Machine Learning with Applications 21 (2025) 100693

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

EmoFusion: An integrated machine learning model leveraging embeddings and lexicons to improve textual emotion classification

Anjali Bhardwaj, Muhammad Abulaish *

Department of Computer Science, South Asian University, New Delhi, India

ARTICLE INFO

Keywords:
Machine learning
Text mining
Emotion AI
Emotion classification
Word embedding
NLP

ABSTRACT

Human emotions are complicated and intertwined with cognitive processes, influencing mental health, learning, and decision-making. The Web 2.0 era has seen a remarkable spike in the number of people sharing their experiences and emotions on online social media, mostly through posts or text messages. Due to inherent challenges associated with textual data, the issue of discovering the intricate relationships between texts and its inherent emotions is still an increasingly prevalent topic in AI and NLP. This paper presents EmoFusion, an integrated machine learning model that improves emotion classification in textual data by integrating pre-trained word embeddings and emotion lexicons. Instead of relying on a single emotion lexicon, EmoFusion integrates multiple emotion lexicons since a single lexicon might not fully cover all possible words or phrases linked with emotions. The proposed approach uses semantically related features to bridge the semantic gap between words and emotions, capturing a wide range of emotional nuances and resulting in better classification performance. The efficacy is further improved by employing emotion-specific pre-processing techniques. EmoFusion is evaluated using three benchmark datasets, namely Google AI GoEmotions, CBET, and TEC. The evaluation results demonstrate a significant improvement compared to six baselines and a state-of-the-art technique using different classifiers.

1. Introduction

Human emotions are integral components of cognitive processes, as changes in physiological aspects often convey emotions that are pertinent to various psychological states occurring spontaneously or in the absence of any conscious efforts. Emotions play a significant role in our lives, aiding in understanding fundamental human behavior and decision-making. These complex reactions can arise from both internal and external events, such as specific topics, events, or people. Emotions are pivotal for human understanding and positively impact learning, mental health, and daily interactions. Emotion classification is a challenging yet well-explored area within artificial intelligence (AI) and natural language processing (NLP), extending beyond sentiment analysis. Unlike sentiment analysis, which focuses on identifying the polarity of text, emotion classification seeks to associate text with the emotion(s) that best represent the author's psychological state (Li et al., 2020). These emotions may range from *anger* and *fear* to *joy*, covering a wide spectrum.

Over the last decade, the proliferation of social media platforms has led to an unprecedented surge in information exchange. During the COVID-19 outbreak, individuals faced unparalleled challenges, leading to diverse and overwhelming emotional responses. This led to a significant increase in the expression of feelings, experiences, and attitudes through microblogs and online social networks. Platforms like Twitter (now "X") accumulate vast volumes of structured, unstructured, and semi-structured data, making text analysis increasingly intricate. User-generated content on social media provides a rich tapestry of opinions and emotions, surpassing the insights derived from factual or professional content. This surge has prompted extensive research and development efforts to create automated tools to classify users' emotions from textual data.

Recognizing emotions from text is crucial for developing applications across diverse domains, including sentiment analysis (Maruf et al., 2024), rumor detection (Haque & Abulaish, 2022), mental health monitoring (Li et al., 2020), and social media analysis (Bhardwaj & Abulaish, 2023; Hu et al., 2013). Recent researches have focused on recognizing emotions through various modalities like gestures, speech, audio, and facial expressions (Deng & Ren, 2023); however, extracting emotions solely from text remains challenging due to its psychological and social nuances. To address this, NLP employs emotion analysis, also known as emotion classification, to identify human emotions and extract valuable insights from user-generated content.

* Corresponding author.
E-mail addresses: anjali_bhardwaj@students.sau.ac.in (A. Bhardwaj), abulaish@sau.ac.in (M. Abulaish).

https://doi.org/10.1016/j.mlwa.2025.100693
Received 21 October 2024; Received in revised form 28 April 2025; Accepted 18 June 2025
Available online 30 June 2025

Existing models for emotion classification predominantly rely on individual words and utilize numerous architectural designs to extract semantic information from sequential inputs, thereby enabling emotion comprehension at the document-level. Given the domain-specific nature of emotion classification, several efforts have been made to enhance the understanding of word-level emotional content through the development and use of several hand-crafted lexicons. These lexicons have demonstrated significant utility in enhancing model performance (Mudinas et al., 2012; Taboada et al., 2011). However, a single emotion lexicon may not comprehensively cover all possible phrases or words associated with emotions.

To address this limitation, we introduce `EmoFusion`, a lightweight and resource-efficient approach that integrates multiple emotion lexicons with contextual linguistic features to capture fine-grained emotional semantics. It leverages four widely used emotion lexicons, grounded in well-established emotion theories, are categorized into discrete and dimensional types (Borod, 2000; Yadollahi et al., 2017). Discrete lexicons, such as EmoSenticNet (Poria et al., 2014) and EmoLex (Mohammad & Turney, 2013), label words with categorical emotions based on Ekman's (1999) basic six emotions. While offering direct mappings, their fixed label sets limit expressiveness and provide no insight into emotion intensity. In contrast, dimensional lexicons, including NRC-VAD (Mohammad, 2018a) and NRC-Emotion Intensity (Mohammad, 2018b), assign continuous scores along emotional dimensions such as valence, arousal, and dominance, enabling a richer affective representation.

Unlike recent approaches that rely on large-scale transformer models (e.g., BERT, GPT), `EmoFusion` adopts a scalable design by integrating pre-trained `GloVe` embeddings with a dynamic multi-lexicon fusion mechanism. This approach reduces computational overhead, supporting real-time and low-resource deployment. Unlike static concatenation methods, our dynamic fusion module contextually synthesizes emotion signals from multi-lexicons, resolving semantic ambiguities and enhance representational quality. The approach also incorporates emotion-specific pre-processing, such as recognizing emojis and emoticons as emotion indicators. Additionally, it aggregates document-level representations that combine linguistic and lexicon-derived semantic features (Zhou et al., 2020). This design not only advances emotion classification but also strengthens the semantic-emotional connection in textual data.

The proposed `EmoFusion` follows our previous work (Bhardwaj et al., 2023). As an expansion, we address the limitations of the previous work in terms of classification accuracy. We also conducted extensive experiments on three emotion benchmark datasets and developed new baseline comparison methods to demonstrate the efficacy of our proposed approach. In summary, the main contributions of `EmoFusion` can be briefly described as follows:

- Incorporating a feature extraction technique in `EmoFusion` that is designed to generate linguistic and semantically-similar feature-based vector representations for document-level emotion classification.
- Incorporating emotion-specific pre-processing steps for effective emotion classification. This includes utilizing various emotion indicators, such as emojis and emoticons present in the text, to improve classification accuracy.
- Utilizing emotion lexicons to identify emotion words by combining semantically-similar words and their sentiment score. This integrated approach guarantees a more precise representation of the emotional context in the text data. The source code and experimental data are publicly available at GitHub.[1]

[1] https://github.com/AnjaliBhardwaj20/EmoFusion

The remainder of this paper is organized as follows. Section 2 provides a concise review of related work concerning emotion classification in textual data, focusing on harnessing the synergy of pre-trained word embedding and emotion lexicon-based feature extraction. Section 3 elaborates our proposed approach, encompassing the problem definition, word-level feature vectors generation, and document-level feature vectors generation for emotion classification. Section 4 details the experimental setup, followed by results and comparative analysis in Section 5. Discussion and limitations are covered in Section 6. Finally, Section 7 summarizes our work and presents future direction of research.

2. Related work

This section presents an overview of the literature focusing on classification approaches that harness the synergy of pre-trained word embedding and emotion lexicon-based feature extraction to discern emotions within textual data. A summary of the related works on emotion classification is presented in Table 1.

Textual emotion classification is vital for effective human-to-human communication and achieving seamless interactions between humans and machines. Such integration enables systems to adapt their responses and behavior patterns based on human emotions, thereby making interaction more natural. In some situations, textual communication may be the best or only means for people to communicate. For instance, individuals with speech impairments rely on text messages to convey their thoughts. Moreover, mental health (Li et al., 2020) experts recommend writing as it can improve mood and manage symptoms of depression, thereby benefiting mental well-being. Furthermore, social media platforms serve as rich sources of emotional data, laying the foundation for emotion-related research.

Emotion classification for textual data is a well-studied problem with several existing datasets, such as GoEmotions (Demszky et al., 2020), TEC (Mohammad, 2012), CBET (Shahraki & Zaïane, 2017), and SemEval-2018 Task 1: Affect in Tweets (AIT) (Mohammad et al., 2018). However, some remaining challenges in the domain, as articulated by Deng and Ren (2023), include fuzzy emotional boundaries, incomplete extraction of emotional information from texts, and a lack of large-scale datasets.

In the last decade, many researchers have contributed to the dynamic and evolving domain of textual emotion classification, resulting in a proliferation of methodologies and techniques (Faruqui et al., 2015; Li et al., 2020; Zahiri & Choi, 2018; Zhou et al., 2020). With remarkable advancements in machine learning, deep learning, and word embedding, the precision and efficacy of emotion classification have significantly improved (Li et al., 2020; Zahiri & Choi, 2018). For instance, Zahiri and Choi (2018) laid the foundation for creating a multi-party dialogue corpus dedicated to text data to categorize emotions.

Within the existing body of literature, several studies have witnessed the enhancement of classifier performance through the integration of external knowledge derived from emotion lexicons and syntactic structures. Researchers have adeptly integrated lexicon-based methodologies with machine learning models, resulting in notable improvements in classification accuracy (Agrawal et al., 2018; Alm et al., 2005; Bandhakavi et al., 2017; Bhardwaj et al., 2023; Faruqui et al., 2015; Hu et al., 2013; Islam et al., 2019; Wasi & Abulaish, 2020). For instance, Li et al. (2020) introduced the EmoChannelAttn Network, leveraging dependency relationships within textual emotion constructs to enhance classification performance. Their study explored a spectrum of 151 finely-grained emotions, utilizing domain-specific knowledge and dimensional lexicon dictionaries. Meanwhile, Faruqui et al. (2015) employed lexical knowledge to enrich existing word embedding, advancing effective representation learning. Islam et al. (2019) introduced a multi-channel convolutional neural network for emotion and sentiment recognition on `Twitter` data, exploring the

Table 1
A summary of related works on emotion classification.

Reference	Summary	Limitations
Hu et al. (2013)	Developed a lexicon-based unsupervised sentiment analysis approach, incorporating emotional signals in Twitter data.	Relies heavily on Twitter data for evaluation, potentially limiting the generalizability of the findings to other social media platforms or text sources.
Faruqui et al. (2015)	Proposed retrofitting word embedding to incorporate semantic lexicons, enhancing word vector quality.	Primarily focused on improving standard lexical semantic tasks, less explored other possible applications or domains where the method could be applied effectively.
Vo and Zhang (2015)	Generated lexicon-based distributed contexts, guiding word vector selection using a lexicon dictionary.	Method dependent on the lexicon's comprehensiveness and accuracy, may not adapt well to evolving language use.
Agrawal et al. (2018)	Employed a method to map emotionally similar words closely in vector spaces and dissimilar ones far apart, enhancing emotion word embedding.	Relies on a categorical lexicon, restricting applicability to predefined emotion categories.
Islam et al. (2019)	Proposed a multi-channel CNN designed for emotion and sentiment recognition on Twitter. It examines the effect of integrating various lexical features into the neural network to enhance emotion and sentiment identification.	Does not address challenges or limitations of the multi-channel CNN implementation or the use of hashtags, emoticons, and emojis as emotion and sentiment indicators.
Wasi and Abulaish (2020)	Proposed a logistic regression-based sentiment classification method using domain knowledge from a lexicon and unlabeled domain data.	Focused on sentiment classification, not specifically tailored for nuanced emotion detection.
Li et al. (2020)	Introduced the EmoChannelAttn Network, using dependency relationships to enhance emotion classification performance.	Existing works often assume independence among emotion labels, neglecting potential relationships among different emotions, which could impact classification accuracy.
Haque and Abulaish (2023)	Introduced emotion-enhanced and psycholinguistic features for rumor detection on social media, analyzing word associations with emotions.	Application-specific to rumor detection, findings may not directly translate to broader emotion classification.
Bhardwaj et al. (2023)	Proposed an emotion classification method that integrates emotion lexicons with pre-trained word embeddings. The authors uses semantically-similar features to reconcile the semantic gap between words and emotions.	Emotionally dissimilar words, like *happy* and *sad* can yield high similarity scores in similar contexts, potentially conveying more similar meanings than emotional meanings.

impact of incorporating various lexical features into the neural network model for improved emotion and sentiment identification.

In the quest to acquire word representations infused with emotional nuances, Agrawal et al. (2018) employed a methodology that mapped emotionally analogous words into close proximity within vector spaces, while distancing emotionally dissimilar words. However, their approach relied on a categorical lexicon, limiting its applicability to a specific set of emotions. Additionally, Alm et al. (2005) incorporated emotional words as salient features for text-based emotion prediction. Moreover, Hu et al. (2013) pioneered a lexicon-based, unsupervised sentiment analysis approach, incorporating emotional signals within Twitter datasets. Furthermore, Vo and Zhang (2015) generated lexicon-based distributed contexts by selecting word vectors guided by a lexicon dictionary.

Haque and Abulaish (2023) proposed an emotion-enhanced and psycholinguistic features for detecting rumors on social media. By analyzing word associations with emotions, the authors identified rumors through lexicons and linguistic features from both posts and comments. Additionally, Wasi and Abulaish (2020) proposed a sentiment classification method based on logistic regression, leveraging pre-existing domain knowledge from a lexicon and unlabeled domain data. Moreover, Bhardwaj et al. (2023) introduced an emotion classification method that integrates emotion lexicons with pre-trained word embeddings.

Although these methods have improved accuracy for tasks such as word similarity, their impact on emotion classification has not been thoroughly explored, resulting in a gap in understanding their effectiveness in emotion-related applications. This study aims to enhance the emotion classification tasks by proposing feature extraction techniques that incorporate linguistic features and semantically-similar features-based vector representations, leveraging pre-trained word embedding and hand-crafted emotion lexicon dictionaries. These techniques include linguistic-based, semantically-similar, and semantically-similar linguistic features.

Table 2
List of notations and their descriptions.

Notation	Description
w_i	ith word in the vocabulary
ew	emotion word
sw	similar word
sf	similar feature word
$\overrightarrow{\mathbb{V}_{WE}}$	Word embedding vector
CS	Cosine similarity score
SS	Sentiment score
$\overrightarrow{\mathbb{V}_{S}}$	The top-m similar feature vector
$\overrightarrow{\mathbb{V}_{SS}}$	Semantically-similar features vector
$\overrightarrow{\mathbb{V}_{L}}$	Linguistic-based features vector
$\overrightarrow{\mathbb{V}_{SL}}$	Similar features-based linguistic vector
$\overrightarrow{\mathbb{V}_{W}}$	Word-level features vector
$\overrightarrow{\mathbb{V}_{D}}$	Document-level features vector

3. Proposed approach

In this section, we describe the functioning details of the proposed integrated machine learning model, EmoFusion, which aims to enhance emotion classification in textual data through a synergistic combination of word embeddings and emotion lexicon-based features. Fig. 1 illustrates the work-flow of the proposed emotion classification approach, encompassing various stages such as data preprocessing, feature extraction (including both word embeddings and emotion lexicon-based features), the training phase, and classification that are described in details in the following sub-sections. Table 2 provides the list of notations and their descriptions.

3.1. Problem definition

Given a corpus $D = \{d_1, d_2, \ldots, d_n\}$ of n documents and a set $K = \{c_1, c_2, \ldots, c_k\}$ of k labels, the problem of emotion classification can be considered as estimating a function $f : D \rightarrow K$ leveraging word embeddings and emotion lexicon-based features that assigns each

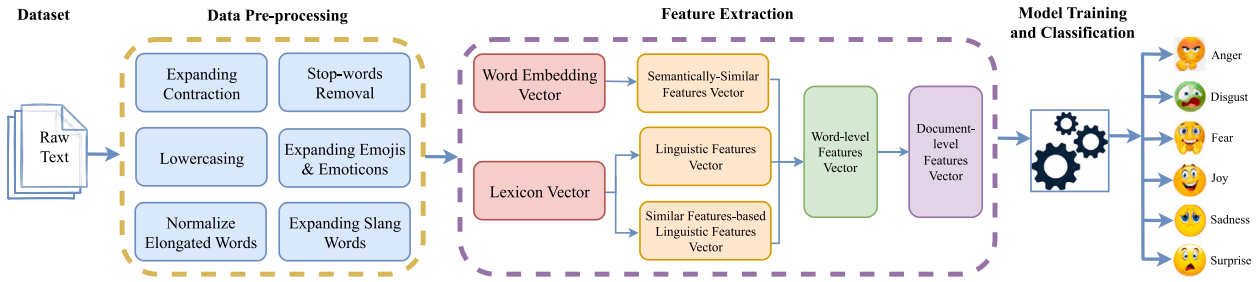


Fig. 1. Work-flow of the proposed approach EmoFusion.

document $d_n \in D$ to one of the emotion labels $c_k \in K$. Formally, the function f can be defined using Eq. (1).

$$f(d_n) = c_k \tag{1}$$

3.2. Word-level features extraction

NLP empowers machines to understand and gain insights from human language. Since machine learning algorithms cannot directly interpret textual semantics, words must be translated into numerical representations.

Each document d_n is decomposed into a vocabulary of words, denoted as $V = \{w_1, w_2, \ldots, w_r\}$, where r is the number of unique tokens. A word is defined as a $h-$dimensional vector, i.e., $w_i \in \mathcal{R}^h$, representing the ith word in the corpus. Given the complexity and significance of feature extraction in emotion classification, our proposed approach leverages pre-trained word embedding and emotion lexicons to enhance performance.

To generate an enriched emotion representation for each word, we utilize several components across different phases of our proposed approach:

- *Word Embedding Model:* We use the pre-trained GloVe (Pennington et al., 2014) vector model to obtain word embedding. This model generates a vector representation for each word w_i.
- *Emotion Lexicon Models:* We use four lexicon models—EmoLex, EmoSenticNet, NRC-VAD, and NRC-EIL, to derive emotion words and representations based on these models. Each instance is classified into six basic emotions as defined by Ekman (1999): *anger, sadness, joy, fear, disgust,* and *surprise*.
- *Cosine Similarity:* This metric measures the similarity between two word vectors.
- *Sentiment Lexicon Model:* We use AFINN-165 (Nielsen, 2011) sentiment lexicon, which provides sentiment scores ranging from −5 (negative) to +5 (positive) for words.
- *Mean Filter:* This process aggregates information by computing the sum of elements and dividing it by the number of elements.
- *Concatenation:* This involves appending features vector to create a new vector.

Sections 3.2.1, 3.2.2, and 3.2.3 demonstrate the methodology of developing three distinct feature vector representations for a word. These representations are concatenated in Section 3.2.4 to produce an emotion-enhanced representation for a word. Fig. 2 provides an illustrative overview of the relationship among different functional modules of our proposed approach, EmoFusion.

3.2.1. Semantically-similar features vector generation

We use GloVe model to extract relevant textual features. For each word w_i, an $l-$dimensional word embedding vector ($\overrightarrow{\mathbb{V}_{WE}}$) is obtained via the GloVe model (G), as shown in Eq. (2).

$$\overrightarrow{\mathbb{V}_{WE}} = G(w_i) \tag{2}$$

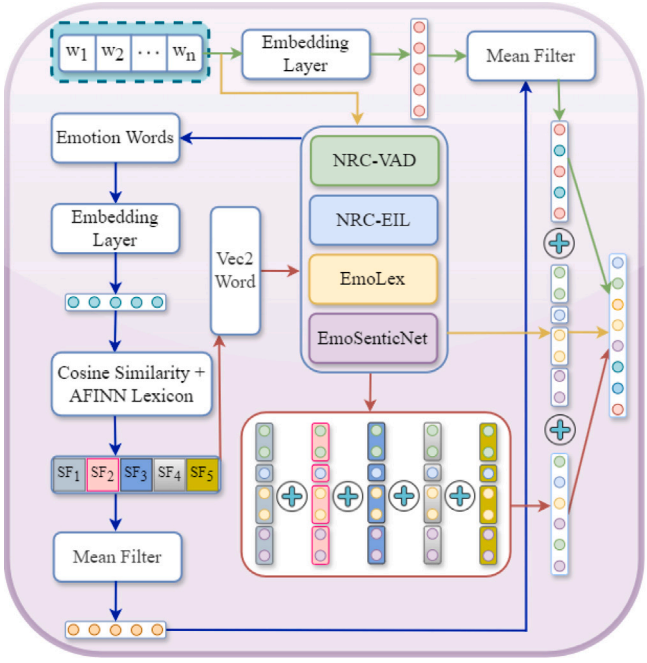


Fig. 2. Overview of the EmoFusion modules, illustrating their relationships: the green line denotes the vector for semantically-similar features, the yellow line represents the vector for linguistic features, and the red line signifies the similar features-based linguistic features.

Each word w_i is passed through four emotion lexicon models to identify emotion words. If the word is present in the lexicons, it is considered an emotion word (ew), and its corresponding word embedding vector $\overrightarrow{\mathbb{V}_{WE}}(ew)$ is retrieved. The cosine similarity between these emotion words is then calculated, as defined in Eq. (3).

$$CS = \frac{\overrightarrow{\mathbb{V}_{WE}}(ew_i) \cdot \overrightarrow{\mathbb{V}_{WE}}(ew_j)}{\|\overrightarrow{\mathbb{V}_{WE}}(ew_i)\| \ \|\overrightarrow{\mathbb{V}_{WE}}(ew_j)\|} \tag{3}$$

In Eq. (3), $\overrightarrow{\mathbb{V}_{WE}}(ew_i)$ and $\overrightarrow{\mathbb{V}_{WE}}(ew_j)$ denote the vectors for the ith and jth emotion words, respectively, and CS represents the cosine similarity score between $\overrightarrow{\mathbb{V}_{WE}}(ew_i)$ and $\overrightarrow{\mathbb{V}_{WE}}(ew_j)$. A higher CS indicates a greater similarity between these two emotion words. This similarity metric is utilized to identify a set of t similar words (sw) that exhibit semantic similarity to an emotion word ew_i based on their vector representations within word embedding space.

However, this method encounters a limitation: emotionally dissimilar words, like *happy* and *sad* might yield high similarity scores when appearing in similar contexts, potentially conveying more similar meanings than emotionally similar words, such as *happy* and *joy*. To address this limitation, we incorporate the manually labeled AFINN-165 sentiment lexicon, recognized as a valuable lexical resource for opinion mining. This lexicon contains over 3300 English words, each associated

with a polarity score, ranging from $\mathbb{A}_{ne} = [-5, 0)$ for negative words to $\mathbb{A}_{po} = (0, 5]$ for positive words.

After identifying sw for a given emotion word ew_i, their sentiment is evaluated using AFINN-165 lexicon. The process begins with calculating the sentiment score $SS(ew_i)$ for ew_i and $SS(sw_i)$ for sw_i within the context of each document in the dataset. Once $SS(ew_i)$ is determined, sw is filtered and sorted based on their SS and similarity score CS. If $SS(ew_i)$ lies within the positive sentiment score range $\mathbb{A}_{po}$ $(0 < SS(ew_i) \leq 5)$, only those sw_i with $SS(sw_i) \in \mathbb{A}_{po}$ are retained. After filtering, these words are sorted based on their CS. Conversely, if $SS(ew_i)$ falls within the negative sentiment score range $\mathbb{A}_{ne}$ $(-5 < SS(ew_i) \leq 0)$, only those sw_i with $SS(sw_i) \in \mathbb{A}_{ne}$ are kept. These words are then sorted based on their highest CS. This filtering ensures that the selected sw share the same sentiment polarity as ew_i, thereby improving the relevance of the results. Finally, the list of top-m similar feature words (sf), obtained from semantic similarity based on their highest scores, comprises emotionally similar words.

This strategic step ensures that the retrieved words convey a similar emotional connotation or sentiment as the emotion word ew_i, thereby enhancing the relevance and applicability of the results. This approach ensures a consistent and systematic method for identifying similar feature words by integrating semantic similarity and sentiment analysis techniques. By leveraging AFINN-165 for sentiment scoring and cosine similarity for semantic similarity computation, the methodology endeavors to provide a holistic understanding of the emotional context and semantic relationships inherent within the text data. This amalgamation of lexical resources with word embedding enhances the relevance and emotional context of the retrieved words.

These similar feature words are represented as $l-$dimensional vector, collectively forming the $\overrightarrow{\mathbb{V}_{WE}}(sf)$ vector. After identifying the top-m similar feature words, denoted by $\{sf_1, sf_2, \ldots, sf_m\}$, we apply the mean filter to derive a singular $l-$dimensional similar feature $(\overrightarrow{\mathbb{V}_S})$ vector, defined in Eq. (4).

$$\overrightarrow{\mathbb{V}_S} = \frac{1}{m}\sum_{j=1}^{m}\overrightarrow{\mathbb{V}_S}(sf_j) \tag{4}$$

Subsequently, a mean filter is applied to merge the word embedding vector $\overrightarrow{\mathbb{V}_{WE}}$ with the aggregated similar feature vector $\overrightarrow{\mathbb{V}_S}$, resulting in an $l-$dimensional semantically-similar features vector $(\overrightarrow{\mathbb{V}_{SS}})$, as described in Eq. (5).

$$\overrightarrow{\mathbb{V}_{SS}} = \frac{1}{2}(\overrightarrow{\mathbb{V}_{WE}} + \overrightarrow{\mathbb{V}_S}) \tag{5}$$

This resultant vector effectively encapsulates both the emotional and semantically-similar features of the words. Fig. 3 displays a word-cloud representation of the keywords associated with emotion words extracted from the emotion lexicons.

3.2.2. Linguistic features generation

After generating the semantically-similar features vector representation $\overrightarrow{\mathbb{V}_{SS}}$, each word is concurrently processed by four renowned emotion lexicon models to obtain its emotion-enriched representation. EmoLex produces an $x-$dimensional vector $(\vec{\mathbb{P}})$, EmoSenticNet produces a $y-$dimensional vector $(\vec{\mathbb{Q}})$, NRC-VAD produces a $z-$dimensional vector $(\vec{\mathbb{R}})$, and NRC-EIL produces a $u-$dimensional vector $(\vec{\mathbb{S}})$. These vectors are concatenated to generate a contextually rich $e-$dimensional linguistic-based features vector $(\overrightarrow{\mathbb{V}_L})$, as defined by Eq. (6), where $\oplus$ represents the concatenation of the vectors.

$$\overrightarrow{\mathbb{V}_L} = \vec{\mathbb{P}} \oplus \vec{\mathbb{Q}} \oplus \vec{\mathbb{R}} \oplus \vec{\mathbb{S}} \tag{6}$$

The vector representation of the word 'robbery' using the word embedding and the aforementioned emotion lexicons is shown in Table 3 for better comprehension.

Fig. 3. Word-cloud displaying keywords associated with emotion words.

3.2.3. Similar features-based linguistic features generation

To derive similar features-based linguistic features, we use the top-m similar features obtained from Eq. (4). These top-m similar features are then mapped back to the corresponding words using the vector to word model. Subsequently, the top-m similar feature words serve as input to the four lexicon models. Each lexicon model generates a vector representation for every similar feature word. Using Eq. (6), these lexicons produce an $e-$dimensional $\overrightarrow{\mathbb{V}_L}(sf)$ vector for each similar feature word. The $e-$dimensional vectors corresponding to each of the top-m similar feature words are concatenated to form the final similar features-based linguistic $(\overrightarrow{\mathbb{V}_{SL}})$ vector, as defined in Eq. (7), with a dimension of $m{\cdot}e$. These features adeptly captures the emotional context embedded within the similar feature words.

$$\overrightarrow{\mathbb{V}_{SL}} = \overrightarrow{\mathbb{V}_L}(sf_1) \oplus \overrightarrow{\mathbb{V}_L}(sf_2) \oplus \cdots \oplus \overrightarrow{\mathbb{V}_L}(sf_m) \tag{7}$$

3.2.4. Features vector generation

The final word-level feature vector provides a comprehensive representation of both the emotional and semantic content of the text. This vector is formed by concatenating three key components: the semantically-similar feature vector $(\overrightarrow{\mathbb{V}_{SS}})$, the linguistic-based features vector $(\overrightarrow{\mathbb{V}_L})$, and the similar features-based linguistic features vector $(\overrightarrow{\mathbb{V}_{SL}})$. The $\overrightarrow{\mathbb{V}_{SS}}$ combines word embeddings with emotional context to capture semantic similarity, while $\overrightarrow{\mathbb{V}_L}$ integrates emotion-specific knowledge from four emotion lexicons, enriching the word representations. Additionally, $\overrightarrow{\mathbb{V}_{SL}}$ captures the shared emotional context among words based on their semantic and emotional similarities. This fusion of components results in a robust $h-$dimensional word-level features vector, as defined in Eq. (8).

$$\overrightarrow{\mathbb{V}_W} = \overrightarrow{\mathbb{V}_{SS}} \oplus \overrightarrow{\mathbb{V}_L} \oplus \overrightarrow{\mathbb{V}_{SL}} \tag{8}$$

3.3. Document-level features vector generation

To construct the document-level features vector, we compute the centroid of the word-level features vector of all words within the document. Let $\overrightarrow{\mathbb{V}_{W_i}}$ denote the $h-$dimensional word-level features vector for the ith word. For a document containing r words, the document-level features vector $\overrightarrow{\mathbb{V}_D}$ is defined in Eq. (9).

$$\overrightarrow{\mathbb{V}_D} = \frac{1}{r}\sum_{i=1}^{r}\overrightarrow{\mathbb{V}_{W_i}} \tag{9}$$

Table 3
Generated features-vector representation for the word 'robbery' using word embedding and emotion lexicons.

Components	Vector						Dimension
Embedding vector	[0	1	⋯		0	1]	100
EmoLex	Anger	Disgust	Fear	Joy	Sadness	Surprise	6
	[1	1	1	0	1	0]	
EmoSenticNet	Anger	Disgust	Fear	Joy	Sadness	Surprise	6
	[0	0	0	0	1	0]	
NRC-VAD	Valence		Arousal		Dominance		3
	[0.133		0.843		0.435]		
NRC-EIL	Sadness						1
	[0.6]						

In Eq. (9), $\overrightarrow{\mathbb{V}_D}$ is the document-level features vector, with a dimension of $r \cdot h$. By combining these word-level features, the document-level feature vector encapsulates the collective emotional and semantic content of the entire text. This aggregated vector serves as the input to machine learning models, enabling the efficient classification of emotions expressed within the document. The proposed approach thus provides a holistic view of emotions at the document-level, enabling a comprehensive understanding of the text's emotional nuances.

4. Experimental setup

In this section, we provide a detailed description of the experimental setup employed in `EmoFusion`.

4.1. Datasets

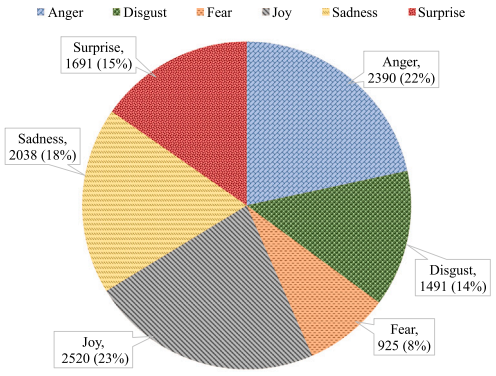
For emotion classification, we conducted experiments on three benchmark datasets: Google AI GoEmotions, CBET, and TEC. Among these, GoEmotions and CBET are multi-label datasets, while TEC is a multi-class dataset. Given the context of our proposed approach in emotion classification, instances with multiple emotion class labels from GoEmotions and CBET were excluded. To the best of our knowledge, Ekman's emotion model is the most broadly used. Therefore, we focused on the Ekman's six basic emotions: *anger, disgust, joy, fear, sadness*, and *surprise*. Fig. 4 shows the distribution of emotions in the datasets used in our experiments.

GoEmotion[2]: The Google AI GoEmotions (Demszky et al., 2020) dataset comprises 58,000 English comments from `Reddit`, which have been manually annotated and labeled with one of 27 emotions or marked as *neutral*.

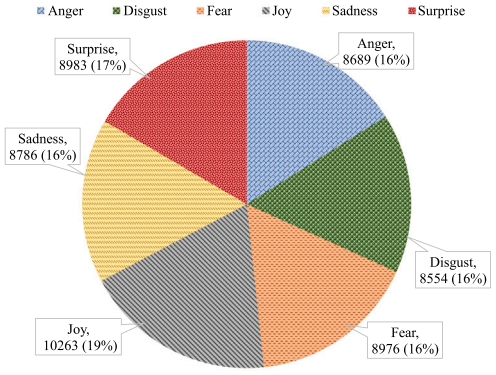
CBET[3]: The Cleaned Balanced Emotional Tweet (CBET) dataset was provided by Shahraki and Zaïane (2017). To the best of our knowledge, this is among the largest publicly available balanced datasets for `Twitter` emotion detection research. It contains 81,163 instances categorized into nine emotional tag categories, including the six basic emotions identified by Ekman.

TEC[4]: The Hashtag Emotion Corpus, referred to as the Twitter Emotion Corpus (TEC), was released by Mohammad (2012). TEC consists of 21,051 tweets collected from `Twitter` using hashtags corresponding to the six basic emotions identified by Ekman.

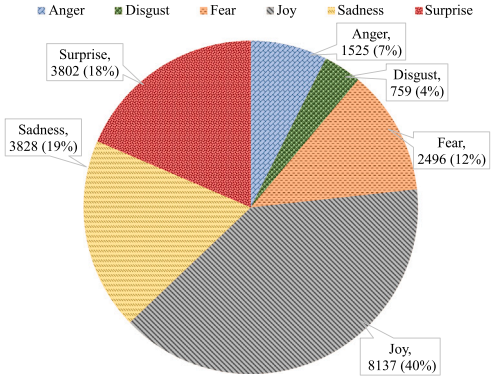
In our approach, each dataset was split into a training set and a test set with a 7:3 ratio. Given that CBET is a balanced dataset, it was randomly sampled into training and test sets. In contrast, GoEmotions and TEC, being unbalanced datasets, were divided using stratified sampling.



(a) GoEmotions dataset



(b) CBET dataset



(c) TEC dataset

Fig. 4. Pie chart representation of emotion quantity description of the datasets.

4.2. Data pre-processing

In the era of social media, individuals effortlessly express their emotions through various forms of content, such as posts, reviews, and comments. However, the unstructured nature of the data presents a challenge for machines in analyzing the underlying emotions. Effective data analysis requires essential pre-processing due to the significant impact of data quality on analytical approaches.

Before using the corpus in our experiments, we conduct emotion-specific pre-processing to eliminate noise from the data. Our initial step involves expanding contractions to improve text clarity; for instance, "can't" becomes "cannot". To achieve this, we employ the `pycontractions` (version 2.0.1) library[5] [e.g., "ain't": "am not/are not/is

[2] https://github.com/google-research/google-research/tree/master/goemotions
[3] https://github.com/chenyangh/CBET-dataset
[4] https://saifmohammad.com/WebDocs/Jan9-2012-tweets-clean.txt.zip
[5] https://github.com/ian-beaver/pycontractions

not/has not/have not"], known for its higher accuracy compared to the standard `Python` contractions library, as it considers the text's grammar. Following this, we convert all characters to lowercase to ensure uniformity and prevent duplication of words with different cases. This step helps in treating words like "Happy" and "happy" as identical, thus facilitating emotion classification. Subsequently, we normalize elongated words by reducing repeated characters to their base word, for example, "haaaapppyy" becomes "happy". This normalization helps in capturing the intended emotion without being biased by word length.

Additionally, we remove English stop-words from the text using the `NLTK` (version 3.8.1) library (Wagner, 2010), followed by lemmatization to reduce words to their base form for contextual meaning. Emojis and emoticons, which have become increasingly prevalent with the rise of social media, serve as valuable context for emotion classification. Emoticons (e.g., *:-(* as sad) are combinations of numbers, punctuation marks, and letters used to create pictorial representations that often directly convey emotions or sentiments, providing valuable context for emotion classification. In contrast, emojis are pictographs of objects, symbols (e.g., ☺, representing a "smiling face"), and faces. The main purpose of both emoticons and emojis is to convey specific emotions and sentiments (Dresner & Herring, 2010). To include these in our analysis, we encoded emoticons and emojis into textual representations using the `emot` (version 3.1) library.[6]

Social media platforms like `Twitter` have a character limit of up to 280, encouraging users to write posts using abbreviations, slang, and acronyms. Since users frequently use abbreviated words on social media, expanding these words for analysis purposes proves beneficial. Therefore, slang words or abbreviations are expanded to their full forms to ensure clarity and consistency in the text; for instance, "lol" becomes "laugh out loud". Resolving abbreviations and contractions aids in accurately interpreting the emotional content of the text. Furthermore, `Twitter` handles are replaced with generic placeholders, denoted as [NAME], to anonymize user information. Instead of entirely removing hashtags, only the symbol "#" is removed, as hashtags often convey intense emotion or the gist of the text. The text is then tokenized, breaking it down into individual words or tokens. Finally, punctuation marks are removed from the text, except for "?" and "!" as they typically indicate intense emotions. Texts containing fewer than three valid English words are discarded. Moreover, non-alphabetic words (including numbers and symbols), HTML tags, URLs, UTF-8, whitespace, special characters like "rt", "\n", and duplicate instances are all eliminated from the text.

4.3. Emotion lexicons

Emotion lexicons play a significant role in detecting emotions using a keyword or lexical approach. The basic idea is to identify emotional clues embedded within the text. This section reviews several commonly used emotion lexicons, emphasizing their contributions to fine-grained emotional understanding in different contexts. The four well-known emotion lexicons discussed below are:

NRC Word-Emotion Association Lexicon[7]: The NRC Word-Emotion Association Lexicon, referred to as EmoLex (Mohammad & Turney, 2013), comprises 14,182 unigrams (words) and about 25,000 word senses for sentiments and emotions, respectively. EmoLex includes a list of words that have been manually annotated with associations to eight emotions (*anticipation, anger, disgust, joy, fear, sadness, trust*, and *surprise*), as well as two sentiments (*positive* and *negative*). Each word in this lexicon is associated with emotion scores indicating the presence or absence of an emotional connection, denoted by binary values (*0* or *1*).

EmoSenticNet[8]: EmoSenticNet (Poria et al., 2014) extends the existing WordNet Affect and SenticNet lexicons by assigning specific emotion labels from WordNet Affect (*joy, sadness, disgust, anger, surprise, fear*) to SenticNet concepts, thus broadening its scope. This lexicon includes 13,189 words, each linked to one or more emotion categories. Like EmoLex, EmoSenticNet provides binary scores (*0* or *1*) that denote whether a word is connected to a particular emotion.

NRC Valence, Arousal, and Dominance (NRC-VAD)[9]: The NRC-VAD lexicon (Mohammad, 2018a) features a list of 20,000 words along with their scores for valence, arousal, and dominance. This lexicon offers real-valued scores ranging from *0* (lowest) to *1* (highest) for each dimension, providing a nuanced representation of each word's emotional connotation.

NRC-Emotion Intensity Lexicon (NRC-EIL)[10]: The NRC-Emotion Intensity Lexicon, also known as the Affect Intensity Lexicon (Mohammad, 2018b), consists of 9921 words annotated with real-valued intensity scores for eight emotions: *anticipation, anger, disgust, joy, fear, sadness, trust*, and *surprise*. This lexicon provides a fine-grained intensity score for each emotion-associated word, allowing for detailed emotional analysis.

These lexicons are utilized to derive emotion words and representations. Each instance is classified into six basic emotions identified by Ekman. Collectively, these lexicons enrich the representation of emotional expressions in text, thereby supporting more precise emotion classification and analysis. By employing these emotion-specific lexicons, we can enhance the representation of words with nuanced emotional content.

4.4. Evaluation metrics

In this section, we present the evaluation metrics employed to assess the efficacy of the proposed approach and it comparison with the baselines and a state-of-the-art technique. Following the state-of-the-art such as Grandini et al. (2020), we have used *accuracy* (Ac), *macro precision* (Pr_{mac}), *macro recall* (Re_{mac}), and *macro f1-score* ($F1_{mac}$). The Ac measures the proportion of correctly classified instances out of the total number of instances in the dataset, as defined in Eq. (10).

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

In Eq. (10), TP stands for True Positive and it represents the number of instances that are correctly predicted to have a specific emotion; TN stands for True Negative and it denotes the number of instances in which the model accurately predicts the absence of a particular emotion; FP stands for False Positive and it denotes the number of instances in which the model incorrectly predicts the presence of a specific emotion; and FN stands for False Negative and it represents the number of instances incorrectly predicted as not conveying a specific emotion.

Pr_{mac} and Re_{mac} are computed as the arithmetic mean of the precision and recall metrics for each class, respectively. They are mathematically defined using Eqs. (11) and (12), respectively.

$$Pr_{mac} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k} \tag{11}$$

$$Re_{mac} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FN_k} \tag{12}$$

Finally, the $F1_{mac}$ calculates the harmonic mean of macro precision and macro recall, resulting in a value between 0 and 1. The $F1_{mac}$ is determined by Eq. (13). This metric assigns equal weight to each class and assesses the classifier's average performance across all classes. It is especially beneficial when handling imbalanced datasets, ensuring that

[6] https://github.com/NeelShah18/emot
[7] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[8] https://www.gelbukh.com/emosenticnet/
[9] http://saifmohammad.com/WebPages/nrc-vad.html
[10] http://www.saifmohammad.com/WebPages/AffectIntensity.htm

each class contributes equally to the final evaluation. A higher $F1_{mac}$ indicates better classification results.

$$F1_{mac} = 2 \times \frac{Pr_{mac} \times Re_{mac}}{Pr_{mac} + Re_{mac}} \tag{13}$$

4.5. Implementation details

We utilized Stanford's pre-trained word embedding GloVe[11] to obtain word representations. Specifically, we employed GloVe-6B (Pennington et al., 2014), which includes 6 billion uncased words from Wikipedia 2014 and Gigaword v5, providing 100−dimensional vectors for each word.

To enrich the word embedding with emotional context, we integrated four renowned emotion lexicons: EmoLex, EmoSenticNet, NRC-VAD, and NRC-EIL. We considered the six basic emotions ($k = 6$). These lexicons produce a 16−dimensional vector for each input word. The emotion-enriched vectors are generated by taking a word w_i and using the lexicons to create a comprehensive representation.

In our approach, we first identified the top-25 semantically-similar words for each input word using cosine similarity of their GloVe vector representations. We then applied sentiment filtering with the AFINN-165 lexicon to refine this list, retaining only the top-5 similar feature words with matching sentiment polarity. The dimensions of the resulting vectors are as follows: $\overrightarrow{\mathbb{V}_{SS}}$ is 100−dimensional, $\overrightarrow{\mathbb{V}_{L}}$ is 16−dimensional, and $\overrightarrow{\mathbb{V}_{SL}}$ is 80−dimensional. The final features vector $\overrightarrow{\mathbb{V}_{W}}$ is a 196−dimensional vector that combines the word embedding and emotion-enriched features.

For implementation, we used a Linux-based HPE ProLiant ML350 server with a Dual Intel Xeon-Bronze 3106 processor (1.7 GHz) and 192 GB of memory for our computation.

Parameter Settings: For model implementation, we optimized the parameters to enhance performance. Our model architecture included a BiLSTM or BiGRU layer with 196 neurons matching the input shape of (max_len, 196). This was followed by a dropout layer with a rate of 0.2 to prevent overfitting. Another BiLSTM or BiGRU layer with 98 neurons was then added, followed by another dropout layer with the same rate. Subsequently, a dense layer with 46 neurons and ReLU activation function was included, followed by another dropout layer with a rate of 0.2. For multi-class classification, we employed the softmax activation function, providing a categorical probability distribution, and used the categorical cross-entropy loss function for training.

To enhance training efficiency and performance, early stopping was implemented, monitoring the validation loss with the 'min' mode. The model was compiled using the Adam optimizer. Training was conducted over 15 epochs with a batch size of 32, and a validation split of 0.1. Early stopping was configured to halt training when no further improvement in validation loss was observed. This comprehensive setup ensured effective and efficient training of our emotion classification model.

5. Results and comparative analysis

This section presents the results of EmoFusion, along with a comparative analysis. The classifiers considered for evaluation are: (i) Support Vector Machine (Boser et al., 1992; Wang et al., 2006); specifically, employed the SVM-OvA (one-vs-all or one-vs-the-rest) strategy, (ii) Gradient Boosting (GB) (Freund et al., 1999), (iii) Naïve Bayes (NB) (Rennie, 2001), (iv) Bidirectional Long Short-Term Memory Network (BiLSTM) (Schuster & Paliwal, 1997), and (v) Bidirectional Gated Recurrent Unit (BiGRU). These classifiers were trained to predict the six basic Ekman emotions: *disgust, anger, fear, sadness, joy*, and *surprise* at the document-level.

We compared our proposed approach against two sets of baselines and a state-of-the-art technique. The first set includes Demszky et al. (2020), Shahraki and Zaïane (2017), which are multi-label classification methods, and Mohammad (2012), which is a multi-class classification method. The second set of baselines, includes Lex@16, $LexSim$@96, and the *Emo-TF–IDF* techniques, which employs semantically-similar features to address the semantic gap between words and emotions in multi-class classification. Selecting appropriate baselines was challenging due to the inadequate number of directly comparable approaches, so previous studies were carefully reviewed to choose suitable baselines.

The following paragraphs briefly describe six baselines and a state-of-the-art technique.

- *Demszky et al. (2020)*: This approach used the BERT-based transfer learning model to evaluate the generalizability of the GoEmotions dataset across different domains and emotion taxonomies.
- *Shahraki and Zaïane (2017)*: This approach proposed lexical and learning-based methods to classify the emotions on Twitter corpus.
- *Mohammad (2012)*: This approach extracted a word-emotion association lexicon from the Twitter corpus, and showed that it significantly performed better in an emotion classification task.
- *Emotion lexicon-based features* (Lex@16): Each document is represented by a vector composed of emotion-related features extracted from EmoLex, EmoSenticNet, NRC-VAD, and NRC-EIL. The culmination of these features results in a 16−dimensional features vector.
- *Combined representation of emotion lexicons with similar features* ($LexSim$@96): Each document uses features associated with emotion lexicons that are merged with similar features to form a vector. This resultant vector comprises a 96−dimensional features vector.
- *Combined representation of TF–IDF with emotion features* (*Emo-TF–IDF*): Each document in the datasets is represented using a unigram TF–IDF vector. This vector is incorporated with emotion-related features to create an emotionalized TF–IDF representation.
- *Bhardwaj et al. (2023)*: This approach integrated emotion lexicons with pre-trained word embeddings. It utilized semantically-similar features to reconcile the semantic gap between words and emotions.

Table 4 presents emotion-wise classification results for the EmoFusion across the three benchmark datasets, evaluated using Pr_{mac}, Re_{mac}, and $F1_{mac}$. For the GoEmotions dataset, EmoFusion demonstrates significant improvements compared to the method presented in Demszky et al. (2020). For *anger*, our proposed approach achieves a notable improvement of 4.8% in the $F1_{mac}$. However, for *disgust*, *fear*, and *joy*, the $F1_{mac}$ decreases by 9.0%, 5.4%, and 6.9%, respectively. Notably, for *sadness*, the proposed approach shows an improvement of 6.4%, while for *surprise*, the $F1_{mac}$ increases by 2.9%. These inconsistencies arise from dataset characteristics: manual inspection reveals noisy samples and ambiguous labeling, particularly for overlapping emotions like *disgust* (often confused with *anger*) and *fear* (frequently overlapping with *sadness*).

For the CBET dataset, EmoFusion demonstrates substantial improvements across multiple emotion categories compared to Shahraki and Zaïane (2017). Specifically, the $F1_{mac}$ scores shows notable gains of 13.8%, 12.5%, 13.7%, 19.6%, 17.4%, and 13.9% are achieved for *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*, respectively. For the TEC dataset, EmoFusion shows significant improvements compared to Mohammad (2012). Specifically, improvements of 4.4%, 5.6%, 6.1%, 7.1%, and 3.5% are observed for *anger*, *disgust*, *joy*, *sadness*, and *surprise*, respectively. However, a slight decrease of 1.4% in the $F1_{mac}$ is observed for *fear*. This aligns with the dataset's sparse representation of *fear*, where limited training examples and subtle linguistic expressions (e.g., implicit *fear* cues like uneasy vs. explicit terms like terrified) reduce

[11] https://nlp.stanford.edu/projects/glove/

Table 4
Emotion category-wise comparison results (in %) for emotion classification.

Emotions		GoEmotions		CBET		TEC	
		Demszky et al. (2020)	EmoFusion	Shahraki and Zaïane (2017)	EmoFusion	Mohammad (2012)	EmoFusion
anger	Pr_{mac}	50.0	**56.5**	39.9	**57.3**	37.3	**50.5**
	Re_{mac}	65.0	**68.1**	36.5	**47.4**	22.3	**23.8**
	$F1_{mac}$	57.0	**61.8**	38.1	**51.9**	27.9	**32.3**
disgust	Pr_{mac}	52.0	43.5	41.2	**59.6**	30.7	**33.6**
	Re_{mac}	53.0	44.6	51.1	**56.7**	13.4	**19.0**
	$F1_{mac}$	53.0	44.0	45.6	**58.1**	18.7	**24.3**
fear	Pr_{mac}	61.0	**62.3**	56.6	**73.1**	59.6	**66.7**
	Re_{mac}	76.0	63.0	57.3	**68.4**	43.9	39.0
	$F1_{mac}$	68.0	62.6	57.0	**70.7**	50.6	49.2
joy	Pr_{mac}	77.0	76.7	48.3	**63.8**	64.5	58.1
	Re_{mac}	88.0	73.7	46.7	**70.7**	60.4	**83.5**
	$F1_{mac}$	82.0	75.2	47.5	**67.1**	62.4	**68.5**
sadness	Pr_{mac}	56.0	**68.4**	36.5	**46.2**	41.9	**46.4**
	Re_{mac}	62.0	**62.7**	28.9	**53.4**	36.0	**45.3**
	$F1_{mac}$	59.0	**65.4**	32.2	**49.6**	38.7	**45.8**
surprise	Pr_{mac}	53.0	**69.6**	48.1	**60.3**	50.6	**60.2**
	Re_{mac}	70.0	59.1	45.2	**60.6**	40.5	**40.7**
	$F1_{mac}$	61.0	**63.9**	46.6	**60.5**	45.0	**48.5**

Table 5
Comparison of *accuracy* Ac (in %) of EmoFusion with baselines and a state-of-the-art technique across three datasets.

Approaches	GoEmotions					CBET					TEC				
	SVM	GB	NB	BiLSTM	BiGRU	SVM	GB	NB	BiLSTM	BiGRU	SVM	GB	NB	BiLSTM	BiGRU
Lex@16	44.1	47.6	35.9	47.3	48.8	36.6	37.4	30.2	42.5	44.8	39.3	44.1	39.6	47.1	44.0
LexSim@96	45.6	49.8	35.8	48.4	49.2	41.3	45.7	32.8	49.2	47.1	45.8	48.0	34.6	49.6	49.2
Emo-TF–IDF	51.1	52.9	29.4	49.5	52.0	49.5	42.2	40.2	51.0	50.2	46.7	46.7	30.8	51.0	51.5
Bhardwaj et al. (2023)	56.7	58.5	42.4	60.2	61.3	50.5	53.5	39.2	57.8	57.7	50.7	50.9	40.9	53.1	52.8
Proposed	**58.2**	**60.3**	**43.0**	**63.6**	**63.0**	**52.0**	**54.7**	**40.7**	**59.8**	**58.5**	**51.0**	**51.8**	**42.6**	**56.0**	**54.3**

Table 6
Comparison of *macro f1-score* ($F1_{mac}$) (in %) of EmoFusion with baselines and a state-of-the-art technique across three datasets.

Approaches	GoEmotions					CBET					TEC				
	SVM	GB	NB	BiLSTM	BiGRU	SVM	GB	NB	BiLSTM	BiGRU	SVM	GB	NB	BiLSTM	BiGRU
Lex@16	34.9	44.5	30.0	42.5	46.7	34.8	36.5	26.6	41.2	44.3	10.2	30.4	29.2	28.5	16.6
LexSim@96	41.1	47.7	31.7	45.6	48.6	40.5	45.1	30.2	48.7	46.1	25.3	34.7	29.2	37.0	37.1
Emo-TF–IDF	48.9	49.8	29.1	49.4	52.9	49.4	41.0	38.4	49.5	54.2	38.4	25.4	27.3	39.3	38.5
Bhardwaj et al. (2023)	54.5	56.1	40.0	58.3	60.5	50.0	53.0	37.7	57.0	56.7	39.1	38.8	**32.6**	**46.2**	37.0
Proposed	**56.1**	**58.3**	**40.7**	**62.2**	**61.3**	**51.6**	**54.2**	**39.3**	**59.6**	**58.6**	**39.3**	**40.7**	32.4	44.8	**42.2**

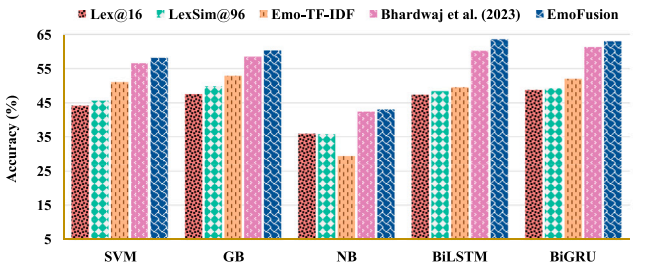
model efficacy. Performance variations across emotions underscore the impact of dataset quality and labeling consistency. For instance, *disgust* and *fear*-emotions often expressed indirectly or contextually-are particularly susceptible to misclassification.

Tables 5 and 6 present the Ac and $F1_{mac}$ scores for various emotion classification approaches using different classifiers. We compare our proposed approach with baseline techniques, namely *Lex*@16, *LexSim*@96, *Emo-TF–IDF*, and a state-of-the-art technique (Bhardwaj et al., 2023). The results indicate that our proposed approach consistently outperforms the baselines and the state-of-the-art technique. Figs. 5 and 6 present bar charts comparing Ac and $F1_{mac}$ across classifiers and datasets.

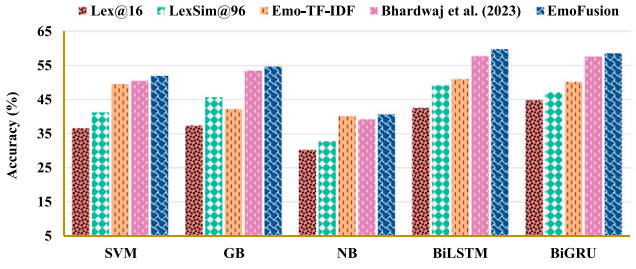
For the GoEmotions dataset, EmoFusion achieves notable improvements with the BiLSTM classifier, showing a 14.1% increase in Ac and a 12.8% increase in the $F1_{mac}$ compared to *Emo-TF–IDF*. In comparison with Bhardwaj et al. (2023), our proposed approach achieves a 3.4% higher Ac and a 3.9% higher $F1_{mac}$. Similarly, with the BiGRU classifier, EmoFusion shows an 11.0% improvement in Ac and an 8.7% improvement in $F1_{mac}$ compared to *Emo-TF–IDF*, while outperforming (Bhardwaj et al., 2023) by 1.7% in Ac and 0.8% in $F1_{mac}$. The improvements are also evident with traditional classifiers. For SVM, GB, and NB, EmoFusion outperforms *Emo-TF–IDF* by 7.1%, 7.4%, 13.7% in Ac, and 7.2%, 8.5%, and, 11.6% in $F1_{mac}$, respectively. Meanwhile, our proposed approach surpasses (Bhardwaj et al., 2023) by 1.5%, 1.8% and 0.6% in Ac and 1.6%, 2.2%, and 0.7% in $F1_{mac}$.

For the CBET dataset, our proposed approach demonstrates superior performance. Using the BiLSTM classifier, EmoFusion achieves an 8.8% increase in Ac and a 10.1% increase in $F1_{mac}$ compared to *Emo-TF–IDF* In comparison with Bhardwaj et al. (2023), our proposed approach outperforms by 2.0% in Ac and 2.6% in $F1_{mac}$. With the BiGRU classifier, EmoFusion surpasses *Emo-TF–IDF* by 8.3% in Ac and 4.4% in $F1_{mac}$, and exceeds (Bhardwaj et al., 2023) by 0.8% and 1.9% in terms of Ac and $F1_{mac}$. The improvements with the SVM and GB classifiers are also significant, with EmoFusion outperforming *Emo-TF–IDF* by 2.5% and 12.5% in Ac, and 2.2% and 13.3% in $F1_{mac}$, respectively. In comparison with Bhardwaj et al. (2023), our proposed approach shows an improvement of 1.5%, 1.2%, and 1.5% in Ac and 1.6%, 1.2%, and 1.6% in $F1_{mac}$.

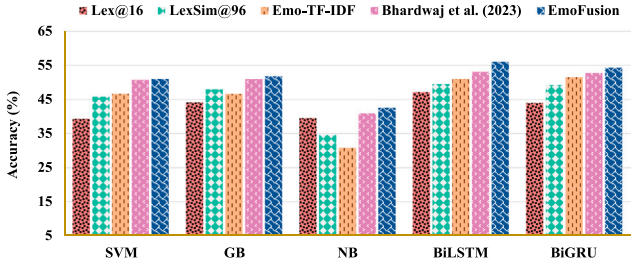
In the TEC dataset, EmoFusion continues to outperform the baseline techniques. Using the BiLSTM classifier, our proposed approach achieves an improvement of 5.0% in Ac and 5.5% in $F1_{mac}$ compared to *Emo-TF–IDF*. When compared to Bhardwaj et al. (2023), our proposed approach shows a 2.9% increase in Ac but a slight decrease of 1.4% in $F1_{mac}$. With the BiGRU classifier, EmoFusion achieves a 2.8% improvement in Ac and a 3.7% improvement in $F1_{mac}$ compared to *Emo-TF–IDF*. It outperforms (Bhardwaj et al., 2023) by 1.5% and 5.2% in terms of Ac and $F1_{mac}$. Additionally, EmoFusion shows improvements with SVM, GB, and NB classifiers by 4.3%, 5.1%, and 10.6% in Ac, and 1.0%, 15.2%, and 6.0% in $F1_{mac}$, respectively, compared to



(a) *Ac* over GoEmotions dataset

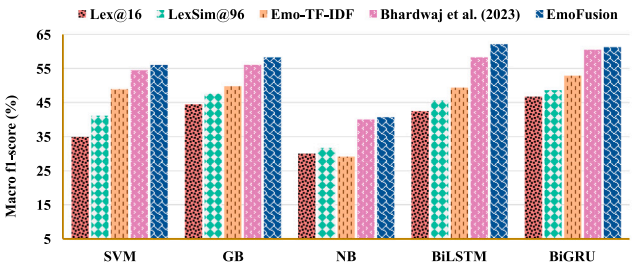


(b) *Ac* over CBET dataset

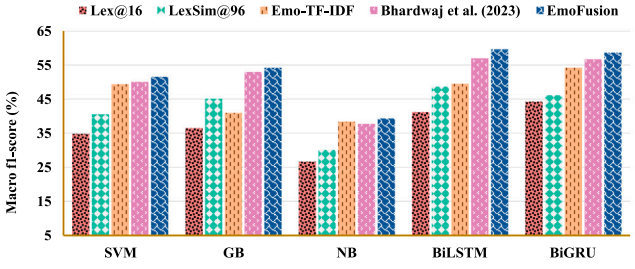


(c) *Ac* over TEC dataset

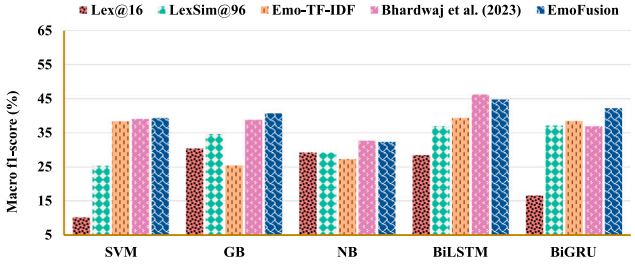
Fig. 5. Visualization of the comparative evaluation results of EmoFusion with baselines and a state-of-the-art technique in terms of *accuracy* (*Ac*).



(a) $F1_{mac}$ over GoEmotions dataset



(b) $F1_{mac}$ over CBET dataset



(c) $F1_{mac}$ over TEC dataset

Fig. 6. Visualization of the comparative evaluation results of EmoFusion with baselines and a state-of-the-art technique in terms of *macro f1-score* ($F1_{mac}$).

Emo-TF–IDF. The proposed approach shows increases of 0.3%, 0.9%, and 1.7% in *Ac* and 0.2%, 1.9%, with a slight decrease of 0.2% in $F1_{mac}$ in comparison with Bhardwaj et al. (2023).

EmoFusion outperforms baseline techniques in all evaluation metrics across various classifiers and datasets. The BiLSTM and BiGRU models, in particular, show the most significant improvement in the proposed approach, achieving the highest accuracy on all three datasets (GoEmotions: 63.6%, CBET: 59.8%, TEC: 56.0%). Our proposed approach performs well on both traditional and advanced classification models, as it captures both emotional and semantic information, leading to better word representation. This highlights the effectiveness of our feature extraction techniques for emotion classification tasks. Moreover, our proposed approach is computationally inexpensive because each document is represented using a 196−dimensional feature vector, making it efficient in dealing with nuanced emotional expressions in text.

5.1. Case study

We conduct a case study to illustrate the emotion classification process of EmoFusion, as shown in Fig. 7. We randomly choose two sample texts from the GoEmotions dataset labeled as *anger* and *joy*, demonstrating how the model integrates linguistic and semantically-similar features for emotion classification. In Fig. 7, words highlighted in blue correspond to emotion words (ew) extracted from the document, while words in green represent similar feature words (sf) found in the cluster associated with ew.

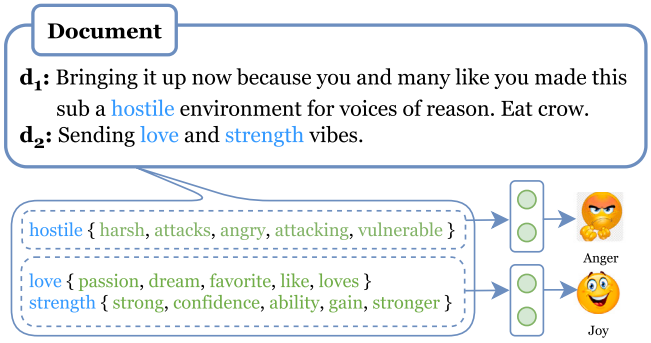


Fig. 7. Case study examples illustrating EmoFusion's emotion classification task. Words in blue are emotion words and words in green are similar feature words.

In document d_1, the word "hostile" is present in the lexicons and explicitly mapped to *anger*, generating strong lexicon-based signals. GloVe embeddings associate similar feature words (sf) such as {harsh, attacks, angry, attacking, and vulnerable} (from the lexicon's hostile cluster) with anger-related contexts. The idiom "Eat crow" (implying humiliation) is absent in the lexicons, forcing reliance on embeddings. However, the fused features prioritize the stronger lexicon signals (hostile), correctly classifying the emotion as *anger*. In document d_2, the words "love" and "strength" (associated with *joy*) are weighted heavily by the fusion mechanism. The word "vibes" lacks direct lexicon entries but is semantically linked to positivity via GloVe embeddings, reinforcing the *joy* classification. More clearly, the approach captures

the emotion and similar feature words which are crucial to identify the emotion.

5.2. Statistical analysis

We performed a paired t-test on the $F1_{mac}$ across all datasets to show the significance of EmoFusion's improvements over the baselines. The t-test is appropriate since the input dataset, classifiers, and experimental conditions remain identical for both approaches, ensuring paired observations. For each baseline, we compared its $F1_{mac}$ against EmoFusion across all classifiers and datasets, computing p-values at a 10% significance level (0.1). Our analysis revealed statistically significant improvements for EmoFusion, with $p < 0.1$ across datasets. While minor performance variations were observed in specific cases (e.g., baselines such as Demszky et al., 2020 and Bhardwaj & Abulaish, 2023 in the TEC dataset), these did not meet the significance threshold. This demonstrates that the proposed approach achieves statistically significant improvements in the majority of comparisons of EmoFusion over the baselines and state-of-the-art.

6. Discussion

This study demonstrates the effectiveness of integrating pre-trained word embeddings, emotion lexicons, and emotion-specific pre-processing for document-level emotion classification. We evaluate the performance of various algorithms on three benchmark datasets. The method constructs a compact 196−dimensional feature vector for each document by combining 100−dimensional GloVe embeddings with features derived from four emotion lexicons. As shown in Table 4 indicate that our approach achieves significant improvements over the baselines Demszky et al. (2020), Shahraki and Zaïane (2017), and Mohammad (2012). By integrating emotional nuances through lexicons, our approach ensures these emotion lexicons play a critical role to understand emotional nuances. Further, we highlight the importance of embeddings model for textual emotion classification. Figs. 5 and 6 reveal that $Lex@16$, which relies solely on linguistic features, performs the worst across all datasets compared to the other baselines and our approach. $LexSim@96$, combining linguistic and semantically-similar features, shows moderate improvement over $Lex@16$. While *Emo-TF–IDF*, using TF–IDF for document representation, outperformed all the baselines, it is computationally expensive due to the high dimensionality. In contrast, EmoFusion significantly surpasses all baselines and a state-of-the-art technique while maintaining computational efficiency. These results collectively demonstrate that our approach show a remarkable improvement over methods that rely on individual features.

While transformer-based embeddings (e.g., BERT, GPT) represent the state-of-the-art for many NLP tasks, their computational overhead and resource-intensive nature limit their applicability in real-time or resource-constrained environments. In contrast, our approach adopts a lightweight architecture based on pre-trained GloVe embeddings, which are computationally efficient and offer stable word representations. This static nature enables direct integration with external emotion lexicons-an advantage over contextual embeddings like BERT, whose token representations vary by input context and complicate lexicon alignment. To construct a compact and expressive feature space, EmoFusion combines 100−dimensional GloVe vectors with features from four emotion lexicons, resulting in a 196−dimensional representation. This reduces memory and computation requirements without compromising classification accuracy. We evaluated this trade-off empirically on the GoEmotions dataset, comparing our approach against a BERT-base model. Our results show that while BERT achieved a marginally accuracy improvement (1.5 − 2.2%), it incurred significantly greater inference time 0.59 s per sample and computational complexity due to its 768−dimensional feature space compared to EmoFusion's 0.37 s and 196−dimensions. These findings, aligned with prior work (Kumar & Raman, 2022), highlight the efficiency of our approach for deployment in real-time or resource-constrained settings.

A limitation of our approach is its inability to fully capture the inherently fuzzy boundaries between overlapping emotions (e.g., *happiness* and *love*), which complicates accurate classification. This challenge is particularly evident in cases like sarcasm or implicit emotional cues, where the model may misinterpret the underlying emotion. For instance, texts like "This is just fantastic... not" can be misclassified as neutral due to the model's reliance on explicit emotional keywords rather than a deeper contextual understanding.

7. Conclusion and future work

In this paper, we have introduced EmoFusion, an integrated machine learning model for text-based emotion classification that uses both semantic and linguistic feature representations. The proposed model converts each document into the appropriate representation by combining linguistic and semantically-similar feature-based vector representations, which include linguistic-based features, semantically-similar features, and semantically-similar linguistic features. The proposed approach is empirically evaluated on three benchmark datasets, such as Google AI GoEmotions, CBET, and TEC to demonstrate its efficacy in emotion classification. The evaluation results show significant gains in classification accuracy when compared to baselines and the state-of-the-art technique. The combination of richer linguistic features and emotion-specific pre-processing techniques aids in comprehending the nuanced emotional content in textual data. In future work, we plan to explore a wider range of general topics and incorporate a broader spectrum of emotions, such as *amusement*, *admiration*, *desire*, *grief*, and *love*, in order to get a more thorough grasp of emotion classification and its applicability across diverse domains. To address current limitations in future work. First, integrating contextual embeddings (e.g., BERT) could improve handling of implicit cues and sarcasm. Second, exploring multi-label, multi-modal data (e.g., combining text with images) may enhance interpretation of ambiguous expressions. Additionally, we could curate or augment datasets to balance the representation of less common emotions, addressing imbalanced distributions while enhancing robustness for context-aware models.

CRediT authorship contribution statement

Anjali Bhardwaj: Data curation, Formal analysis, Methodology, Validation, Writing – original draft, Review. **Muhammad Abulaish:** Conceptualization, Supervision, Validation, Editing, Review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Agrawal, A., An, A., & Papagelis, M. (2018). Learning emotion-enriched word representations. In *Proceedings of the 27th international conference on computational linguistics* (pp. 950–961). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Al Maruf, A., Khanam, F., Haque, M. M., Jiyad, Z. M., Mridha, F., & Aung, Z. (2024). Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*, *12*, 18416–18450.

Alm, R. D., Ovesdotter, C., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 579–586). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Bandhakavi, A., Wiratunga, N., Padmanabhan, D., & Massie, S. (2017). Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, *93*, 133–142.

Bhardwaj, A., & Abulaish, M. (2023). MINDS: A Multi-label emotion and sentiment classification dataset related to COVID-19. In *WNLPe-health 2022: the first workshop on context-aware NLP in eHealth* (pp. 3–13). IIIT Delhi, India: CEUR Workshop Proceedings.

Bhardwaj, A., Wasi, N. A., & Abulaish, M. (2023). Unlocking emotions in text: A fusion of word embedding and lexical knowledge for emotion classification. In *Proceedings of the 20th international conference on natural language processing* (pp. 766–772). Goa University, Goa, India: Association for Computational Linguistics.

Borod, J. C. (2000). The neuropsychology of emotion. Affective Science.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152). Association for Computing Machinery.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A. S., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL* (pp. 4040–4054). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.372.

Deng, J., & Ren, F. (2023). A survey of textual emotion recognition and its challenges. *IEEE Trans. Affect. Comput.*, *14*(1), 49–67. http://dx.doi.org/10.1109/TAFFC.2021.3053275.

Dresner, E., & Herring, S. C. (2010). Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory*, *20*(3), 249–268.

Ekman, P. (1999). Basic emotions. *Handbook of Cognition and Emotion*, *98*(16), 45–60.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1606–1615). Denver, Colorado: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/N15-1184.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society for Artificial Intelligence*, *14*(771–780), 1612.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Haque, A., & Abulaish, M. (2022). A graph-based approach leveraging posts and reactions for detecting rumors on online social media. In *Proceedings of the 36th Pacific Asia conference on language, information and computation* (pp. 533–544).

Haque, A., & Abulaish, M. (2023). An emotion-enriched and psycholinguistics features-based approach for rumor detection on online social media. In *Proceedings of the 11th international workshop on natural language processing for social media* (pp. 28–37). Bali, Indonesia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.socialnlp-1.4.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on world wide web* (pp. 607–618). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2488388.2488442.

Islam, J., Mercer, R. E., & Xiao, L. (2019). Multi-channel convolutional neural network for Twitter emotion and sentiment recognition. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 1355–1365). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1137.

Kumar, P., & Raman, B. (2022). A BERT based dual-channel explainable text emotion recognition system. *Neural Networks*, *150*, 392–407.

Li, Z., Chen, X., Xie, H., Li, Q., & Tao, X. (2020). EmoChannelAttn: Exploring emotional construction towards multi-class emotion classification. In *2020 IEEE/WIC/aCM international joint conference on web intelligence and intelligent agent technology (WI-IAT)* (pp. 242–249). IEEE, http://dx.doi.org/10.1109/WIIAT50758.2020.00036.

Mohammad, S. (2012). #Emotional tweets. In ** SEM 2012: the first joint conference on lexical and computational semantics–volume 1: proceedings of the main conference and the shared task, and volume 2: proceedings of the sixth international workshop on semantic evaluation (semEval 2012)* (pp. 246–255).

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. *1*, In *Proceedings of the 56th annual meeting of the association for computational linguistics, ACL* (pp. 174–184). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1017.

Mohammad, S. M. (2018). Word affect intensities. In *Proceedings of the eleventh international conference on language resources and evaluation, LREC*. Miyazaki, Japan: European Language Resources Association (ELRA).

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). New Orleans, Louisiana: Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465. http://dx.doi.org/10.1111/j.1467-8640.2012.00460.x.

Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining* (pp. 1–8).

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP* (pp. 1532–1543). Doha, Qatar: ACL, http://dx.doi.org/10.3115/v1/d14-1162.

Poria, S., Gelbukh, A. F., Cambria, E., Hussain, A., & Huang, G. (2014). EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, *69*, 108–123. http://dx.doi.org/10.1016/j.knosys.2014.06.011.

Rennie, J. D. (2001). Improving multi-class text classification with naive Bayes. URL https://dspace.mit.edu/handle/1721.1/7074.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Shahraki, A. G., & Zaïane, O. R. (2017). Lexical and learning-based emotion mining from text. In *Proceedings of the international conference on computational linguistics and intelligent text processing*.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307. http://dx.doi.org/10.1162/COLI_a_00049.

Vo, D.-T., & Zhang, Y. (2015). Target-dependent Twitter sentiment classification with rich automatic features. In *Twenty-fourth international joint conference on artificial intelligence* (pp. 1347–1353).

Wagner, W. (2010). Steven Bird, Ewan Klein and Edward Loper: Natural language processing with Python, analyzing text with the natural language toolkit. *vol. 44*, (4), (pp. 421–424). Beijing: O'Reilly Media, Inc., http://dx.doi.org/10.1007/s10579-010-9124-x,

Wang, Z.-Q., Sun, X., Zhang, D.-X., & Li, X. (2006). An optimal SVM-Based text classification algorithm. In *2006 international conference on machine learning and cybernetics* (pp. 1378–1381). IEEE, http://dx.doi.org/10.1109/ICMLC.2006.258708.

Wasi, N. A., & Abulaish, M. (2020). Document-level sentiment analysis through incorporating prior domain knowledge into logistic regression. In *Proceedings of the 19th IEEE/WIC/aCM international joint conference on web intelligence and intelligent agent technology (WI-IAT)* (pp. 969–974). IEEE, http://dx.doi.org/10.1109/WIIAT50758.2020.00148.

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, *50*(2), 1–33. http://dx.doi.org/10.1145/3057270.

Zahiri, S. M., & Choi, J. D. (2018). Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Zhou, D., Wu, S., Wang, Q., Xie, J., Tu, Z., & Li, M. (2020). Emotion classification by jointly learning to lexiconize and classify. In *Proceedings of the 28th international conference on computational linguistics* (pp. 3235–3245).