligent models. Opinion or sentiment analysis has proven to be an effective tool within

(NLP) techniques, designed to identify emotions within texts [(5)]. It is widely used in business to classify customers' sentiments and opinions towards specific products or services. As a subset of data mining, machine learning, and artificial intelligence, sentiment analysis allows for the extraction of valuable information from various text sources like emails, blogs, websites, chatbots, and customer reviews ([6]. The strategies for sentiment analysis are categorized into three primary types: rule-based systems, which use lexicons and predefined rules to determine sentiment polarity; Automated Systems, which apply machine learning algorithms to detect patterns; and Hybrid Systems that combine both approaches [(7)]. Opinion mining, a subdomain of sentiment analysis, goes beyond mere polarity detection, also identifying the subject of the opinion, the opinion holder, the intensity of sentiment (positive or negative), and the type of emotion expressed [(8)]. Sentiment Analysis (SA)is applied across various domains, including product reviews, politics, healthcare, financial markets, blogging, Facebook fan pages, and even identifying issues such as human trafficking, with analysis applicable at different levels, ranging from document-level to sentence, phrase, or sub-phrase levels [(9).] Machine learning (ML) based on classification techniques has shown strong performance, often surpassing traditional methods in fields like image classification, text categorization, and sentiment analysis [(10)]. Recent trends indicate that sentiment analysis has become one of the most popular areas in machine learning, offering superior effectiveness compared to traditional approaches [(11).]

Notwithstanding the significant progress in sentiment analysis applications, most studies focus on commercial or general political contexts, with limited exploration of the analysis of racist discourse and sentiment on social media platforms, particularly Twitter. There is a clear lack of research in this area, in different language contexts, where a structured, data-driven approach to racial sentiment analysis has not been adequately addressed. This study aims to develop an analytical framework that utilizes advanced machine learning models to analyze sentiment in tweets, addressing a gap in previous research. The contributions of this work include: (1) the creation of a large dataset composed of diverse tweets and comments on racial topics; (2) annotating this dataset under the "TruthSeeker" project using classification tools to assign sentiment labels (positive, negative) based on polarity; (3) application and refinement of several widely-used machine learning models, including Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB) K-Nearest Neighbors (KNN), and Stochastic Gradient Descent (SGD); (4) use of common feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW); and (5) evaluation of the models' performance using four key metrics: Accuracy, Precision, Recall, and F1-measue. This paper is organized into five main sections. Section one presents the introduction, theoretical background, and research motivation. Section two reviews relevant previous studies. Section three outlines the methodology. Section four presents and analyzes the results, comparing them across models. Section Five is Conclusions.

RELATED WORK

As for the traditional machine learning approach, it has also been applied to carry out sentiment analysis in recent years for which different performances have been recorded., [(12)] applied Support Vector Machines (SVMs) to Twitter data, achieving notable results but struggled with mixed sentiments and ambiguous expressions [(13)]. used ensemble methods on social media platforms, although their approach was computationally expensive and difficult to interpret [(14)] combined XGBoost and Random Forest on Twitter data, showing high accuracy, though XGBoost required extensive tuning and was computationally intensive [(15)] utilized Random Forest for movie reviews, achieving good results but struggled with overfitting due to noisy data and large feature sets [16] applied Naive Bayes to social media data, but its assumption of feature independence limited its effectiveness in more complex sentiment data [17] used Naive Bayes for IMDB reviews, but noted difficulties in dealing with complete sentence structures and context-dependent sentiments [18] combined Random Forest and SVM on movie reviews, with Random Forest being slow for large datasets and SVM not scaling well for larger datasets [19] employed Decision Trees on social media data, though the model was prone to overfitting with too many features or insufficient data [20] utilized Naive Bayes for Twitter, but the method's simplicity and assumption of feature independence were limitations[21] applied SVM to Twitter data, which required careful parameter tuning to avoid underfitting or overfitting [22] used K-Nearest Neighbors (KNN) on reviews but faced challenges with computational expense for large datasets and sensitivity to irrelevant features [23] employed Logistic Regression and Naive Bayes on reviews, with Logistic Regression struggling with non-linear data relationships [24] used Naive Bayes and SVM on social media data, but found the models less accurate for complex sentiment tasks in noisy environments [25] applied Logistic Regression (LR) to product reviews, but the method had limitations in capturing complex sentiment without extensive feature engineering [26] utilized Naive Bayes (NB) on news articles, but the method's assumption of feature independence limited its ability to handle complex data. They also used Random Forest on online forums, but the model struggled with very large or high-dimensional data [27] employed Logistic Regression, Decision Tree, Naïve Bayes, and SVM in Intrusion Detection Systems. These techniques, including SVM and Naïve Bayes, have also proven effective in other domains, such as sentiment analysis, where they demonstrated their capability in processing and classifying textual data. However, these methods have certain weaknesses, such as Naïve Bayes' reliance on simple assumptions that limit its ability to handle complex data, and the difficulty SVM and Decision Tree face when dealing with high-dimensional data or large datasets [28] applied Random Forest (RF) to social media data, facing the risk of overfitting if not carefully tuned [29] used Decision Trees (DT) for e-commerce reviews, but the model was prone to overfitting on small or noisy datasets [30] applied Naive Bayes(NB) and SVM to product feedback, noting that ensemble methods could become computationally demanding. They also used Random Forest on reviews, which required careful feature engineering [31] employed KNN for movie reviews, but KNN struggled with large datasets and irrelevant features [32] used XGBoost on movie reviews, highlighting the sensitivity to parameter tuning and the model's limitations with small datasets. Finally, [33] applied Decision Tree(DT) to customer reviews, but noted the model's struggles with unstructured or highly imbalanced data

METHODOLOGY

Dataset

Experiments on fake news detection were conducted using Twitter data. The dataset, TruthSeeker, includes over 180,000 English-language tweets labeled as real or fake to study misinformation on social media. Tweets were labeled using a three-factor active learning approach, with contributions from 456 skilled Amazon Mechanical Turk annotators. It also includes bot, credibility, and influence scores for each user to analyze user behavior. The dataset supports various analyses like deep learning-based fake news detection, clustering-based event detection, and the study of tweet labels about user characteristics. Implementation of our proposed model in Python programming Language, libraries used: Pandas, scikit-learn, NumPy, nltk, seaborn. The dataset was processed and converted into CSV format, It is available at: [TruthSeeker Twitter Dataset 2023 | Kaggle].

Preprocessing

Natural processing language (NLP) is applied to the data to remove noise and prepare for analysis, which improves text quality and enhances machine learning performance. which is subdivided into the following steps [34].

Tokenization: This process converts text to the abbreviation of 'bag of words' in which text consists of tokens (it could be words or letters) [35]. Initialize a bag of words with the count of zero and then process each token, add the count of tokens which are in the given dictionary [36]. This leads to the vector for the text in terms of Tokens, where the value of Term is substantial depending on the Token.

Stop-word Removal: an elimination of unnecessary words or words which add no value to the document known as stop words is done. These are words such as 'and," "the,'' 'and,' among others. This way the text is much more valuable for analysis since the indicated words do not have any research value [37].

Stemming: Stemming eliminates words to their root word status and involves the elimination of every prefix and suffix added to a word. For example, where we have "running" it is shortened to "run" or "runner". This process is used in normalizing related words into a common form and at the same time the stem may not always be a proper stem [38].

Normalization: Normalization entails applying a common case to all text tokens in that they are either in uppercase or lowercase. This makes it easier to process all tokens and maintains homogeneity as each token is always in one format. For instance, converting all tokens to lower cases is essential when used when comparing words without considering their case [39].

Feature Extraction

Feature extraction in SA involves feature discovery feature discovery which detects terms of sentiment and is the same as feature extraction in identification [40]. These features are rather important to define the type of sentiment in the text: is it negative or positive and that is why they have a strong influence on the orientation of the given text. Taken from the family of TF-IDF (Term Frequency-Inverse Document Frequency), the significance of the terms was mapped with reference to their frequently too in the document and their rarity in all the documents[41].

$Term\ Frequency\ (TF)$:

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

……Equation (1)

This measures how frequently a term appears in a specific document.

Inverse Document Frequency (IDF):

$$IDF(t,D) = log\ \left(\frac{Total\ number\ of\ documents\ |D|}{Number\ of\ documents\ containing\ term\ t}\right)$$

…Equation (2)

This measures how rare or common a term is across the entire dataset.

TF-IDF:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

……Equation (3)

This value indicates the importance of a term in a document, with higher values showing greater significance.

Machine Learning Model

A branch of artificial intelligence, machine learning enables systems to learn and improve from data [42]. After pre-processing textual and feature extraction The model training by supervised learning is currently the most common sub-branch of machine learning (ML), as Fig (1). It is supervised by utilizing labeled datasets to train and prepare calculations that classify information or foresee results precisely. The label of the dataset consists of positive, negative. The dataset training model is 70%, and 30% for testing [43]. Commonly applied machine learning algorithms include [44]:

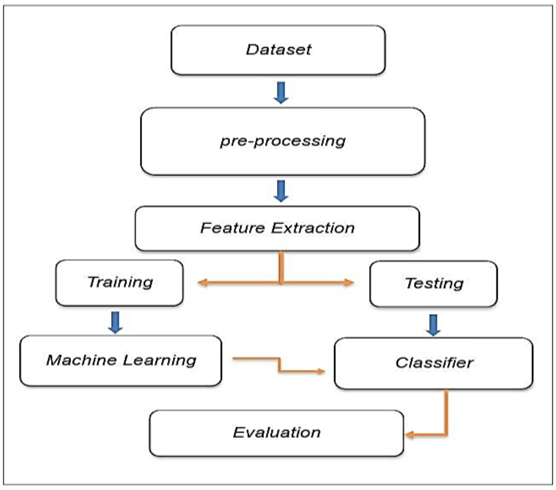

Fig. 1: The proposed sentiment model.

Random Forest (RF)

is a machine learning method that increases accuracy by combining the results of multiple decision trees [45]. Each tree is trained on a random subset of the data, and at each step, only a random set of features is used. This randomness makes the trees diverse, which helps reduce errors and prevents overfitting. For classification, the final prediction is based on majority voting, meaning the class predicted by most trees is chosen. For regression, the final prediction is the average of the predictions from all the trees.

$\hat{y}(x) = argmax_k \sum_{b=1}^{B} \mathrm{II}(T_b(x) = k)$ ……Equation (4)

For regression, average s the tree outputs:

$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$ ……Equation (5)

Where $\hat{y}(x)$is predicated class input x ,k is total number of classes, $T_b(x)$ predication of the tree (return class label k), $\mathrm{II}(.)$, is indicator function (1 if true, 0 otherwise).

Logistic Regression Algorithm

Logistic Regression Algorithm is an algorithm used in machine learning to identify a desirable result from various attributes in enclosed data. This begins with looking into the individual attribute contribution to the prediction followed by figuring out class distribution. Using these insights, it uses logistic regression model to assign a class label to an attribute value. The algorithm then measures the level of error with the rules derived, then pick the level of error that is low, meaning the best predictor [46]. The above selected rule is applied in making some predictions on the new data. In general, Logistic Regression repeatedly refines rules to increase their accuracy in the output and performs greatly for tasks such as sentiment analysis.

$P(y = 1|x) = \sigma(w^T x + b) = \frac{1}{1 + e - (w^T x+b)}$ ……Equation (6)

Where $\sigma(w^T x + b)$ is sigmoid function, x is input of the feature vector, w is weight vector, b is the bias term and $P(y = 1|x)$ is a probability that the output class y is 1 (positive), $(1 - P(y = 1|x)$ or $P(y = 0|x))$ is a probability that the output class y is 0 (negative).

Decision Tree Algorithm

Decision Tree is an algorithm that constructs a tree structure from data set with attributes and then classify the data [47]. It begins by verifying whether all the instances in the same node are of the same class; otherwise, a new node is created alongside this entire node. If the dataset is empty or has got only a single attribute, it generates a terminal node with the most popular class label. Then it chooses the best attribute to divide the dataset normally according to information gain ratio and puts a decision node in it. The end data is split according to attribute values and the algorithm is re-implemented on each of the partitions. This process continues until all nodes are created; at the end of it; you are left with a decision tree model [48].

$Entropy = -\sum_{i=0}^{c-1} p(i|t)\ log\ 2\ p(i|t)$ ……Equation (7)

Where c is the number of classes, $p(i \mid t)$ denotes the fraction of records belonging to class i at a given node t.

K-NN Algorithm

The K-Nearest Neighbors (KNN) clustering method is the technique of identifying data's closest neighbors in the feature space[49]. This basic implementation of the KNN algorithm uses the training set, a single test object and k as parameter.

DOI: https://doi.org/10.25130/tjps.v30i4.1916

It then works out the distance between the test object and all the training samples and then chooses k of the closest neighbors. Then it determines which of these neighbors belongs to the most commonly flooding category. Last but not the least, it determines the category of the test object. KNN is another simple and versatile algorithm as a result of which it can classify a given problem with the help of machine learning concepts [(50)]. There are several distance metrics used in the KNN algorithm; the best metrics are Euclidean distance, and the Manhattan distance, and the Equation (8) shows Euclidean Distance.

$d = \sqrt{(x_1 - x_2)^2 + (x_1 - y_2)^2 + \cdots}$ ……Equation (8)

Where d is a distance x_1, y_1 are represented as a first point, x_2, y_2 are represented as the second point.

NB Algorithm

Naive Bayes (NB) is one of the most tended and used probabilistic classifier and adopted for range of text classification tasks including Sentiment analysis of the text [(51)]. As an example, NB uses Baye's Theorem to compute a set of probabilities given frequency and combination of values in a data set. The algorithm identifies the chance that a fixed feature set corresponds to a specific label, allowing documents to be classified into specific classes.

$P(A|B) = P^{c^{\frac{P(A|B)*P(A)}{P(B)}}}$ ……Equation (9)

Where $(A|B)$ are events1, $P(A)$ and $P(B)$1are the probabilities1of A and B1without regard for each other $P(A|B)$ is the conditional1probability, the probability1of A given that B is true ,$P(B|A)$ is the probability1of B given1that A is true [(52)].

Stochastic Gradient Descent (SGD)

This algorithm includes important steps that can improve a set of parameters to a model: starts with the selection of random parameters for a model [(53)]. The learning rate which defines the size of each update step, is set to fixed value. In each pass through the training data, the data set may be shuffled in an attempt to gain robustness during the process, an algorithm then cycles through the training case, using the current model coefficients to compute the expected value. The gradient of the loss function it's calculated based on the prediction error and input features [(54)]. This gradient is used to modify the model parameters in a direction that helps to minimize a loss. This is done a fixed number of epochs to go through this process of continually updating the estimates of the model parameters. Some of them are used for the initialization of model parameters and others are used for making predictions with regard to the current parameters [(55)].

SGD equation given by:

$\theta_{t+1} = \theta_t - \eta . \nabla J(\theta_t; x_i, y_i)$ ……Equation (10)

Where θ_t represents weight of models at time step t, η is learning rate , J is the loss function, and (x_i, y_i) are single training samples used in stochastic update [(56)].

Evaluation Metrics

Standardize classification metrics used — Accuracy, Precision, Recall, and F1 measure to evaluate the performance of the models. These metrics give a thorough assessment of the predictive abilities of the classifiers [(57)]:

Accuracy :is the fundamental metric for a text classification model. respectively defined as:

$Accuracy = \frac{\text{TP+TN}}{\text{N}}$ ……Equation (11)

precision: is the ratio of true positives to the sum of the true positives and false positives, which is the total predictions that were made as positive.

$\text{precision} = \frac{TP_i}{TP_i+FP_i}$ ……Equation (12)

Recall measures: the ratio of correctly predicted positive observations to all actual positive observations.

$\text{Recall} = \frac{TP_i}{TP_i+FN_i}$ ……Equation (13)

F1 measure: is the harmonic mean of precision and recall, giving equal weight to both:

$F1 = \frac{2\times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} = \frac{2TP}{2TP+FP+FN}$ ……Equation (14)

Where True Positive (TP) is positive correctly classified, False Positive (FP) is negative reviews that are incorrectly classified as positive, True Negative (TN) is Negative reviews are correctly classed and False Negative (FN) is Positive reviews that incorrectly classified as negative.

RESULT

In this section, we present the evolution result of six supervised machine learning algorithms applied to the sentiment analysis tasks. The model was trained and tested on a labeled data set using consistent.TF-IDF Feature extraction method. To ensure object compression, the performance of each algorithm was assessed using 4 standardizing evaluation metrics. Accuracy, precision, Recall, and f1-measure.

Table 1: Results table.

Algorithm	Accuracy	Precision	Recall	F1-measure
RF (Random Forest)	93%	93%	93%	93%
NB (Naive Bayes)	89%	89%	89%	89%
LR (Logistic Regression)	92%	92%	92%	92%
KNN(K-Nearest Neighbors)	66%	88%	66%	70%
DT (Decision Tree)	89%	89%	89%	89%
SGD(Stochastic Gradient Descent)	92%	92%	92%	92%

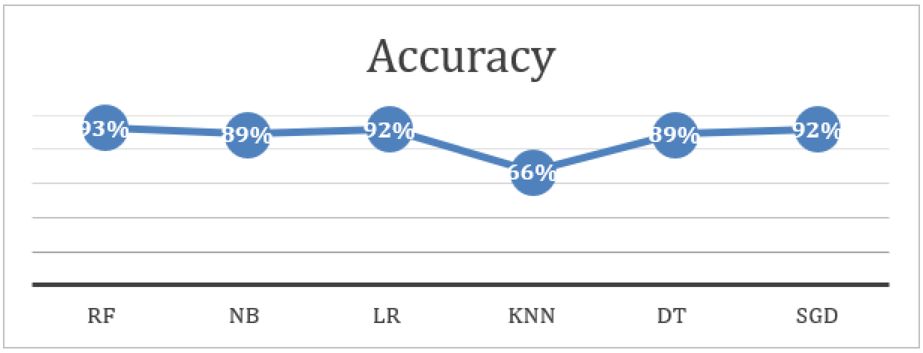

Fig. 2: Accuracy result.

As shown in Table 1, Random Forest (RF) algorithms outperformed all other models, achieving the highest accuracy and consistency across all metrics, at 93%. LR and SGD also showed strong performance across all metrics. In contrast, KNN showed limited performance in high feature spaces generated by TF-IDF. The DT and NB lost average performance as showed in Fig (2), reflecting their impact of the data distribution and their prior assumptions.

The system achieves significant improvements by applying pre-processing techniques to the text data and utilizing an advanced feature extraction method. Unlike other models that work directly with raw text, our approach streamlines the process by removing noise and irrelevant information, enhancing the quality of the input data. This optimization leads to more efficient processing and better model performance. The success of the proposed system is heavily reliant on the extraction of meaningful features, which are then used as inputs for the machine learning classification model. When evaluating the model's accuracy, precision, recall, and F-measure against other text classification methods, it is clear that our model outperforms the competition. In comparison to other methods presented in Table 2 these comparisons highlight the superior capabilities of the recommended approach, showcasing its potential to set a new standard in text analysis and classification within the field of machine learning.

Table 2: Comparison between proposed model and related works.

Ref	Technique	Dataset	Accuracy	Precision	Recall	F-measure
(12)	Support Vector Machines (SVMs)	Twitter Dataset	76.92%	77%	100%	87%
(13)	Various Machine Learning Algorithms	Social Media	85%	83%	87%	85%
(14)	XGBoost and Random Forest	Twitter	88%	86%	89%	87%
(15)	Random Forest	Movie Reviews	88%	86%	89%	87%
(16)	Naive Bayes	Social Media	89%	88%	89%	89%
	Decision Trees and Random Forests		86%	84%	87%	85%
(17)	Logistic Regression	Reddit	85%	83%	87%	85%
(18)	Naive Bayes	IMDB Reviews	84%	82%	85%	83%
	Random Forest and SVM	Movie Reviews	85%	84%	86%	85%
(19)	Decision Trees	Social Media	86%	84%	88%	86%
(20)	Naive Bayes	Twitter	88%	87%	89%	88%
(21)	Support Vector Machines (SVMs)	Twitter	87%	85%	88%	86%
(22)	K-Nearest Neighbors (KNN)	Reviews	82%	80%	82%	81%
(23)	Logistic Regression and Naive Bayes	Reviews	85%	83%	86%	84%
(24)	Naive Bayes and SVMs	Social Media	84%	82%	85%	83%
(26)	Random Forest	Online Forums	89%	87%	90%	89%
(29)	Decision Trees	E-commerce Reviews	88%	86%	89%	87%
(30)	Naive Bayes and SVM	Product Feedback	90%	89%	91%	90%
	Multi-Channel CNNs (Shallow)	Twitter	88%	87%	88%	88%
	Feature Engineering with Random Forest	Reviews	86%	84%	87%	85%
(31)	K-Nearest Neighbors (KNN)	Movie Reviews	85%	83%	86%	84%
(32)	XGBoost	Movie Reviews	87%	85%	88%	86%
(33)	Machine Learning (e.g., SVM, RF)	Twitter Sentiment	88%	85%	87%	86%
Our proposed model			**93%**	**93%**	**93%**	**93%**

CONCLUSION

This study illustrates the efficacy of machine learning algorithms, specifically ensemble models such as random forest, in comprehending feelings articulated on social media platforms, specifically Twitter. The results indicate that the random forest (RF) model surpasses other models in accuracy and dependability, making it the optimal choice for sentiment categorization in social data contexts, This area is crucial for comprehending the societal responses and facilitating public opinion analysis, hence, enhancing decision-making across several sectors, additionally, these technologies enhance experts and the scholars ability to analyses textual material more accurately and effectiveness of combining natural language processing techniques with supervised machine learning algorithms in classifying different sentiment in social media. The random forest (RF) algorithms achieved the highest performance, indicating that ensemble methods are particularly suitable for this type of application. Future exploration of deep learning and sentiment analysis techniques for multiple languages is suggested.

REFERENCES

1. Bessarab A, Mitchuk O, Baranetska A, Kodatska N, Kvasnytsia O, Mykytiv G. Social networks as a phenomen on of the information society. Journal of Optimization in Industrial Engineering. 2021;14(29):17-24. https://doi.org/10.22094/JOIE.2020.677811

2. Preis J, Klika D. Sustainability as a Message on Social Media: A Case Study of the World Economic Forum's Twitter (Now X) Account. Problemy Ekorozwoju. 2024;19(2):126-38. https://doi.org/10.35784/preko.5673

3. Liu B. Sentiment analysis and opinion mining. Cham: Springer Nature; 2022.

https://doi.org/10.1007/978-3-031-02145-9

4. Olakangil A, Wang C, Nguyen J, Zhou Q, Jethwa K, Li J, et al., editors. Exploring Embeddings for Measuring Text Relatedness: Unveiling Sentiments and Relationships in Online Comments. 2023 Second International Conference on Informatics (ICI); 2023: IEEE. https://doi.org/10.1109/ICI60088.2023.10421308

5. Tabassum A, Patil RR. A survey on text pre-processing & feature extraction techniques in natural language processing. International Research Journal of Engineering and Technology (IRJET). 2020;7(06):4864-7.

6. Diwali A, Saeedi K, Dashtipour K, Gogate M, Cambria E, Hussain A. Sentiment analysis meets explainable artificial intelligence: a survey on explainable sentiment analysis. IEEE Transactions on Affective Computing. 2023. : https://doi.org/10.1109/TAFFC.2023.3296373

7. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN computer science. 2021;2(3):160. https://doi.org/10.1007/s42979-021-00592-x

8. Maruthupandi J, Sivakumar S, Kumar VS, Srikaanth PB. Sem-AI: A Unique Framework for Sentiment Analysis and Opinion Mining Using Social Network Data. SN Computer Science. 2025;6(2):99. https://doi.org/10.1007/s42979-024-03628-0

9. Žitnik S, Blagus N, Bajec M. Target-level sentiment analysis for news articles. Knowledge-Based Systems. 2022;249:108939. https://doi.org/10.1016/j.knosys.2022.108939

10. Sharma S, Mehra R. Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. Journal of digital imaging. 2020;33(3):632-54. https://doi.org/10.1007/s10278-019-00307-y

11. Tan KL, Lee CP, Lim KM. A survey of sentiment analysis: Approaches, datasets, and future research. Applied Sciences. 2023;13(7):4550. https://doi.org/10.3390/app13074550

12. Han K-X, Chien W, Chiu C-C, Cheng Y-T. Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. Applied Sciences. 2020;10(3):1125. https://doi.org/10.3390/app10031125

13. Chen J, Becken S, Stantic B. Lexicon based Chinese language sentiment analysis method. Computer Science and Information Systems. 2019;16(2):639-55. https://doi.org/10.2298/CSIS181015013C

14. Huang L, Liu Y, Huang W, Dong Y, Ma H, Wu K, et al. Combining random forest and XGBoost methods in detecting early and mid-term winter wheat stripe rust using canopy level hyperspectral measurements. Agriculture. 2022;12(1):74. https://doi.org/10.3390/agriculture12010074

15. Liu J, Cao M, Bai D, Zhang R, editors. Solar radiation prediction based on random forest of feature-extraction. IOP Conference Series: Materials Science and Engineering; 2019: IOP Publishing. https://doi.org/10.1088/1757-899X/658/1/012006

16. Li F. Novel Data-Driven Machine Learning Models for Heating Load Prediction: Single and Optimized Naive Bayes. International Journal of Advanced Computer Science & Applications. 2024;15(8). https://doi.org/10.14569/IJACSA.2024.0150866

17. Sundararajan SK, editor Detection of conjunctivitis with deep learning algorithm in medical image processing. 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC); 2019: IEEE. https://doi.org/10.1109/I-SMAC47947.2019.9032705

18. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence. 2022;115:105151. https://doi.org/10.1016/j.engappai.2022.105151

19. Soni T, Gupta D, Dutta M, editors. Machine Learning in Healthcare: Decision Trees for Asthma Risk Prediction. 2024 4th International Conference

on Sustainable Expert Systems (ICSES); 2024: IEEE. https://doi.org/10.1109/ICSES63445.2024.10763341

20. Qian H, Pan Y, Wang X, Li Z. Research on the optimization of belief rule bases using the Naive Bayes theory. Frontiers in Energy Research. 2024;12:1396841. https://doi.org/10.3389/fenrg.2024.1396841

21. Cano A. Social media and machine learning2020. https://doi.org/10.5772/intechopen.78089

22. Alabdulatif A, Thilakarathne NN, Aashiq M. Machine Learning Enabled Novel Real-Time IoT Targeted DoS/DDoS Cyber Attack Detection System. Computers, Materials & Continua. 2024;80(3). https://doi.org/10.32604/cmc.2024.054610

23. Agerri R, Rigau G. Language independent sequence labelling for opinion target extraction. Artificial Intelligence. 2019;268:85-95. ttps://doi.org/10.48550/arXiv.1901.09755. https://doi.org/10.48550/arXiv.1901.09755

24. Siddiqui MK, Morales-Menendez R, Huang X, Hussain N. A review of epileptic seizure detection using machine learning classifiers. Brain informatics. 2020;7(1):5. https://doi.org/10.1186/s40708-020-00105-1

25. Lee CS, Cheang PYS, Moslehpour M. Predictive analytics in business analytics: decision tree. Advances in Decision Sciences. 2022;26(1):1-29. https://doi.org/10.47654/v26y2022i1p1-30

26. Patel TS, Patel DP, Sanyal M, Shrivastav PS. Prediction of heart disease and survivability using support vector machine and Naive Bayes algorithm. bioRxiv. 2023:2023.06. 09.543776. https://doi.org/10.1101/2023.06.09.543776

27. Mohammed MS, Talib HA. Using machine learning algorithms in intrusion detection systems: A review. Tikrit Journal of Pure Science. 2024;29:3. https://doi.org/10.25130/tjps.v29i3.1553

28. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review. 2020;53(6):4335-85. https://doi.org/10.1007/s10462-019-09794-5

29. Sharma NA, Ali AS, Kabir MA. A review of sentiment analysis: tasks, applications, and deep learning techniques. International journal of data science and analytics. 2024:1-38. https://doi.org/10.1007/s41060-024-00594-x

30. Zhang D, Wu C, Liu J. Ranking products with online reviews: A novel method based on hesitant fuzzy set and sentiment word framework. Journal of the Operational Research Society. 2020;71(3):528-42. https://doi.org/10.1080/01605682.2018.1557021

31. Singh H, Sharma V, Singh D. Comparative analysis of proficiencies of various textures and geometric features in breast mass classification using k-nearest neighbor. Visual Computing for Industry, Biomedicine, and Art. 2022;5(1):3. https://doi.org/10.1186/s42492-021-00100-1

32. Osman AIA, Ahmed AN, Chow MF, Huang YF, El-Shafie A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal. 2021;12(2):1545-56. https://doi.org/10.1016/j.asej.2020.11.011

33. Gomez-Adorno H, Bel-Enguix G, Sierra G, Barajas J-C, Álvarez W, editors. Machine learning and deep learning sentiment analysis models: Case study on the sent-covid corpus of tweets in mexican spanish. Informatics; 2024: MDPI. https://doi.org/10.3390/informatics11020024

34. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review. 2022;55(7):5731-80. https://doi.org/10.1007/s10462-022-10144-1

35. Juluru K, Shih H-H, Keshava Murthy KN, Elnajjar P. Bag-of-words technique in natural language processing: a primer for radiologists. RadioGraphics. https://doi.org/10.1148/rg.2021210025

36. Erkan A, Güngör T. Analysis of Deep Learning Model Combinations and Tokenization

Approaches in Sentiment Classification. IEEE Access. 2023;11:134951-68. https://doi.org/10.1109/ACCESS.2023.3337354

37. Kalaivani ER, Marivendan ER. The effect of stop word removal and stemming in datapreprocessing. Annals of the Romanian Society for Cell Biology. 2021;25(6):739-46.

38. Jabbar A, Iqbal S, Tamimy MI, Rehman A, Bahaj SA, Saba T. An analytical analysis of text stemming methodologies in information retrieval and natural language processing systems. IEEE Access. 2023;11:133681-702. https://doi.org/10.1109/ACCESS.2023.3332710

39. Ehrmanntraut A. Historical German text normalization using type-and token-based language modeling. arXiv preprint arXiv:240902841. 2024. https://doi.org/10.48550/arXiv.2409.02841

40. Jader R, Aminifar S. An Intelligent Gestational Diabetes Mellitus Recognition System Using Machine Learning Algorithms. Tikrit Journal of Pure Science. 2023;28(1):82-8. https://doi.org/10.25130/tjps.v28i1.1269

41. Jena D, Rautaray J, Mishra P, editors. Summarization of document using feature selection method: TF-IDF. International Conference on Artificial Intelligence and Data Science Applications - 2023, ICAIDSC2023; 2025. https://doi.org/10.5120/icaidsc202409

42. Paramesha M, Rane NL, Rane J. Big data analytics, artificial intelligence, machine learning, internet of things, and blockchain for enhanced business intelligence. Partners Universal Multidisciplinary Research Journal. 2024;1(2):110-33. https://doi.org/10.5281/zenodo.12827323

43. Luengo J, García-Gil D, Ramírez-Gallego S, García S, Herrera F. Big data preprocessing. Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-39105-8

44. Mahesh B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR)[Internet]. 2020;9(1):381-6. https://doi.org/10.21275/ART20203995

45. Hussein DM, Beitollahi H. A hybrid deep learning model to accurately detect anomalies in online social media. Tikrit Journal of Pure Science. 2022;27(5):105-16. https://doi.org/10.25130/tjps.v27i5.24

46. Huang FL. Alternatives to logistic regression models in experimental studies. The Journal of Experimental Education. 2022;90(1):213-28. https://doi.org/10.1080/00220973.2019.1699769

47. Priyanka, Kumar D. Decision tree classifier: a detailed survey. International Journal of Information and Decision Sciences. 2020;12(3):246-69. https://doi.org/10.1504/IJIDS.2020.108141

48. Hashim EKM. Arabic sentiment analysis for determining terrorism supporters on Twitter using data mining techniques [Master's thesis]: University of Babylon, College of Information Technology, Iraq; 2019.

49. Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal. 2022;3:100071. https://doi.org/10.1016/j.dajour.2022.100071

50. Halder RK, Uddin MN, Uddin MA, Aryal S, Khraisat A. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Journal of Big Data. 2024;11(1):113. https://doi.org/10.1186/s40537-024-00973-y

51. RAGHUNANDA S, ALEX SA, KANAVALLI DA. Error Classification based on Multinomial NB and Random Forests. https://doi.org/10.36872/LEPI/V51I2/301127

52. Mansoor HH. Classification technique design for spam SMS [Master's thesis]. Baghdad, Iraq: Iraqi Commission for Computers and Informatics, Informatics Institute for Postgraduate Studies; 2019.

53. Al-Anzi FS. An effective hybrid stochastic gradient descent arabic sentiment analysis with partial-order microwords and piecewise

differentiation. Fractals. 2022;30(08):2240222. https://doi.org/10.1142/S0218348X22402228

54. Ekici B. Deterministic and Stochastic Schemes for Unconstrained Optimization. 2023.

55. Tikosi K. Convergence results regarding stochastic gradient descent methods for dependent data streams [PhD thesis]: Central European University; 2021.

56. Hassan AF, Bhaya WS. Analysis of BBC News by Applying Classification Algorithms. Journal of Advanced Research in Dynamical and Control Systems. 2020;12(1):148-52. https://doi.org/10.5373/JARDCS/V12I1/20201023

57. Li Q, Peng H, Li J, Xia C, Yang R, Sun L, et al. A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology (TIST). 2022;13(2):1-41. https://doi.org/10.1145/3495162