Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19032-y

HDEL: a hierarchical deep ensemble approach for text-based emotion detection

Shivani Vora[1] · Rupa G. Mehta[2]

Received: 20 October 2023 / Revised: 7 February 2024 / Accepted: 22 March 2024

Abstract

Emotion detection from social media data plays a crucial role in studying societal emotions concerning different events, aiding in predicting the reactions of specific social groups. However, it is complex to automatically extract implicit emotional information from noisy social media text data. This study introduces the Hierarchical Deep Ensemble Learning (HDEL) system to identify emotions in text data. The proposed HDEL model utilizes BiLSTM (Bidirectional Long Short-Term Memory), CNN (Convolutional Neural Network), BiGRU (Bidirectional Gated Recurrent Unit), and RCNN (Recurrent Convolutional Neural Network) in the first level of its hierarchy. The predicted probabilities of the four models are embedded with input data to prepare the intermediate hybrid data. This hybrid data is fed to the next layer of the proposed system, which utilizes a Random Forest (RF) algorithm to predict the emotion. The proposed approach is tested using three emotion datasets: the WASSA-2017 Emotion Intensity (EmoInt) dataset, the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset, and the CrowdFlower (CF) dataset. EmoInt and ISEAR are clean and balanced, while CF is noisy and imbalanced. The results are compared with various state-of- the-art Machine Learning models. The outperforming results depict the superiority of the proposed approach.

Keywords Emotion Classification · Deep Learning · Ensemble Learning · Random Forest

1 Introduction

Emotions are part of human life and play an essential role in decision-making. Emotion classification can significantly contribute to medicine, sociology, psychology, and more creative areas such as human–computer interaction. It has evolved as the complex

✉ Shivani Vora
shivani.vora@ckpcet.ac.in

Rupa G. Mehta
rgm@coed.svnit.ac.in

[1] Information Technology Department, C.K.Pithawala College of Engineering and Technology, Surat, Gujarat, India

[2] Department of Computer Science and Engineering, SVNIT, Surat, Gujarat, India

Published online: 17 April 2024

problem in Natural Language Processing (NLP) applications. This study addresses the intricate challenge of unimodal emotion classification from textual data, a task that extends beyond the realm of sentiment analysis [1] to encompass a more detailed analysis of emotions [2]. Despite advancements in other modalities like speech and facial expression, text-based emotion identification remains a compelling area of research due to machines' struggle with interpreting context, particularly in contrast to human capabilities. Even in recent times, identification of emotion in a text has gained popularity due to its numerous promising applications in Artificial intelligence [3], Political science [4], Leveraging psychology and emotion detection to personalize recommendations based on user reviews [5], Human–computer interaction [6], Suicide prevention or evaluating the well-being of a community [7], prediction of stock price [8], and many more.

Emotion identification from text is often formulated as to find emotion of different categories ('anger', 'joy','fear','sadness' etc.) and solved using lexicon-based, machine learning, deep learning or hybrid approaches. Study showed that lexicon-based processes depend on linguistic features such as dictionaries, a bag of words, ontologies, and linguistic rules. In contrast, Machine Learning (ML) strategies use algorithms such as support vector machines, Naive Bayes classifier, logistic regression, and artificial neural networks, among others. The limitations of lexicon-based methods concerning scalability and domain customization can be overcome by ML approaches. Moreover, it can also learn implicit signals of emotions. The conventional ML algorithms required heavy feature engineering whereas deep learning algorithms learn high-level features from data in an incremental manner. This eliminates the need for domain expertise and hard-core feature extraction.

Recent studies indicate a growing use of Ensemble Learning (EL) methods [9–11] and Deep Learning (DL) algorithms [12–14] to enhance emotion detection tasks. EL is a technique that combines multiple machine learning models to improve the generalization performance of the overall system whereas Deep learning (DL) technique is a powerful subset of machine learning that automatically learns and extract complex features from the input data. Moreover, studies have shown that ensemble learning methods derive benefits from a degree of classifier heterogeneity. This diversity aids in reducing variance-error while maintaining a low bias-error, ultimately enhancing the overall performance of the ensemble system. The efficacy of these methods relies on employing diverse classifiers to produce uncorrelated errors, thereby improving prediction accuracy. Achieving optimal predictive performance requires sufficient diversity among the involved classifiers. This diversity primarily stems from the influence of both the classification algorithm and the training data used to create each classifier. Therefore, diversification can be achieved by varying training samples, utilizing different classification algorithms, exploring diverse network topologies, or tuning hyperparameters within neural networks.

The primary objective of this research is to develop a novel approach termed Hierarchical Deep Ensemble Learning (HDEL), leveraging a hierarchy of deep learning models and ensemble techniques for accurate emotion classification.

This study positions itself within the evolving landscape of emotion identification research, distinguishing itself by focusing on text-based content. As we delve into the proposed framework, through rigorous experimentation and hyperparameter tuning across diverse deep learning models, we have identified and select four distinct deep neural network architectures, including Bidirectional LSTM (Long Short Term Memory), 1D-CNN (Convolutional Neural Network), Bidirectional GRU (Gated Recurrent Unit), and RCNN (Recurrent Convolutional Neural Network). All models are trained on a dataset using the GloVe pre-trained word embedding model that is a pre-trained word embedding model with 200 dimensions and is learned from tweets [15]. Post-training, all four heterogeneous models produced probability scores for each emotion class during predictions on the test dataset. Incorporating these probabilities as additional features augmented the original test dataset, enriching it with insights into the model's confidence levels in its predictions. The extended test dataset served as input for the Random Forest ensemble classifier for prediction of emotions in the given text.

Figure 1 provides a visual representation of the HDEL system, illustrating the integration of four hyper-tuned deep neural network architectures.

The proposed work makes below major contributions:

- The research introduces a novel hierarchical approach to emotion classification that effectively captures complex emotional relationships in text. By combining hierarchical modeling with ensemble techniques, it achieves impressive f-scores on multiple datasets, outperforming existing methods.
- This approach leverages the power of hierarchical modeling to handle non-linear emotional dependencies, alleviates class imbalance issues, and excels in multi-class emotion classification. It shows promise for improving real-world applications dependent on accurate emotion identification from text.

This study demonstrates the effectiveness of our HDEL model in capturing complex emotional relationships and achieving high f-scores. However, further research could explore ways to address potential biases in the pre-trained embeddings and optimize the hyperparameters of the individual deep learning models within the ensemble. Investigating novel optimization frameworks suggested in [16–18] and hybrid approaches with complementary machine learning paradigms could potentially lead to even greater accuracy and robustness for emotion analysis tasks.

The remainder of the paper is structured as follows: The second section explores the relevant research literature. The proposed hierarchical deep ensemble learning approach is described in the third section. In the fourth section, we discussed an empirical comparison of the proposed HDEL method to individual machine learning and deep learning models along with a comparative and error analysis. The fifth section ends with a conclusion and discussion of future research, followed by a list of citations.

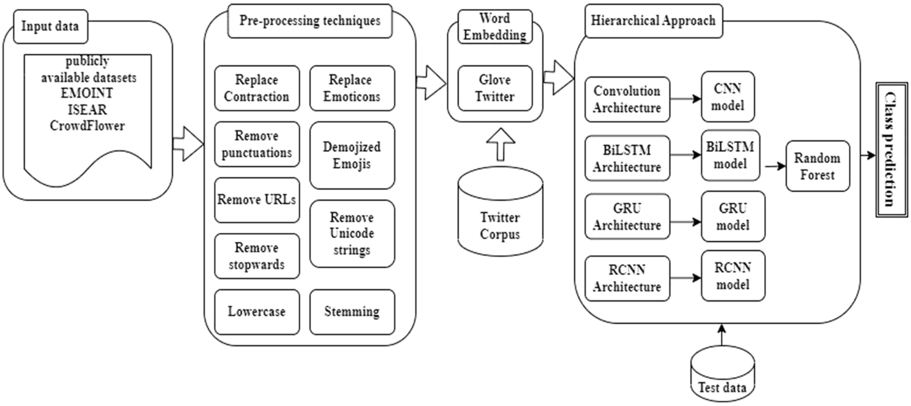

Fig. 1 High level Overview of proposed HDEL hierarchical deep learning model for text classification

2 Related work

The field of text-based emotion detection has evolved through various approaches, including lexicon-based methods (e.g., keyword-based, ontology-based, linguistic rule-based, statistical techniques), machine learning-based methods (both supervised and unsupervised), deep learning-based methods and hybrid techniques. The categorization of computational methods in the literature for detecting emotion in text is presented through graphical representation. Figure 2 depicts the visualization of these computational methods along with their respective advantages and disadvantages. Lexicon-based methods, as employed by the authors in [19–21], rely on predefined emotional lexicons. However, they face challenges related to generalization and context.

Machine learning methods, especially supervised learning, have shown promise in achieving accurate results but require substantial annotated data. Canales et al. [22] demonstrated that supervised ML algorithms have been widely used in text-based emotion detection, often outperforming unsupervised algorithms.

Various authors have utilized ML approaches in their research [23–26]. Lexicon-based methods are effective yet limited in scalability, while machine learning approaches overcome these limitations. However, they demand feature engineering and

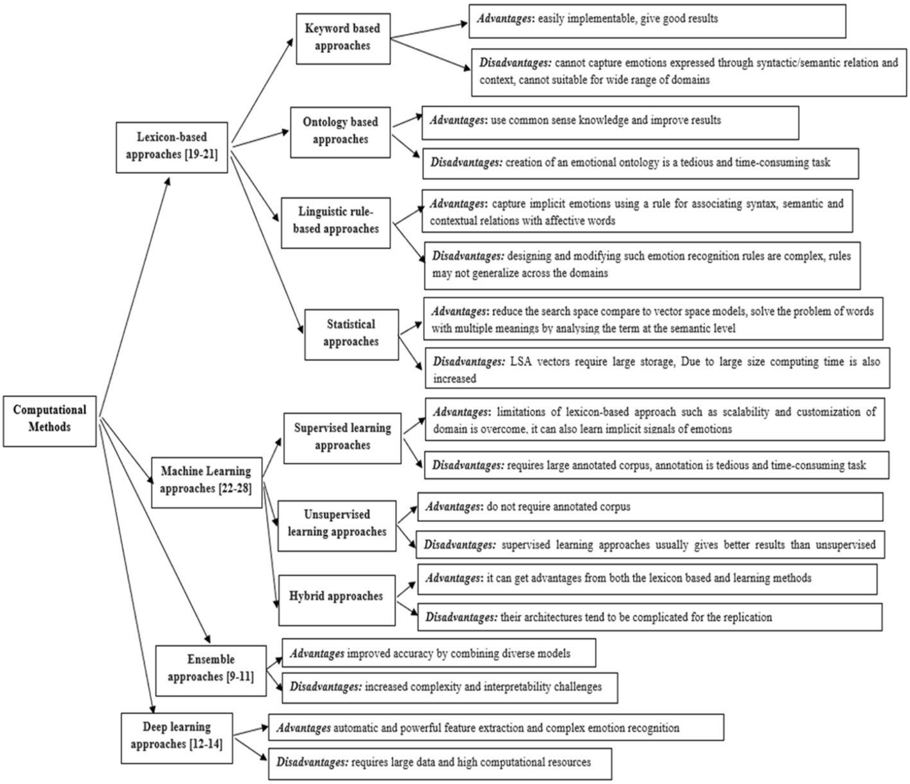
Computational Methods
Lexicon-based approaches [19-21]
Keyword based approaches
Advantages: easily implementable, give good results
Disadvantages: cannot capture emotions expressed through syntactic/semantic relation and context, cannot suitable for wide range of domains
Ontology based approaches
Advantages: use common sense knowledge and improve results
Disadvantages: creation of an emotional ontology is a tedious and time-consuming task
Linguistic rule-based approaches
Advantages: capture implicit emotions using a rule for associating syntax, semantic and contextual relations with affective words
Disadvantages: designing and modifying such emotion recognition rules are complex, rules may not generalize across the domains
Statistical approaches
Advantages: reduce the search space compare to vector space models, solve the problem of words with multiple meanings by analysing the term at the semantic level
Disadvantages: LSA vectors require large storage, Due to large size computing time is also increased
Machine Learning approaches [22-28]
Supervised learning approaches
Advantages: limitations of lexicon-based approach such as scalability and customization of domain is overcome, it can also learn implicit signals of emotions
Disadvantages: requires large annotated corpus, annotation is tedious and time-consuming task
Unsupervised learning approaches
Advantages: do not require annotated corpus
Disadvantages: supervised learning approaches usually gives better results than unsupervised
Hybrid approaches
Advantages: it can get advantages from both the lexicon based and learning methods
Disadvantages: their architectures tend to be complicated for the replication
Ensemble approaches [9-11]
Advantages improved accuracy by combining diverse models
Disadvantages: increased complexity and interpretability challenges
Deep learning approaches [12-14]
Advantages automatic and powerful feature extraction and complex emotion recognition
Disadvantages: requires large data and high computational resources

Fig. 2 Computational approaches for affective computing

domain expertise. Hybrid approaches, as adopted by the authors in [27, 28], seek to strike a balance between accuracy and complexity.

The studies have shown an increasing trend in the utilization of EL [9–11] and DL algorithms [12–14] to enhance the performance of emotion detection tasks. Above graphical visualization (Fig. 2) summarizes computational approaches for affective computing along with its advantages and disadvantages. These studies emphasize the combination of both techniques as a means to improve the accuracy of sentiment and emotion analysis.

In their research published in the IEEE Computational Intelligence Magazine in 2020, Akhtar, Ekbal, and Cambria proposed a multi-layer perceptron (MLP) ensemble technique for two tasks: emotion analysis and fine-grained sentiment analysis. They aimed to predict the intensities of emotions and sentiments, using generic tweets for emotion analysis and financial text for sentiment analysis. Initially, they developed a feature-driven system based on support vector regression (SVR) and three deep learning systems: CNN, LSTM network, and GRU network for intensity prediction. In the next step, they combined these system outputs using an MLP network, resulting in improved performance compared to individual models. To enhance the quality of the text data, they applied normalization heuristics, increasing the readability and improving the representativeness of word embeddings [29].

In the work by Araque et al., the authors conducted sentiment analysis classification using ensemble methods on seven publicly available datasets from microblogging and movie reviews domains. They employed various ensemble techniques, including voting and meta-learning methods and demonstrated that these ensemble techniques outperformed the baseline DL models. This suggests that the ensemble approaches can be highly effective in improving the accuracy of sentiment analysis across diverse datasets [30].

In Akhtar et al.'s study, they proposed a multi-task ensemble framework for emotion, sentiment, and intensity prediction. The framework employed three deep learning models (CNN, LSTM, GRU) to learn representations, which were then combined with handcrafted features using an MLP network. This approach aimed to make multiple predictions all at once. The multi-task approach outperformed single-task methods in the experiments, showing its effectiveness. Although multi-label emotion classification was not evaluated due to dataset limitations [31].

In the paper, Jain P et al. discusses the development of an ensemble system for predicting emotion intensity in tweets, focusing on emotions such as anger, fear, joy, and sadness. They employ three distinct deep neural network models: a feed-forward neural network, a multitask deep learning model, and a sequence modeling approach using CNNs and LSTMs. The models utilize various input features, including word embeddings and lexicon-based features, to capture tweet sentiment. They fine-tune multiple architectural parameters for each model, optimizing them using cross-validation. The ensemble system combines the predictions from these three models, assigning weights based on cross-validation results. Experimental results show significant improvements in predicting emotion intensity compared to a baseline model. The ensemble model achieves a substantial increase in performance, outperforming individual models by at least 2% on various emotions [32].

In their work, Haralabopoulos et al. developed a multi-label ensemble method for performing multilabel binary classification of user-generated content. They applied this approach to two datasets: Toxic comments and Semeval-2018-Task. To build the ensemble, they utilized various baseline deep learning models, including Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), LSTM, RCNN, and GRU. The ensemble approach enhances classification accuracy across multiple labels [33]. Table 1 summarizes

Table 1 A Review of Recent Studies and Their Key Findings

Study	Task	Algorithms Used	Features Utilized	Advantages	Limitations	Datasets	Accuracy/F1-Score
Akhtar et al. (2020) [29]	Emotion & Sentiment Intensity Prediction	MLP Ensemble	Word embeddings, sentiment features	Improved performance, readability, representativeness	Generalizability to other domains	Generic tweets, financial text	85.5% accuracy on tweets, 82.4% F1-score on financial text
Araque et al. (2017) [30]	Sentiment Analysis Classification	Voting, Meta-learning ensembles	Textual features	Improved accuracy across datasets	Effectiveness for specific tasks	7 microblogging & movie review datasets	85%-95% accuracy across datasets
Akhtar et al. (2022) [31]	Multi-Task Emotion, Sentiment, Intensity Prediction	CNN, LSTM, GRU, MLP	Word embeddings, sentiment features	Outperformed single-task methods	Multi-label evaluation not done	Multiple emotion and sentiment datasets	83% accuracy on emotion, 81% F1-score on sentiment
Jain et al. (2017) [32]	Emotion Intensity Prediction (Anger, Fear, Joy, Sadness)	Feed-forward NN, Multitask NN, CNN-LSTM	Word embeddings, lexicon-based features	Significant improvement over baseline	Limited scope of emotions	Tweets dataset	75% accuracy, 0.4 RMSE
Haralabopoulos et al. (2020) [33]	Multi-label Binary Classification	DNN, RNN, LSTM, RCNN, GRU	Textual content	Enhanced classification accuracy	Complexity, computational cost	Toxic comments, Semeval-2018	88% F1-score on Toxic, 75% AUC on Semeval

related work through the lens of research task, algorithms employed, training features, utilized datasets, model performance, and strengths and limitations of each study.

Above existing emotion classification models face limitations when it comes to generalizability and effectiveness. Some excel on specific datasets or tasks, struggling to adapt to different domains or emotion types. Others claim high accuracy but suffer from complexity and a black-box nature, making it difficult to understand how emotions are being classified. Additionally, many limits their scope to particular emotions or intensity prediction, hindering their wider applicability.

The proposed HDEL system tackles these limitations by combining the strengths of diverse algorithms within a robust, layered structure. By integrating deep learning probabilities with non-linear classifiers, HDEL achieves remarkable performance across varied datasets and a broad range of emotions. Its transparency makes it easier to understand the factors influencing emotion classification, while its potential for multi-task learning and domain adaptation opens doors for even more comprehensive and flexible applications. In essence, HDEL offers a versatile and accurate solution for emotion classification in the real world, overcoming the shortcomings of its predecessors.

Our work is more similar to work of Akhtar et al. [31], in which they performed a multitask ensemble framework that learns an understanding of several related problems of EA and SA. The ensemble model uses a manual feature representation and the features learned from three DL models (i.e., LSTM, GRU and CNN) to make predictions. A multitask framework is used to address four challenges of EA and SA: "valence and arousal for the sentiment," "emotion classification and intensity," "5-class ordinal and 3-class classification for the sentiment, and "valence, arousal, and dominance for emotion."

Differences between our proposed model and the model proposed by Akhtar et al. [31] are mentioned below.

i. The proposed HDEL approach is unimodal (Text based), Emotion recognition model whereas they proposed multitask ensemble framework.
ii. We utilize complex features from four deep learning models such as BiLSTM, BiGRU, CNN and RCNN. The features are aggregated and combined with original test data and fed to random forest tree-based learners instead of MLP based learning.

In summary, using a hierarchical approach that combines dL approach with Random Forest in the proposed HDEL framework offers many advantages. These include better accuracy, protection against overfitting, improved understanding of data, handling complex relationships, and dealing with imbalanced data, making it a promising choice for emotion classification.

3 Proposed framework

Our proposed framework, HDEL, is intricately designed to tackle the challenges of text-based emotion classification. The framework consists of sections describing the methodology and datasets. The Methodology section of the research paper outlines the data-preparation and model-preparation processes. The data-preparation phase employs state-of-the-art pre-processing techniques to refine the input data. Subsequently, the model-preparation section utilizes a variety of deep learning models, with hyperparameter tuning for optimal performance.

The model-preparation incorporates feature augmentation and ensemble learning, enriching the dataset for robust predictions. Hyperparameter tuning ensures the fine-tuning of learning parameters, enhancing adaptability across diverse datasets. Rigorous evaluation is conducted on three datasets: EmoInt [34], ISEAR [35], and CF [36], encompassing balanced and imbalanced scenarios.

3.1 Methodology

Next subsection discusses the data preparation techniques used to prepare all three datasets such as EmoInt, ISEAR and CF datasets followed by methodology.

3.1.1 Data preparation

Social media data (EmoInt twitter data and CrowdFlower (CF)) commonly considered as short text that includes noisy elements like special characters, symbols, and hyperlinks. These noisy components are eliminated using regular expressions. During the preprocessing phase, the text is segmented into tokens, specifically individual words, to facilitate further analysis. In the preprocessing of text data, we perform several essential tasks: replacing contraction words with their full forms, removing punctuation, numbers, and URLs, reducing extra line spaces and white spaces, substituting emoticons with relevant words, and converting emojis in tweets into text using Python's 'emot' library. Figure 3 depicts a visual representation of this process.

3.1.2 Model preparation

The subsequent phases involve the three key functions to create the system architecture. Key functions are base models' selection and critical hyperparameter tuning, feature augmentation, and linear and non-linear model selection for final classification tasks.

The foundation of HDEL involves the selection of distinct deep neural network architectures. Through extensive experimentation and optimization across diverse deep learning architectures, we meticulously identify and select four deep learning models demonstrating superior performance such as Bidirectional LSTM, 1D-CNN, Bidirectional GRU, and RCNN. To fine-tune the models, we employ Hyperopt [37] for hyperparameter optimization, ensuring an optimal combination of parameters for efficient learning. Table 2 gives details regarding hyper-parameters for training deep neural networks. In our 1D-CNN model, a single convolutional layer is followed by maximum-pooling layers (Conv-pool). The convolution layer incorporates 64 filters sliding across three words.

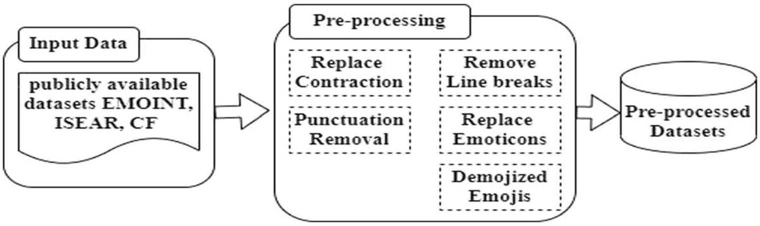

Fig. 3 Pre-processing operations performed on EmoInt, ISEAR and CrowdFlower (CF) datasets

Table 2 Hyper-parameters for training DL algorithms

Parameter	EmoInt, ISEAR and CF
Loss	Cross-Entropy
Hidden Activations	ReLU [38]
Output Activations	Softmax [40]
Shared Layers	CNN—1 (conv-pool) LSTM – bidirectional LSTM (64 neurons each) GRU – bidirectional GRU (64 neurons each) RCNN – RNN layer-bidirectional GRU (64 neurons each) followed by CNN-1 (conv-pool)
Fully-connected Layer	64 neurons, L2 regularization [39]
Convolution Filters	64 filters of size 3,4 and 5
Batch	64
Epochs	51 (with checkpoint option)
Dropout [41]	25%
Optimizer	Adam [42]

The architecture of Layer 1 with the CNN model for the proposed HDEL model is illustrated in Fig. 4 as one of the layers in our proposed model. Following the design of Layer 1, the other three layers incorporate different deep learning architectures for the proposed HDEL model.

For LSTM and GRU models, bidirectional LSTM and bidirectional GRU layers with 64 neurons per layer are employed. In the RCNN model, the RNN layer consists of 64-neuron bidirectional GRU layers, followed by a convolutional-max-pooling layer (Conv-pool). The convolution layer comprises 64 filters sliding across three words. Similarly, for LSTM and GRU models, 64-neuron bidirectional LSTM and bidirectional GRU layers are utilized. Features extracted from CNN, LSTM, GRU, and RCNN layers are fed into a fully connected layer, and subsequently to the output layer.

The fully connected layer consists of 64 nodes for all models, with the number of output layer nodes proportional to the defined labels in the dataset. The rectified linear activation function [38] is applied in the fully connected layer, and the L2 regularization function [39] serves as the loss function. For classification tasks, the output layer activation employs softmax [40]. To introduce regularization, a 25% dropout [41] is applied to the fully

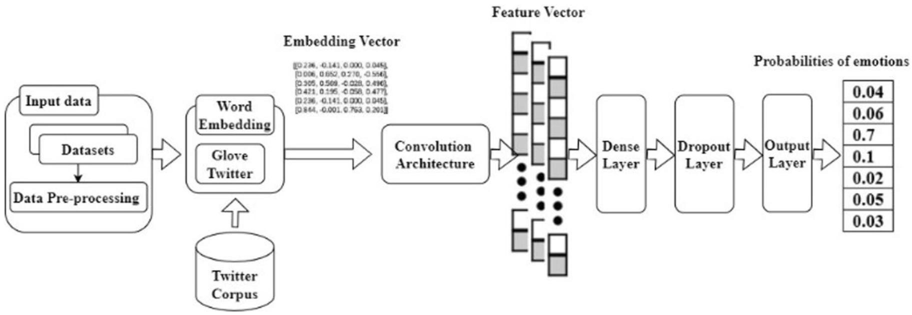

Fig. 4 The architecture of Layer 1 with the CNN model for the proposed HDEL model

connected layer. Gradient-based training utilizes the Adam optimizer [42] with its default parameters. The incorporation of GloVe pre-trained word embeddings, derived from a vast dataset of 2 billion tweets, 27 billion tokens, and 1.2 million vocabularies, enhances the contextual understanding of the language during training.

Building upon the foundation of our base models, the framework incorporates feature augmentation as a key component. An ensemble of diverse models produces probability scores for each emotion class during predictions on the test dataset. These probability scores are then appended to the original test dataset, enriching it with insights into the models' confidence levels in their predictions. This augmented dataset serves as the input for our subsequent classification task.

Another subsequent key function of our framework involves the careful selection of linear or non-linear models for the final classification task. Features extracted from the diverse deep learning models are combined to construct a hybrid dataset. This dataset is then subjected to various classification algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), XGBoost (XGB), and Random Forest (RF). After rigorous evaluation, we opt for the Random Forest (RF) algorithm as the final model for classification.

The overall architecture of HDEL, as depicted in Fig. 5, embodies a hierarchical approach, leveraging deep learning models with random forest algorithms. This proposed framework integrates the power of deep learning, ensemble techniques, and feature augmentation to advance the accuracy and reliability of text-based emotion classification.

3.2 Datasets

We examine our HDEL framework utilizing (EmoInt) (Mohammad S et al., 2018), (ISEAR) (Wallbott et al., 1986) and CrowdFlower (CF) as benchmark datasets.

(i) The "EmoInt" dataset originates from the "8th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA-2017)" shared task on emotion intensity. It is designed for analyzing emotion intensity in text data. The EmoInt-2017 dataset consists of tweets that express four different emotions: anger, fear, joy, and sadness. There are 3,613 tweets for training, 347 for validation, and 3,142 for testing in this dataset.

(ii) The ISEAR dataset consists of 7,666 sentences expressing different emotions like joy, fear, disgust, guilt, sadness, anger, and shame. These sentences were gath-

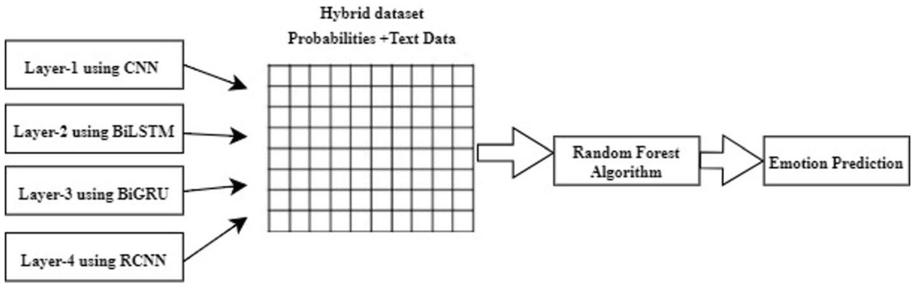

Fig. 5 HDEL: A Proposed framework of Hierarchical Deep Ensemble Learning based Emotion Detection

Table 3 Details of datasets

Datasets	Type of class	#Classes	Train	Validation	Test	Total
EmoInt	balanced	4	3,613	347	3,142	7,102
ISEAR	balanced	7	6132	307	1227	7,666
CF	imbalanced	13	9581	480	1916	11,977

ered from over 1,096 individuals with diverse cultural backgrounds who answered questions about these emotions. An equal number of sentences are available for each emotion. For research purposes, the dataset is divided into training, validation, and testing sets in a ratio of 80% for training, 4% for validation, and 16% for testing.

(iii) The "CrowdFlower (CF)" dataset has 40,000 tweets, each labeled with one of 13 different emotions. This is a highly imbalanced dataset. To address this, we used a method called "proportionate stratified sampling." This method ensures that we select tweets in a way that's proportional to the number of tweets in each emotion category. We only took a fraction of the tweets, about 0.3, using this method, so we ended up with 11,977 tweets that properly represent each emotion category for our experiments. For research, we split the dataset into training (80%), validation (4%), and testing (16%) sets.

Both EmoInt and ISEAR are balanced datasets whereas the CF dataset is an imbalanced dataset. Table 3 depicts a description of the datasets.

Tables 4 illustrate a few example challenges associated with the emotion analysis of the datasets. In the first instance shown in Table 4, the phrase “so pleasing” produces the strong emotion "joy". In comparison, the second expresses the emotion of "anger" with considerably less intensity. Similarly, examples of ISEAR and CrowdFlower (CF) are listed in Table 4 with its emotion.

4 Experiment setup and result analysis

This section presents our experimental setup and a detailed analysis of the results. We evaluate hypotheses through rigorous experiments, utilizing state-of-the-art algorithms on various datasets. We conducted experiments in the following settings to evaluate the hypotheses, exploring the effectiveness of our proposed emotion detection model for text data.

Table 4 Examples of dataset

Dataset	Text	Emotion
EmoInt	Being in the countryside all day was so pleasing	Joy
	Happiness is the best revenge	Anger
ISEAR	Cueing for a bus and the drivers having long dinner	Disgust
	When I unjustly accused a person of my family of something, she á didn't really do	Shame
CF	Wants to hang out with friends soon	Enthusiasm
	Happy mothers day	Love

Table 5 Performance comparison of ML, Ensemble, and DL algorithms on unembedded original datasets

Models	f-score (%)		
	EmoInt	ISEAR	Crowd-Flower (CF)
Naïve Bayes (NB)	67.0	56.8	29.82
Logistic regression (LR)	76.7	58.7	33.16
Support Vector Classifier (SVC)	80.6	57.5	29.92
Random Forest (RF)	76.6	54	31.84
XGBoost (XGB)	81.9	54.1	31.85
Artificial Neural Network (ANN)	75.9	56.8	23.76
Bidirectional GRU (BiGRU)	83	60.3	24.23
Bidirectional LSTM (BiLSTM)	83.2	58.7	30.39
1D-CNN	85.2	60.2	33.89
Recurrent Convolutional Neural Network (RCNN)	83.1	60.1	32.64

4.1 Experiment setup

> Hypothesis 1: What is the performance of State-of-the-art ML algorithms, ensemble algorithms, and DL-based algorithms on balanced and imbalanced datasets such as EmoInt, ISEAR, and CF textual datasets?

Experiment 1: We trained State-of-the-art ML algorithms such as Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Artificial Neural Network (ANN), ensemble algorithms like Random Forest (RF), XGBoost (XGB), and deep learning algorithms such as Long-short term (LSTM), Gated Recurrent Unit (GRU), Bidirectional GRU (BiGRU), Bidirectional LSTM (BiLSTM), 1D-CNN, and Recurrent Convolutional Neural Network (RCNN) on balanced and imbalanced textual datasets using Python Scikit-learn and TensorFlow libraries. We compared the performance of these models in terms of f-score (refer to Table 5). We used a support vector machine with a radial basis function kernel. This experiment was conducted on balanced datasets (EmoInt and ISEAR) and an imbalanced (CF) dataset. Deep learning algorithms were hyper-tuned using the Hyperopt optimization algorithm, a form of Bayesian optimization enabling the selection of optimal parameters for a given model. After comparing performance, we selected four deep learning models (Bidirectional GRU (BiGRU), Bidirectional LSTM (BiLSTM), 1D-CNN, and Recurrent Convolutional Neural Network (RCNN)) to create the proposed ensemble system. An ensemble of diverse models produced probability scores for each emotion class during predictions on the test dataset. The probability scores were integrated into the original dataset, resulting in a novel hybrid dataset. This hybrid dataset was then used to train diverse linear and non-linear classification algorithms, establishing an effective and resilient hierarchical methodology.

> Hypothesis 2: What is the performance of linear and non-linear algorithms on the novel hybrid dataset?

Experiment 2: The experiments involved the careful selection of linear or non-linear models for the final classification task. The hybrid dataset was subjected to various classification algorithms, including linear algorithms like Logistic Regression (LR) and non-linear algorithms namely Support Vector Machine (SVM), XGBoost (XGB), and Random Forest (RF). After evaluation and comparison, we opted for the Random Forest (RF) algorithm as the hierarchical model for classification.

For the Random Forest algorithm, we selected 25 trees as the optimal number to balance accuracy and computational efficiency, thereby enhancing text classification. We used the f-score as the main performance metric for comparison, as it measures the balance between precision and recall, reducing both false-positive and false-negative errors.

We used Python-based libraries, Keras, and Scikit-learn for implementation. For prediction, we used softmax for the classification task.

4.2 Result analysis

We conducted experiments to assess the effectiveness of a hierarchical ensemble technique that combines deep learning algorithms with the Random Forest algorithm. The study also includes a comparative analysis of the proposed emotion detection model against other models in the literature. Additionally, we performed an analytical breakdown of results of the proposed model on the EmoInt, ISEAR, and CF datasets.

Experiment 1: The detailed results of consecutive Experiment 1 are presented in Table 5. The findings from Table 5 offer a comprehensive view of the performance of various models across three distinct datasets: EmoInt, ISEAR, and CF. In the realm of state-of-the-art ML algorithms, Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Classifier (SVC) exhibit varying degrees of effectiveness across the datasets, with each algorithm encountering challenges in specific emotional contexts. The ensemble methods, Random Forest (RF) and XGBoost (XGB), show competitive performance, although with variations in their adaptability to different emotional patterns. Artificial Neural Network (ANN) struggles, especially in handling the complexities of the CrowdFlower (CF) dataset.

Deep learning models, such as Bidirectional GRU (BiGRU), Bidirectional LSTM (BiLSTM), 1D-CNN, and Recurrent Convolutional Neural Network (RCNN), consistently outperform traditional ML algorithms.

These models exhibit an average improvement in f-scores of 7%, 3.5%, and 0.16% on the EmoInt, ISEAR, and CrowdFlower (CF) datasets, respectively. This performance shows their effectiveness in capturing subtle emotional nuances across diverse datasets. Based on the analysis of performance of various models, we select four deep learning models namely as Bidirectional GRU (BiGRU), Bidirectional LSTM (BiLSTM), 1D-CNN, and Recurrent Convolutional Neural Network (RCNN) to construct ensemble system for the proposed HDEL framework. This ensemble generated probability scores for each emotion class during predictions on the test dataset. These probability scores were integrated into the original dataset, creating a novel hybrid dataset. Subsequently, this hybrid dataset was employed to train diverse linear and non-linear classification algorithms, aiming to establish a resilient hierarchical methodology.

Experiment 2: The outcomes of successive Experiment 2 are illustrated in Table 6. Table 6 summarizes the performance of linear and non-linear classification algorithms within the proposed Hierarchical Deep Ensemble Learning (HDEL) framework across EmoInt, ISEAR, and CF datasets. Notably, Random Forest consistently outperformed other models, achieving exceptional f-scores of 98.6%, 99.4%, and 99.7% for EmoInt, ISEAR, and CF datasets, respectively. This indicates its robust capability in capturing subtle

Table 6 Results of linear and non-linear classification algorithms on embedded hybrid datasets

Model	f-score (%)		
	EmoInt	ISEAR	Crowd-Flower (CF)
Logistic Regression (LR)	85.2	61.5	34.5
Support Vector Machine (SVM)	85.4	62.3	34.8
XGBoost (XGB)	90.1	87.4	62.2
Random Forest (RF)	**98.6**	**99.4**	**99.7**

emotional nuances. XGBoost also demonstrated strong performance, confirming its effectiveness in emotional pattern recognition. Based on results, the Hierarchical deep learning model with Random Forest algorithm emerges as the most appropriate and robust choice for all three evaluated datasets.

We compare our proposed HDEL system to the system proposed by Akhtar et al., a multi-task ensemble framework that learns numerous related tasks concurrently. The model tries to leverage a manual feature representation and the features learned from the deep learning models viz., LSTM, GRU and CNN for predictions. They address four problems of EA and SA using a multi-task framework, namely "valence, arousal, and dominance for emotion", "emotion classification and intensity", "3-class categorical and 5-class ordinal classification for the sentiment", and "valence and arousal for the sentiment" [34]. The authors classified emotions using the EmoInt benchmark dataset. The system proposed by Akhtar et al., reported an f-score of 89.3%.

Our obtained results are compared to the work of Bostan et al. [43]. The authors performed both cross-corpus and in-corpus classification experiments on various emotion datasets as part of a comprehensive study. For the EmoInt dataset, they reported an f-score of 88% for emotion classification. Many research efforts have been made for the EmoInt dataset. Still, most work is related to identifying the intensity of tweets and not emotion classification, so we cannot compare our results with those systems. DeepEmotex by Hasan M et al. [44] researched the EmoInt dataset. They employed DeepEmotex as an (ESTL) Effective Sequential Transfer Learning method for detecting emotion in textual content. They conducted a study using benchmark data sets and curated Twitter data. Their models correctly classify 73% of the instances in the EmoInt benchmark dataset. The system DeepEmotex [44] identifies only three emotion classes, unlike the four emotion classes we recognize. Table 7 depicts the results of the proposed system by Akhtar et al. [34] and the system DeepEmotex [44] with our proposed HDEL system.

The system proposed by Kratzwald et al. [45] is compared to the results of the proposed HDEL model for the ISEAR dataset. A proposed Sent2affect model for affective computing is a form of transfer learning. Their bidirectional LSTM layer is pre-trained for a different task (i.e. SA), and the output layer is then fine-tuned to investigate the emotion identification task. On six benchmark datasets, the resultant performance is tested. They claimed that the f-score for the system [45] on the ISEAR dataset was 56.9%. The approach described in system [43] achieved an f-score of 64% on the ISEAR dataset.

DeepMoji by Felbo et al. [46] model is also compared to the proposed model. A million emojis are present on social media platforms. An author trained neural models to interpret emotional context representations using these emojis as noisy labels. Two bidirectional

Table 7 Comparative results of HDEL system vs. existing proposed models for EmoInt dataset

Models	f-score (%)
System [34]	89.3
System [43]	88
System [44] (with 3 emotion category)	73
Proposed HDEL with RF	**98.6**

LSTM layers with 1024 units (512 in each direction) and an attention layer that accepts input from all LSTM levels via skip connections were utilized to capture the context of each phrase. Using a single pre-trained model, they achieved state-of-the-art performance on eight benchmark datasets for sentiment, emotion, and sarcasm detection tasks.

DeepMoji achieved an f-score of 57% on the ISEAR dataset. Our HDEL system achieves a 99.4% f-score for ISEAR complex dataset. Table 8 depicts the results of system [45], system [43] and system [46] with our proposed HDEL system.

We compare our proposed HDEL approach with the system proposed by Youngquist, O. [47], which is a novel ensemble neural network architecture that is capable of classifying the emotional context of short sentences. The model consists of three distinct branches, each composed of a combination of recurrent, convolutional, and pooling layers to capture the emotional context of the text. For emotion classification authors used five distinct datasets. The system [47] reported that the model achieved an average f-score of 38.0% for the CrowdFlower dataset. Our proposed HDEL model with RF gives better results.

Seyeditabari et al. [48] introduced a novel network based on a bidirectional GRU model, highlighting its ability to capture more meaningful information from text, which in turn led to significant performance enhancements for these models. Their study primarily focused on assessing the f-score for six emotions within the CrowdFlower dataset and comparing their results with Boston's work. It's important to note that our proposed HDEL approach does not directly compare with the findings of system [48] since our model is designed to classify all 13 imbalanced classes within the CrowdFlower dataset. Table 9 shows the result of system [47], the system [48] and system [43] with our proposed HDEL system for CrowdFlower (CF) complex and imbalanced dataset.

4.3 Analytical breakdown of results

In the evaluation of emotions using the EmoInt, ISEAR and CF datasets, our proposed approach encounters specific challenges.

Table 8 Comparative results of HDEL system vs. existing proposed models for ISEAR dataset

Models	f-score (%)
System [45]	56.9
System [43]	64
System [46]	57
Proposed HDEL with RF	**99.4**

Table 9 Comparative results of HDEL system vs. existing proposed models for CrowdFlower (CF) dataset

Models	f-score (%)
System [47]	38.0
System [48] (with 6 emotion category)	63.2
System [43] (with 6 emotion category)	32.0
Proposed HDEL with RF	**99.7**

4.3.1 EmoInt misclassifications analysis

For the EmoInt dataset, it tends to mix up class labels between the fear and sadness categories in the EmoInt dataset. To elaborate, out of 995 tweets expressing fear, 1.8% (18 tweets) are incorrectly classified as sadness, and for anger, 1.6% (12 tweets) are misidentified as sadness. These issues are further detailed in Table 10 of the confusion matrix. Additionally, during qualitative analysis, we observe that the proposed system faces difficulties in certain scenarios.

Implicit emotion with negation Implicit emotions can lead to misclassifications in the model's output. For example, consider the sentence: "Remember, your journey is unique. Do not get discouraged because you are comparing your journey to someone else's. You will get there." In this case, the actual emotion is 'fear,' but the model predicts 'sadness.'

Strong expressions Powerful phrases within sentences can influence the model's predictions. For instance, in the sentence "not by wrath does one kill but by laughter," the true emotion is 'anger,' yet our proposed model incorrectly predicts 'joy' because of the presence of the word 'laughter,' which appears to be the misleading factor.

Sentences with idioms Emotions can be challenging to detect when sentences contain idioms. Take, for instance, the sentence: "the pout tips me over the edge." In this case, the real emotion is 'anger,' but the model predicts 'sadness.' Table 11 provides a synopsis of the most common error scenarios.

4.3.2 ISEAR misclassifications analysis

For the ISEAR dataset, the confusion matrix suggests a total of 7 misclassified sentences. These are all mapped to the emotion of 'shame'.

Upon conducting the qualitative analysis of test cases, it has been observed that all these misclassified sentences are just one sentence with various emotions, and that is "no response". Table 12 represents the confusion matrix.

Table 10 Confusion matrix for emotion classification on EmoInt dataset

	Anger	Fear	Joy	Sadness
Anger	747	0	1	12
Fear	0	977	0	18
Joy	0	1	712	1
Sadness	6	4	0	663

Table 11 Synopsis of the common error scenarios

	Text	Actual Emotion	Predicted Emotion	Possible reason
EmoInt	Remember your journey is unique do not get discouraged because you are comparing your journey to someone else’s	Fear	Sadness	Implicit emotion with negation
	Not by wrath does one kill but by laughter	Anger	Joy	Strong expression
	The pout tips me over the edge	Anger	Sadness	Sentence with Idioms

Table 12 Confusion matrix for emotion classification on ISEAR dataset

	Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame
Anger	178	0	0	0	0	0	1
Disgust	0	185	0	0	0	0	3
Fear	0	0	172	0	0	0	0
Guilt	0	0	0	160	0	0	2
Joy	0	0	0	0	163	0	0
Sadness	0	0	0	0	0	168	1
Shame	0	0	0	0	0	0	194

4.3.3 CrowdFlower misclassifications analysis

For the emotion classification problem, we analyze the confusion matrix for the CrowdFlower dataset. For qualitative analysis we export the actual and predicted emotion with tweets to the excel sheet. For the CF dataset, the confusion matrix suggests that there are a total of 9 misclassified sentences. Upon conducting the qualitative analysis of test cases, it has been observed that there are 2 similar tweets with different emotions and that is “thank you”. Same way another two tweets “happy mothers day” have labeled with different emotions but both predicted as “love”.

Upon manual observation of the dataset, certain trends have emerged. In the training dataset, approximately 70 tweets containing the phrase "happy mothers day" are labeled with the ’love’ emotion, leading our model to predict ’love’ even when the labeled emotion is ’worry.’ Additionally, tweets like "good morning," which are labeled as ’love,’ are often predicted as ’neutral’ by the model.

Furthermore, in cases where slang words like "zwarte maillot" are present, the model tends to predict ’neutral’ emotion despite the label being ’boredom.’ This discrepancy arises because the model struggles to recognize such slang terms.

For a more comprehensive view of these issues, you can refer to the confusion matrix in Table 13 and a summary of frequent error cases in Table 14. The confusion matrix, as shown in Table 13, covers 13 distinct emotions, each represented by an abbreviation:

Table 13 Confusion matrix for emotion classification on CF dataset

	A	B	E	EN	F	HN	H	L	N	R	SN	SU	W
A	8	0	0	0	0	0	0	0	0	0	0	0	0
B	0	9	0	0	0	0	0	0	1	0	0	0	0
E	0	0	46	0	0	0	0	0	0	0	0	0	0
EN	0	0	0	39	0	0	0	0	0	0	0	0	0
F	0	0	0	0	79	0	0	0	0	0	0	0	0
HN	0	0	0	0	0	246	0	1	0	0	0	0	0
H	0	0	0	0	0	0	70	0	0	0	0	0	0
L	0	0	0	0	0	0	0	190	1	0	0	0	0
N	0	0	0	0	0	2	0	1	431	0	0	0	0
R	0	0	0	0	0	0	0	0	1	64	0	0	0
SN	0	0	0	0	0	0	0	0	0	0	262	0	0
SU	0	0	0	0	0	0	0	0	0	0	0	102	0
W	0	0	0	0	0	0	0	2	0	0	0	0	361

Table 14 Summary of the frequent error cases

	Text	Actual Emotion	Predicted Emotion	Possible reason
CrowdFlower (CF) Dataset	thank you	Neutral	Love	Similar tweet with different emotion
	thank you	Happiness	Love	Similar tweet with different emotion
	zwarte maillot	Boredom	Neutral	Slang word
	btw i still cannot believe how awesome the newjabbakidz performance was u in the masks i screamed at my pc	Neutral	Happiness	Strong expression “awesome”
	happy mothers day	Worry	Love	Similar tweet with different emotion
	happy mothers day	Worry	Love	Similar tweet with different emotion

A-Anger, B-Boredom, E-Empty, EN-Enthusiasm, F-Fun, HN-Happiness, H-Hate, L-Love, N-Neutral, R-Relief, SN-Sadness, SU-Surprise, W-Worry.

This research introduces a novel approach, the HDEL framework, for unimodal emotion classification from text data. This approach addresses class imbalance using the Random Forest algorithm, outperforming state-of-the-art methods on diverse datasets. The proposed framework has a wide range of practical applications, including sentiment analysis and mental health monitoring. It distinguishes itself from existing work through its unique architecture, superior performance, and focus on specific challenges such as class imbalance.

5 Conclusion and future work

This paper presents a novel approach called HDEL for unimodal emotion classification from text data. By combining the strengths of deep learning models and ensemble learning, the proposed approach achieves remarkable f-scores of 98.6%, 99.4%, and 99.7% on the EmoInt, ISEAR, and CrowdFlower (CF) datasets, respectively. The HDEL framework proves its adaptability and compatibility with other State-of-the-art emotion recognition systems, making it suitable for diverse applications.

However, the error analysis reveals that the proposed model encounters challenges in handling strong expressions, implicit emotions with negation, and idiomatic expressions, leading to misclassifications. These challenges provide opportunities for future research to improve the performance and capture the full range of emotions in text data.

Despite achieving exceptional performance, the proposed framework requires substantial computational resources due to the incorporation of deep learning models and Random Forest. Future scope includes multimodal emotion classification supported by multiple languages. Integrating the HDEL model into practical applications such as chatbots, virtual assistants, or mental health support systems to help users better understand and manage their emotions.

Data Availability 1 EmoInt Data openly available in a public repository that issues datasets with DOIs: The data that support the findings of this study are openly available in SemEval-2018 task 1: Affect in tweets. at http://doi.org/https://doi.org/10.18653/v1/S18-1001, Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In: Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 1–17

2 ISEAR Data openly available in a public repository that issues datasets with DOIs**:** The data that support the findings of this study are openly available in ISEAR. at https://doi.org/https://doi.org/10.1177/053901886025004001, Wallbott, H.G., Scherer, K.R. (1986). How universal and specific is emotional experience? Evidence from 27 countries on five continents. Social Science Information, 25(4), 763–795

3 CrowdFlower (CF) Data openly available in a public repository that does not issue DOIs: The data that support the findings of this study are openly available in CrowdFlower (CF) at http://www.crowdflower.com/wp-content/uploads/2016/07/text_emotion.csv, CrowdFlower. Sentiment Analysis: Emotion in Text., (2016)

Declarations

Conflict of interest We have no conflicts of interest to disclose.

References

1. Liu B (2012) Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers

2. Picard RW (1997) Affective Computing. MIT Press Cambridge MA
3. Damani S, Raviprakash N, Gupta U, Chatterjee A, Joshi M, Gupta K, Narahari KN, Agrawal P, Chinnakotla MK, Magapu S, Mathur A (2018) Ruuh: A deep learning based conversational social agent. ArXiv abs/1810.12097
4. Ansari MZ, Aziz MB, Siddiqui MO, Mehra H, Singh KP (2020) Analysis of political sentiment orientations on twitter. Procedia Comput Sci 167:1821–1828. https://doi.org/10.1016/j.procs.2020.03.201
5. Yang C, Chen X, Liu L, Sweetser P (2021) Leveraging semantic features for recommendation: 38 Sentence-level emotion analysis. Inf Process Manage 58(3):102543
6. Chowdary MK, Nguyen TN, Hemanth DJ (2021) Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications:1–18
7. Babu NV, Kanaga EGM (2022) Sentiment analysis in social media data for depression detection using artificial intelligence: A Review. SN COMPUT SCI 3:74. https://doi.org/10.1007/s42979-021-00958-1
8. Vijh M, Chandola D, Tikkiwal VA, Kumar A (2020) Stock closing price prediction using machine learning techniques. Procedia Comput Sci 167:599–606. https://doi.org/10.1016/j.procs.2020.03.326
9. Lin SY, Kung YC, Leu FY (2022) Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis. Inf Process Manage 59(2):102872
10. Kazmaier J, Vuuren JHv (2022) The power of ensemble learning in sentiment analysis. Expert Syst Appl 187:115819. https://doi.org/10.1016/j.eswa.2021.115819
11. Briskilal J, Subalalitha CN (2022) An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. Inf Process Manage 59(1):102756. https://doi.org/10.1016/j.ipm.2021.102756
12. Colnerič N, Demšar J (2020) Emotion recognition on twitter: Comparative study and training a unison model. IEEE Trans Affect Comput 11(3):433–446
13. Bharti SK, Varadhaganapathy S, Gupta RK, Shukla PK, Bouye M, Hingaa SK, Mahmoud A (2022) Text-Based Emotion Recognition Using Deep Learning Approach. Comput Intell Neurosci 2022:2645381. https://doi.org/10.1155/2022/2645381
14. Behera RK, Jena M, Rath SK, Misra S (2021) Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Inf Process Manage 58(1):102435
15. Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
16. Zhang Z, Lu Y, Zheng L, Li S, Yu Z, Li Y (2018) A new varying-parameter convergent-differential neural-network for solving time-varying convex qp problem constrained by linear-equality. IEEE Trans Autom Control 63(12):4110–4125. https://doi.org/10.1109/TAC.2018.2810039
17. Zhang Z, Zheng L, Weng J, Mao Y, Lu W, Xiao L (2018) A new varying-parameter recurrent neural-network for online solution of time-varying sylvester equation. IEEE Trans Cybern 48(11):3135–3148. https://doi.org/10.1109/TCYB.2017.2760883
18. Zhang Z et al (2018) A varying-parameter convergent-differential neural network for solving joint-angular-drift problems of redundant robot manipulators. IEEE/ASME Trans Mechatron 23(2):679–689. https://doi.org/10.1109/TMECH.2018.2799724
19. Shang L, Xi H, Hua J, Tang H, Zhou J (2023) A lexicon enhanced collaborative network for targeted financial sentiment analysis. Inf Process Manage 60(2):103187. https://doi.org/10.1016/j.ipm.2022.103187
20. Sykora MD, Jackson TW, Elayan S (2013) Emotive ontology: extracting fine-grained emotions from terse, informal messages. IADIS Intl J Comput Sci Inform Syst 8(2):106–118
21. Bandhakavi A, Wiratunga N, Padmanabhan D, Massie S (2017) Lexicon based feature extraction for emotion text classification. Pattern Recogn Lett 93:133–142
22. Canales L, Martínez-Barco P (2014) Emotion detection from text: a survey. Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days, 37-–43; ACM
23. Hasan M, Rundensteiner E, Agu E (2019) Automatic emotion detection in text streams by analyzing twitter data. Int J Data Sci Anal 7(1):35–51
24. Suhasini M, Srinivasu B (2020) Emotion detection framework for twitter data using supervised classifiers. Springer, New York, NY, pp 565–576
25. Singh L, Singh S, Aggarwal N (2019) Two-stage text feature selection method for human emotion recognition. Proceedings of 2nd international conference on communication, computing and networking, lecture notes in networks and systems, vol 46. Springer, Singapore, 531–538
26. Chowanda A, Sutoyo R, Meiliana TS (2021) Exploring text-based emotions recognition machine learning techniques on social media conversation. Procedia Comput Sci 179:821–828. https://doi.org/10.1016/j.procs.2021.01.099
27. Amelia W, Maulidevi NU (2016) Dominant emotion recognition in short story using keyword spotting technique and learning-based method. 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA), 1–6
28. Shaheen S, El-Hajj W, Hajj H, Elbassuoni S (2014) Emotion recognition from text based on automatically generated rules. IEEE Intl Conf Data Mining Workshop (ICDMW) 2014:383–392

29. Akhtar MS, Ekbal A, Cambria E (2020) How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]. IEEE Comput Intell Mag 15(1):64–75. https://doi.org/10.1109/MCI.2019.2954667
30. Araque O, Corcuera-Platas I, Sánchez-Rada JF, Iglesias CA (2017) Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Syst Appl 77:236–246. https://doi.org/10.1016/j.eswa.2017.02.002
31. Akhtar MS, Ghosal D, Ekbal A, Bhattacharyya P, Kurohashi S (2022) All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. IEEE Trans Affect Comput 13(1):285–297. https://doi.org/10.1109/TAFFC.2019.2926724
32. Goel P, Kulshreshtha D, Jain P, Shukla KK (2017) Prayas at EmoInt 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 58–65. https://doi.org/10.18653/v1/W17-5207
33. Haralabopoulos G, Anagnostopoulos I, McAuley D (2020) Ensemble deep learning for multilabel binary classification of user-generated content. Algorithms 13(4):83
34. Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S (2018) SemEval-2018 task 1: Affect in tweets. In: Proceedings of the 12th International Workshop on Semantic Evaluation, pp 1–17
35. Wallbott HG, Scherer KR (1986) How universal and specific is emotional experience? Evidence from 27 countries on five continents. Soc Sci Inf 25(4):763–795
36. CrowdFlower (2016) Sentiment analysis: Emotion in Text
37. Bergstra J, Yamins D, Cox DD (2013) Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. ICML'13, 115–123. JMLR.org
38. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 315–323. JMLR Workshop and Conference Proceedings
39. Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning (Vol. 4). Springer, Singapore
40. Goodfellow I, Bengio Y, Courville A (2016) 6.2. 2.3 softmax units for multinoulli output distributions. Deep Learning, 180
41. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 15(56):1929–1958
42. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. CoRR abs/1412.6980
43. Bostan L-AM, Klinger R (2018) An analysis of annotated corpora for emotion classification in text. Proceedings of the 27th International Conference on Computational Linguistics, 2104–2119. Association for Computational Linguistics, Santa Fe, New Mexico, USA. https://aclanthology.org/C18-1179
44. Hasan M, Rundensteiner E, Agu E (2021) Deepemotex: Classifying emotion in text messages using deep transfer learning. IEEE Intl Conf Big Data (Big Data) 2021:5143–5152. https://doi.org/10.1109/BigData52589.2021.9671803
45. Kratzwald B, Ili´c S, Kraus M, Feuerriegel S, Prendinger H (2018) Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems, 115, 24–35
46. Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1615–1625. Association for Computational Linguistics, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1169
47. Youngquist O (2020) An Ensemble neural network for the emotional classification of text. The Thirty-Third International FLAIRS Conference
48. Seyeditabari A et al (2019) Emotion detection in text: focusing on latent representation