IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 January 2023, accepted 18 February 2023, date of publication 23 February 2023, date of current version 1 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3248506

RESEARCH ARTICLE

Emotion Detection From Micro-Blogs Using Novel Input Representation

FAHIM ANZUM, (Member, IEEE), AND MARINA L. GAVRILOVA, (Senior Member, IEEE)
Biometric Technologies Laboratory, Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada

Corresponding author: Fahim Anzum (fahim.anzum@ucalgary.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, in part by the NSERC Strategic Partnership Grant (SPG), and in part by the Innovation for Defense Excellence and Security Network (IDEaS).

ABSTRACT Emotion is a natural intrinsic state of mind that drives human behavior, social interaction, and decision-making. Due to the rapid expansion in the current era of the Internet, online social media (OSM) platforms have become popular means of expressing opinions and communicating emotions. With the emergence of natural language processing (NLP) techniques powered by artificial intelligence (AI) algorithms, emotion detection (ED) from user-generated OSM data has become a prolific research domain. However, it is challenging to extract meaningful features for identifying discernible patterns from the short, informal, and unstructured texts that are common on micro-blogging platforms like Twitter. In this paper, we introduce a novel representation of features extracted from user-generated Twitter data that can capture users' emotional states. An advanced approach based on Genetic Algorithm (GA) is used to construct the input representation which is composed of stylistic, sentiment, and linguistic features extracted from tweets. A voting ensemble classifier with weights optimized by a GA is introduced to increase the accuracy of emotion detection using the novel feature representation. The proposed classifier is trained and tested on a benchmark Twitter emotion detection dataset where each sample is labeled with either of the six classes: sadness, joy, love, anger, fear, and surprise. The experimental results demonstrate that the proposed approach outperforms the state-of-the-art classical machine learning-based emotion detection techniques, achieving the highest level of precision (96.49%), recall (96.49%), F1-score (96.49%), and accuracy (96.49%).

INDEX TERMS Affective computing, emotion detection, ensemble classifier, genetic algorithm, machine learning, natural language processing, online social media, social behavior.

I. INTRODUCTION

As technology continues to advance, the proliferation of user-generated content on online social media (OSM) platforms has made opinion mining an important domain of research. Nowadays, individuals are increasingly influenced by the innovative features and trends introduced by different OSM platforms. Over the last few years, these factors resulted in a dramatic increase in users' interest in different social media platforms [1]. According to the Digital 2022 Global Statshot Report [2], there were a staggering 4.62 billion active social media users globally in January 2022, a 10.10% increase from the previous year. Online social media platforms have become such a dominant force that, along with traditional face-to-face and electronic media communication, people's interactions are profoundly shaped by sharing their opinions, thoughts, and stories about global events online. In addition, these platforms provide access to a plethora of user-generated data that are leveraged by practitioners across various sectors for effective decision-making in business and technological intervention.

Online social media provides users with platforms and opportunities to express, communicate, and share their opinions, views, and thoughts. The user-generated social media content reveals valuable insights into people's emotional states that can inform a wide range of behavioral and psychological stances of an individual [3]. Emotion is essential in every aspect of a person's life, influencing decision-making, social relationships, and behavior. As a result, there has been a surge in research on the application of artificial intelligence

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

(AI) and machine learning (ML) techniques to infer users' emotional states from their social media content [4], [5]. Automatic emotion detection (ED) involves using natural language processing (NLP) techniques and machine learning algorithms to decipher an individual's emotional states [6]. In online social media platforms, people share content in various data formats such as text, image, video, graphics interchange format (GIF), and others that inform users' emotional, psychological, and personality traits. While emotion detection from facial expression [7], [8], and more recently gait [9] have been widely explored, detecting emotions from textual data is still an emerging research area.

In text-based emotion detection, natural language processing (NLP) techniques are employed to extract meaningful patterns from the text data, that are leveraged by ML algorithms to infer the users' emotions. Text-based emotion detection has important applications in various fields, including customer service, mental health support, opinion mining, and personalization [4]. It enables machines to recognize and understand the emotions expressed in text, which can improve communication and has the potential to enhance user experience. However, automatic emotion detection from OSM textual data is challenging due to the scarcity of publicly available datasets labeled with emotion categories, the unstructured nature of the user-generated data, and the need for user's privacy [5]. Moreover, emotion recognition is especially critical for user-generated tweets. They consist of short, informal, and unstructured text; incomplete, misspelled, and slang words; abbreviations, acronyms, and special characters, all of which require extensive text preprocessing and accurate feature extraction.

The previous research demonstrated that emotions can be predicted by training classical machine learning models with different language-based features extracted from the OSM text data [10]. The features can be extracted using bag of words (BoW) [11], N-gram [12], term frequency-inverse document frequency (TF-IDF) [13], Word2Vec [14], and GloVe [15], among others. Language-based features tend to capture information from the texts that is useful for inferring emotional cues [16]. While the majority of emotion detection systems utilize different linguistic characteristics, other features for determining emotions remain largely unexplored [17]. Moreover, to the best of our knowledge, the significance of stylistic markers and sentiment-related features has not been previously investigated in the context of social media text emotion detection. This paper aims to address the above gaps by introducing the following research questions:

1) Can a system based on NLP techniques be developed to accurately recognize emotions in short, informal, and unstructured texts?
2) Besides linguistic features, can any other features in the tweets show discernible patterns which can be used to infer users' emotional states?
3) Can the combination of distinctive feature types create a novel input feature representation that can accurately predict users' emotions?

The objective of this research is to design a novel system that combines linguistic (L), stylistic (S), and sentiment (SE)-based features extracted from tweets to accurately predict users' emotions. Thus, the paper makes the following contributions:

1) Evaluating the effectiveness of different text-based feature categories such as stylistic features, sentiment features, and linguistic features (e.g., BoW, TF-IDF, and Word Embeddings) in detecting emotions from tweets.
2) Proposing a novel input feature representation SSEL by utilizing the stylistic (S), sentiment (SE), and linguistic (L) features that are combined and compressed by employing a genetic algorithm (GA).
3) Proposing an ensemble prediction system consisting of XGBoost, random forest (RF), and support vector machine (SVM) classifiers for accurate emotion detection using the newly introduced stylistic-sentiment-linguistic (SSEL) features.
4) Demonstrating the superior performance of the proposed approach on tweets when compared against the other state-of-the-art ML-based emotion detection systems using a publicly available multi-class emotion detection dataset.

The rest of the paper is organized as follows. Section II presents an overview of the latest research in the domain of emotion detection from social networks. Section III details the proposed ED methodology based on an ensemble of XGBoost, RF, and SVM classifiers using the SSEL features. In Section IV, the performance of the proposed system is measured and compared against the state-of-the-art systems. The conclusion and future direction of this research are discussed in Section V.

II. LITERATURE REVIEW

An emotion detection system can either determine discrete emotion categories or interpret multi-dimensional emotion characteristics [18]. Using discrete emotion detection techniques, fine-grained emotion categories such as fear, anger, joy, sadness, disgust, surprise, depression, love, etc. are determined. Two of the most commonly used discrete emotion models are the Paul Ekman model [19] and the Robert Plutchik model [20]. The Paul Ekman model categorizes emotions based on six distinct classes, whereas the Robert Plutchik model differentiates emotions into eight primary emotions. On the contrary, multi-dimensional emotion models are used to understand the valence, arousal, and power of emotions [21]. These emotion models consider different emotions as dependent and associated with each other. Polarity, activation/deactivation state, and degree of emotions can be studied using multi-dimensional emotion models. Some of the commonly used multi-dimensional emotion models are Russell's 2D circumplex model [22], Plutchik's 2D wheel of emotion models [20], and Russell's 3D model [23].

Recently, text-based discrete emotion classification using machine learning has been widely explored by researchers. Sundaram et al. [25] proposed a classical ML-based emotion

TABLE 1. Summary of recent research discussed in Section II.

Paper	Year	Dataset	Proposed Methods
[24]	2022	Twitter	Trained an ensemble classifier (EC) - AdaBoost using the features extracted using TF-IDF and BoW. The tweets were categorized into five (happy, extremely happy, sad, extremely sad, and neutral) classes.
[25]	2021	Twitter	Proposed multi-class ML-based ED models using TF-IDF features where the tweets were categorized by six emotion classes. The proposed approach did not perform hyperparameter optimization of the classifiers.
[26]	2020	Twitter	A voting classifier (LR-SGD) was proposed for classifying emotions from textual tweets. The classifier was trained using TF-IDF features. However, the authors considered only happy and unhappy as emotion types.
[27]	2020	Twitter	Trained supervised ML models for classifying tweets into four emotions. The emotion classes were constructed based on only happy and unhappy classes and did not consider the other potential emotion categories.
[28]	2020	RAMAS [29]	Extracted the features from speech transcripts using BoW, Word2Vec, FastText, and BERT methods and trained classical ML models for predicting texts into four emotions, namely, happy, sad, angry, and neutral.
[30]	2015	Buscape [31]	Extracted 93 stylistic features and compared them with TF-IDF and Delta TF-IDF to build a sentiment classifier.

detection technique from tweets where the text features were extracted using TF-IDF. The authors used a publicly available emotion dataset where the tweets are labeled with six discrete emotion classes [32]. After extracting the features from the texts, the authors trained RF and SVM classifiers for emotion recognition. They also demonstrated that their proposed classical ML algorithm trained on TF-IDF features performed better when compared with the ontology-based approach, and deep learning-based ED system. Discrete emotion detection using classification ML-based approaches was further explored by Yousaf et al. [26]. In this paper, a voting classifier was designed that combined logistic regression (LR) and stochastic gradient descent (SGD). A similar feature representation was utilized by Suhasini and Srinivasu [27] where Naïve Bayes (NB) and k-nearest neighbor (KNN) algorithms were trained for ED from tweets. In this paper, the tweet messages were classified into four emotion categories, namely, happy-active, happy-inactive, unhappy-active, and unhappy-inactive. However, the dataset included only two major classes and disregarded the other potential emotion categories. While the above-mentioned works have made significant strides in identifying strategies for detecting discrete emotions using TF-IDF, the authors did not determine the impact of other language-based features.

In recent studies, researchers have explored the potential of using various combinations of TF-IDF and other linguistic features to extract contextual information from texts. Kavitha et al. [24] proposed an ensemble classifier (EC) based ED model and trained it using the features extracted by TF-IDF and BoW. In another research, Dvoynikova et al. [28] extended the feature space by comparing BoW, Word2Vec, FastText, and BERT methods on RAMAS corpus [29]. Although TF-IDF is a widely used method for extracting features from text and has shown promising results in recent emotion detection research, it has limitations in its ability to capture the semantic similarity between words [33]. Recently, transformer-based models have shown improved performance in detecting emotion from texts due to their ability to effectively capture semantic relationships between words and phrases. There also has been emerging research focused on the application of multimodal data for emotion detection in web-based contexts [34].

In addition to the language-based features, users' emotions can be conveyed through stylistic patterns such as the inclusion of special characters, emoticons, and punctuation [17]. Anchiêta et al. [30] extracted 93 features related to users' writing styles and compared these with TF-IDF and Delta TF-IDF to build a sentiment classification system. Using the extracted features, three classical ML models namely, SVM, Naïve Bayes, and J48 were trained for a binary (positive and negative) classification task. Other stylistic markers introduced in the recent literature are the length of the tweets in terms of characters and words, frequency of hashtags, frequency of shared hyperlinks, emoticons, ellipses, question marks, and exclamation marks [35], [36]. The recent research discussed above is summarized in Table 1.

In summary, unlike the previously developed methodologies, our proposed emotion detection approach introduces a novel feature representation that combines linguistic, stylistic, and sentiment features. This representation enables the detection of discernible patterns from the Twitter dataset and improves the performance of the trained ensemble prediction system compared to the existing classical ML-based ED systems. This research demonstrates the potential to enhance our understanding of emotions expressed through short, informal, and instructed tweets.

III. PROPOSED METHODOLOGY

In this paper, an ensemble prediction system is proposed for detecting users' emotions from their tweets using the novel SSEL feature representation. A tweet can be up to 280 characters in length and may include spaces, hyperlinks, hashtags, emoticons, and other elements [37]. Before extracting linguistic features, tweets are pre-processed by

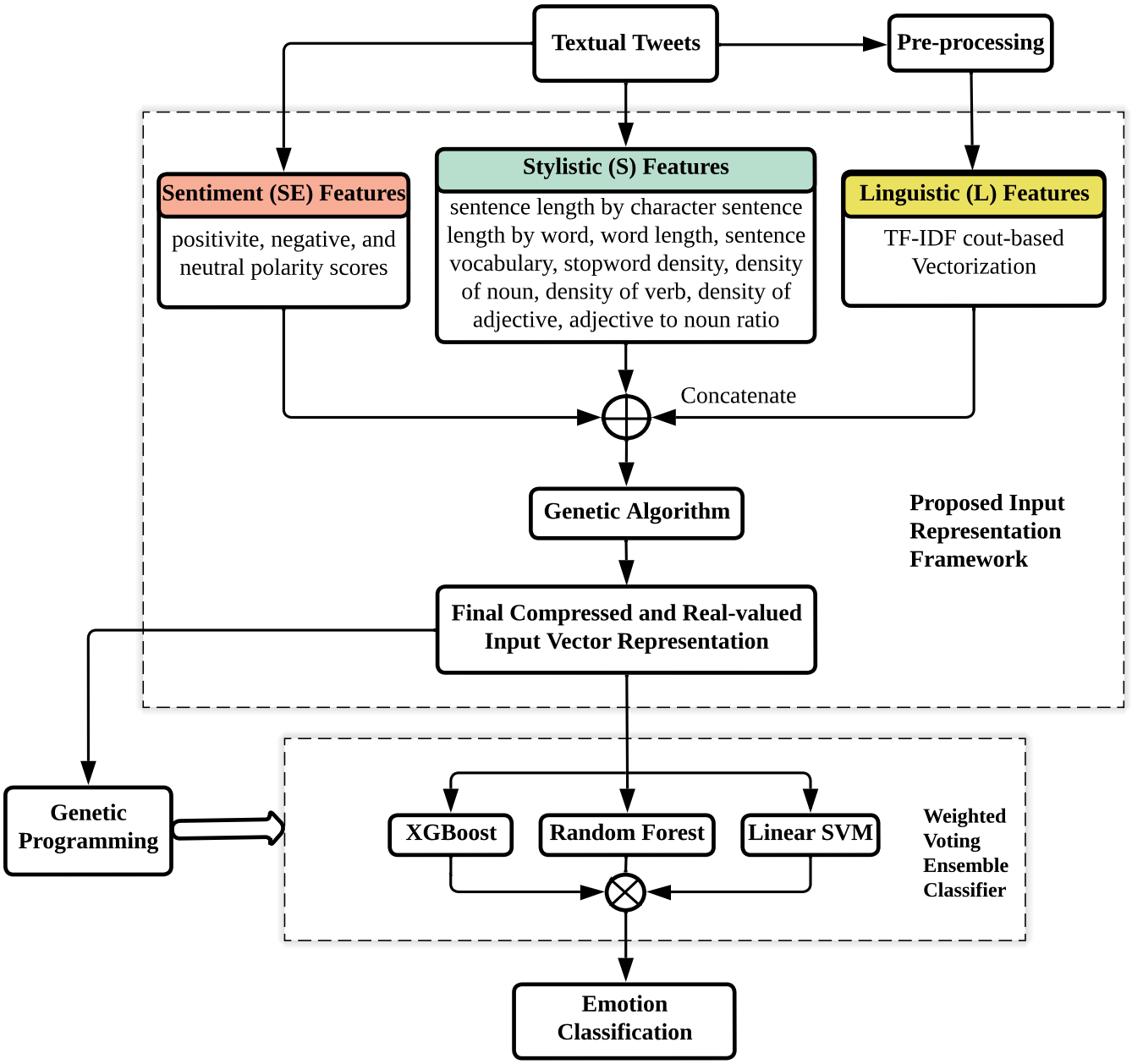

FIGURE 1. Proposed ensemble prediction system for tweet emotion detection using the SSEL input representation.

tokenization, lemmatization, lowercasing, and removing contents that do not contribute value to emotion detection. The pre-processed tweets are then converted into a real-valued vector representation using a count-based representation called term frequency-inverse document frequency (TF-IDF), which forms the linguistic (L) features. In addition, stylistic (S) and sentiment (SE) features are extracted from the raw tweets before the pre-processing step. These three feature representations are combined using a genetic algorithm (GA) and the reduced feature set is used as input for the proposed ensemble ED model. The major components of the proposed system are illustrated in Figure 1.

A. PRE-PROCESSING AND FEATURE EXTRACTION

During pre-processing, the data is first divided into smaller units called tokens through a process called tokenization [38]. As part of this process, it is common to remove stop words, which are frequently used words in a text document. However, research has shown that removing stop words can decrease the classification accuracy of sentiment and emotion detection models [39]. Therefore, in this work, we adapt the standard pre-processing pipeline and do not remove stop words during tokenization. Following that, the tokens are converted to their root form using the lemmatization process [40]. Furthermore, we remove hashtags, URLs, and punctuation

since these do not contribute value to the linguistic features for emotion detection [10].

To train a supervised ML classification model with texts, the data needs to be converted into a meaningful numerical representation which is known as vectorization [41]. To extract the linguistic features from the dataset, we use TF-IDF count-based vectorization technique to represent each tweet as a feature vector of normalized TF-IDF scores of the unigrams (sequence of one word in a sentence), bi-grams (sequence of two words in a sentence), and tri-grams (sequence of three words in a sentence). TF-IDF is a representation that captures the importance of a word in a document corpus [37]. It is a widely used language-based feature extraction technique in opinion mining, information retrieval, and text classification, among others. The TF-IDF is composed of two metrics: term frequency (TF) and inverse document frequency (IDF). TF computes the number of times term t appears in document d. IDF measures how common or rare a word is in the entire document set. The TF-IDF term is calculated as follows:

$$tf - idf(t, d) = tf(t, d) \times idf(t) \tag{1}$$

$$tf(t, d) = \frac{\text{Frequency of term } t \text{ in document } d}{\text{Total terms in document } d} \tag{2}$$

$$idf(t) = \log \frac{n}{df(t)} + 1 \tag{3}$$

where n is the total number of documents, $tf(t, d)$ is the total number of times t appears in document d, and $df(t)$ is the number of documents containing t.

Given the significance of different stylistic markers in capturing users' unique behavioral patterns, a set of stylistic features is extracted to construct the proposed feature representation as illustrated in Figure 1. Initially, from each tweet sample, the set of extracted stylistic features includes the sentence length (by character); sentence length (by word); word length; sentence vocabulary representing the ratio of different words to all words; the density of stop words; the density of noun; the density of verb; the density of adjective; the adjective to noun ratio; the density of different punctuation such as comma, period, colon, and semicolon; the density of question marks, and the density of exclamation marks. The occurrences of these stylistic markers are counted using regular expressions.

Furthermore, we extract the sentiment features from the tweets using a text sentiment analysis model known as VADER (Valence Aware Dictionary for sEntiment Reasoning) [42]. VADER is an OSM-oriented lexicon and rule-based sentiment analysis tool that maps lexical features to sentiment scores. The sentiment score of a text is obtained by summing up the intensity of each word in the text. The sentiment score is measured on a scale from -4 to $+4$, where -4 is the most negative and $+4$ is the most positive score. Zero is considered to be a neutral sentiment. However, the normalized sentiment score lies within the range of -1 to $+1$, from most negative to most positive. Moreover, VADER returns a compound score by normalizing the summation of positive, negative, and neutral sentiment scores. Therefore, we extract four sentiment-based features, namely, positive polarity score, negative polarity score, neutral polarity score, and compound polarity score from the tweets. Overall, the initial feature sets are the linguistic features (15247 dimensions), stylistic features (15 dimensions), and sentiment features (4 dimensions).

B. FEATURE COMBINATION AND DIMENSIONALITY REDUCTION USING GENETIC ALGORITHM

To combine the stylistic (S), sentiment (SE), and linguistic (L) features by reducing the dimension of the feature vector, a genetic algorithm (GA) is employed [17]. It provides a framework for combining distinct feature sets into one unified representation. The genetic algorithm is initialized with a set of parameters that control the behavior of the algorithm. The parameters include the population size (the number of feature sets that are evaluated in each generation), the crossover rate (the probability of combining two feature sets to form a new one), the mutation rate (the probability of introducing random changes to a feature set), and the selection method (the strategy used to choose which feature sets will be used to create the next generation). An initial population of feature sets is generated by randomly selecting a subset of the available features. The initial population can be represented by a matrix X of size $n \times m$, where n is the population size and m is the number of features. Each row of X represents a feature set, and each element $x[i, j]$ is a binary value that indicates whether feature j is included (1) or excluded (0) in feature set i. The fitness of each feature set in the population is then evaluated using a fitness criterion. We use the macro average F1-score of the emotion detection model as the fitness criterion for the feature selection. Therefore, the fitness criterion ($f(s)$) represents the macro average F1-score of the model using the selected feature set s. The best-performing feature sets from the current population are selected to form the next generation. In the next step, the selection, crossover, and mutation operations on the initial population of feature sets are performed to create a new generation of solutions. This process is repeated until the termination criterion of the genetic algorithm is met. After terminating the algorithm, the best-performing feature set from the final generation is obtained. This combines the stylistic (S), sentiment (SE), and linguistic (L) features and is referred to as the proposed SSEL features. The final input representation has the following dimensions: linguistic features (5000 dimensions), stylistic features (9 dimensions), and sentiment features (3 dimensions), which is approximately 67.17% smaller than the initially extracted feature size.

C. PROPOSED EMOTION DETECTION MODEL

After the final input feature representation is created, we train different machine learning models using the proposed SSEL features. In this research, genetic programming (GP) based approach [43] is leveraged to efficiently discover

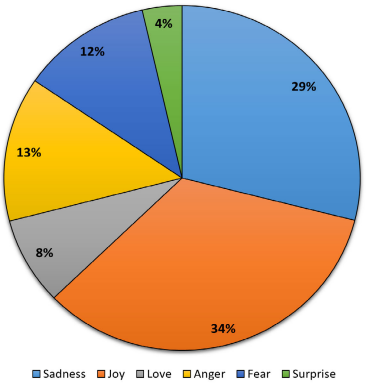

FIGURE 2. Distribution of emotion categories in the dataset.

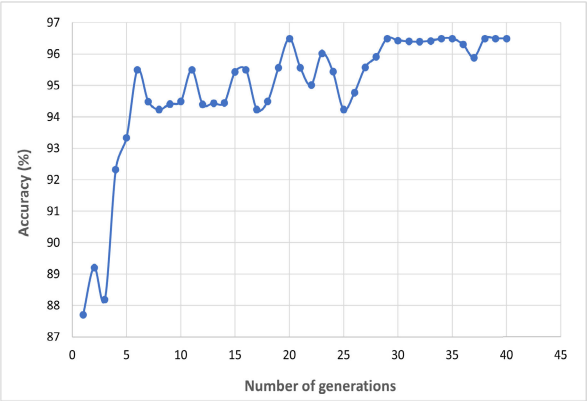

FIGURE 3. Comparing the accuracy of predictions using GA across different numbers of generations.

a top-performing classification model. The top-performing models identified by the GP technique are random forest (RF), extreme gradient boosting (XGBoost), and linear support vector machine (SVM) classifiers. Furthermore, an ensemble classifier is developed for the tweet emotion classification which is composed of RF, XGBoost, and linear SVM classifiers. Since the target output is a discrete emotion category, the ensemble uses a hard-voting strategy to determine the final prediction, which is computed from the prediction with the highest number of votes. An ensemble-based prediction system ensures multiple learners contribute to the overall accuracy through diversity when compared against a single classifier. The configurations of the classifiers' hyperparameters in the ensemble system are fine-tuned using grid search. Since the performance of each individual classification model differs depending on certain factors such as the distribution, the decision weight of each classifier can be tuned to produce a robust average ensemble. We tune the decision weights of each classifier in the ensemble model using GA. Unlike in the feature dimensionality reduction using GA, here the chromosomes are the weights of each component classifier in the ensemble. The GA is run through 40 generations with a population of 50, a crossover rate of 0.7, and a mutation rate of 0.2, which were empirically found to reach the best ensemble weight.

D. EVALUATION METRICS

To evaluate the performance of the proposed ED system, the commonly used performance evaluation metrics such as accuracy, precision, recall, and F1-score are used. These metrics are based on how well the trained ML model is able to predict the actual class of a given sample. If the model correctly predicts the positive class, this is called a true positive (TP). If the actual class is positive but the model predicts it as negative, this is called a false negative (FN). If both the actual and predicted classes are negative, this is a true negative (TN). If the actual class is negative but the model predicts it as positive, this is a false positive (FP). Accuracy measures the overall proportion of correct predictions. Precision is the percentage of positively labeled predictions that are actually positive, while recall is the percentage of positive classes that the model correctly detects. Although precision and recall provide useful insights, it is convenient to express the balance between them by using the F1-score, which is the weighted harmonic mean of precision and recall. Accuracy, precision, recall, and F1-score are computed as follows [26]:

$$Accuracy = \frac{\text{No. of correctly classified predictions}}{\text{Total predictions}} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

IV. EXPERIMENTS AND RESULTS

All the experiments were conducted on a Windows-based operating system, with an Intel Core-i7-10750H CPU, 16GB of memory, and an NVIDIA GeForce RTX 3060 GPU.

A. DATASET DESCRIPTION

In this research, a publicly available Twitter emotion dataset is used that contains 20,000 tweet samples, and each is labeled with one of the six discrete emotion classes: sadness, joy, love, anger, fear, and surprise [32]. The dataset is also available for emotion detection research on Kaggle [44]. As shown in Figure 2, the dataset is highly imbalanced and requires pre-processing for effective learning. Therefore, we use SMOTE (synthetic minority oversampling) technique to balance the classes using the combination of oversampling and undersampling [45]. The dataset is then divided into 80:20 for training and testing, following a stratified sampling approach to preserve the proportions of different classes in the sample.

B. HYPERPARAMETER SETTINGS

To implement the GA, we use a crossover rate of 0.70 and a mutation rate of 0.20 that were empirically determined to converge to a stable set of solutions. Figure 3 depicts that

TABLE 2. Experimental results demonstrating the effectiveness of different types of features extracted from tweets for emotion detection.

Input Representation	Classifier	Precision	Recall	F1-Score	Accuracy
TF-IDF Vectorization	Random Forest	0.87	0.81	0.84	0.85
	XGBoost	0.87	0.87	0.87	0.87
	SVM	0.87	0.71	0.81	0.80
Bag of Words	Random Forest	0.81	0.80	0.80	0.80
	XGBoost	0.80	0.80	0.80	0.80
	SVM	0.78	0.79	0.79	0.79
Word2Vec	Random Forest	0.78	0.79	0.78	0.79
	XGBoost	0.77	0.78	0.77	0.78
	SVM	0.77	0.78	0.78	0.79
Stylistic features	Random Forest	0.50	0.50	0.50	0.50
	XGBoost	0.43	0.43	0.43	0.43
	SVM	0.42	0.43	0.43	0.43
Sentiment features	Random Forest	0.43	0.44	0.44	0.44
	XGBoost	0.42	0.45	0.43	0.43
	SVM	0.41	0.42	0.42	0.42
Stylistic + Sentiment features	Random Forest	0.68	0.68	0.68	0.68
	XGBoost	0.69	0.69	0.69	0.69
	SVM	0.69	0.69	0.69	0.69
TF-IDF + Stylistic + Sentiment features	Random Forest	0.90	0.91	0.90	0.90
	XGBoost	**0.91**	**0.90**	**0.90**	**0.90**
	SVM	0.89	0.90	0.89	0.89

the iterative algorithm makes reliable predictions when the generation count reaches 20. The proposed ED system is an ensemble of XGBoost, RF, and SVM classifiers. For each classifier, the hyperparameters are determined using the grid search method. The number of estimators of 350 was used for both the XGBoost and RF classifiers. For the linear SVM classifier, we used the regularization parameter C = 10.

C. EXPERIMENTAL RESULTS WITH DIFFERENT INPUT REPRESENTATIONS

In the first set of experiments, we evaluate the effectiveness of different language-based feature extraction techniques: TF-IDF vectorization, bag of words (BoW), and Word2Vec in the emotion detection system. These are some of the most widely used and emerging NLP techniques to extract language-based features from texts for opinion-mining research. Subsequently, we examine the influence of the extracted stylistic and sentiment-based features both independently and combined on the performance of the explored ED models. Furthermore, we investigate if combining the initially extracted linguistic, stylistic, and sentiment features leads to an improvement in the model's performance. Using these different input representations, we train some classical ML models and identify that the XGBoost and RF classifiers performed reasonably well. The primary objectives of these experiments are to identify the significance of different types of features independently and to gain insight into the classification models.

Based on the experimental result, TF-IDF is determined to be the most effective technique to extract the discernible linguistic patterns from the dataset. While TF-IDF reflects the importance of each term in the tweets representing emotions, BoW representation only checks whether a term is present in a corpus or not and assigns equal weight to each term. Therefore, the emotion detection models using TF-IDF input representation outperform the models trained using BoW representation. Unlike TF-IDF and BoW, Word2Vec generates dense and low-dimensional vectors (also called ''embeddings'') for words in a large text dataset that capture the semantic relationships between words. However, since the dataset is relatively small, Word2Vec cannot learn effective representations of the words in the dataset. On the other hand, TF-IDF does not necessarily require a large dataset to learn effective representations. Moreover, depending on the nature of the emotion detection task, it is more important to capture the importance of individual words rather than the relationship between them. Therefore, our experimental results demonstrate that the emotion detection models trained with TF-IDF input representation outperform the models trained with Word2Vec representations. The experimental results also show that although the stylistic and sentiment-based features are not effective independently for emotion detection, the combination of these features achieves performance improvement. Finally, since TF-IDF is determined as the best-performing linguistic feature representation, a combination of the linguistic, stylistic, and sentiment-based features is investigated which improves the overall performance of the experimented emotion detection models. Note, that the initially extracted features suffer from high dimensionality, which is mitigated in the proposed system by employing a genetic algorithm. The experimental results with different input representations are presented in Table 2.

D. PERFORMANCE OF THE PROPOSED MODEL

As discussed in Section III, the novel input feature representation is generated using a GA-based approach. The 15,266-D input vector is transformed into a 5012-D vector, reducing the input feature size by approximately 67.17%. Table 3 shows the list of representative stylistic features selected by the algorithm from the initially extracted 15 stylistic features.

TABLE 3. List of selected stylistic features after employing the feature reduction technique.

Feature Name	Description	Selected
Sentence Length (by character)	Calculates the average length of a tweet in characters	Yes
Sentence Length (by word)	Calculates the average length of a tweet in words	Yes
Word Length	Calculates the average characters per word	Yes
Density Comma	Calculates the ratio of comma (,) in tweets	No
Density Period	Calculates the ratio of period (.) in tweets	No
Density Colon	Calculates the ratio of colon (:) in tweets	No
Density Semi-color	Calculates the ratio of semi-colon (;) in tweets	No
Density Question Marks	Calculates the ratio of question marks (?) in tweets	No
Density Exclamation Marks	Calculates the ratio of exclamation (!) in tweets	No
Vocabulary Sentence	Calculates the ratio of different words to all words in tweets	Yes
Density Stopwords	Calculates the ratio of stopwords to all words in tweets	Yes
Density Noun	Calculates the ratio of parts of speech (noun) to all words in tweets	Yes
Density Adjective	Calculates the ratio of parts of speech (adjective) to all words in tweets	Yes
Density Verb	Calculates the ratio of parts of speech (verb) to all words in tweets	Yes
Adjective to Noun	Calculates the ratio of adjective to noun in tweets	Yes

TABLE 4. Performance comparison of the proposed emotion detection model with different ML classifiers using the SSEL input representation.

Classifier	Precision	Recall	F1-Score	Accuracy
KNN	38.96%	38.91%	38.93%	38.91%
DT	86.69%	86.65%	86.66%	86.65%
RF	91.47%	91.38%	91.40%	91.39%
XGBoost	93.00%	93.01%	93.00%	93.01%
SVM	93.00%	93.98%	93.49%	93.20%
Proposed model: RF, XGBoost, and SVM	**96.49%**	**96.49%**	**96.49%**	**96.49%**

We observe that certain stylistic features such as the number of question marks and exclamation marks; the density of punctuation do not provide high predictive value for emotion detection. Moreover, the selected sentiment-based features from the initially extracted features are the normalized positive, negative, and neutral polarity scores.

The performance of our proposed emotion detection model using the novel input feature representation is compared against different ML models: XGBoost, RF, SVM, decision tree (DT), and K-nearest neighbors (KNN). Table 4 shows that our proposed approach achieves the highest performance in terms of precision, recall, F1-score, and accuracy. The proposed ensemble of XGBoost, RF, and SVM achieves the highest precision (96.49%), recall (96.49%),

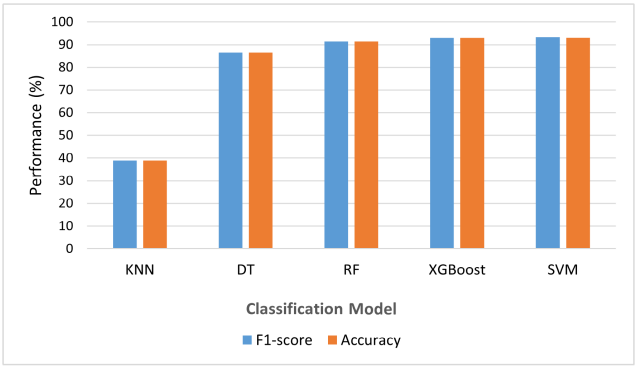

FIGURE 4. Performance of the experimented ML classifiers for emotion detection using the SSEL input representation.

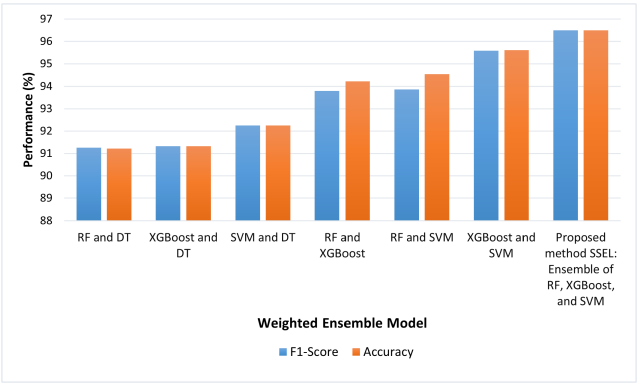

FIGURE 5. Performance comparison of the proposed emotion detection model with weighted ensemble classifiers using the SSEL input representation.

F1-score (96.49%), and accuracy (96.49%) followed by the 93% precision, 93.98% recall, 93.49% F1-score, and 93.20% accuracy achieved by the SVM classifier. The XGBoost classifier trained on the proposed input features achieves 93% precision, 93.01% recall, 93% F1-score, and 93.01% accuracy. Among the compared ML models, KNN achieves the lowest precision (38.96%), recall (38.91%), F1-score (38.93%), and accuracy (38.91%). While RF achieves reasonably well precision, recall, F1-score, and accuracy by over 91%, the decision tree (DT) classifier achieves 86.69% precision, 86.65% recall, 86.66% F1-score, and 86.65% accuracy. Figure 4 illustrates the emotion detection performance achieved by the experimented machine learning models in terms of F1-score and accuracy.

We also compare our proposed ensemble model's performance against other ensemble classifiers as presented in Table 5. According to the insights from genetic programming and the performance comparison of individual classifiers, the best-performing classifiers are SVM, RF, and XGBoost. Overall, the proposed weighted ensemble of RF, XGBoost, and SVM classifiers outperforms the explored ensemble classifiers. Performance comparison of our proposed emotion

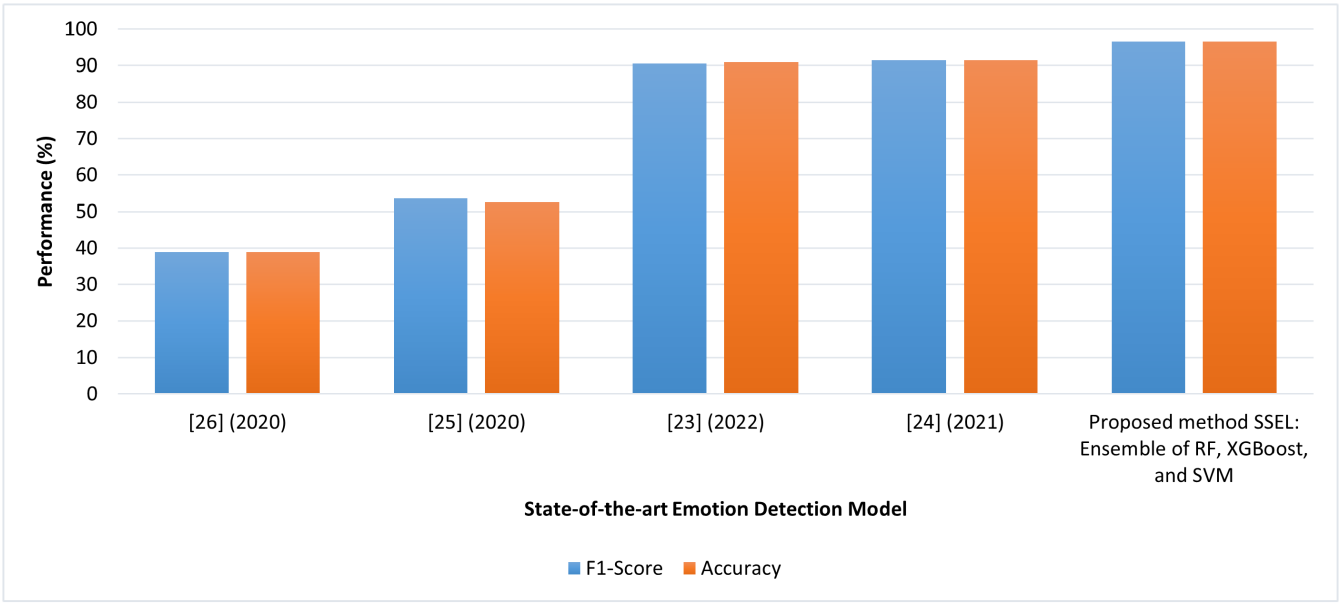

FIGURE 6. Performance comparison of the proposed emotion detection model with other state-of-the-art methods using the SSEL input representation.

TABLE 5. Performance comparison of the proposed emotion detection model with weighted ensemble classifiers using the SSEL input representation.

Ensemble Model	Precision	Recall	F1-Score	Accuracy
RF and DT	91.27%	91.24%	91.25%	91.22%
XGBoost and DT	91.32%	91.32%	91.32%	91.32%
SVM and DT	92.25%	92.25%	92.25%	92.25%
RF and XGBoost	94.27%	93.20%	93.79%	94.22%
RF and SVM	94.40%	93.33%	93.86%	94.55%
XGBoost and SVM	95.58%	95.58%	95.58%	95.61%
Proposed model: RF, XGBoost, and SVM	**96.49%**	**96.49%**	**96.49%**	**96.49%**

TABLE 6. Performance comparison of the proposed emotion detection model with other state-of-the-art methods using the SSEL input representation.

ED Methods	Precision	Recall	F1-Score	Accuracy
[27] (2020)	38.96%	38.91%	38.93%	38.91%
[26] (2020)	75.30%	41.47%	53.48%	52.47%
[24] (2022)	90.51%	90.31%	90.44%	90.89%
[25] (2021)	91.47%	91.38%	91.40%	91.42%
Proposed model: RF, XGBoost, and SVM	**96.49%**	**96.49%**	**96.49%**	**96.49%**

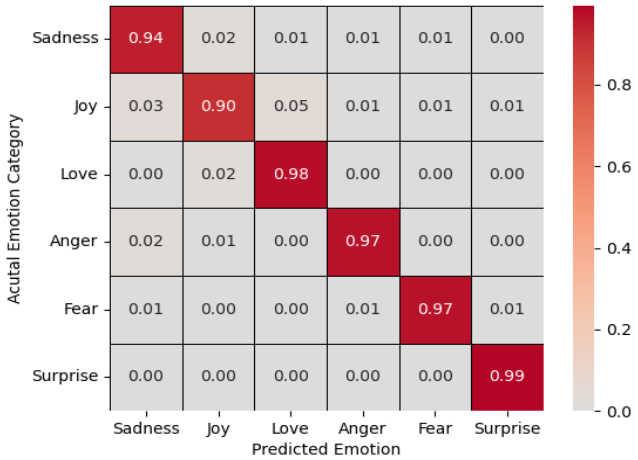

FIGURE 7. Confusion matrix showing the emotion prediction performance of the proposed emotion detection model.

detection model with the experimented ensemble classifiers is illustrated in Figure 5.

Table 6 and Figure 6 show the performance comparison of our proposed system with other leading works in this domain. We re-implemented the existing methods by training on the same dataset that our proposed model is trained on for a valid and fair comparison. From this comparison, we show that the proposed ensemble classifier trained on our novel combination of stylistic, sentiment, and linguistic features outperforms the current emotion detection methods on a benchmark Twitter dataset. Moreover, the proposed ED model's emotion prediction performance on the test data is shown in Figure 7 by illustrating the confusion matrix in the form of a heatmap. The confusion matrix demonstrates that the proposed model performs very well in predicting emotions from unseen data which further validates the reliability of the proposed system.

In summary, through the series of experiments, we establish that the proposed SSEL input feature representation, which is composed of linguistic features calculated by the TF-IDF vectorization, stylistic features, and sentiment-based features, can serve as discriminative feature sets for automatic detection of emotions from users' tweets. Moreover, a genetic

algorithm (GA) is proposed to combine the distinct dataset to be used as input to the ensemble classifier. Experimental results indicate that the proposed weighted average ensemble technique, using the linear support vector machine, XGBoost, and random forest classifiers can be successfully used to learn discernible patterns from the extracted features and can accurately detect emotions from tweets.

V. CONCLUSION AND FUTURE WORK

This paper proposed a novel input representation SSEL by combining the stylistic (S), sentiment (SE), and linguistic (L) features extracted from tweets for representing users' emotional states. A genetic algorithm was leveraged to combine and compress the distinct feature sets into one unified representation. This paper also presented a novel combination of linear support vector machine, XGBoost, and random forest as a weighted average voting classifier for detecting emotions by classifying the tweets into six independent categories using the proposed input representation. This research showed that the stylistic and sentiment attributes when combined with the language-based input representation can capture discernible patterns in the tweets that are highly predictive for emotion detection.

The proposed emotion detection approach was compared with five independent classical ML classifiers, six different combinations of weighted ensemble voting classifiers, and four recent state-of-the-art ML-based ED techniques by employing the proposed input representations extracted from a publicly available Twitter emotion detection dataset. The experimental results show that our proposed ED system outperforms all the recent approaches considering each of the performance evaluation metrics and establishes a new performance benchmark for the experimented dataset.

In future work, we plan to explore the use of different categorical and multi-dimensional emotion models to capture a larger emotion spectrum, as well as investigate the performance of our approach on different user groups.

REFERENCES

[1] P. A. Rauschnabel, P. Sheldon, and E. Herzfeldt, "What motivates users to hashtag on social media?" *Psychol. Marketing*, vol. 36, no. 5, pp. 473–488, May 2019.

[2] S. Kemp. (2022). *Digital 2022: Digital Adoption Doubled Over the Past Decade*. [Online]. Available: https://datareportal.com/reports/digital-2022-digital-adoption-doubled-over-the-past-decade

[3] S. Kusal, S. Patil, K. Kotecha, R. Aluvalu, and V. Varadarajan, "AI based emotion detection for textual big data: Techniques and contribution," *Big Data Cognit. Comput.*, vol. 5, no. 3, p. 43, Sep. 2021.

[4] M. L. Gavrilova, F. Anzum, A. Hossain Bari, Y. Bhatia, F. Iffath, Q. Ohi, M. Shopon, and Z. Wahid, "A multifaceted role of biometrics in online security, privacy, and trustworthy decision making," in *Breakthroughs in Digital Biometrics and Forensics*. Cham, Switzerland: Springer, 2022, pp. 303–324.

[5] F. Anzum, A. Z. Asha, and M. L. Gavrilova, "Biases, fairness, and implications of using AI in social media data mining," in *Proc. Int. Conf. Cyberworlds (CW)*, Sep. 2022, pp. 251–254.

[6] N. Andalibi and J. Buss, "The human in emotion recognition on social media: Attitudes, outcomes, risks," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–16.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022.

[8] Y. Luo, M. L. Gavrilova, and P. S. P. Wang, "Facial metamorphosis using geometrical methods for biometric applications," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, no. 3, pp. 555–584, May 2008.

[9] Y. Bhatia, A. H. Bari, and M. Gavrilova, "A LSTM-based approach for gait emotion recognition," in *Proc. IEEE 20th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI*CC)*, Oct. 2021, pp. 214–221.

[10] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–19, Dec. 2021.

[11] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS ONE*, vol. 15, no. 5, May 2020, Art. no. e0232525.

[12] J. Fürnkranz, "A study using *n*-gram features for text categorization," *Austrian Res. Inst. Artif. Intell.*, vol. 3, no. 1998, pp. 1–10, 1998.

[13] S. Qaiser and R. Ali, "Text mining: Use of TF-IDF to examine the relevance of words to documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, Jul. 2018.

[14] K. W. Church, "Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017.

[15] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[16] V. Anusha and B. Sandhya, "A learning based emotion classifier with semantic text processing," in *Advances in Intelligent Informatics*. Springer, 2015, pp. 371–382.

[17] K. P. Kumar and M. L. Gavrilova, "Latent personality traits assessment from social network activity using contextual language embedding," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 2, pp. 638–649, Apr. 2022.

[18] P. S. Sreeja and G. Mahalakshmi, "Emotion models: A review," *Int. J. Control Theory Appl.*, vol. 10, no. 8, pp. 651–657, 2017.

[19] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*. Hoboken, NJ, USA: Wiley, 1999, pp. 45–60.

[20] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories Emotion*. Amsterdam, The Netherlands: Elsevier, 1980, pp. 3–33.

[21] D. Kollias and S. Zafeiriou, "Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the OMG in-the-wild dataset," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 595–606, Jul. 2021.

[22] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.

[23] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.

[24] D. Kavitha, P. P. Reddy, and K. V. Rao, "Emotion recognition in tweets using optimized ensemble classifiers," in *Proc. 7th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jun. 2022, pp. 1728–1731.

[25] V. Sundaram, S. Ahmed, S. A. Muqtadeer, and R. Ravinder Reddy, "Emotion analysis in text using TF-IDF," in *Proc. 11th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2021, pp. 292–297.

[26] A. Yousaf, M. Umer, S. Sadiq, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Emotion recognition by textual tweets classification using voting classifier (LR-SGD)," *IEEE Access*, vol. 9, pp. 6286–6295, 2021.

[27] M. Suhasini and B. Srinivasu, "Emotion detection framework for Twitter data using supervised classifiers," in *Proc. 3rd Data Eng. Commun. Technol. (ICDECT-2K19)*. Singapore: Springer, 2020, pp. 565–576.

[28] A. Dvoynikova, O. Verkholyak, and A. Karpov, "Emotion recognition and sentiment analysis of extemporaneous speech transcriptions in Russian," in *Proc. Int. Conf. Speech Comput.* Cham, Switzerland: Springer, 2020, pp. 136–144.

[29] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian multimodal corpus of dyadic interaction for affective computing," in *Proc. Int. Conf. Speech Comput.* Cham, Switzerland: Springer, 2018, pp. 501–510.

[30] R. T. Anchiêta, F. A. R. Neto, R. F. D. Sousa, and R. S. Moura, "Using stylometric features for sentiment classification," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, 2015, pp. 189–200.

[31] N. Hartmann, L. Avanço, P. B. Filho, M. S. Duran, M. D. G. V. Nunes, T. Pardo, and S. Aluísio, "A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 3865–3871.
[32] E. Saravia, H.-C.-T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3687–3697.
[33] M. Kumari, A. Jain, and A. Bhatia, "Synonyms based term weighting scheme: An extension to TF. IDF," *Proc. Comput. Sci.*, vol. 89, pp. 555–561, 2016.
[34] A. Kumar and G. Garg, "Sentiment analysis of multimodal Twitter data," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 24103–24119, Sep. 2019.
[35] G. Roffo, C. Segalin, A. Vinciarelli, V. Murino, and M. Cristani, "Reading between the turns: Statistical modeling for identity recognition and verification in chats," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 99–104.
[36] S. Spina, "Role of emoticons as structural markers in Twitter interactions," *Discourse Process.*, vol. 56, no. 4, pp. 345–362, May 2019.
[37] S. N. Tumpa and M. L. Gavrilova, "Score and rank level fusion algorithms for social behavioral biometrics," *IEEE Access*, vol. 8, pp. 157663–157675, 2020.
[38] A. Amin, I. Hossain, A. Akther, and K. M. Alam, "Bengali VADER: A sentiment analysis approach using modified VADER," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6.
[39] H. Saif, M. Fernández, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 810–817.
[40] Y. Bao, C. Quan, L. Wang, and F. Ren, "The role of pre-processing in Twitter sentiment analysis," in *Proc. Int. Conf. Intell. Comput.* Cham, Switzerland: Springer, 2014, pp. 615–624.
[41] A. K. Singh and M. Shashi, "Vectorization of text documents for identifying unifiable news articles," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 1–6, 2019.
[42] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and VADER sentiment," in *Proc. Int. Multiconf. Eng. Comput. Scientists*, vol. 122, 2019, p. 16.
[43] J. Ma and X. Gao, "Designing genetic programming classifiers with feature selection and feature construction," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106826.
[44] P. Pandey. (2021). *Emotion Dataset for Emotion Recognition Tasks*. [Online]. Available: https://www.kaggle.com/datasets/parulpandey/emotion-dataset
[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 28, pp. 321–357, Jun. 2006.

FAHIM ANZUM (Member, IEEE) received the B.Sc. degree in computer science and engineering (CSE) from the Military Institute of Science and Technology (MIST), Dhaka, Bangladesh, and the M.Sc. degree in computer science from the University of Calgary (UCalgary), Calgary, AB, Canada, in 2021, where he is currently pursuing the Ph.D. degree in computer science. During his master's study, he achieved the Mitacs-Accelerate Graduate Research Scholarship and worked with Suncor Energy Inc., Canada, as a Research Intern. He conducted research under the supervision of Dr. Mario Costa Sousa and Dr. Usman Alim with UCalgary. He is also an Alberta Innovates and Eyes High Doctoral Scholar and conducting research under the supervision of Dr. Marina L. Gavrilova with the Biometric Technologies Laboratory (BT Lab). Before joining UCalgary, he was a Lecturer (currently on study leave) with the Department of Computer Science and Engineering, United International University (UIU), Bangladesh. His research interests include the application of machine learning (ML) and deep learning (DL) in areas, such as emotion detection and user-behavior analysis in online social media, natural language processing, and data mining.

MARINA L. GAVRILOVA (Senior Member, IEEE) is currently a Full Professor with the University of Calgary and an international expert in the areas of biometric security, machine learning, pattern recognition, and information fusion. She directs the Biometric Technologies Laboratory and has published over 300 books, conference proceedings, and peer-reviewed articles. Her professional excellence was recognized by the Canada Foundation for Innovation, the Killiam Foundation, U Make a Difference Award, and the Order of the University of Calgary. She is the founding Editor-in-Chief of *Transactions on Computational Sciences* (Springer). She serves on the Editorial Board for the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SCIENCES, IEEE ACCESS, *The Visual Computer*, *Sensors*, and the *International Journal of Biometrics*.