Enhancing Emotion Detection in Textual Data: A Comparative Analysis of Machine Learning Models and Feature Extraction Techniques

Wedad Q. A. Saif
Faculty of Engineering and Information Technology, Taiz University, Taiz, Yemen
wedadalshameri49@gmail.com

Majid Khalaf Alshammari
School of Educational Studies, Universiti Sains Malaysia, 11800 Penang, Malaysia
m.alnehait@hotmail.com

Badiea Abdulkarem Mohammed
College of Computer Science and Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia
b.alshaibani@uoh.edu.sa

Amer A. Sallam
Faculty of Engineering and Information Technology, Taiz University, Taiz, Yemen
amer.sallam@taiz.edu.ye (corresponding author)

Received: 12 May 2024 | Revised: 25 July 2024 and 26 July 2024 | Accepted: 30 July 2024

 DOI: https://doi.org/10.48084/etasr.7806

ABSTRACT

The digital age has resulted in a massive increase in the amount of available textual data, including articles, comments, texts, and updates on social networks. The value of analyzing such a large volume of data extends to many other industries and applications, as it provides important insights into the perspectives of customers, strategic decision-making, and market demands. Detecting emotions in texts faces challenges due to linguistic patterns and cultural nuances. This study proposes a system capable of accurately identifying emotions expressed in text using a variety of machine learning models, including logistic regression, extra randomized tree, voting, SGD, and LinearSVC. It also employs different feature extraction techniques, such as TF-IDF, Bag-of-Words, and N-grams, comparing their performance in these models. An evaluation was carried out using two English emotion datasets, namely ISEAR and AIT-2018, using F1 score, accuracy, recall, and precision. The findings demonstrate the ability and effectiveness of the system to detect emotions conveyed within texts. The LinearSVC model with N-grams achieved the highest accuracy of 88.63% on the ISEAR dataset, while the extra randomized tree classifier with N-grams achieved 89.14% accuracy on the AIT-2018 dataset. Furthermore, the SGD model with TF-IDF achieved 88.18% and 84.54% accuracy on the ISEAR and the AIT-2018 datasets, respectively.

Keywords-emotion detection; textual data; machine learning; data encoding; feature extraction

I. INTRODUCTION

Emotion analysis in text is vital across diverse domains, including social media, product reviews, scientific research, and social studies. By examining user comments and understanding their perspectives, emotion analysis helps monitor brand and product reputations, improve consumer engagement, and develop products and services that align with user needs [1, 2]. Its impact extends to public opinion, media analysis, and human-automated system interactions, enhancing the comprehension of public reactions and communication effectiveness [3-5].

Despite its significance, identifying emotions in text presents challenges due to linguistic variability and cultural nuances. Balancing verbal and emotional analysis remains difficult, but advances in natural language processing and machine learning models can address these issues [7]. This study introduces a system designed to accurately analyze and identify emotions in texts using various learning models and feature extraction methods. The system incorporates Logistic Regression (LR), Extra Randomized Tree (ERT), Voting Classifier (VC), SGD, and LinearSVC, along with TF-IDF, Bag-of-Words (BoW), and N-grams for feature extraction.

II. RELATED WORKS

Many studies have investigated emotion recognition from text using various methods and datasets. The ISEAR dataset, which includes emotions such as joy, shame, fear, sadness, and guilt, has been widely used, along with the AIT-2018 dataset. In [8], an emotion recognition method was proposed using LR, achieving an F1 score of 85%. In [9], emotions were analyzed using VSM and STASIS on the ISEAR dataset, achieving average accuracies of 0.53 and 0.50, respectively. This approach involved preprocessing, such as removing unnecessary words and using Word Sense Disambiguation (WSD), to enhance the accuracy of STASIS. In [10], a system was proposed to classify emotions on social media using natural language processing techniques and machine learning algorithms. This system achieved 91.7% accuracy with SMO and 85.4% with J48 on a dataset of 13,000 tweets classified into six emotions. In [11], Random Forest (RF), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM) classifiers were used to perform sentiment analysis on scientific articles. The results showed F1 scores of 88% with SVM and 87% with NB.

In [12], WordNet-Affect and EmoSenticNet were used for emotion classification on the AIT-2018 dataset, achieving 88.23% accuracy with EmoSenticNet. In [13], 64.08% accuracy was achieved using multinomial NB on the ISEAR dataset. In [14], the ISEAR dataset was processed to simplify syntax and expand words using synonyms, achieving 65% accuracy. In [15], LSTM and nested SVM were employed, and LSTM achieved 94.15% accuracy. In [16], SVM achieved 83.31% accuracy in airline sentiment analysis. In [17], NB and KNN were used to analyze tweets, achieving 72.60% accuracy with NB. In [18], various machine and deep learning techniques were employed on the SemEval2018 dataset, achieving up to 91.90% accuracy with LSTM. In [19], various classifiers were used on the AffectiveTweets dataset, achieving up to 90.2% accuracy with NB. In [20], LSTM achieved 91.90% accuracy. In [21], a hybrid approach, combining CNN, Bi-GRU, and SVM, achieved 80.11% accuracy. In [22], CNN-LSTM with multiple embeddings achieved an accuracy up to 90.4%. In [23], a system using BiGRU and BiLSTM achieved 87.66% accuracy. In [24], different SVM kernels were employed on the ISEAR dataset, achieving 61.8% accuracy with linear kernels. In [25], a sequential neural network model on Twitter data achieved 89.98% accuracy.

This study introduces a system for emotion detection utilizing machine learning models, including Logistic Regression, Extra Trees, Voting Classifier, SGD, and LinearSVC, along with feature extraction techniques such as TF-IDF, Bag-of-Words (BoW), and N-grams.

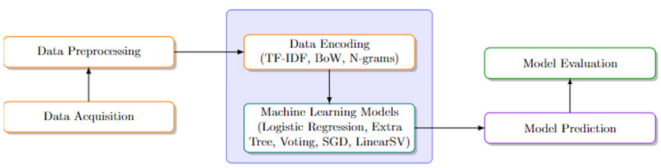

Fig. 1. The architecture of the proposed method.

III. METHODOLOGY

A. Data Collection

This study focuses on the comprehensive analysis of sentiment and emotion across two essential datasets, as shown in Figure 2.

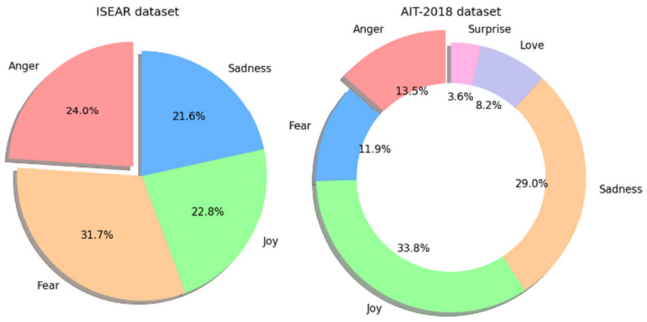

Fig. 2. Emotion analysis across the datasets.

The first dataset used is the ISEAR dataset [26], well-regarded in the domain of English language research. Comprising 7,102 texts, this dataset meticulously classifies emotions into four distinct classes: fear, joy, sadness, and anger. Additionally, the SemEval-2018 Affect In Tweets distant supervision corpus was used, commonly known as the AIT-2018 dataset [27]. This extensive corpus encompasses 20,000 texts, representing a broad spectrum of emotions, including anger, sadness, love, fear, joy, and surprise. These datasets served as the foundation of this investigation, offering a diverse and comprehensive repository to analyze and evaluate emotion recognition and sentiment analysis within textual data.

B. Data Preprocessing

Preprocessing is a crucial set of steps applied to the dataset before feature extraction and model training to enhance results. In this study, preprocessing involved eliminating stop-words, special characters, and non-English words, such as quotation marks, parentheses, and apostrophes. Additionally, numerical values were removed and text was converted to lowercase to prevent redundancy. Finally, stemming was used to reduce words to their root form, facilitating more effective analysis and comparison.

C. Data Encoding

1) Term Frequency-Inverse Document Frequency (TF-IDF)

This is a method to determine a word's relevance in a document based on how frequently it appears. This method consists of two parts to extract information from the text, [28, 29]:

- Term Frequency (TF): The following equation is used to calculate each document's unique TF, which expresses the frequency of a particular phrase relative to the total number of words in the document.

$$TF(i,j) = \frac{frequency\ of\ item\ i\ in\ document\ j}{Total\ word\ count\ in\ the\ J\ document} \quad (1)$$

- Inverse Document Frequency (IDF) is used to reduce the weight of phrases that appear frequently in a group of documents and increase the weight of uncommon terms in all documents, as shown in

$$IDF(i) = \log\left(\frac{total\ number\ of\ documents}{documents\ containing\ the\ word\ i}\right) \quad (2)$$

Finding terms that are more significant than others is made easier by looking at their TF-IDF values, which are higher for words that occur frequently in the document and less frequently in other documents. TF and IDF are combined in (3) to accomplish this.

$$TF - IDF(i,j) = TF(i,j) * IDF(i) \quad (3)$$

2) Bag of Words (BoW)

Regardless of word order or grammatical relationships, texts are mathematically processed as a set of utilized words using the BoW, also known as the Bag of Features (BoF), algorithm [30]. The great results of this algorithm can be attributed to its ease of use in performing mathematical operations. The text is divided into a collection of words, and the frequency of occurrence and presence of each word in the text is calculated.

3) N-grams

N-grams provide additional information about word associations and allow texts to be represented in a way that takes context and order into consideration, improving the performance of machine learning models. N-grams can be referred to by various names [31], including uni-grams, bi-grams, tri-grams, and four-grams. This study used both uni-grams and bi-grams. Natural language processing algorithms use this algorithm to extract features from texts as contiguous sequences.

D. Machine Learning Models

Models were trained on the extracted features to analyze text sentiments. This study focused on five classifier algorithms.

1) Logistic Regression (LR)

LR is one of the machine learning models used for binary or multiple classification of extracted data. The LR classifier trains on the features to find the optimal weighted parameters that lower the cost function. The wider the vocabulary size n, the longer it will take to train and produce predictions. The probability value can range from 0 to 1 [31, 32].

2) Extremely Randomized Trees (ERT) Classifier

Every Decision Tree (DT) is built from the initial training sample [33], and at each test node, it is given a random sample to select the best feature to partition the data according to a mathematical criterion. This study used the criterion 'gini' [34], denoted as in (4), to represent the criteria function used for evaluating the quality of the split. Since the ERT classifier randomly selects the value by which the features will be divided and subnodes will be generated, it is considered one of the machine learning models that trains multiple DTs and compiles their results to produce a forecast.

$$I_G = 1 - \sum_{j=1}^{c} P_j^2 \quad (4)$$

where P_j represents the probability of samples falling into class C for a given node.

3) Voting Classifier (VC)

This is a machine learning model that makes predictions about a class or output by looking at the class with the highest chance of being selected as an output [35, 36]. It performs the classification process by aggregating the results of every classifier that is fed to it, as opposed to creating custom models. Three models, LR, DT, and Support Vector Classifier (SVC), were integrated into soft voting [37]. To obtain the final classification, these models compute the expected probabilities for every category and combine them.

4) Stochastic Gradient Descent (SGD)

This model is widely applied to obtain the best agreement between expected and actual results [38], minimize the cost function in machine learning projects, and determine the model parameters that yield the highest accuracy in both training and testing data. Instead of using the entire dataset as a single batch during training for each epoch in SGD, small random chunks of data, referred to as batches, are used to compute the gradient and update the parameters.

5) Linear Support Vector Classification (LinearSVC)

Based on SVM, this is a linear classification model for binary or multiple data classification. The objective of this model is to determine the ideal lines that divide the different input features [39]. These locations are called support vectors, and the best hyperplane is the one with the largest margin. The objective is to increase the margin. The distance between the support vectors is called the margin. The number of hyperplanes varies with the diversity of classes [40].

E. Model Prediction

The models were trained on the training dataset following the completion of the data preprocessing and representation steps. The test dataset is used to confirm that the models can accurately predict text sentiment after training. The test dataset is subjected to the same preprocessing and representation techniques as the training dataset. The performance of the models was then assessed using widely adopted performance metrics. It should be mentioned that each model has a different architecture that affects its performance and results.

F. Model Evaluation

Models for text classification are frequently evaluated based on how accurate they are. The model's accuracy measures how effectively it predicts the right classification given the inputs.

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) * 100 \quad (5)$$

Precision assesses how well the model can classify positive cases. Regardless of whether the texts were correctly or wrongly classified, it computes the ratio of correctly classified positive texts to all positively classified texts.

$$Precision = \left(\frac{TP}{TP + FP}\right) * 100 \quad (6)$$

Conversely, recall measures how well the model can recognize real positive examples. Equation 7 computes the proportion of texts that were accurately categorized as positive to all texts that are actually positive.

$$Recall = \left(\frac{TP}{TP + FN}\right) * 100 \quad (7)$$

F1 score is widely used to balance detecting strength and accuracy, combining recall and precision into a single metric.

$$F1\ score = \left(2 * \frac{Precision * Recall}{Precision + Recall}\right) * 100 \quad (8)$$

The number of positive texts that were correctly classified is denoted by TP, the number of negative messages that were accurately classified is represented by TN, FP denotes the quantity of true negative texts that were classified as positive, and FN is the number of genuine positive texts that were classified as negative.

IV. EXPERIMENTAL RESULTS

This study employed two distinct datasets, ISEAR and AIT-2018, which cover a broad spectrum of subjects and contexts. Each dataset was divided into 80% for training and 20% for testing. Before analysis, the data were preprocessed to reduce noise and extract relevant text segments reflecting emotions. The features were then extracted from these text segments using the TF-IDF, BoW, and N-grams algorithms. Subsequently, classification algorithms such as LR, ERT, VC, SGD, and LinearSVC were employed for emotion detection. Table I and Figure 3 show the results using the AIT-2018 dataset. The ERT model using N-grams to extract the features had the best results, with an accuracy of 89.14%. The other models also achieved good results. With TF-IDF, BoW, and N-grams, the SGD classifier achieved fairly similar results (84.54%, 84.14%, and 84.39%, respectively). The lowest accuracy obtained using BoW was 79.17% with the ERT, and the lowest accuracy obtained using TF-IDF was 79.27% with LR. As for N-grams, it achieved good results, with the lowest accuracy being 77.31% with LR.

TABLE I. MODELS' OUTPUT ON THE AIT-2018 DATASET

Model name	Feature extraction	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	TF-IDF	79.27	78.82	66.34	70.47
	BoW	83.94	80.49	78.13	79.13
	N-grams	77.31	80.07	62.4	67.03
ERT	TF-IDF	86.38	83.35	81.01	81.98
	BoW	86.76	82.42	82.87	82.53
	N-grams	89.14	85.97	84.18	84.96
VC	TF-IDF	84.46	79.96	80.92	80.41
	BoW	79.17	74.55	74.04	74.21
	N-grams	87.56	83.57	84.53	83.99
SGD	TF-IDF	84.54	81.2	80.07	80.47
	BoW	84.14	79.83	80.97	80.26
	N-grams	84.39	82.74	78.08	79.8
LinearSVC	TF-IDF	84.29	81.3	79.62	80.36
	BoW	83.53	78.98	79.22	79.07
	N-grams	84.99	81.87	80.34	81.06

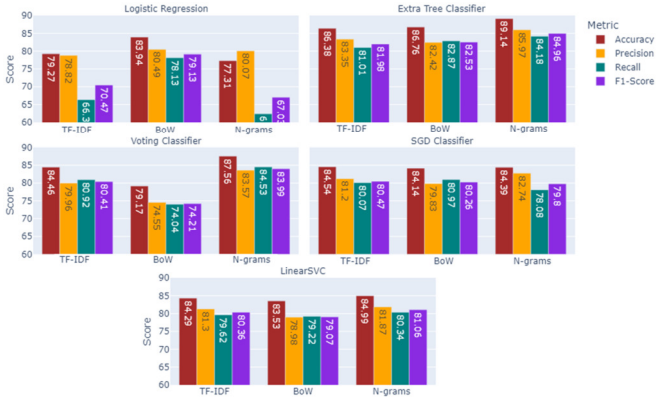

Fig. 3. Models' performance using the AIT-2018 dataset.

In the ISEAR dataset, the SGD model with TF-IDF achieved an accuracy of 88.18%, while the LinearSVC model with N-grams achieved the best accuracy of 88.63%. With TF-IDF, BoW, and N-grams, 87.39%, 87.05%, and 87.56% accuracy was achieved, respectively, with the ERT model. The lowest accuracy obtained using BoW, TF-IDF, and N-grams was 84.63%, 84.52%, and 86.15%, respectively, with the ERT classifier. Table II and Figure 4 show the results achieved with the five models on the ISEAR dataset. These results show that the LinearSVC model using N-grams achieved the highest accuracy (88.63%), suggesting the importance of considering sequential word arrangements for emotion classification. Additionally, the SGD model with TF-IDF features performed well (88.18%), demonstrating its effectiveness in capturing meaningful patterns in text. The ERT model demonstrated robustness across different feature extraction methods, achieving accuracies ranging from 87.05 to 87.56%. However, the ERT classifier achieved lower accuracies, indicating limited improvement over individual classifiers. These findings emphasize the need to explore diverse models and techniques to optimize emotion classification performance.

TABLE II. MODELS OUTPUT ON THE ISEAR DATASET

Model name	Feature extraction	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	TF-IDF	85.64	86.34	85.28	85.7
	BoW	85.02	85.56	84.92	85.19
	N-grams	87.05	87.39	86.87	87.1
ERT	TF-IDF	87.39	88.37	86.93	87.46
	BoW	87.05	87.81	86.69	87.13
	N-grams	87.56	87.85	86.69	87.14
VC	TF-IDF	84.52	84.74	84.42	84.5
	BoW	84.63	84.93	84.51	84.67
	N-grams	86.15	86.51	86	86.19
SGD	TF-IDF	88.18	88.37	88.15	88.25
	BoW	84.68	85.36	84.4	84.8
	N-grams	86.43	87.08	86.12	86.53
LinearSVC	TF-IDF	87.95	88.13	87.88	87.99
	BoW	85.53	86.09	85.46	85.73
	N-grams	88.63	88.83	88.52	88.65

The findings of this study were compared with those of [8] on the ISEAR dataset. The LR model in [8] achieved an F1 score of 85%. This study achieved similar results with LR, utilizing both TF-IDF and BoW techniques, achieving 85.7%

and 85.19% F1 scores, respectively. Using the LR model with N-grams surpassed the results of [8], attaining 87.1% F1 score. Table III provides a comparative overview of the results of this study alongside those of prior studies. Additionally, in [23] BoW with LR and RF achieved 81.34% and 81.54% accuracy, respectively. However, this study demonstrated enhanced performance, achieving 85.02% accuracy with BoW and LR.

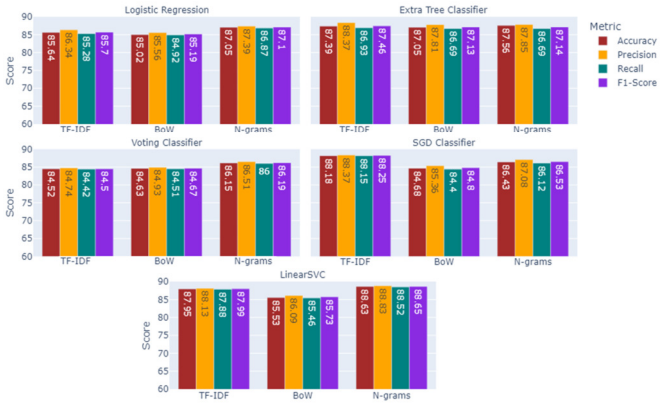

Fig. 4. Models' performance using the ISEAR dataset.

In [8], the method employed for feature extraction remained unspecified. In [23], preprocessing steps included lemmatization and stemming, whereas this study adopted only stemming. When using a combination of stemming and lemmatization, there can be conflicting results and interference between the processes of stemming and converting words to their infinitive forms (lemmatization). This inconsistency can sometimes lead to a loss of linguistic accuracy. In addition, efforts were made to expand abbreviations to their full forms. Despite these discrepancies, the proposed study consistently yielded remarkable results, with performance metrics ranging from 84.63% to 88.63% across the five models evaluated on the ISEAR data set. A comparative analysis with the findings of [18] and [12] underscores the robustness of this study, particularly evident in the superior results achieved on the AIT-2018 dataset. The variance with [18] and [12] can be attributed to the use of different classification algorithms, the difference in the number of emotions, and the size of the dataset.

As shown in Table III, various methods were employed in previous studies. Most studies used different classification algorithms from this study, except [8] and [23], which aimed to detect emotions conveyed through textual data using the same datasets using LR. To our knowledge, the ERT, VC, SGD, and LinearSVC classifiers have not been evaluated in any previous study in this context. This study used TF-IDF, BoW, and N-grams techniques to extract features from the datasets. Furthermore, although previous studies predominantly focused on both TF-IDF and BoW for feature extraction, they neglected the utilization of N-grams, whose efficacy was demonstrated in this study.

TABLE III. COMPARISON WITH PREVIOUS STUDIES

Ref.	Dataset	#Tweets/Texts	#Emotions	Data encoding methods	Machine learning models	Performance	
						Accuracy (%)	F1 (%)
[18]	AIT-2018	10983	8	TF-IDF, BoW, and Word2vec	Naïve Bayes,	80.9	76.2
				TF-IDF, BoW, and Word2vec	Random Forest	81.9	79.4
				TF-IDF, BoW, and Word2vec	Support Vector Machine	81.5	79.8
[12]	AIT-2018	4000	4	TF-IDF, and WordNet	Support Vector Machine	49.23	-
				TF-IDF, and WordNet	Decision Tree	45.64	-
				TF-IDF, and WordNet	Multinomial Naive Bayes	47.17	-
				TF-IDF, and EmoSenticNet	Support Vector Machine	86.42	-
				TF-IDF, and EmoSenticNet	Decision Tree	80.09	-
				TF-IDF, and EmoSenticNet	Multinomial Naive Bayes	88.23	-
[8]	ISEAR	-	5	-	Logistic Regression	-	85
[23]	Dataset	7102	4	BoW	Logistic Regression	81.34	-
				BoW	Random Forest	81.54	-
This study	ISEAR	7102	4	TF-IDF	Logistic Regression	85.46	85.7
				BoW	Logistic Regression	85.02	85.19
				N-grams	Logistic Regression	87.05	87.1
				N-grams	LinearSVC	88.63	88.65
				TF-IDF	SGD	88.18	88.25
	AIT-2018	20000	6	N-grams	Extra Tree	89.14	84.96
				BoW	Extra Tree	86.76	82.53
				N-grams	Voting	87.56	83.99

V. DISCUSSION

The findings demonstrate that feature extraction techniques and classification models greatly influence performance, underscoring the critical role of method choice in emotion classification tasks. For the AIT-2018 dataset, the ERT classifier with N-grams achieved the highest accuracy of 89.14%, highlighting the effectiveness of N-grams in capturing sequential text dependencies essential for nuanced emotion detection. Although the SGD classifier produced consistent results across different feature extraction methods, ERT showed the lowest accuracy with BoW at 79.17%. This suggests that the ERT classifier might be less effective in aggregating predictions from multiple classifiers for this dataset, possibly due to its limitations in leveraging the nuanced features captured by other extraction methods. In contrast, in the ISEAR dataset, the LinearSVC model with N-

grams attained the highest accuracy of 88.63%, with the SGD model using TF-IDF also performing well at 88.18%. These results emphasize the strength of TF-IDF in extracting significant text features. The ERT model proved robust across various feature extraction methods, with accuracies between 87.05% and 87.56%, while VC again underperformed with accuracies ranging from 84.63% to 86.15%. This pattern suggests that, although ensemble methods such as VC may be advantageous, they may not always surpass individual models in every scenario. Comparing the results with those of [8] and [23], LR with N-grams resulted in an F1 score of 87.1%, surpassing previous results and demonstrating that N-grams can provide a more nuanced representation of text for improved classification. Similarly, while in [23] LR and RF with BoW achieved 81.34% and 81.54% accuracy, respectively, this study achieved superior results with 85.02% accuracy using BoW and LR, further validating the effectiveness of this approach.

Methodologically, this study used stemming alone, avoiding lemmatization, which contrasts with some previous studies that used both processes. This choice likely contributed to more consistent results by preventing potential conflicts between stemming and lemmatization. Additionally, abbreviation expansion possibly improved the clarity and completeness of the textual data used for emotion detection. The findings also underscore the importance of feature extraction methods. Despite the predominant focus of previous research on TF-IDF and BoW, this study highlights the effectiveness of N-grams, which, although underutilized in previous studies, demonstrated consistent high performance in capturing critical sequential patterns in text.

VI. CONCLUSION

In fields such as politics, social sciences, and marketing, obtaining meaningful insights from human responses is crucial. Analyzing sentiment and emotions in text poses significant challenges due to the complexity of human language. This study addressed these challenges by employing five machine learning models, LR, ERT, VC, SGD, and LinearSVC, along with various feature extraction techniques, including TF-IDF, BoW, and N-grams, to detect emotions in text. These methods were tested on two datasets, ISEAR and AIT-2018. The results indicated that the ERT model with N-grams achieved the highest accuracy of 89.14% on the AIT-2018 dataset, while the LinearSVC model with N-grams reached an accuracy of 88.63% on the ISEAR dataset. The SGD model with TF-IDF also demonstrated strong performance, with an accuracy of 88.18% on ISEAR. These findings highlight the superior ability of N-grams to capture sequential text patterns and improve classification accuracy. In contrast, the VC, particularly with BoW, showed consistently lower performance, underscoring its limitations in this context.

This study underscores the importance of selecting the appropriate feature extraction methods and models for effective emotion detection. The notable performance of N-grams in combination with ET and LinearSVC models illustrates their potential to improve classification results. Although ensemble methods, such as VC, have their advantages, they may not always exceed the effectiveness of individual models. The novelty of this work lies in the comparative evaluation of these techniques and models, providing new insights into their relative effectiveness. Future research should investigate additional models, including deep learning approaches, and leverage larger, multilingual datasets to further enhance emotion detection accuracy and capabilities.

REFERENCES

[1] A. R. Abas, I. Elhenawy, M. Zidan, and M. Othman, "BERT-CNN: A Deep Learning Model for Detecting Emotions from Text," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 2943–2961, 2022, https://doi.org/10.32604/cmc.2022.021671.

[2] B. A. Mohammed *et al.*, "Hybrid Techniques of Analyzing MRI Images for Early Diagnosis of Brain Tumours Based on Hybrid Features," *Processes*, vol. 11, no. 1, Jan. 2023, Art. no. 212, https://doi.org/10.3390/pr11010212.

[3] R. Piryani, D. Madhavi, and V. K. Singh, "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015," *Information Processing & Management*, vol. 53, no. 1, pp. 122–150, Jan. 2017, https://doi.org/10.1016/j.ipm.2016.07.001.

[4] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, Nov. 2020, https://doi.org/10.1016/j.future.2020.05.034.

[5] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 113–126, Jan. 2022, https://doi.org/10.1515/jisys-2022-0001.

[6] A. Athar, "Sentiment analysis of scientific citations," University of Cambridge, Computer Laboratory, UCAM-CL-TR-856, 2014. https://doi.org/10.48456/tr-856.

[7] A. M. Abubakar, D. Gupta, and S. Palaniswamy, "Explainable Emotion Recognition from Tweets using Deep Learning and Word Embedding Models," in *2022 IEEE 19th India Council International Conference (INDICON)*, Kochi, India, Nov. 2022, pp. 1–6, https://doi.org/10.1109/INDICON56171.2022.10039878.

[8] F. M. Alotaibi, "Classifying Text-Based Emotions Using Logistic Regression," *VAWKUM Transactions on Computer Sciences*, vol. 7, no. 1, pp. 31–37, Apr. 2019, https://doi.org/10.21015/vtcs.v16i2.551.

[9] F. Mozafari and H. Tahayori, "Emotion Detection by Using Similarity Techniques," in *2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, Bojnord, Iran, Jan. 2019, pp. 1–5, https://doi.org/10.1109/CFIS.2019.8692152.

[10] B. Gaind, V. Syal, and S. Padgalwar, "Emotion Detection and Analysis on Social Media." arXiv, Jun. 12, 2019, https://doi.org/10.48550/arXiv.1901.08458.

[11] H. Raza, M. Faizan, A. Hamza, A. Mushtaq, and N. Akhtar, "Scientific Text Sentiment Analysis using Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 157–165, 2019, https://doi.org/10.14569/IJACSA.2019.0101222.

[12] F. M. Shah, A. S. Reyadh, A. I. Shaafi, S. Ahmed, and F. T. Sithil, "Emotion Detection from Tweets using AIT-2018 Dataset," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, Dhaka, Bangladesh, Sep. 2019, pp. 575–580, https://doi.org/10.1109/ICAEE48663.2019.8975433.

[13] A. F. A. Nasir *et al.*, "Text-based emotion prediction system using machine learning approach," *IOP Conference Series: Materials Science and Engineering*, vol. 769, no. 1, Oct. 2020, Art. no. 012022, https://doi.org/10.1088/1757-899X/769/1/012022.

[14] D. Seal, U. K. Roy, and R. Basak, "Sentence-Level Emotion Detection from Text Based on Semantic Rules," in *Information and Communication Technology for Sustainable Development*, 2020, pp. 423–430, https://doi.org/10.1007/978-981-13-7166-0_42.

[15] M. Karna, D. S. Juliet, and R. C. Joy, "Deep learning based Text Emotion Recognition for Chatbot applications," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Tirunelveli, India, Jun. 2020, pp. 988–993, https://doi.org/10.1109/ICOEI48184.2020.9142879.

[16] A. I. Saad, "Opinion Mining on US Airline Twitter Data Using Machine Learning Techniques," in *2020 16th International Computer Engineering Conference (ICENCO)*, Cairo, Egypt, Dec. 2020, pp. 59–63, https://doi.org/10.1109/ICENCO49778.2020.9357390.

[17] M. Suhasini and B. Srinivasu, "Emotion Detection Framework for Twitter Data Using Supervised Classifiers," in *Data Engineering and Communication Technology*, 2020, pp. 565–576, https://doi.org/10.1007/978-981-15-1097-7_47.

[18] D. Kher and K. Passi, "Multi-label Emotion Classification using Machine Learning and Deep Learning Methods," in *Proceedings of the 18th International Conference on Web Information Systems and Technologies*, Valletta, Malta, 2022, pp. 128–135, https://doi.org/10.5220/0011532400003318.

[19] A. Chowanda, R. Sutoyo, Meiliana, and S. Tanachutiwat, "Exploring Text-based Emotions Recognition Machine Learning Techniques on Social Media Conversation," *Procedia Computer Science*, vol. 179, pp. 821–828, Jan. 2021, https://doi.org/10.1016/j.procs.2021.01.099.

[20] M. Krommyda, A. Rigos, K. Bouklas, and A. Amditis, "An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media," *Informatics*, vol. 8, no. 1, Mar. 2021, Art. no. 19, https://doi.org/10.3390/informatics8010019.

[21] S. K. Bharti *et al.*, "Text-Based Emotion Recognition Using Deep Learning Approach," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, 2022, Art. no. 2645381, https://doi.org/10.1155/2022/2645381.

[22] L. Khan, A. Amjad, K. M. Afaq, and H. T. Chang, "Deep Sentiment Analysis Using CNN-LSTM Architecture of English and Roman Urdu Text Shared in Social Media," *Applied Sciences*, vol. 12, no. 5, Jan. 2022, Art. no. 2694, https://doi.org/10.3390/app12052694.

[23] M. M. Rahman and S. Shova, "Emotion Detection From Social Media Posts." arXiv, Feb. 11, 2023, https://doi.org/10.48550/arXiv.2302.05610.

[24] R. Ramanda and M. Affandes, "Emotion Classification Using Support Vector Machine," *Appissode: Application, Information System and Software Development Journal*, vol. 1, no. 1, pp. 15–19, Dec. 2023, https://doi.org/10.20823/vsx33w33.

[25] M. Dai, "Machine Learning Based Sentiment Analysis of Message on Twitter," *Highlights in Science, Engineering and Technology*, vol. 38, pp. 942–948, Mar. 2023, https://doi.org/10.54097/hset.v38i.5980.

[26] H. G. Wallbott and K. R. Scherer, "How universal and specific is emotional experience? Evidence from 27 countries on five continents," *Social Science Information*, vol. 25, no. 4, pp. 763–795, Dec. 1986, https://doi.org/10.1177/053901886025004001.

[27] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," in *Proceedings of the 12th International Workshop on Semantic Evaluation*, New Orleans, LA, USA, Mar. 2018, pp. 1–17, https://doi.org/10.18653/v1/S18-1001.

[28] H. E. Wynne and Z. Z. Wint, "Content Based Fake News Detection Using N-Gram Models," in *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, Munich, Germany, Dec. 2019, pp. 669–673, https://doi.org/10.1145/3366030.3366116.

[29] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," *ICML*, vol. 97, pp. 143–151, 1997.

[30] W. A. Qader, M. M. Ameen, and B. I. Ahmed, "An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges," in *2019 International Engineering Conference (IEC)*, Erbil, Iraq, Jun. 2019, pp. 200–204, https://doi.org/10.1109/IEC47844.2019.8950616.

[31] M. A. Kausar, S. O. Fageeri, and A. Soosaimanickam, "Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10849–10855, Jun. 2023, https://doi.org/10.48084/etasr.5854.

[32] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, Jan. 2011, https://doi.org/10.1504/IJDATS.2011.041335.

[33] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr. 2006, https://doi.org/10.1007/s10994-006-6226-1.

[34] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." arXiv, Feb. 08, 2020, https://doi.org/10.48550/arXiv.1909.11942.

[35] C. Y. Suen and L. Lam, "Multiple Classifier Combination Methodologies for Different Output Levels," in *Multiple Classifier Systems*, Cagliari, Italy, 2000, pp. 52–66, https://doi.org/10.1007/3-540-45014-9_5.

[36] B. Parhami, "Voting algorithms," *IEEE Transactions on Reliability*, vol. 43, no. 4, pp. 617–629, Sep. 1994, https://doi.org/10.1109/24.370218.

[37] A. Özçift, "Medical sentiment analysis based on soft voting ensemble algorithm," *Yönetim Bilişim Sistemleri Dergisi*, vol. 6, no. 1, pp. 42–50, Jun. 2020.

[38] L. Bottou, "Stochastic Gradient Descent Tricks," in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Springer, 2012, pp. 421–436.

[39] T. Gunasekaran and S. Kumar, "Data Classification Using Support Vector Machine," *Journal of Theoretical and Applied Information Technology*, vol. 12, no. 1.

[40] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 1, no. 10, pp. 185–189, Dec. 2012.