sustainability

MDPI

Article

Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages

Zahra Ahanin [1], Maizatul Akmar Ismail [1,*], Narinderjit Singh Sawaran Singh [2,*] and Ammar AL-Ashmori [3]

[1] Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia; z.ahanin@siswa.um.edu.my
[2] Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Malaysia
[3] Department of Computer and Information Sciences, University Technology PETRONAS, Seri Iskandar 32610, Malaysia; ammar_18003398@utp.edu.my
* Correspondence: maizatul@um.edu.my (M.A.I.); narinderjits.sawaran@newinti.edu.my (N.S.S.S.)

Abstract: Emotions are vital for identifying an individual's attitude and mental condition. Detecting and classifying emotions in Natural Language Processing applications can improve Human–Computer Interaction systems, leading to effective decision making in organizations. Several studies on emotion classification have employed word embedding as a feature extraction method, but they do not consider the sentiment polarity of words. Moreover, relying exclusively on deep learning models to extract linguistic features may result in misclassifications due to the small training dataset. In this paper, we present a hybrid feature extraction model using human-engineered features combined with deep learning based features for emotion classification in English text. The proposed model uses data augmentation, captures contextual information, integrates knowledge from lexical resources, and employs deep learning models, including Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Encoder Representation and Transformer (BERT), to address the issues mentioned above. The proposed model with hybrid features attained the highest Jaccard accuracy on two of the benchmark datasets, with 68.40% on SemEval-2018 and 53.45% on the GoEmotions dataset. The results show the significance of the proposed technique, and we can conclude that the incorporation of the hybrid features improves the performance of the baseline models.

Keywords: emotion classification; feature extraction; natural language processing; neural networks; word embeddings; social media

check for updates

Citation: Ahanin, Z.; Ismail, M.A.; Singh, N.S.S.; AL-Ashmori, A. Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages. *Sustainability* **2023**, *15*, 12539. https://doi.org/10.3390/su151612539

Academic Editors: Muhammad Syafrudin, Ganjar Alfian, Norma Latif Fitriyani and Muhammad Anshari

Received: 12 February 2023
Revised: 16 April 2023
Accepted: 24 April 2023
Published: 18 August 2023

1. Introduction

Understanding people's opinions and emotions has always been important for governments and non-governmental organizations. While mastering natural language is easy for humans, it is still unobtainable by computers. Interacting with computers in natural language and identifying sensitive information is an enormous challenge in the field of computer science, then, and this area is known as Natural Language Processing (NLP). Artificial Intelligence (AI) is merging with the physical lives of humans, and it is going to change the ways that we live, work, and interact [1]. Today, social media platforms and microblogging services are extremely popular sources of information dissemination, which enable millions of users to create new content and share their opinions, thoughts, and emotions. Twitter, Facebook, and YouTube are among the most popular social media platforms in the world, accounting for massive amounts of digital data. With over 322 million registered users and over 500 million messages, Twitter is of immense importance to researchers in NLP. Sentiment analysis is a sub-field of NLP that uses machine learning techniques to determine polarity (positive, negative, neutral), emotions (disgust, love, joy, etc.), and even intentions (interested vs. not interested) in a chunk of text. Determining sentiment in social media data is a difficult task because human language is generative; words can be combined in various orders, with infinite grammatical variations, misspellings, and other challenges that are unique to each social media platform. For instance, Facebook allows users to decorate messages with stickers, and users on Twitter usually use short text messages (Tweets) that incorporate Twitter-specific abbreviations as well as slang, hashtags, emojis, and mentions of other users. Twitter text constructs such as hashtags and emojis play an important role in expressing emotions.

Sentiment and emotion analysis have been widely researched in neuroscience, psychology [2], and behavioral science because they are an important element of human nature. Emotions are one of the most valuable pieces of information for human communication as they are undisputed parts of our day-to-day life. Researchers have introduced several emotion models, such as Ekman's six basic emotions, which are joy, sadness, anger, fear, disgust, and surprise [3]. The Parrot emotion model introduces six primary emotions: love, joy, surprise, anger, sadness, and fear [4]. Plutchik's Wheel of Emotion, meanwhile, specifies eight basic emotions: anger, disgust, fear, sadness, anticipation, joy, surprise, and trust [5]. Finally, the Hourglass of Emotions includes 24 emotion categories, including joy, fear, calmness, and eagerness [6,7].

In computer science, the task of understanding emotions has piqued the interest of many researchers, especially in the field of Human–Computer Interactions (HCI). Sentiment and emotion analysis is beneficial for organizations and industries where the potential of customer feedback and queries is recognized as providing a tailored experience based on the customers' needs. This includes, for example, understanding tourist complaints in the tourism industry [8], or gaining an understanding of public emotions by the government on distributed palliatives during the COVID-19 pandemic [9], or describing the current learning state of students in distance learning applications [10], or identifying emotional barriers toward sustainable behaviors [11]. Detecting the emotions associated with sustainable behaviors can help organizations to design more effective communication strategies that resonate with the emotional responses of people. Similarly, it is useful in social media marketing strategies to influence customers and leverage their businesses. Sentiment and emotion analysis determines an individual's attitude towards a particular topic, event, product, etc., and it provides valuable insight into the market, which plays a vital role for companies in the decision-making process. Therefore, researchers and companies are always seeking better approaches to solving challenges via understanding of human emotions.

Feature selection and feature extraction are crucial in analyzing data. Feature selection refers to selecting the most relevant subset of features without modifying them, and feature extraction refers to transforming the raw data into a set of relevant features. Extracting the appropriate features from the textual information leads to better interpretability and also improves the performance of machine learning models [12–14]. Traditional approaches mostly rely on handcrafted features and lexicons (e.g., a list of words and their corresponding emotions). This is not only a time-consuming and expensive process that requires extensive domain knowledge but also relies heavily on keywords, where the emotion in sentences without the keywords is not detectable. Moreover, it ignores linguistic information, such as syntax structures and semantics, which can result in misclassification.

Recently, complex machine learning algorithms such as deep learning-based approaches have been shown to be effective in computer vision [15] and speech recognition [16,17] as well as NLP [13,18,19]. In comparison to conventional methods, deep learning automatically learns feature representation from the data. Feature extraction in textual data is vital in text classification that directly affects the accuracy of the learning algorithm. As a result, utilizing deep learning is of the utmost importance in emotion classification.

Despite the advantages of deep learning in NLP, a massive amount of data is required for the model to learn. To compensate for this drawback, researchers often utilize pre-trained word embeddings in the deep learning models [20–22]. Word2Vec and GloVe (Global Vectors for Word Representation) are the two main unsupervised learning algorithms for developing word embeddings. These algorithms capture the semantic and syntactic information of a word based on its surrounding words. The GloVe word embeddings developed by Stanford [23] uses billions of words from Wikipedia and Twitter to develop vectors. The Word2Vec pre-trained word embedding [24] is trained on the Google News dataset with about 100 billion words. Word2Vec can be obtained using two methods based on the Neural Network, including Skip-Gram and the Common Bag of Words (CBOW). A study by Baziotis et al. [25] collected an unlabeled dataset of 550 million English Tweets and trained the Word2Vec algorithm with the Skip-Gram model to develop Twitter-specific word embeddings—-namely, NTUA. The aforementioned pre-trained word embeddings are limited when interpreting context, though. For example, word position in the sentence is ignored. Thus, Devlin et al. [26] introduced Bidirectional Encoder Representations from a Transformers (BERT) framework, which uses the Transformer and Attention models to calculate the weights for each word. BERT is pre-trained on text from Wikipedia and has shown itself to be effective in question-and-answer datasets [27,28]. Therefore, such models require a significant amount of data, which limits the possibility of detection of linguistic features. For example, to detect negations or sentiment shifts, prior knowledge of the negation cues is invariably required [29].

Several researchers have used deep learning and machine learning for the task of emotion classification. Jianqiang et al. [30] implemented binary classification on a Tweet dataset using GloVe and Convolutional Neural Network (CNN). First, Tweets were preprocessed by removing URLs, numbers, stopwords, and non-English words. They then handled negation by transforming "n't" into "not" and replacing acronyms and slang with their full word form. Furthermore, the emoticons and emojis were replaced with their origin text from Emoticon Dictionary (https://en.wikipedia.org/wiki/List_of_emoticons, accessed on 28 January 2023). Based on the results, word embeddings and deep learning-based methods outperformed traditional linear-based methods. Li et al. [31] incorporated psychological domain knowledge and pretrained BERT, and they proposed emotional word embedding to enhance the sentiment feature. The authors used implicit (identify intensifier and negation shift) and explicit (using CNN to model the shifting) approaches, and they compared the proposed approaches with deep learning-based models such as Bi-LSTM and CNN-LSTM. Results have shown implicit features, explicit features, and a combination of both features to be effective on different datasets. Huang et al. [32] proposed hierarchical LSTMs for contextual emotion detection combined with the BERT pre-trained model to classify emotions into three categories (angry, sad, and happy). The pre-processing included misspellings and normalizing tokens. In addition, the emojis from Tweets were extracted and converted into texts corresponding to their name (https://pypi.org/project/emoji, accessed on 28 January 2023) in order to keep the semantic meanings of the emojis. According to the results, the combination of LSTM and BERT could slightly improve the macro-F1 scores.

However, existing approaches for emotion detection in online social media mainly focus on single emotion classification and ignore the cooccurrence of multiple emotions in a sentence or a Tweet. Based on these recent studies [13,33], users often express more than one emotion in a Tweet in different forms, such as text or an emoji. Thus, we consider the emotion classification on Twitter as a multi-label emotion classification problem, since the existing single-label emotion classification approaches are not suitable for this problem. Furthermore, relying exclusively on deep learning models to extract linguistic features may result in misclassifications due to the small training dataset. The aim of our research is to improve multi-label emotion classification accuracy by developing a model to extract features using a combination of deep learning-based features and human-engineered features. Researchers and practitioners can use this model in intelligent systems to understand emotions in various fields (including but not limited to affective HCI) and social data mining in order to provide personalized solutions and effective communication strategies.

This paper proposes and compares two deep learning-based methods for emotion classification in social media messages. The main contributions of this study are summarized below:

- Extract features based on the linguistic context of a word in a sentence, semantic, or syntactic similarity by employing Word2Vec embedding features.
- Propose a model that incorporates hybrid features, including Word2Vec based word embedding, human-engineered features, and Twitter specific features (emoji and hashtag), and that deploys a deep learning algorithm (Bi-LSTM) and a Transformer model (BERT) to classify emotions using context information.
- The proposed model was evaluated through various extensive experiments on two benchmark datasets, and, according to the results, the proposed technique that incorporates hybrid features outperformed the baseline models for emotion classification.

The remainder of this paper is organized as follows. Section 2 presents the literature survey. The proposed model for emotion classification is provided in Section 3. Section 4 discusses the experiments and provides discussion about the performance results. Finally, Section 5 concludes the article with suggestions for future work.

2. Related Work

Sentiment analysis and emotion classification are a growing research area. Different methods have been employed, including lexicon-based approach, rule-based approach, conventional machine learning, and deep learning techniques. The target of the lexicon-based method is to identify emotions by making use of well-known dictionaries, such as NRC, EmoSenticNet, SentiWordNet (SWN), and Linguistic Inquiry and Word Count (LIWC) [34]. However, the performance of lexicon-based techniques relies on the quality and coverage of dictionaries, since words or phrases that are not in the dictionaries are not identifiable. Such dictionaries are widely used in rule-based approaches to improve the accuracy of classification. Asghar et al. [35] proposed a rule-based classification with an enhanced lexicon and four different classifiers, including Emoticon Classifier (EC), Modifier and Negation Classifier (MNC), SentiWordNet Classifier (SWNC), and Domain Specific Classifier (DSC). The study manually created a negation list that includes all possible negation terms and an emoticon dictionary that includes emoticons and their corresponding labels (positive or negative) by three human annotators. They have also used the SentiWordNet (SWN) lexicon to retrieve the sentiment score of each word. For evaluation purposes, precision, recall, accuracy, and f-measure were performed on three datasets in order to classify the text into positive, negative, or neutral classes.

Machine learning approaches, such as supervised approaches, automatically recognize emotions in which the model is trained based on a labeled training set. Human annotators can annotate datasets, which is a time-consuming and labor-intensive process, or datasets can be annotated automatically based on Twitter properties such as hashtags or emoticons [36,37]. Although these studies have highlighted the importance of hashtags and emoticons in determining the emotion of a Tweet, focusing only on these features may cause possible emotive information in a sentence to be overlooked. In order to determine the emotion of the text, a variety of feature extractions, feature selection, and feature representations have been used with classifiers. Several studies have used machine learning to classify emotions in textual data. Ameer et al. [38] proposed a content-based method that deployed TF-IDF (Term Frequency–Inverse Document Frequency) and a variety of features, such as word n-grams, character n-grams, and their combinations, which were used for Multi-label Emotion Classification. They performed two machine learning classifiers, including RF (Random Forest) and DT (DecisionTable), which perform classification for each emotion label. Such methods require multi-label classifiers, including BinaryRelevance (BR), BPNN, Classifier Chain (CC), and Label Combination (LC), to transform the task of single-label classification to multi-label classification. The results have shown that Binary Relevance and Random Forest (BR + RF) with unigram word features provided the best Jaccard accuracy. Flor et al. [39] applied several ML classification methods, including Support Vector Machine (SVM), Logistic Regression (LR), Multilayer Perceptron (MLP), and Naive Bayes (NB), to classify emotions in one of the emotion classes: anger, fear, sadness, or joy. They employed affective lexical features in the Spanish language, such as the SEL, the iSOL, and the NRC emotion lexicon, and, according to the results, the SVM was found to have the best F1-score while the MLP recorded the lowest F1-score.

In recent studies, deep learning methods are found to provide reliable results in speech recognition, image classification, and natural language processing tasks. Deep learning-based methods such as Long Short-Term Memory networks (LSTM), Bidirectional Long Short Term Memory networks (Bi-LSTM), and Gated Recurrent Unit (GRU) automatically extract high-level features from raw data and model sequential information.

Jabreel and Moreno [40] have presented a deep learning approach for the multi-label classification problem. They aimed to propose a model that learns associations between class labels and words in a sentence based on their semantic similarity by using an attention model. The attention model was designed based on a pre-trained word embedding known as Stanford's GloVe (Global Vectors for Word Representation) and a trainable word embedding to find associations between words and each emotion class. The resultant word embedding was fed into the encoder module, which consisted of a Bidirectional Gated Recurrent Unit (Bi-GRU), and, finally, they combined the results of the emotion classification on the dataset of SemEval-2018 Task1. The results show the effectiveness of the proposed method, with a Jaccard accuracy of 59%. Ahanin and Ismail [13] created a method based on Pointwise Mutual Information (PMI) to model the association between an emoji and each emotion class, and to classify the emoji into one or more emotion classes. The resultant emoji embedding was used as a feature to extend the existing pre-trained word embeddings (e.g., Stanford's GloVe and NTUA), which are deployed in deep learning models such as LSTM or GRU. NTUA has 300-dimensional word embeddings, which are trained on 330 million Twitter messages using word2vec. According to the results, the combination of emoji embedding and NTUA provided better accuracy, which could be due to the fact that Stanford's GloVe [23] embedding does not incorporate a newer popular Unicode emoji.

The study by Alhuzali and Ananiadou [41] proposed a model called "SpanEmo", in which a BERT encoder receives emotion classes and an input sentence, allowing the encoder to interpolate between emotion classes and all words in the input sentence. Zygadło, Kozłowski, and Janicki [42] proposed an emotion recognition approach in the context of a therapeutic chatbot by creating a dedicated dataset and employing various classifiers (Naïve Bayes, Support Vector Machines, and BERT). They obtained the best scores for the BERT-based model. Similarly, Demszky et al. [43] performed a BERT-based model on their manually annotated dataset, which outperformed the Bi-LSTM model. Ameer et al. [18] used transformed models such as RoBERTa (Robustly Optimized BERT Approach) [44], XLNet [45], and DistilBERT [46]. They added a multiple-attention layer to the output of the Transformer models and fine-tuned the final layer on the multi-label emotion classification dataset, and XLNet, DistilBERT, and RoBERTa achieved an accuracy of 59.4%, 60.3%, and 61.2%, respectively. Silva Barbon and Akabane [47] employed BERT and DistilBERT for text classification, in which BERT obtained slightly higher accuracy in classifying blog corpus. Thus, BERT has been widely adopted and can achieve state-of-the-art results on various NLP tasks, including emotion classification, by capturing complex patterns and relationships within the unstructured data.

Alswaidan and Menai [48] proposed a hybrid feature model in Arabic text, which consists of a human-engineered feature-based model (e.g., lexical features, linguistic features, and syntactic and semantic features), and a deep feature-based model (e.g., pre-trained word embedding). These features are then concatenated and fed into a dense neural network (DNN) with a rectified linear unit (ReLU) activation function. Despite the use of human-engineered and deep learning-based features, the proposed model had difficulty in recognizing the trust emotion, and it failed to recognize the surprise emotion. Gee and Wang [49] pre-trained a model on a sentiment dataset to transfer knowledge, since the sentiment dataset is semantically similar to the emotion dataset. This process is known as Transfer Learning (TR), which helps to learn the weights of the LSTM networks and to identify words in a Tweet that contribute to the task of emotion classification. Five components have been designed: (1) LSTM and NTUA word embedding as input; (2) LSTM with attention mechanism and NTUA word embedding as input; (3) Bi-LSTM and a lexicon vector by Weka's TweetToLexiconFeatureVector; (4) a lexicon vector by using TweetToLexiconFeatureVector of the entire Tweet; and (5) the four components mentioned above are concatenated and fed into a 5-layer DNNs. According to the results, the combination of four components performed better. In Table 1, we summarized the emotion classification methods that are the most influential in solving our problem. The attention mechanism has shown itself to be effective in extracting more relevant words by giving more weight to the keywords [13,49,50].

Table 1. Related works.

Author	Year	Language	Features/Model	Evaluation
Ameer et al. [18]	2023	English	RoBERTa is used and a multiple-attention mechanism is added to the output of the RoBERTa	Micro F1-score, Macro F1-score, Jaccard Index score
Ahanin and Ismail [13]	2022	English	Word embedding (NTUA) and emoji embedding are used as features to the encoder module (Bi-LSTM + attention mechanism)	Micro F1-score, Macro F1-score, Jaccard index score
Alhuzali and Ananiadou [41]	2021	English	Emotion classes and input sentence are fed to encoder BERT	Micro F1-score, Macro F1-score Jaccard Index score
Demszky et al. [43]	2020	English	Created the largest human annotated dataset on Reddit comments, and employed BERT-based model	Precision, Recall, Macro average F1-score
Alswaidan and Menai [48]	2020	Arabic	Combined human-engineered features and deep learning features	Micro F1-score, Macro F1-score, Jaccard Index score
Ameer et al. [38]	2020	English	Unigram word features are used as input in BR and RF classifiers	Micro F1-score, Macro F1-score, Hamming Loss, Accuracy
Jabreel and Moreno [40]	2019	English	Pre-trained word embedding (GloVe) and trainable word embedding are used to create word representations and Bi-GRU is used in encoder module	Micro F1-score, Macro F1-score, Jaccard Index score
Gee and Wang [49]	2018	English	Pre-trained word embedding, and lexicon vectors are the inputs to the sub-models (LSTM, Bi-LSTM + attention mechanism)	Micro F1-score, Macro F1-score, Jaccard Index score
Baziotis et al. [25]	2018	English	Pre-trained word embedding (NTUA) was incorporated with Bi-LSTM and a multi-layer self-attention mechanism	Micro F1-score, Macro F1-score, Jaccard Index score

3. Methodology

We proposed two models for emotion classification in the Twitter dataset. The first model integrates Word2Vec-based word embedding, human-engineered features, emoji embedding, hashtag, and mood features. The generated features are employed to train and test the model using a deep learning algorithm (Bi-LSTM). The second method consists of a transformer learning (BERT) model that captures the context in emotion text, and a deep learning algorithm (Bi-LSTM). The pre-processing is similar in both models.

3.1. Data Preprocessing

Data preprocessing starts with tokenization. In tokenization, each text (Tweet) is separated into words. First, hashtags are extracted, which will then be used in the classification model. Next, considering that Twitter includes misspellings and abbreviations, we handled misspellings using Python natural language toolkit (NLTK) and normalized the following terms: URL, number, email, money, time, and date. Moreover, we manually created a custom dictionary to handle the abbreviations, and we eliminated noisy data that is not useful in the emotion classification such as stop words, mentions (e.g., @user), and punctuation. Special characters (excluding exclamation marks) are removed using the Python Regular Expression Regex. Any words that are in a language other than English are removed. The data is further normalized using lemmatization and converted to lowercase. Table 2 demonstrates a small sample of Tweets before and after pre-processing.

Table 2. Small sample of Tweets before and after pre-processing and data augmentation.

Dataset	Pre-Processing Status	Sentence
SemEval-2018	Before Pre-processing	This man doesn't even look for his real family 🙄😂
	After Pre-processing	this man does not even look for his real family 🙄😂
	Data augmentation	this man does not even await for his real family 🙄
GoEmotions	Before Pre-processing	Hardly surprising. Now they also lost their coach . . .
	After Pre-processing	hardly surprising now they also lost their coach
	Data augmentation	long finished now have also lost our coach

3.2. Data Augmentation

Machine learning and deep learning-based methods achieve better performance when a substantial amount of annotated data is available to train the model. Since providing annotated data is expensive and time-consuming, researchers opted for data augmentation techniques. Augmentation is commonly used in Computer Vision in which a synthetic dataset is generated by altering (e.g., flipping, rotating, etc.) the training images [51]. However, in the field of NLP, complexity of language posed a certain degree of challenges to augment text. Data augmentation in sequential data can be achieved through (i) modifying the input sequence by randomly deleting, swapping, or inserting words, or replacing words with one of their synonyms [52], or (ii) employing a neural network to replace or increase the number of words in a sentence [53]. In this research, we used two approaches: (1) WordNet, as a lexical database, which randomly replaces a word in a sentence with its synonym, and (2) BERT, as Contextualized Word Embeddings to generate training data (https://github.com/makcedward/nlpaug, accessed on 28 January 2023). This model generated words based on the words around it, with the aim of inserting a suitable word for augmentation.

In the next section, textual data is converted into array representation of features as inputs in the deep learning model.

3.3. Human-Engineered Features (HEF)

After preprocessing, we extracted the features that represent different aspects of the text such as syntactic features and domain-specific features.

3.3.1. Sentiment of Sentence

Syntactic analysis identifies the grammatical rules for a group of words. It is an important process in making computers understand human language. In this study, syntactic features refer to Part of Speech (POS) features, where a role is tagged to each component of a sentence. A word can be tagged as verb, noun, adverb, adjective, etc., and we concentrated on particles (such as 'not'), adjectives, and verbs. First, POS tags are assigned to each sentence in the Tweet, and then a rule-based strategy is designed to highlight more affective words, such as adjectives and verbs. Therefore, particle-verb and particle-adjective pairs are extracted. Next, the SentiWordNet lexicon was employed to determine the sentiment score (positive and negative) of each word in the pair.

Afterwards, a polarity score of -1, +1, or 0 is given based on the positive score and negative score of the word (Equation (1)). The positive and negative scores are also used to calculate the sentiment range score, which is based on the absolute value of subtracting the positive score and the negative score of each word in the pair (Equation (2)). Instead of simply using −1, +1, and 0, we used range, since there is multiple emotion classes and the range between positive and negative emotions might differ in each emotion class.

Considering the sentiment values of words are the opposite when negative adverbs are before these words, an inverse score is calculated to handle the polarity shift, where w_x refers to the first word (particle) and w_{x+1} refers to the second word (verb, or adjective) in the pair (Equation (3)).

$$polscore(w) = \begin{cases} -1, & if\ sentscore(w)_{pos} < sentscore(w)_{neg} \\ 1, & if\ sentscore(w)_{pos} > sentscore(w)_{neg} \\ 0, & else \end{cases} \tag{1}$$

$$sentrange_{score}(w) = \text{abs}\Big(sentscore(w)_{pos} - sentscore(w)_{neg}\Big) \tag{2}$$

$$invscore = sentrange_{score}(w_{x+1}) \times polscore(w_{x+1}) \times polscore(w_x) \tag{3}$$

Moreover, we calculated a sentiment of adjectives, nouns, and verbs in the sentence, which are not preceded by a particle, using Equations (1)–(3) (where $polscore(w_x)$ is equal to 1). Finally, the summation of all scores of each pair and word in one or more sentences in the Tweet gives a total sentiment score of the whole Tweet (Equation (4)). The positive score indicates a positive sentiment, and the negative inverse score indicates a negative sentiment.

$$\begin{aligned} sent_{score}(sentence) = \sum_i invscore(adjective)_i + invscore(verb)_i + invscore(noun)_i \\ invscore(particle\_adjective)_i + invscore(particle\_verb)_i \end{aligned} \tag{4}$$

To reduce the complexity, the score is normalized before utilizing it in the proposed Bi-LSTM-based model. Therefore, instead of using a floating number, a symbol is given that is the result of normalizing inverse score between range of [−n, +n], which we call the Sentiment Contrast Score (SCS) (Equation (5)). In this study, n is set to 5.

$$\text{normalize sentiment contrast score } \in \mathbb{Z} \tag{5}$$

The main reason for extracting particle-verb and particle-adjective pairs is that users mention positive words in negative sentences, but the word is preceded by a negative word such as 'not' (e.g., not good). Algorithm 1 shows the pseudocode for generating SCS.

Algorithm 1: Rule-based conversion for Sentiment Contrast Score (SCS) generation

```
Input: A sentence with M words, {$w_1$, $w_2$, . . . , $w_M$ },
POS tags $POS_w$,
Sentiment values $s_w$ using SentiWordNet.
Output: Sentence-level sentiment contrast score (SCS)
Initialization: Vector of $p_s$, constant c.
for each affective word in the sentence do
    if a particle appears then
        $w_x$ = particle word
        $w_{x+1}$ = word that comes after particle
        $pol\_score_{w\_x}$ = polarity ($s_{w\_x}$) ϵ −1,0,+1
        if $POS_{w\_x+1}$ = verb or adj then
            $pol\_score_{w\_x+1}$ = polarity ($s_{w\_x+1}$) ϵ −1,0,+1
        score = abs($sent^+ - sent^-$)× $pol\_score_{w\_x+1}$ × $pol\_score_{w\_x}$
            $p_s$+ = score
    if $POS_w$ = verb or adj or noun, and no particle appears before it then
        $pol\_score_w$ = polarity ($s_w$) ϵ −1,0,+1
        score = abs($sent^+ - sent^-$) × $pol\_score_w$
        $p_s$+ = score
end for
SCS ← Normalize ($p_s$);
return SCS
```

Since deep learning methods, such as LSTM and BERT, which are used in this study, require prior knowledge about negation cues, they only partially detect the polarity inference. Our method identifies the negation shift without requiring prior knowledge about negation cues.

3.3.2. Syntactic Feature

In this research, interjection words are used as a syntactic feature. Interjection is commonly used in informal text or speech to convey emotions, such as joy, disgust, or sudden bursts of feelings. Such emotional cues might be neglected in the sentence. To detect the interjections in natural language, POS tagging is employed that helps to count the total number of interjections in a sentence such as cool, oh!, wow, etc.

3.3.3. Domain-Specific Features

The mood tag (e.g., calmness, joy) of each word in the sentence is retrieved from SenticNet [54]. The mood tags are generated based on the Hourglass of Emotions model [7]. The TF-IDF is then calculated.

3.4. Deep Learning-Based Features (DLF)

In this study, word embedding representation is employed to obtain similar representation for words with similar meaning. The most popular embedding methods are GloVe, BERT, and Word2Vec.

3.4.1. Word2Vec Word Embedding

Word2vec is a two-layer neural network that uses back-propagation to efficiently learn word embeddings from large datasets. Word embeddings capture the semantic similarity of the texts, and derive dense vector representations for words, where the similar words have a similar value in the vector space. This paper utilized a pre-trained word embedding based on Word2Vec model because it has proven itself to be effective in several NLP tasks.

3.4.2. Hashtag Feature Embedding

Hashtag is a catchphrase that is widely used to depict the content of a Tweet. A hashtag can contain invaluable information about the actual emotion of the user. For example, in the Tweet "this broadband is shocking regretting signing up now #angry", the hashtag #angry expresses the user's emotion. In this research, the hashtag features are extracted, and then segmentation is employed to breakdown the hashtag into meaningful words. For example, the hashtag #notgood is transformed to "not good". Afterwards, all the hashtags are transformed to the word vector form using pre-trained word embeddings. In this section, the pre-trained word embedding by Baziotis et al. [25] is used, which is based on Word2Vec.

3.4.3. Emoji Feature Embedding

Several researchers have relied on emotion symbols such as emoticons and emoji ideograms to classify Twitter messages [36,37]. Emojis indicate the tone of the message or the emotions in a Tweet [22]. Therefore, these features are a rich source of information in emotion classification. We used the emoji embedding proposed by Ahanin and Ismail [13], which includes a vector representation of each emoji after they are classified into one or more emotion classes.

3.5. Emotion Classification Based on Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language representation model introduced by Devlin et al. [26], with advanced state-of-the-art results in NLP tasks such as Named Entity Recognition (NER), Question Answering, and Sentiment Analysis. The BERT model is based on a multi-layer bidirectional Transformer encoder with bidirectional self-attention. As opposed to the directional models that read the text input from a single direction (left to right or right to left), the Transformer encoder reads the entire sequence of words at once. Thus, every token can attend to its context to the right and the left. BERT is trained using a document-level corpus based on BooksCorpus (800 M words) and English Wikipedia (2500 M words). BERT$_{\text{base}}$ has a vocabulary of 30 k tokens, and each token has 768 features in its embedding.

In our emotion classification, contextual information is first obtained from the pre-trained BERT layer, and then the BERT model is fine-tuned on the annotated dataset. Fine-tuning is important because the BERT model is pre-trained on a large corpus of general text in Wikipedia and Book Corpus. To use BERT model on specific tasks, neural network layers can be attached to the model to train it on a labelled dataset for emotion classification.

3.6. Classification Using Long-Short Term Memory (Bi-LSTM)

Recurrent Neural Networks (RNNs) obtain the flow of information in a sequence in short contexts. RNNs are not suitable for long data sequences as they suffer from a vanishing gradient problem, where gradients can vanish or explode. When the gradient vanishes, or becomes smaller, the updates in RNN parameters become insignificant and no real learning is achieved.

LSTM has an advantage over the RNN because of its property of selectively remembering patterns for long durations of time. An LSTM network has an input vector $[h_{t-1}, x_t]$ at time step t. Each LSTM unit consists of three gates, which are input gate, forget gate, and output gate. These gates are designed to update and control the unit state by deciding which information is important to keep, or which fraction of information is irrelevant and needs to be discarded.

Bi-LSTM concatenates a sequence of forward hidden states as well as sequence of backward hidden states, as given below:

$$h_t = \left[\overrightarrow{h_t};\ \overleftarrow{h_t}\right] \tag{6}$$

Lastly, attention mechanism [55,56] can be used to focus on the most relevant words:

$$s_i = \sum_t a_{it} h_{it} \tag{7}$$

where,

$$a_{it} = \frac{exp\left(u_{it}^{\intercal} u_w\right)}{\sum_t exp\left(u_{it}^{\intercal} u_w\right)} \tag{8}$$

$$u_{it} = tanh(W_w h_{it} + b_w) \tag{9}$$

where u_{it} is the hidden representation of h_{it}, which is the result of non-linear activation function ($tanh$) on the Bi-LSTM output (Equation (9)). The attention similarity score of a word is calculated based on the similarity of the context vector u_w with u_{it}. The weights a_{it} are computed and normalized by a Softmax given by Equation (8). Finally, the summation of the outputs of the attention layer are sent through a Sigmoid operation to obtain the probability distribution of the emotion classes for the task of multi-label classification (Equation (7)).

Figure 1 shows the classification model. It takes the text (Tweet) as input, and then it performs data cleaning, extracts hashtag and mood features, calculates sentiment contrast score (SCS), and obtains the total number of interjection (INTJ) features in each Tweet. The text features and hashtag features are then concatenated to return a single output. The output of the concatenation was fed into Bi-LSTM with 64 units with a ReLU activation function. To avoid overfitting, a dropout layer was added. An attention model was then designed to give more weight to important words. The attention features, mood features, SCS, and INTJ features are concatenated, and then fed into dense neural networks (DNNs).

Finally, the sigmoid activation function was added as an output layer for the classification of the Tweets into one or more emotion classes (Figure 1).

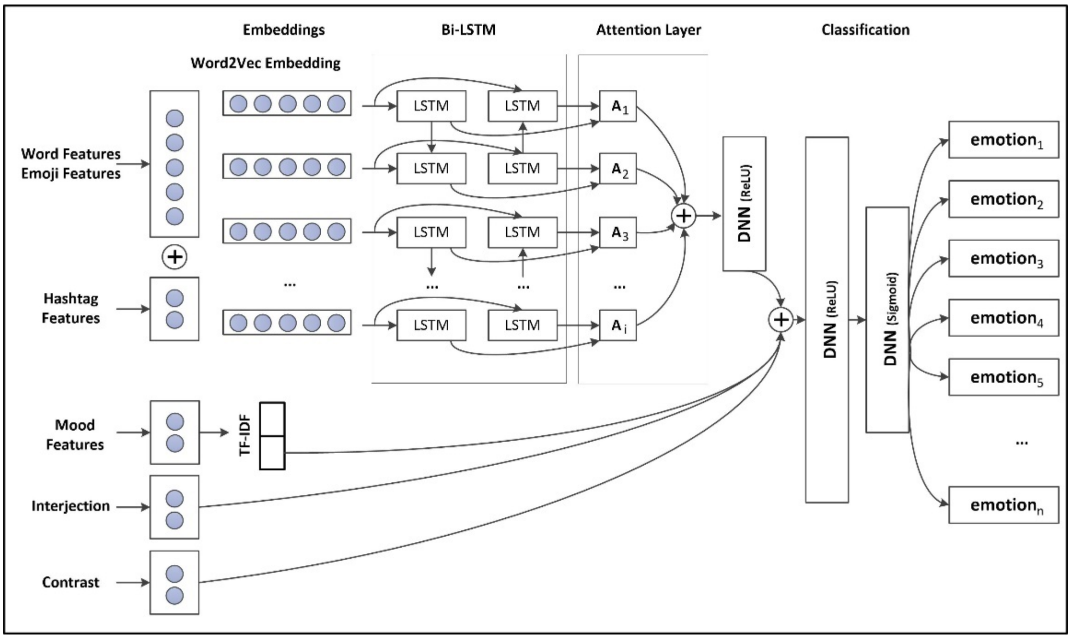


Figure 1. Proposed emotion classification using DLF, HEF, and Bi-LSTM with attention model (Bi-LSTM + DLF + HEF).

In addition to the aforementioned features, BERT contextual embeddings can be used in emotion classification. Contextual embeddings are the extracted activations from one or more of the last layers of the BERT pre-trained model. The features in these activations are more complex, and they include more context. In this case, the context of the inputs is incorporated in the representation. These activations (contextual embeddings) can be used as input to another model, such as LSTM (Figure 2).

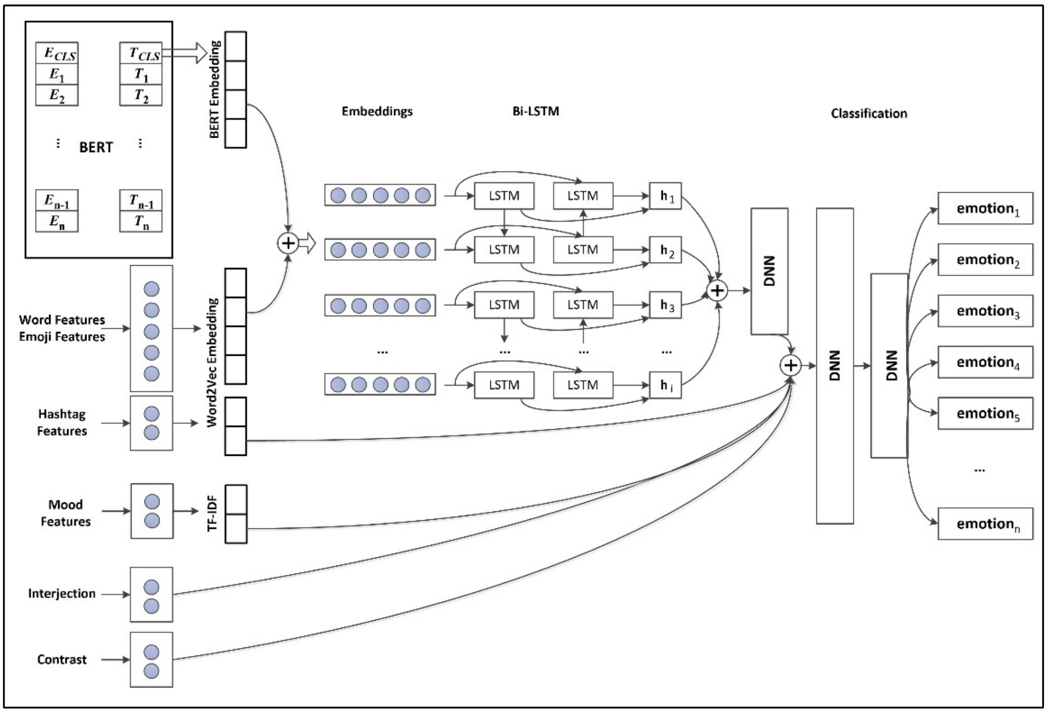


Figure 2. Proposed emotion classification using BERT, DLF, HEF, and Bi-LSTM (BERT + DLF + HEF).

4. Experiments and Results

We implemented the proposed emotion classification models in Python, using the Keras Library and Tensorflow (https://www.tensorflow.org, accessed on 28 January 2023) frameworks. The deep learning models were performed on the Google Collaboratory (https://colab.research.google.com/notebooks/welcome.ipynb, accessed on 28 January 2023) platform on a 16-GB GPU. Table 3 shows the hyperparameter values of our system. During training, we used the Adam optimizer [57] with binary cross-entropy loss.

Table 3. Parameter values.

Parameter	Bi-LSTM + DLF + HEF	BERT$_{\text{base}}$	Bi-LSTM + Att [13]
Batch size	32	8 (semEval-2018) 16 (GoEmotions)	32
Epochs	30	8	30
Learning rate	0.001	3×10^{-5}	0.001
Loss function	BinaryCrossentropy	BinaryCrossentropy	BinaryCrossentropy
Optimizer	Adam	Adam	Adam
Dropout rate	0.3		0.3

4.1. Datasets

This section presents details of the datasets used to evaluate performance of the proposed emotion classification models. The number of instances in each emotion label and the distribution percentages of instances in the SemEval-2018 dataset and the GoEmotions dataset is provided in Tables 4 and 5. Each of these datasets provided separate training, development, and test datasets.

Table 4. Total number of instances in dataset.

Dataset	Number of Instances
SemEval-2018 Task 1: E-C	Train (6838), Development (886), Test (3259)
GoEmotions	Train (43,410), Development (5426), Test (5427)

Table 5. Data statistics in Semval-2018 Task 1 dataset and GoEmotions dataset (Dev stands for development).

Dataset	Emotion	Number of Instances				Emotion	Number of Instances			
		Train	Dev	Test	Total		Train	Dev	Test	Total
Semval-2018	Anger	2544	315	1101	3960	Optimism	1984	307	1143	3434
	Anticipation	978	124	425	1527	Pessimism	795	100	375	1270
	Disgust	2602	319	1099	4020	Sadness	2008	265	960	3233
	Fear	1242	121	485	1848	Surprise	361	35	170	566
	Joy	2477	400	1442	4319	Trust	357	43	153	553
	Love	700	132	516	1348					
GoEmotions	Admiration	4130	488	504	5122	Fear	596	90	78	764
	Amusement	2328	303	264	2895	Gratitude	2662	358	352	3372
	Anger	1567	195	198	1960	Grief	77	13	6	96
	Annoyance	2470	303	320	3093	Joy	1452	172	161	1785
	Approval	2939	397	351	3687	Love	2086	252	238	2576
	Caring	1087	153	135	1375	Nervousness	164	21	23	208
	Confusion	1368	152	153	1673	Optimism	1581	209	186	1976
	Curiosity	2191	248	284	2723	Pride	111	15	16	142
	Desire	641	77	83	801	Realization	1110	127	145	1382
	Disappointment	1269	163	151	1583	Relief	153	18	11	182
	Disapproval	2022	292	267	2581	Remorse	545	68	56	669
	Disgust	793	97	123	1013	Sadness	1326	156	143	1625
	Embarrassment	303	35	37	375	Surprise	1060	141	129	1330
	Excitement	853	96	103	1052	Neutral	14,219	1787	1766	17,772

SemEval-2018 Task 1: E-C [58]: This dataset consists of English Tweets and each Tweet is labeled as neutral (no emotion) or as one or more of eleven emotions: anger, anticipation, disgust, fear, happiness, love, optimism, pessimism, sadness, surprise, and trust.

GoEmotions [43]: This dataset includes 58,000 English Reddit comments, which are manually annotated as neutral or one or more of 27 emotion categories. We chose this dataset because of its similarities with Tweets (includes short messages and supports emoji).

The statistics of the training dataset after performing data augmentation are depicted in Table 6. Data augmentation is applied to emotion classes with the lowest number of instances.

Table 6. Data statistics in the Semval-2018 train dataset and the GoEmotions train dataset after data augmentation.

SemEval-2018		**GoEmotions**					
Emotion	**Train**	**Emotion**	**Train**	**Emotion**	**Train**	**Emotion**	**Train**
Anger	3502	Admiration	4167	Disgust	812	Realization	1135
Anticipation	2485	Amusement	2352	Embarrassment	914	Relief	308
Disgust	3743	Anger	1586	Excitement	861	Remorse	563
Fear	2983	Annoyance	2500	Fear	672	Sadness	1412
Joy	4195	Approval	2959	Gratitude	2682	Surprise	1073
Love	1613	Caring	1111	Grief	235	Neutral	14,277
Optimism	3442	Confusion	1377	Joy	1473		
Pessimism	1927	Curiosity	2218	Love	2094		
Sadness	3287	Desire	643	Nervousness	497		
Surprise	967	Disappointment	1311	Optimism	1613		
Trust	969	Disapproval	2039	Pride	224		

4.2. Evaluation Measures

To obtain a greater insight into the performance of the proposed classification models, we used Jaccard accuracy, F1-score, precision, and recall. Precision measures the quality of a classifier. Higher precision indicates less false positives (FP), while lower precision means more false positives. Moreover, higher recall means less false negatives (FN), which indicates that more truly relevant results are returned.

The true positives (TP), false positives (FP), false negative (FN), and true negative (TN) values for each emotion class are defined as TP occurring when the classification correctly predicts the emotion labels, and, therefore, the predicted labels are in the gold labels. FP occurs when the classification incorrectly predicts emotion labels that are not in the gold labels. FN occurs when emotion labels are in the gold labels but classification falsely predicted emotion labels. TN occurs when classification correctly predicts the emotion labels that are not in the gold labels.

$$Jaccard\ accuracy = \frac{1}{|T|} \sum_{t \in T} \frac{G_t \cap\ P_t}{G_t \cup\ P_t} \tag{10}$$

where G_t is the set of the gold labels for Tweet t, P_t is the set of the predicted labels for Tweet t, and T is the set of Tweets. Moreover, we used a micro-averaged F-score and macro-averaged F-score, which take into account both precision and recall.

To calculate micro-averaged results, TP, FP, and FN for each emotion label are summed up, and the average is taken:

$$Precision_{micro} = \frac{\sum_{e \in E} TP}{\sum_{e \in E} TP + \sum_{e \in E} FP} \tag{11}$$

$$Recall_{micro} = \frac{\sum_{e \in E} TP}{\sum_{e \in E} TP + \sum_{e \in E} FN} \tag{12}$$

$$F1_{micro} = \frac{2 \times Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{13}$$

For macro-averaged results, the precision and recall are calculated independently for each emotion label e, and then the average is taken:

$$Precision_e = \frac{TP_e}{TP_e + FP_e} \tag{14}$$

$$Recall_e = \frac{TP_e}{TP_e + FN_e} \tag{15}$$

$$F1_{macro} = \frac{1}{|E|} \sum_{e \in E} \frac{2 \times Precision_e \times Recall_e}{Precision_e + Recall_e} \tag{16}$$

4.3. Results and Discussion

In this section, we discuss the experimental results obtained from the proposed models on the two benchmark datasets. Precision, Recall, F1-score, and Jaccard accuracy are utilized to evaluate the performance of the proposed models with the baseline approaches for emotion classification. In our proposed model, we incorporated hybrid features (HEF+DLF) with a deep learning model (Bi-LSTM) or transformer-based model (BERT).

Our experiments examined the performance of the proposed models, and Tables 7–9 show the comparison results on the SemEval-2018 task 1: E-c dataset and GoEmotions dataset. In our recent study [13], we performed the Bi-LSTM + Att model only on Tweets that included emojis in order to examine the effectiveness of the emoji feature in emotion classification. In the current study, we performed the same model on the original dataset and compared the results with the proposed model that incorporated HEF and DLF features.

Table 7. Comparison results of the proposed models with state-of-the-art models on the SemEval-2018 task 1:E-C dataset.

Model	Accuracy (Jaccard)	$F1^{macro}$	$F1^{micro}$
Proposed model (Bi-LSTM + DLF + HEF)	**68.40**	**65.77**	**78.55**
Bi-LSTM	57.30	52.25	68.53
BERT	57.21	53.28	69.16
BERT + Bi-LSTM	56.88	53.71	69.05
Ahanin and Ismail (Bi-LSTM + Att) [13]	60.53	56.4	71.1
Ameer et al. (RoBERTa + MA) [18]	62.4	60.3	74.2
Alhuzali and Ananiadou (BERT) [41]	60.1	57.8	71.3
Ameer et al. (RF + BR) [38]	45.2	55.9	57.3
Jabreel and A. Moreno (GRU) [40]	59.0	56.4	69.2
Baziotis et al. (Bi-LSTM) [25]	58.8	52.8	70.1

Table 8. Comparison results of the proposed models with baselines on the SemEval-2018 task 1: E-C dataset.

Model	Precision	Recall	F1-Score
Proposed model (Bi-LSTM + DLF + HEF)	79.11	83.34	80.49
Bi-LSTM	72.79	62.70	66.04
BERT	71.92	64.26	67.03
BERT + Bi-LSTM	70.66	65.34	67.31
Ahanin and Ismail (Bi-LSTM + Att) [13]	70.22	70.31	69.41

Table 9. Comparison results of the proposed models with baselines on the GoEmotions dataset.

Model	Accuracy (Jaccard)	Precision^{macro}	Recall^{macro}	F1^{macro}
Proposed model (BERT + DLF + HEF)	53.45	54%	46%	49%
Bi-LSTM	46.34	61%	27%	31%
BERT + Bi-LSTM	49.87	51%	36%	41%
Demszky et al. (BERT) [43]		40%	63%	46%
Ahanin and Ismail (Bi-LSTM + Att) [13]	50.2	63%	35%	41%

We compared the results of the proposed model with existing studies using F1^{macro}, F1^{micro}, and Jaccard accuracy (Table 7). The results indicate that our proposed model performed better than the other models, and the BERT and Bi-LSTM models yielded similar values. It also shows that out of the baselines, the RF+BR approach by Ameer et al. [38] performed least in terms of Jaccard accuracy and F1^{micro}. This could be because the semantic similarities between the tokens were not captured properly. The proposed model based on RoBERTa and the multiple-attention mechanism by Ameer et al. [18] outperformed other baselines in terms of Jaccard accuracy by attaining an accuracy of 62.40%. The Bi-LSTM and Attention model by Ahanin and Ismail [13] and the BERT model by Alhuzali and Ananiadou [41] attained the Jaccard accuracy of 60.53% and 60.1%, respectively. The BERT model [41] captured emotion correlations by including label information to the input sentence, and this might be why they achieved higher results compared to the baseline BERT model (57.21%). The proposed model outperformed the model of Ameer et al. [18], achieving improvements of 6%, 5%, and 4% in Jaccard accuracy, F1^{macro}, and F1^{micro}, respectively. According to the results, the models that used the attention mechanism achieved higher accuracy.

Table 8 demonstrates the performance of the proposed model together with baselines under each metric. On average, compared with Bi-LSTM, the proposed emotion classification model achieved +20.64% recall and +14.45% F1-score improvement. The Bi-LSTM with attention mechanism [13] attained a precision of 70.22% and a recall of 70.31%. BERT and Bi-LSTM models appeared to have a similar performance, and a combination of the two models (BERT + Bi-LSTM) slightly improved recall but achieved lower precision. The higher recall value shows that all the models have a lower FN rate. The proposed method resulted in higher precision, which shows that the classifier predictions are indeed legitimate. Moreover, the higher F1-score in the proposed method indicates the higher accuracy of the classifier. Therefore, it can be said that incorporating more features such as syntactic features, sentiment features, domain-specific features (mood tags), hashtags, and emoji assist classifiers enable the best performance.

The predictive results obtained on the GoEmotions dataset are presented in Table 9. In this section, due to the competitive results generated by Demszky et al. [43], we combined the hybrid features (DLF + HEF) with BERT. Demszky et al. [43] employed a BERT-based model and applied fine-tuning by adding a dense output layer on top of the pretrained model. Based on the results presented in Table 9, our proposed model (BERT + DLF + HEF) improved precision (54%) when compared to the models of Demszky et al. [43] (40%) and BERT + Bi-LSTM (51%). On the same note, the model of Ahanin and Ismail [13] recorded the highest precision (63%). The lowest F1macro belongs to the Bi-LSTM model with 31%, and the highest F1macro is achieved by the proposed model (49%), followed by the models of Demszky et al. [43] (46%), Ahanin and Ismail [13] (41%), and Bi-LSTM (41%). Similarly, the proposed model outperformed the other models and obtained the highest Jaccard accuracy (53.45%), while Bi-LSTM achieved the lowest Jaccard accuracy (46.34%). We observed an improvement when combining BERT and Bi-LSTM models (BERT + Bi-LSTM) compared to the Bi-LSTM model using the GoEmotions dataset.

The previous studies have utilized pre-trained word embeddings for the emotion classification problem. These pre-trained models already retain most of the semantics of the terms present in a sentence, thus decreasing the need for very large, supervised training data. However, these pre-trained models are mainly trained on general datasets, which ignores the nuances of human emotion in text. Fine-grained emotions (e.g., joy, sadness, disgust, etc.) are often mixed and ambiguous, which adds complexity to the process of emotion detection. Therefore, we used augmentation to generate a larger training dataset from existing data in the domain of emotion classification and to increase the number of training samples for emotion classes containing the lowest number of instances. Data augmentation generates more emotional related features, reduces overfitting, and enhances performance. However, none of the previous studies (see discussion in Section 2) have applied data augmentation on these multi-label emotion datasets. Moreover, since this dataset is dealing with multilabel emotion classification, instead of using −1, 0, +1 for sentiment polarity, we used a SCS to generate a range, which is then transformed into one-hot encoding vectors. According to the results, extracting hybrid features such as semantic features, interjections, mood tags, hashtags, emojis, and sentiment scores improved the performance of the model, and especially the emotion classes with lower number of instances.

Figure 3 illustrates the value of the F1-score in each emotion class on the SemEval-2018 dataset, in which the proposed model showed better performance in almost all emotion classes. Based on the results presented in Figure 3, the proposed model improved the F1-score in emotion classes with the smallest number of instances, such as trust, surprise, and pessimism. Moreover, emotion classes with higher numbers of instances such as joy and love achieved the highest F1-score results. The Bi-LSTM + Att model proposed by Ahanin and Ismail [13] achieved the second-best performance results, followed by the BERT and Bi-LSTM models.

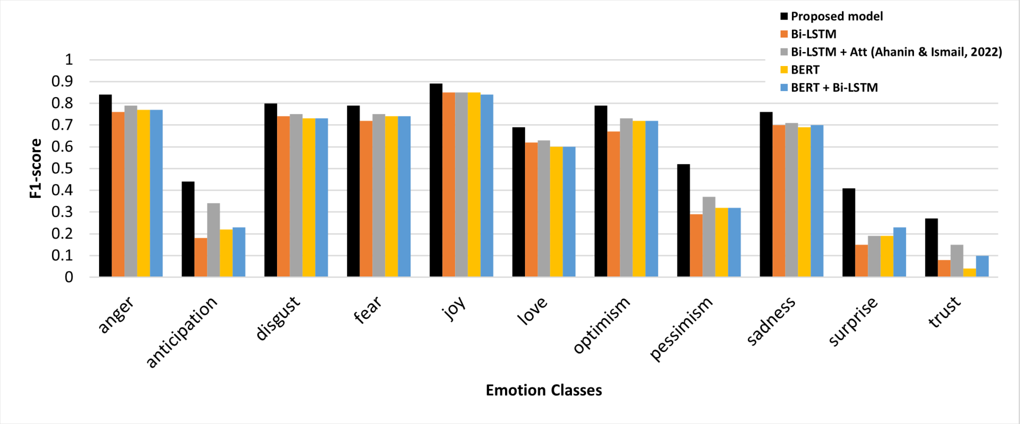


Figure 3. F1-score in each emotion class on the SemEval-2018 task 1: E-C dataset. Proposed model refers to Bi-LSTM + HEF + DLF. Results of Bi-LSTM + Att model is based on research by Ahanin and Ismail [13].

Figure 4 depicts the results on the GoEmotions dataset. Compared to other state-of-the-art models, the proposed model improved the performance of emotion classification significantly, particularly in emotion classes with a smaller number of instances. All of the classification models failed to detect the emotion of grief, but our proposed model achieved an F1-score of 22%. Similarly, the proposed model recorded the highest F1-score in amusement, annoyance, confusion, excitement, joy, love, pride, relief, and sadness compared to other models. Table 10 provides details of precision, recall, and F1-score for each emotion class.

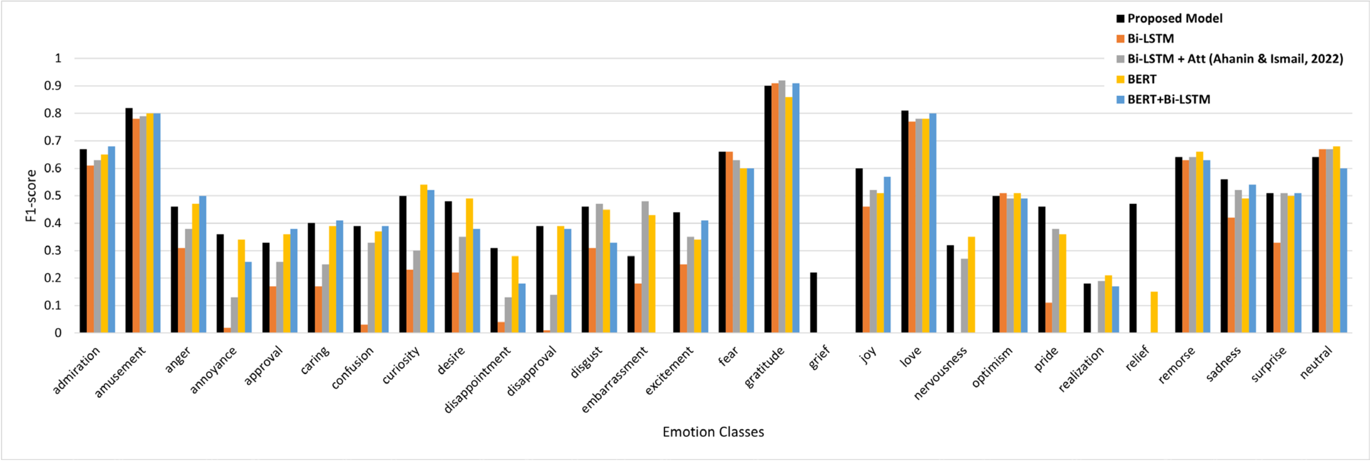


Figure 4. Result of Emotion Classification (F1-score) on GoEmotions dataset. Proposed model refers to BERT + HEF + DLF. Results of Bi-LSTM + Att model is based on research by Ahanin and Ismail [13].

Table 10. Result of proposed model (BERT + DLF + HEF) on GoEmotions dataset.

Emotion	Precision	Recall	F1-Score	Emotion	Precision	Recall	F1-Score
Admiration	65%	68%	67%	Fear	68%	64%	66%
Amusement	79%	84%	82%	Gratitude	92%	88%	90%
Anger	51%	42%	46%	Grief	33%	17%	22%
Annoyance	39%	34%	36%	Joy	64%	57%	60%
Approval	33%	34%	33%	Love	78%	83%	81%
Caring	43%	39%	40%	Nervousness	4%	26%	32%
Confusion	36%	42%	39%	Optimism	54%	46%	50%
Curiosity	51%	49%	50%	Pride	60%	38%	46%
Desire	57%	41%	48%	Realization	24%	14%	18%
Disappointment	38%	26%	31%	Relief	67%	36%	47%
Disapproval	41%	37%	39%	Remorse	59%	7%	64%
Disgust	52%	41%	46%	Sadness	62%	51%	56%
Embarrassment	40%	22%	28%	Surprise	57%	46%	51%
Excitement	51%	39%	44%	Neutral	66%	61%	64%

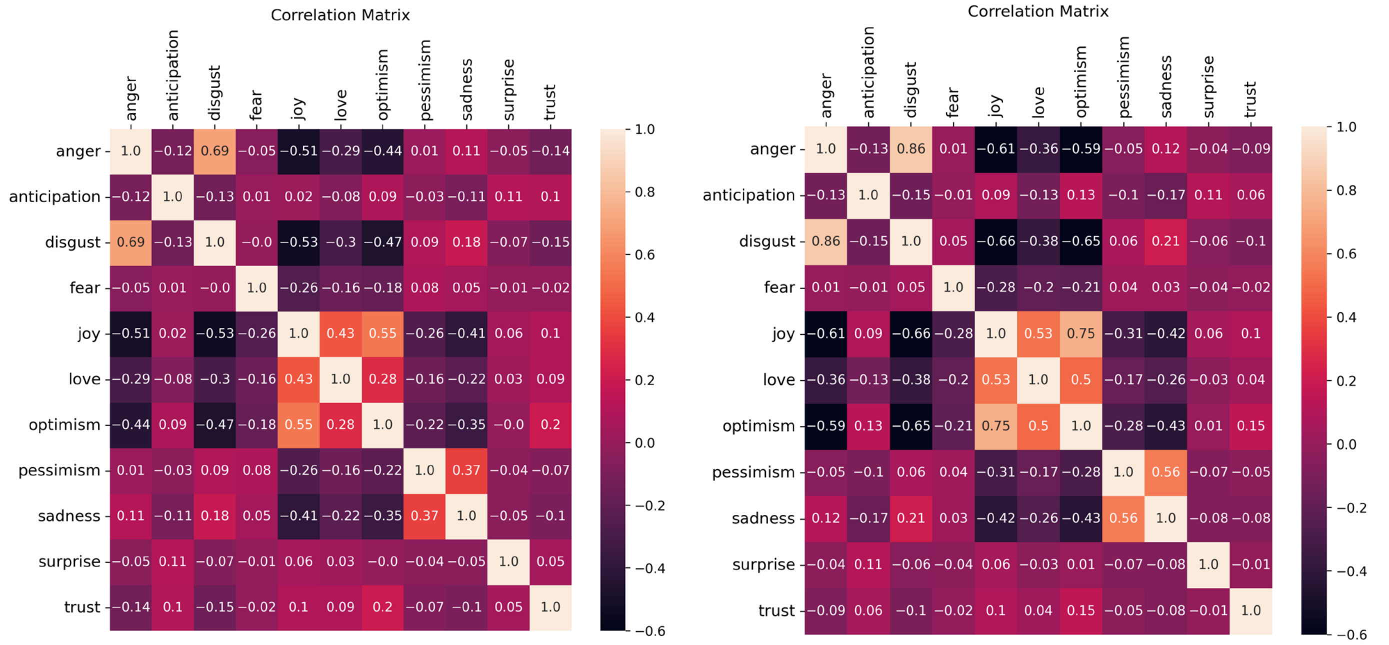
It can be observed from Table 10 that the proposed model attained the best F1-score on the emotions of gratitude (90%), amusement (82%), and love (81%), which are positive emotions with more explicit emotional words. The lowest F1-score is recorded for realization (18%), grief (22%), and embarrassment (28%).

According to the results, certain emotions are simpler to recognize by classifier than others, regardless of their frequency of samples. In the SemEval-2018 dataset, emotion trust and emotion surprise have almost the same number of train and test samples. Despite that, the F1-score for emotion surprise (41.13%) is almost double that of the emotion trust (27.45%). Emotion anticipation with 2485 samples achieved a 44.2% accuracy, which is lower than the emotion of love (69.28%) with 1613 instances. Other emotions, including anger, disgust, fear, optimism, joy, and sadness, achieved similar accuracy.

In the GoEmotions dataset, emotion amusement with 2352 instances, joy with 1473 instances, and love with 2094 instances achieved 82%, 60%, and 81%, respectively, while the emotion of admiration with 4176 instances obtained an F1-score of 67%. Moreover, the emotion of annoyance with 2500 samples attained a lower F1-score (36%) compared to fear with 672 samples (66%). Similarly, the emotion of remorse achieved a higher F1-score, despite having a smaller number of instances. Thus, some emotions, such as joy, love, remorse, sadness, fear, anger, or amusement, have characteristics and indicators that are simpler for the classifier to grasp. For example, people tend to use more emojis when expressing joy, love, sadness, and anger. Emojis, hashtags, and affective features are among the indicators that can improve the accuracy of a classifier.

4.4. Correlation Analysis

From a confusion matrix of our model (Figure 5), it is notable that the model captured relations among the emotion labels. The correlation scores between pairs of emotions in predicted labels are very similar to ground truth, and there are only a small margin of misclassifications among the emotion labels. For instance, relationships between the emotions of fear and pessimism are successfully captured.



(**a**) (**b**)

Figure 5. Correlation matrices of emotion labels (SemEval-2018 dataset). (**a**) Emotion Correlations: Ground Truth. (**b**) Emotion Correlation: Prediction.

We further generated an accuracy score obtained by the proposed model for each emotion label. The results for the SemEval-2018 dataset and the GoEmotions dataset are presented in Tables 11 and 12, respectively.

Table 11. Accuracy for each emotion class in SemEval-2018 Task-1: E-C dataset.

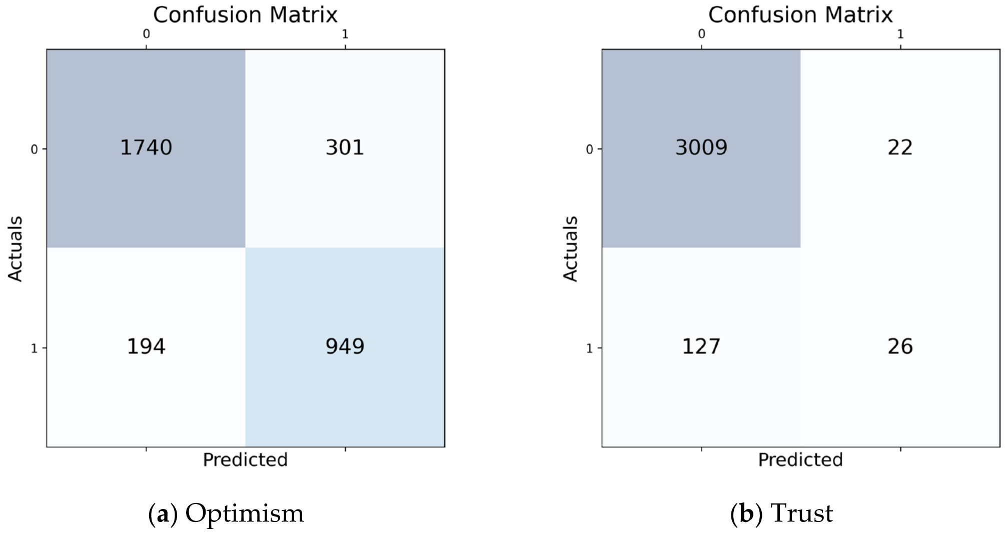
Emotion	**Accuracy**
Anger	89.82
Anticipation	88.69
Disgust	86.93
Fear	94.69
Joy	90.70
Love	89.91
Optimism	84.45
Pessimism	90.10
Sadness	87.53
Surprise	95.54
Trust	95.32

Table 12. Accuracy for each emotion class in GoEmotions dataset.

Emotion	Accuracy	Emotion	Accuracy
Admiration	82.68	Fear	97.24
Amusement	90.46	Gratitude	87.80
Anger	93.55	Grief	99.83
Annoyance	89.79	Joy	94.56
Approval	87.64	Love	91.39
Caring	95.38	Nervousness	99.30
Confusion	94.03	Optimism	93.94
Curiosity	90.44	Pride	99.52
Desire	97.44	Realization	95.80
Disappointment	95.41	Relief	99.69
Disapproval	91.10	Remorse	97.83
Disgust	96.04	Sadness	94.80
Embarrassment	98.99	Surprise	95.38
Excitement	96.72	Neutral	58.02

According to the results, our model correctly classified the emotion of trust with an accuracy of 95.32%. This is followed by the emotions of surprise, fear, joy, pessimism, and love. Despite recording a very high accuracy for the emotion of trust, it appears that the model is not skilled enough to detect Tweets that evoke this emotion. As was discussed earlier, there could be less emotional indicators for emotion trust compared to other emotion classes such as joy and love. Therefore, the recall is lower, which results in a lower F1-score (27%). Compared to the recent study by Ameer et al. [18], our model scored a better accuracy for the emotion of trust.

We illustrated the confusion metrices (Figure 6) for the emotions of trust and optimism on the SemEval-2018 dataset, which scored the highest and lowest accuracy, respectively.



(**a**) Optimism (**b**) Trust

Figure 6. Confusion matrices obtained for the emotions Trust and Optimism on SemEval-2018 dataset, where 1 indicates that the Tweet is labeled with the specified emotion and 0 means that the Tweet is not labeled. The highlights represent the proportion that each entry value in matrix takes up relative to all of the entries.

Beside the size of the emotion class in the dataset, we believe that the Out-Of-Vocabulary (OOV) issue and the misspellings added too much complexity and too many challenges for the model to generate accurate results.

We provide a deeper insight into the effects of considering hybrid features (DEF and HEF) in emotion classes. Table 13 presents cases that are correctly classified by the proposed model but that failed to be classified correctly by the baseline models, Bi-LSTM or BERT. In contrast, Table 14 presents cases that are correctly classified by either of the baseline models. It is evident that the proposed model is more effective at identifying non-dominant emotions such as fear or trust than the baseline models. Features such as emojis, interjections, syntactic features, moods, or hashtags are the main ones that are found to be effective in determining the correct emotion classes. The results suggest that the proposed features can mitigate the overfitting problem, and can consequently make the model more robust.

Table 14 presents special situations where the proposed model cannot classify Tweets correctly. It can be observed that despite failing to produce precise predictions, the predicted emotion classes were reasonable. For instance, in Cases #6, #7, and #9, the proposed model predicted additional class labels that are correlated to the actual emotion classes (e.g., sadness and pessimism, joy and optimism). In cases #8 and #10, the words in the Tweet indicate negative emotions, and so the predicted emotion is one of the negative emotion classes (sadness or anger).

Table 13. Case studies where the proposed model correctly classified the Tweets, yet was incorrectly classified by the baseline models.

Sentence	Label(s)	Prediction		
		Ours	Baseline (Bi-LSTM)	Baseline (BERT)
[Case #1] #Deppression is real. Partners w/ #depressed people truly dont understand the depth in which they affect us. Add in #anxiety & makes it worse	Fear, Sadness	Fear, Sadness	Sadness	Joy
[Case #2] Some people just need to learn how to #smile 😁 and #laugh 😂😂 #live #life	Joy, Love, optimism	Joy, Love, Optimism	Joy, Love, Optimism	Anticipation, Optimism,
[Case #3] We may have reasons to disagree but that in itself is no reason to hate. Our resentment is more a reflection of ourselves than others.	Anger, Optimism, Trust	Anger, Optimism, Trust	Anger, Disgust, Sadness	Anticipation
[Case #4] Life is too short to hide your feelings. Don’t be afraid to say what you feel.	Fear, Optimism	Fear, Optimism	Fear	Sadness
[Case # 5] Can’t sleep!! Maybe #worry !!!!!! Or Maybe I need to Chang my pillow !! Maybe..! 😏	Fear, Sadness	Fear, Sadness	None	Sadness

Table 14. Case studies where the proposed model failed to classify the Tweets correctly.

Sentence	**Label(s)**	**Prediction**		
		Ours	**Baseline (Bi-LSTM)**	**Baseline (BERT)**
[Case #6] Comparing yourself to others is one of the root causes for feelings of unhappiness and depression.	Sadness	Sadness, Pessimism	Sadness	Anger, Disgust
[Case #7] Scared to leave the routine but excited to break out the mould 😖 #scared #confused #happy #undecided #excited	Fear, Joy	Fear, Joy Optimism	Fear, Joy	Fear, Sadness
[Case #8] You hold my every moment you calm my raging seas you walk with me through fire and heal all my disease ❤❤❤🖤🖤	Joy, Love, Optimism	Anger, Disgust, Love	Joy, Love, Optimism	Anger, Disgust
[Case #9] What makes you feel #joyful?	Joy	Joy, Love, Optimism	Joy, Love, Optimism	Joy
[Case #10] suddenly I want to be in the middle of chaos, feel the #wonderful sense of sound; my feet are tired of these long stretched silences	Joy, Optimism	Sadness	Sadness	Joy, optimism

5. Conclusions

Detecting emotions has been an appealing research topic, and emotion classification aims to detect emotions in Natural Language Processing applications. The accurate identification of emotion can enhance Human-Computer Interaction (HCI) systems. In this paper, we proposed hybrid features (DLF + HEF) focusing on the word embedding feature, emoji feature, hashtag feature, semantic similarities, syntactic features, and domain-specific features. Moreover, data augmentation is performed to compensate for emotion classes with a smaller number of instances, since both benchmark datasets were imbalanced. Two benchmark datasets, SemEval-2018 and GoEmotions, were utilized for classification using two learning models. The first model employs hybrid features incorporated with a deep learning model (Bi-LSTM) with the Attention mechanism in order to give the keywords that have been learnt more weight, and the second model is based on integrating the Transformer model (BERT) with Bi-LSTM. The conducted experiments have shown that incorporating the hybrid features with deep learning models delivered a significant improvement to the results. Furthermore, it enhanced the prediction of the emotion labels with small numbers of instances, such as surprise, fear, and trust. Additionally, hashtags and emojis, which are more likely to be used with emotions such as love or sadness, have been employed with SentiWordNet lexicon and syntactic features to improve the recognition of other emotions, which lack distinct characteristics and indicators such as pride, realization, and surprise. The Bi-LSTM+DLF+HEF model performed better in the SemEval-2018 dataset and achieved the highest F1^{macro} (65.77%) and Jaccard accuracy (68.40%), while the BERT+DLF+HEF model attained the highest F1^{macro} (49%) and Jaccard accuracy (53.45%) in the GoEmotions dataset. Comparing the obtained results with the baselines indicated that the proposed model outperformed all of the baseline approaches in terms of $F1^{macro}$ and Jaccard accuracy, which shows the significance of the proposed technique for emotion classification. Researchers and practitioners can use this method in intelligent systems to understand emotions in various fields, such as affective HCI and social data mining, to provide personalized solutions. Despite the use of hybrid features, the model has limitations in recognizing slang or idioms, such as *be on cloud nine*. In the future, we plan to experiment with the hybrid features with other deep learning models that have delivered promising results in several NLP tasks. Moreover, we will focus on extracting more emotion-enriched features to improve the robustness of the model.

Author Contributions: Conceptualization, Z.A., A.A.-A. and M.A.I.; methodology, Z.A., M.A.I. and N.S.S.S.; software, Z.A.; validation, Z.A. and M.A.I.; formal analysis, Z.A., M.A.I. and N.S.S.S.; investigation, Z.A., M.A.I. and N.S.S.S.; resources, Z.A.; data curation, Z.A. and A.A.-A.; writing—original draft preparation, Z.A. and A.A.-A.; writing—review and editing, Z.A., M.A.I. and N.S.S.S.; visualization, Z.A.; supervision, M.A.I. and N.S.S.S.; funding acquisition, N.S.S.S. and A.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The SemEval-2018 dataset provided by Mohammad et al. [58] as a part of the competition SemEval-2018 Task 1: Affect in Tweets (task E-C), available at https://competitions.codalab.org/competitions/17751#learn_the_details-datasets (accessed on 28 January 2023). The GoEmotions dataset is provided by Demszky et al. [43], and it is available at https://github.com/google-research/google-research/tree/master/goemotions/data (accessed on 28 January 2023).

Acknowledgments: The authors wish to extend their sincere thanks to the support of the Faculty of Data Science and Information Technology, INTI International University, Malaysia for providing the state-of-the-art research support to carry on this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahanin, E.; Bakar Sade, A.; Tat, H.H. Applications of Artificial Intelligence and Voice Assistant in Healthcare. *Int. J. Acad. Res. Bus. Soc. Sci.* **2022**, *12*, 2545–2554. [CrossRef] [PubMed]
2. Fernández, A.P.; Leenders, C.; Aerts, J.M.; Berckmans, D. Emotional States versus Mental Heart Rate Component Monitored via Wearables. *Appl. Sci.* **2023**, *13*, 807. [CrossRef]
3. Ekman, P.; Friesen, W.V. Hand Movements. *J. Commun.* **1972**, *22*, 353–374. [CrossRef]
4. Parrott, W.G. *Emotions in Social Psychology: Essential Readings*; Psychology Press: East Sussex, UK, 2001; ISBN 0863776833.
5. Plutchik, R.; Kellerman, H. *Emotion, Theory, Research, and Experience*; Academic Press: Cambridge, MA, USA, 1980.
6. Cambria, E.; Livingstone, A.; Hussain, A. The Hourglass of Emotions. In *Proceedings of the Cognitive Behavioural Systems*; Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 144–157.
7. Susanto, Y.; Livingstone, A.G.; Ng, B.C.; Cambria, E. The Hourglass Model Revisited. *IEEE Intell. Syst.* **2020**, *35*, 96–102. [CrossRef]
8. Ren, G.; Hong, T. Investigating Online Destination Images Using a Topic-Based Sentiment Analysis Approach. *Sustainability* **2017**, *9*, 1765. [CrossRef]
9. Adamu, H.; Lutfi, S.L.; Malim, N.H.A.H.; Hassan, R.; di Vaio, A.; Mohamed, A.S.A. Framing Twitter Public Sentiment on Nigerian Government COVID-19 Palliatives Distribution Using Machine Learning. *Sustainability* **2021**, *13*, 3497. [CrossRef]
10. Huang, Y.; Bo, D. Emotion Classification and Achievement of Students in Distance Learning Based on the Knowledge State Model. *Sustainability* **2023**, *15*, 2367. [CrossRef]
11. Zhang, X.; Yan, Z.; Wu, Q.; Wang, K.; Miao, K.; Wang, Z.; Chen, Y. Community Governance Based on Sentiment Analysis: Towards Sustainable Management and Development. *Sustainability* **2023**, *15*, 2684. [CrossRef]
12. Liang, K.; He, J.; Wu, P. Trust Evaluation Method of E-Commerce Enterprises with High-Involvement Experience Products. *Sustainability* **2022**, *14*, 15562. [CrossRef]
13. Ahanin, Z.; Ismail, M.A. A Multi-Label Emoji Classification Method Using Balanced Pointwise Mutual Information-Based Feature Selection. *Comput. Speech Lang.* **2022**, *73*, 101330. [CrossRef]
14. Liu, Y.; Li, P.; Hu, X. Combining Context-Relevant Features with Multi-Stage Attention Network for Short Text Classification. *Comput. Speech Lang.* **2022**, *71*, 101268. [CrossRef]

15. Mustafa Hilal, A.; Elkamchouchi, D.H.; Alotaibi, S.S.; Maray, M.; Othman, M.; Abdelmageed, A.A.; Zamani, A.S.; Eldesouki, M.I. Manta Ray Foraging Optimization with Transfer Learning Driven Facial Emotion Recognition. *Sustainability* **2022**, *14*, 14308. [CrossRef]
16. Kumar, L.A.; Renuka, D.K.; Rose, S.L.; Priya, M.C.S.; Wartana, I.M. Deep Learning Based Assistive Technology on Audio Visual Speech Recognition for Hearing Impaired. *Int. J. Cogn. Comput. Eng.* **2022**, *3*, 24–30. [CrossRef]
17. Weng, Z.; Qin, Z.; Tao, X.; Pan, C.; Liu, G.; Li, G.Y. Deep Learning Enabled Semantic Communications with Speech Recognition and Synthesis. *IEEE Trans. Wirel. Commun.* **2023**. [CrossRef]
18. Ameer, I.; Bölücü, N.; Siddiqui, M.H.F.; Can, B.; Sidorov, G.; Gelbukh, A. Multi-Label Emotion Classification in Texts Using Transfer Learning. *Expert Syst. Appl.* **2023**, *213*, 118534. [CrossRef]
19. Eke, C.I.; Norman, A.A.; Shuib, L. Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model. *IEEE Access* **2021**, *9*, 48501–48518. [CrossRef]
20. Waheeb, S.A.; Khan, N.A.; Shang, X. An Efficient Sentiment Analysis Based Deep Learning Classification Model to Evaluate Treatment Quality. *Malays. J. Comput. Sci.* **2022**, *35*, 1. [CrossRef]
21. Priyadarshini, I.; Cotton, C. A Novel LSTM-CNN-Grid Search-Based Deep Neural Network for Sentiment Analysis. *J. Supercomput.* **2021**, *77*, 13911–13932. [CrossRef]
22. Ahanin, Z.; Ismail, M.A. Feature Extraction Based on Fuzzy Clustering and Emoji Embeddings for Emotion Classification. *Int. J. Technol. Manag. Inf. Syst.* **2020**, *2*, 102–112.
23. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
24. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781. [CrossRef]
25. Baziotis, C.; Athanasiou, N.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; Potamianos, A. Ntua-Slp at Semeval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive Rnns and Transfer Learning. *arXiv* **2018**, arXiv:1804.06658. [CrossRef]
26. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
27. Sakata, W.; Tanaka, R.; Shibata, T.; Kurohashi, S. FAQ Retrieval Using Query-Question Similarity and BERT-Based Query-Answer Relevance. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; Association for Computing Machinery, Inc.: New York, NY, USA, 2019; pp. 1113–1116.
28. Qu, C.; Yang, L.; Qiu, M.; Croft, W.B.; Zhang, Y.; Iyyer, M. BERT with History Answer Embedding for Conversational Question Answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; ACM: New York, NY, USA, 2019; pp. 1133–1136.
29. Singh, P.K.; Paul, S. Deep Learning Approach for Negation Handling in Sentiment Analysis. *IEEE Access* **2021**, *9*, 102579–102592. [CrossRef]
30. Jianqiang, Z.; Xiaolin, G.; Xuejun, Z. Deep Convolution Neural Networks for Twitter Sentiment Analysis. *IEEE Access* **2018**, *6*, 23253–23260. [CrossRef]
31. Li, Z.; Xie, H.; Cheng, G.; Li, Q. Word-Level Emotion Distribution with Two Schemas for Short Text Emotion Classification. *Knowl. Based Syst.* **2021**, *227*, 107163. [CrossRef]
32. Huang, C.; Trabelsi, A.; Zaïane, O.R. ANA at SemEval-2019 Task 3: Contextual Emotion Detection in Conversations through Hierarchical LSTMs and BERT. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; Association for Computational Linguistics (ACL): Minneapolis, MI, USA, 2019; pp. 49–53.
33. Zhang, X.; Li, W.; Ying, H.; Li, F.; Tang, S.; Lu, S. Emotion Detection in Online Social Networks: A Multilabel Learning Approach. *IEEE Internet Things J.* **2020**, *7*, 8133–8143. [CrossRef]
34. Bollen, J.; Mao, H.; Pepe, A. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In Proceedings of the International AAAI Conference on Web and Social Media, Barcelona, Spain, 17–21 July 2011; pp. 450–453.
35. Asghar, M.Z.; Khan, A.; Bibi, A.; Kundi, F.M.; Ahmad, H. Sentence-Level Emotion Detection Framework Using Rule-Based Classification. *Cogn. Comput.* **2017**, *9*, 868–894. [CrossRef]
36. Ileri, I.; Karagoz, P. Detecting User Emotions in Twitter through Collective Classification. In Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), Porto, Portugal, 9–11 November 2016; Science and Technology Publications: North York, ON, Canada, 2016; pp. 205–212.
37. Tocoglu, M.A.; Ozturkmenoglu, O.; Alpkocak, A. Emotion Analysis from Turkish Tweets Using Deep Neural Networks. *IEEE Access* **2019**, *7*, 183061–183069. [CrossRef]
38. Ameer, I.; Ashraf, N.; Sidorov, G.; Adorno, H.G. Multi-Label Emotion Classification Using Content-Based Features in Twitter. *Comput. Y Sist.* **2020**, *24*, 1159–1164. [CrossRef]
39. Miriam, P.-A.F.; Martín-Valdiviaa, M.T.; Ureña-Lópeza, L.A.; Mitkov, R. Improved Emotion Recognition in Spanish Social Media through Incorporation of Lexical Knowledge. *Future Gener. Comput. Syst.* **2020**, *110*, 1000–1008.
40. Jabreel, M.; Moreno, A. A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets. *Appl. Sci.* **2019**, *9*, 1123. [CrossRef]

41. Alhuzali, H.; Ananiadou, S. SpanEmo: Casting Multi-Label Emotion Classification as Span-Prediction. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Kyiv, Ukraine, 19–23 April 2021; pp. 1573–1584.
42. Zygadło, A.; Kozłowski, M.; Janicki, A. Text-Based Emotion Recognition in English and Polish for Therapeutic Chatbot. *Appl. Sci.* **2021**, *11*, 10146. [CrossRef]
43. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4040–4054.
44. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V.; et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
45. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Advances in Neural Information Processing Systems: Vancouver, BC, Canada, 2019.
46. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108. [CrossRef]
47. Barbon, R.S.; Akabane, A.T. Towards Transfer Learning Techniques-BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors* **2022**, *22*, 8184. [CrossRef]
48. Alswaidan, N.; El, M.; Menai, B. A Survey of State-of-the-Art Approaches for Emotion Recognition in Text. *Knowl. Inf. Syst.* **2020**, *62*, 2937–2987. [CrossRef]
49. Gee, G.; Wang, E. PsyML at SemEval-2018 Task 1: Transfer Learning for Sentiment and Emotion Analysis. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 369–376.
50. Qin, Y.; Shi, Y.; Hao, X.; Liu, J. Microblog Text Emotion Classification Algorithm Based on TCN-BiGRU and Dual Attention. *Information* **2023**, *14*, 90. [CrossRef]
51. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
52. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 6382–6388.
53. Kobayashi, S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Association for Computational Linguistics (ACL), New Orleans, LA, USA, 6 June 2018; Volume 2, pp. 452–457.
54. Cambria, E.; Poria, S.; Hazarika, D.; Kwok, K. SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: New Orleans, LA, USA, 2018; Volume 32, pp. 1795–1802.
55. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishanathan, S., Garnett, R., Eds.; Neural Info Process: San Diego, CA, USA, 2017; pp. 5998–6008.
56. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
57. Kingma, D.P.; Ba, J.L. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
58. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. Semeval-2018 Task 1: Affect in Tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 1–17.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.