Comparative Analysis of Transformer-Based and Neural Network Models for Emotion Detection in Tweets

Bin Zhang[1], Xi Yang[2*], Chen Zhang[3], Tangsen Huang[1]
[1]School of Information Engineering, Hunan University of Science and Engineering, Yongzhou 425199, Hunan, China
[2]College of Intelligent Manufacturing, Hunan University of Science and Engineering, Yongzhou 425199, Hunan, China
[3]School of International Education, Guangdong University of Technology, Guangzhou 510006, Guangdong, China
E-mail:yangxi_0423@163.com
[*]Corresponding author

Keywords: transformer-based models, natural language processing (NLP), neural networks, tweet classification

Received: November 9, 2024

This study assesses the effectiveness of transformer-based and neural network models for detecting emotions in tweets. Five models are evaluated: two transformer-based frameworks (DistilBERT and ALBERT), two neural network architectures (CNN and 3CNN-3LSTM), and a hybrid model (3CNN-3LSTM-GloVe 300x). The models are evaluated based on accuracy, precision, recall, and F1-score. The findings indicate that ALBERT attains the maximum accuracy at 86.38%, succeeded by DistilBERT with an accuracy of 84.35%. The 3CNN-3LSTM model exhibits an accuracy of 83.79%, whilst the CNN model demonstrates the lowest performance at 65.37%. The hybrid 3CNN-3LSTM-GloVe 300x model exhibits a performance of 75.61%. The results demonstrate that transformer-based models surpass neural network models in emotion recognition, especially in recognizing subtle emotional expressions. Nonetheless, transformer-based models demonstrate increased computational expenses, highlighting the necessity for optimization in real-time applications. This study enhances the domain of emotion detection by a comparative comparison of diverse models, emphasizing the benefits of transformers while acknowledging the computational difficulties. The results indicate significant implications for marketing, mental health, and digital communication, highlighting the need for further enhancement of transformer models for effective implementation.

Povzetek: Študija primerja različne modele za zaznavanje čustev v tvitih in ugotavlja, da so transformerji (zlasti ALBERT) najbolj natančni, vendar zahtevajo več računske moči.

1 Introduction

The swift expansion of social media enables individuals to readily articulate their views and emotions. This offers a significant opportunity to examine emotional expressions in real-time, as networks such as Twitter function as extensive warehouses of user-generated content. This study examines emotion analysis and detection, applying a comprehensive methodology that integrates diverse features and unique approaches from machine learning and deep learning, employing pre-trained frameworks.

Our research will consider six core emotions: anger, fear, happiness, love, sadness, and surprise. Each of these core emotions represents an essential aspect of human sentiment. These are bound to affect the behavior of people and hence help mold mass opinions and trends on social media. We seek to investigate how these six core emotions are manifested through tweets with the hope that there might come to light some hidden patterns, which could eventually lead us to further insights into emotional communication in digital spaces.

This was achieved through a broad set of features extracted from the textual data of Twitter. In this proposal, we will further improve the efficiency and robustness of the ED framework by integrating linguistic cues, sentiment scores, and contextual elements. By utilizing both traditional ML strategies and innovative DL frameworks, we will be able to compare their effectiveness and determine which are the most promising strategies that have been used for emotion classification.

This lack of integration between theoretical insights and practical application in ED also provides a backdrop against which our study is placed. We want to demonstrate the usability of the frameworks in a way that hopes to add new, useful tools for researchers, marketers, and mental health experts who are trying to harness emotional data for their various purposes. In this manner, the present research moves the field of emotion analysis one step further and paves the ground for further investigation into the intricacies of human emotions in the digital era.

1.1 Literature review

The concept of ED and sentiment analysis is a rapidly evolving field that has gained significant interest from researchers. Consequently, numerous studies and frameworks have been proposed to advance this area of research.

Acheampong et al. [1] delve into transformer-based frameworks for natural language processing tasks, examining their advantages and disadvantages. The study includes an analysis of frameworks such as GPT and its variants, Transformer-XL, XLM, and BERT. A particular emphasis is placed on BERT's efficiency in text-based emotion identification. The paper reviews a variety of BERT-centered frameworks, discussing their contributions, outcomes, limitations, and the datasets employed. Additionally, the authors propose future research directions to boost TBED.

Mathew and Bindu [2] introduced Efficient Transformer-based Sentiment Classification (ETSC) models to mitigate the drawbacks of big transformer models, such as substantial hardware requirements and prolonged training durations. By reducing model configurations, randomizing datasets, and altering training data, their models attained enhanced performance without compromising accuracy. The ETSC models surpassed current transformer-based sentiment categorization methods in both speed and accuracy.

In another study by Acheampong et al. [3], their article reviews emotion detection (ED) from texts, focusing on the basic strategies investigators use to develop TBED systems. It explores recent innovative proposals, detailing their contributions, frameworks, datasets, results, strengths, and weaknesses. Additionally, the article provides a list of emotion-labeled data sources for newcomers and discusses open issues and future research directions in TBED.

Zad et al. [4] Text-Based Emotion Detection: The paper reviews a very important emerging field in NLP, which has focused on classifying text into emotion categories defined by psychological frameworks. It presents how ML strategies automate the process of emotion extraction and puts forward applications of this, such as document analysis related to terrorist activities, historical corpora, and product reviews for customer satisfaction. The work reviews existing literature on TBED and associated psychological frameworks.

Guo [5] Introduces DL Assisted Semantic Text Analysis, DLSTA, that can detect human emotions in massive datasets. The approach leveries NLP strategies and word embeddings toward the implementation of functions such as Sentiment Analysis. This can, therefore, enrich traditional learning-based methods with the inclusion of semantic and syntactic attributes in DLSTA. It documents a 97.22% human ED rate and 98.02% accuracy of classification, which outperforms state-of-the-art innovative frameworks. Hence, this work suggests that at least partial further improvement is achieved by the application of more emotive word embeddings.

Seal et al. [6] Therefore, this work proposes an effective ED strategy that identifies emotional words from a predefined keyword dataset. The proposed approach by them estimates emotional words and phrasal verbs along with negation, outperforming some of the recent strategies for ED.

Rashid et al. [7] Aimen's system, which detects emotions within social media dialogues, specifically identifies happiness, sadness, and anger. The system makes use of an LSTM framework along with word2vec and doc2vec embeddings. Results show major improvement in f-scores above baseline frameworks to 0.7185, therefore proving effective for ED.

Dang et al. [8] It describes the importance of sentiment analysis on public opinion in social networks such as Twitter and Facebook. They also note some challenges in NLP, which are considered the reason for reducing the accuracy of sentiment analysis. The paper reviews recent developments in DL frameworks to address these issues, especially in sentiment polarity, by using term frequency-inverse document frequency (TF-IDF) and word embeddings.

In a recent investigation, Zhu et al. [9] Conducted sentiment analysis: This is a technology applied for the assessment of attitude in view of several entities. Its applications are to give directions on product reviews, analyses of public opinions, psychological analyses, and risk assessment. Current approaches are bound by traditional frameworks that rely solely on texts, thereby critical contextual information can be lost due to figurative languages such as irony. Multimodal sentiment analysis, therefore, merges visual and acoustic information to enhance the detection of sentiments. Some challenges that are faced while integrating cross-modal data, a review of various fusion frameworks, have been discussed in this paper. It also covers the state-of-the-art in multimodal sentiment analysis, focusing on popular datasets, strategies for feature extraction, applications, and challenges yet to be overcome, to stir researchers into developing workable frameworks.

Murthy et al. [10] Overview ED involves a survey of the difficult regions in NLP. These authors explore the study of emotional states from textual data, along with facial and audio modalities. Their significance is highlighted using studies from neuroscience, psychology, and human-computer interaction. Authors based on text from social media, blogs, and customer reviews outline existing approaches and architectures, data sets, lexicons, measures, and their limitations in ED which can enable the researchers with valuable insights.

Madhuri & Lakshmi's study [11] Review emotion analysis in natural text and indicate that the majority of AI research does not try very hard to explain the reasons behind the misclassifications. They find defects in traditional frameworks due to weak emotion correlations and propose a model that associates emotion recognition with correlation mining. Their approach employs ML and DL strategies using features representing anger, fear, and joy emotions. Experiments with the Kaggle dataset were also performed to appraise the accuracy of different frameworks in ED. It is followed by a short description of frameworks and frameworks presented in this paper.

Shahzad et al. [12] reviewed recent studies on emotion recognition, highlighting the role of physiological signals like respiration patterns. They noted challenges with traditional methods, such as facial recognition and speech analysis, including the use of invasive sensors and limited study diversity. Despite progress, they concluded that a universal solution for emotion detection remains unachieved. This aligns with the current study, which addresses similar challenges in emotion detection from tweets using deep learning models. Khan et al. [13] examined facial expression analysis methodologies, observing that current models encounter difficulties with authentic datasets such as FER-2013. They proposed a CNN-based model that enhanced accuracy, attaining 74.92% on FER-2013 and over 99% on CK+ and FERG, with increased feature extraction and diminished computing complexity. This paper investigates deep learning models for emotion recognition in tweets, focusing on performance and efficiency difficulties using real-world data. Koufakou et al. [14] highlighted the necessity for emotion-annotated corpora by compiling a comprehensive dataset from diverse sources since 2018. They illustrated the influence of varied datasets on emotion classification through the fine-tuning of a pretrained model. Their research addressed a deficiency in emotion detection studies by offering a consistent resource for benchmarking and comparison. Plaza-del-Arco et al. [15] conducted a survey of 154 NLP papers on emotion analysis (EA), examining critical inquiries on task definitions, emotion frameworks, and the influence of demography and culture. They found four deficiencies: insufficient consideration of cultural diversity, inadequate alignment of emotional categories, absence of standardized emotional assessment terminology, and restricted multidisciplinary research. Their research highlights the need for a more comprehensive approach to representing emotions within NLP.

1.2 Research gaps and novelties

Although considerable advancements have been made in emotion detection via transformer-based models such as DistilBERT and ALBERT, current research has predominantly emphasized their advantages, especially their elevated accuracy, while neglecting to properly examine their computing inefficiencies. Although these models demonstrate superior performance, they are computationally intensive, rendering them less appropriate for real-time applications. Moreover, several contemporary methodologies depend on a singular dataset, such as tweets, which engenders biases that undermine the models' generalizability across other emotional circumstances and platforms. This paper rectifies these deficiencies by offering a comparative examination of transformer models and conventional neural network topologies, highlighting their different advantages and disadvantages in emotion recognition.

This work's originality resides in its balanced methodology, showcasing a hybrid model that amalgamates convolutional neural networks, long short-term memory networks, and pre-trained GloVe embeddings, therefore efficiently reconciling performance with computational efficiency. This approach decreases the computational expense of transformer models while preserving competitive accuracy in emotion categorization. This research provides novel insights into the practical use of emotion recognition by integrating linguistic and contextual information and examining six fundamental emotions across multiple models. This study reframes the contribution of transformer models, which consistently surpass simpler topologies, by emphasizing the essential requirement for optimization in practical applications, so presenting the work as a significant comparison rather than a methodological innovation. This study examines three principal research inquiries: What is the comparison between transformer-based models (DistilBERT, ALBERT) and neural network models (CNN, LSTM) in the context of emotion recognition from tweets? What difficulties emerge in representing emotional complexity across six categories: anger, fear, happiness, love, sadness, and surprise? What are the trade-offs between computational efficiency and precision in real-time emotion recognition on social media? These inquiries direct the examination of model efficacy and operational viability.

Table 1 encapsulates recent studies on emotion recognition, contrasting approaches, datasets, and performance indicators. It delineates various methodologies and datasets employed, offering a succinct reference for contemporary trends and benchmarks in the domain.

Table 1: Summary of methodologies, datasets, and key performance metrics in recent emotion detection studies.

Study	Methodology	Datasets	Key Performance Metrics
Acheampong et al. [1]	Transformer-based frameworks for NLP tasks	GPT, Transformer-XL, XLM, BERT	Efficiency in emotion identification
Mathew & Bindu [2]	Efficient Transformer-based Sentiment Classification (ETSC)	Randomized datasets	Speed, accuracy
Acheampong et al. [3]	Emotion detection using various TBED strategies	Emotion-labeled datasets	Contributions, strengths, weaknesses
Zad et al. [4]	Machine learning strategies for emotion extraction	Historical corpora, product reviews	Emotion classification accuracy
Guo [5]	DL-assisted Semantic Text Analysis (DLSTA)	Large-scale datasets	97.22% human ED rate, 98.02% accuracy

Seal et al. [6]	Emotional word identification from keyword datasets	Predefined keywords dataset	Accuracy, F1-score
Rashid et al. [7]	LSTM framework with word2vec and doc2vec embeddings	Social media dialogues	F-scores: 0.7185
Dang et al. [8]	Sentiment analysis using TF-IDF and word embeddings	Twitter, Facebook	Sentiment polarity accuracy
Zhu et al. [9]	Multimodal sentiment analysis (text, visual, acoustic)	Multimodal datasets	Sentiment classification accuracy
Murthy et al. [10]	Survey of emotion detection across modalities	Social media, blogs, customer reviews	Emotion detection insights
Madhuri & Lakshmi [11]	Emotion recognition with correlation mining	Kaggle dataset	Accuracy of emotion detection
Shahzad et al. [12]	Physiological signal-based emotion recognition	Face and speech datasets	General emotion detection efficiency
Khan et al. [13]	CNN-based facial expression recognition	FER-2013, CK+, FERG	Accuracy: 74.92% (FER-2013), >99% (CK+, FERG)
Koufakou et al. [14]	Fine-tuning pretrained models for emotion detection	Diverse datasets since 2018	Impact of dataset variation on classification
Plaza-del-Arco et al. [15]	Survey on emotion analysis in NLP	154 NLP papers	Task definitions, frameworks, demographic factors

2 Materials & methods

This study presents and compares five different ML-based frameworks using various performance metrics. The frameworks include pre-trained transformers such as ALBERT and Distil BERT, neural network frameworks like CNN and 3CNN-3LSTM, and a hybrid model combining neural networks with pre-trained embeddings, specifically 3CNN-3LSTM-GloVe 300x.

The transformer-based models, ALBERT and DistilBERT, necessitate comprehensive elucidations about their structures, tokenization methodologies, and the application of pre-trained weights. ALBERT, a streamlined variant of BERT, diminishes the parameter count by factorized embedding parameterization and cross-layer parameter sharing, all while preserving superior performance. It employs the WordPiece tokenizer and utilizes pre-trained weights from BERT for subsequent jobs. DistilBERT, a more compact and efficient variant of BERT, maintains 97% of BERT's performance while utilizing 60% fewer parameters through knowledge distillation. It employs the WordPiece tokenizer and pre-trained weights from BERT, with modifications implemented during the distillation process to improve efficiency.

A brief description of the ML frameworks and proposed frameworks is provided below, followed by an explanation of the methodology. Before this, a detailed description of the data and preprocessing steps is presented in the next section.

2.1 Data description and preprocessing

The dataset underwent preprocessing, which involved the elimination of stop words, punctuation, and non-English characters to achieve consistency and clarity. These preprocessing processes were uniformly implemented across all models, encompassing the transformer-based models (ALBERT and DistilBERT) and the neural network models (CNN and 3CNN-3LSTM). This standardization guarantees uniform, clean input data for all models, facilitating an equitable assessment of their performance. Furthermore, statements exceeding 200 characters were omitted from the dataset to diminish complexity and concentrate on more typical, succinct samples.

The report presents figures on emotion distribution, sentence length, and word clouds, although it omits information regarding the dataset's size, collection methods, and annotation process. The dataset comprises tweets from Twitter, meticulously annotated by a team adhering to explicit rules. Annotator biases may have affected classification, and the dataset is uneven, with a predominance of "happy" and "sad" tweets. To ensure repeatability, the data were divided into training (65%), validation (15%), and test (20%) sets; nonetheless, additional clarity regarding the dataset's size, collecting methodology, and potential biases would enhance the evaluation of generalizability.

The data includes various textual inputs collected from the Twitter (X) social network. The collected data is categorized into several classes with specific labels, which are depicted in Fig. 1.

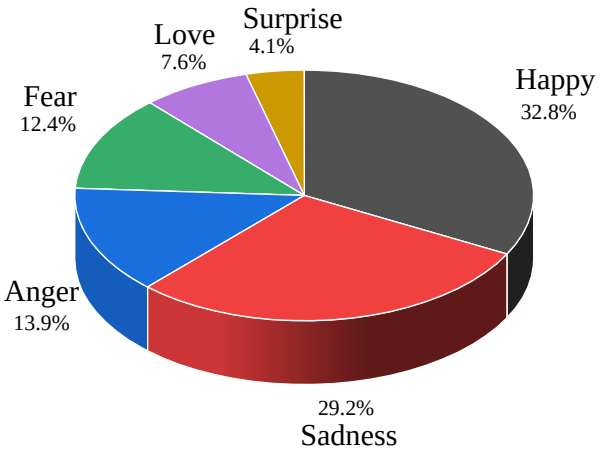

Figure 1: Data classification into specific labels

Fig. 1 demonstrates the distribution of data across each class. The labels for happiness and sadness account for the majority of the dataset, while only 4.1% of the data is classified under the label of surprise. Each label represents specific sentiments and includes sentences related to that sentiment.

The length of the sentences in each label is presented in Fig. 2:

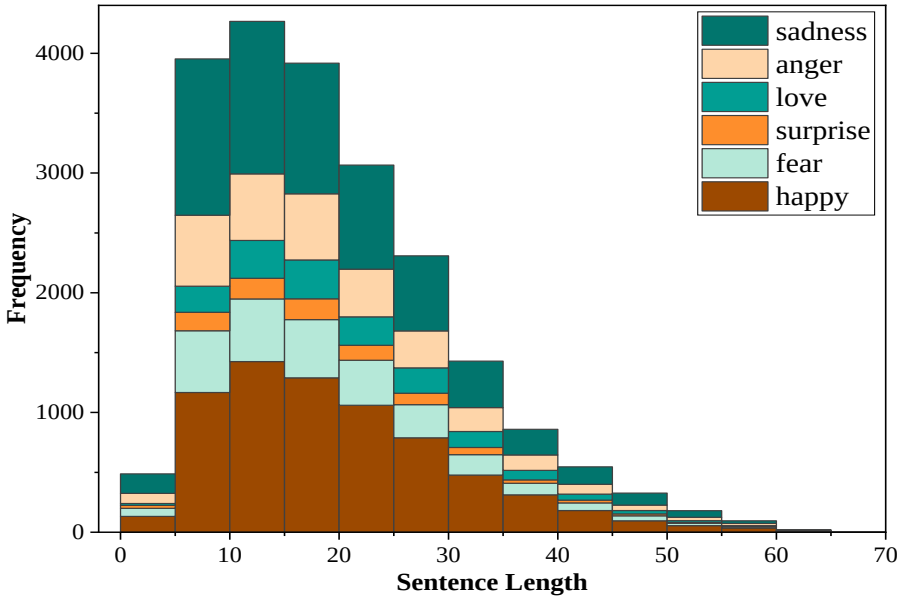

Figure 2: The length of the sentences on each label

Fig. 2 illustrates the frequency of the sentences with the specified length. As shown, the sentences with sad content tend to be longer, and the sentences containing anger are the second longest class of sentences. Also, according to the graphs, the sentences in the class of happy tend to be the shortest.

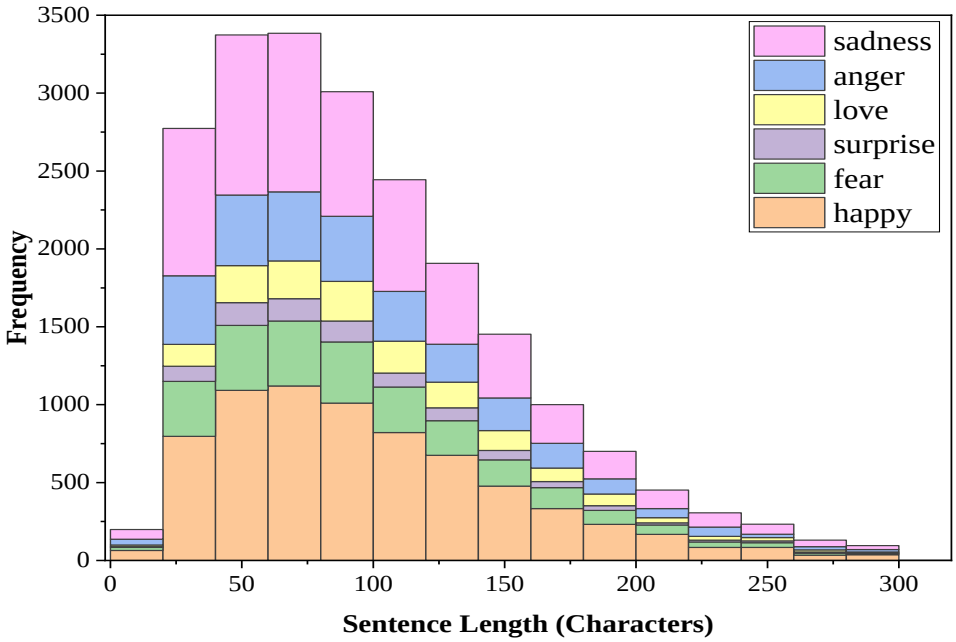

Figure 3: The length of the sentences in each label is based on the character length

Fig. 3 indicates that sentences containing the sadness sentiments tend to have longer characters, while on the other hand, sentences including happy content have shorter characters. Furthermore, it can be derived that the shorter characters (length of 50 to 100) are less frequent than the longer characters (length of 200 or more).

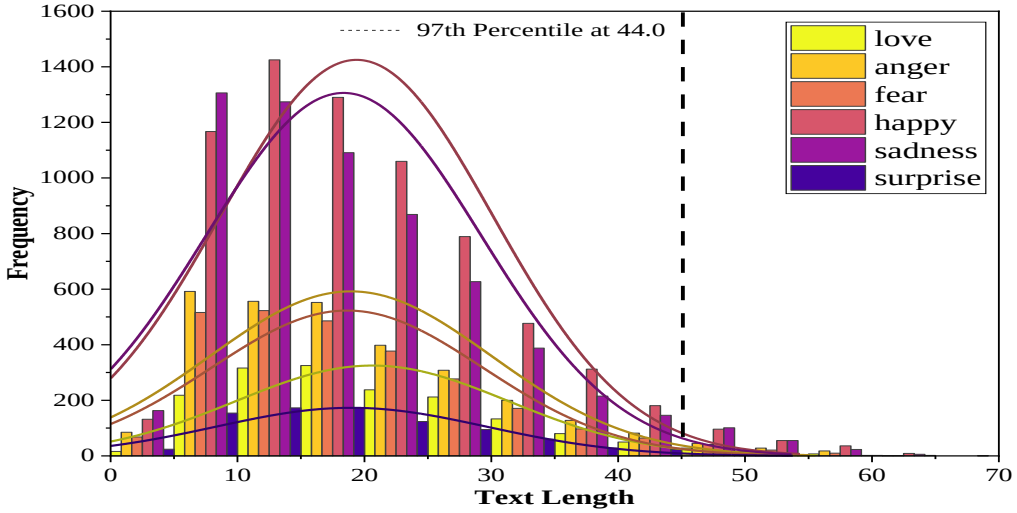

Figure 4: The frequency of the texts based on the length

As it is shown in Fig. 4, about 97 percent of the textual samples have a length of 44 or less. This study aims to use these samples, and the longer texts are ignored. This is because it saves a lot of information and reduces the complexity of the frameworks.

The Figs. 5 to 10 demonstrate the word cloud of each class:

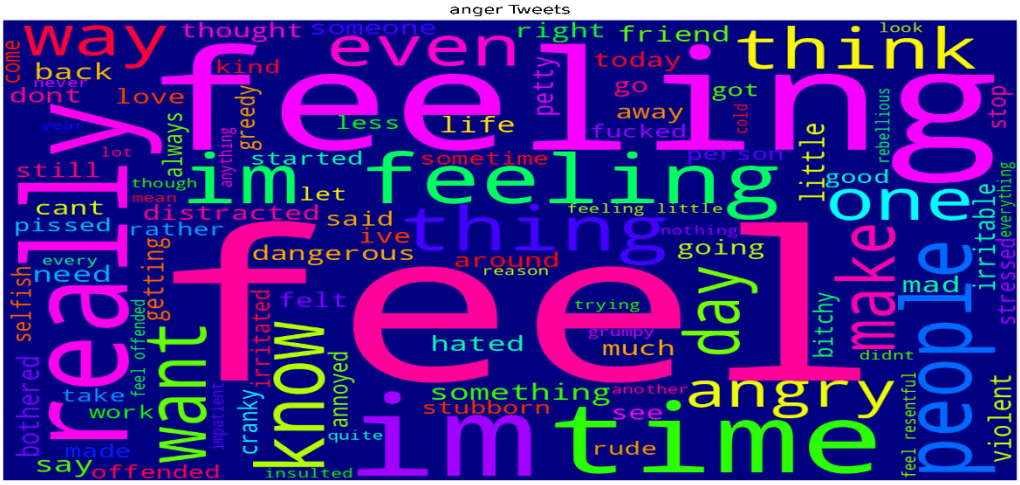

Figure 5: The word cloud of the label of anger

According to Fig. 5, the tweets with anger sentiment contain words such as feel, feeling, really, and time, more frequently. It denotes that user usually talk about their feelings while being angry.

This word cloud emphasizes the predominant phrases in tweets conveying rage, with words such as "feeling," "want," and "know," which signify emotional irritation. Terms such as "angry," "irritable," and "offended" denote significant emotional intensity, yet words like "people," "time," and "day" imply social interactions and circumstances that provoke anger. This picture elucidates the expression of anger on social media platforms.

Figure 6: The word cloud of the label of fear

Fig. 6 also shows that words such as feel, feeling, I'm, anxious, and know are more frequent in the tweets of the label fear. It indicates that sentences with fear sentiment and anger sentiment share almost similar words.

This word cloud illustrates the most prevalent terms in tweets concerning fear. Terms such as "feeling," "really," and "know" denote particular emotional states, whereas words like "anxious," "scared," and "nervous" emphasize the severity of fear. The visualization emphasizes themes of vulnerability and uncertainty, providing insights into the expression of fear on social media.

Figure 7: The word cloud of the label happy

Fig. 7 demonstrates that, besides words such as feel and feeling that are common in most of the classes, words like happy, love, time, people, and excited have been repeated more.

This word cloud highlights the dominant terms in tweets regarding happiness. Terms such as "feel," "truly," and "joyful" denote authentic elation, whereas words like "existence," "day," and "duration" imply that happiness is associated with quotidian events. Social relationships are highlighted by concepts such as "family," "work," and "friend," offering insights into the expression of happiness on social media.

Figure 8: The word cloud of the label of love

According to Fig. 8, words such as feel, feeling, love, really, time, sweet, and want are more frequent in the tweets, including the love content.

This word cloud depicts prevalent terms in tweets concerning love. Terms such as "feel," "love," and "sweet" denote profound affection, but words like "tender," "caring," and "loyal" emphasize emotional bonds. Terms such as "romantic" and "beloved" underscore ideas of intimacy and connection, providing insight into the expression of love on social media.

Figure 9: The word cloud of the label of sadness

Fig. 9 illustrates that after words like feel and feeling, words such as know, really, time, thing, day, and life are among the most frequent words.

This word cloud emphasizes prominent phrases in tweets conveying grief, with words such as "feel," "really," and "sad," which indicate profound inner turmoil. Terms such as "alone," "hurt," and "depressed" underscore the severity of negative feelings, whereas "time," "life," and "year" imply associations with personal experiences. The prevalence of "people," "family," and "friend" signifies that relationships are crucial in articulations of grief.

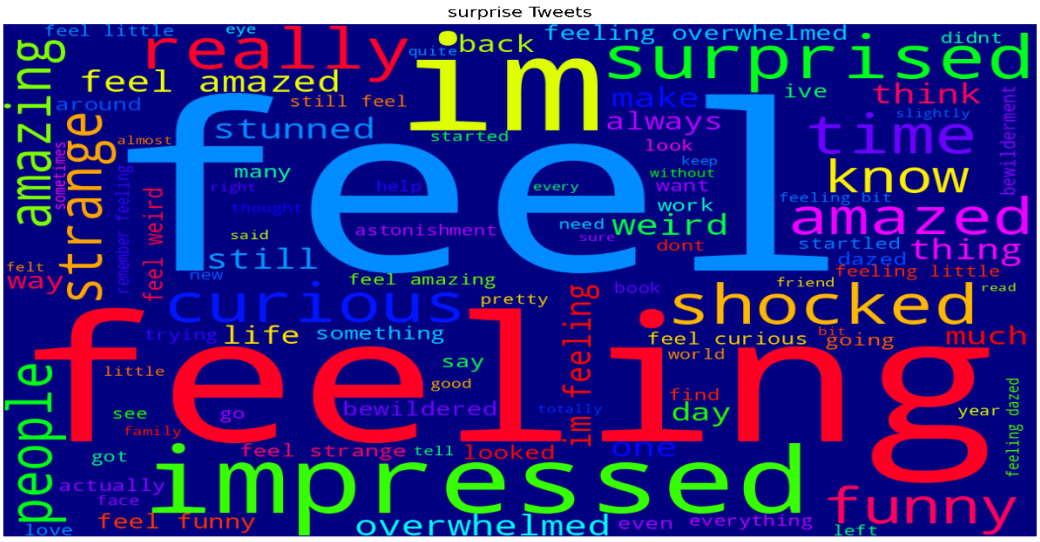

Figure 10: The word cloud of the label of surprise

Fig. 10 denotes that tweet in the label of surprise are mostly comprised of words such as feel, feeling, impressed, surprised, funny, curious, amazing, and shocked. This word cloud depicts the most prevalent phrases in tweets concerning surprise, using words such as "feel," "amazed," and "surprised," which signify profound emotional responses. Terms like "strange," "weird," and "curious" indicate that surprise is frequently elicited by atypical events, whereas "people" and "life" emphasize the social and situational circumstances of surprise.

Since this study seeks to classify and detect the emotion from the tweets, it can be seen in the Figs. 5 to 10 that users tend to write about their feelings in tweets containing sentiments such as anger, fear, love, etc.

The preprocessing in this study includes the removal of stop words and punctuation to clean the tweet data. Non-English characters were then eliminated to ensure uniformity and readability. Any instances with missing values were removed to maintain data integrity. The cleaned data were then segmented into three sets: 65% for training, 15% for validation, and 20% for testing, facilitating robust model development and assessment.

These figures illustrate the ROC curves for five emotion recognition models, utilizing the area under the curve (AUC) to assess model performance. The curves illustrate the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for each model, where a larger AUC signifies superior performance. DistilBERT exhibits the greatest AUC of 0.9814, signifying exceptional performance, succeeded by ALBERT with an AUC of 0.9777. The 3CNN-3LSTM model attains an AUC of 0.9604, whilst the 3CNN-3LSTM+Glove300x

and CNN models exhibit AUCs of 0.9283 and 0.8879, respectively. The findings indicate that transformer-based models such as DistilBERT and ALBERT excel in emotion recognition, particularly regarding sensitivity (true positive rate), whereas CNN models have inferior performance.

Sentences above 200 characters were omitted from the dataset to minimize noise and complexity. Extended sentences frequently have superfluous elements that could hinder the model's generalization capabilities, whereas concise sentences better reflect standard tweet durations. This exclusion enhances computational efficiency, accelerating the training and evaluation processes. Concentrating on concise words enables the model to more effectively discern pertinent patterns for emotion recognition.

Example of Preprocessing:

- Before Preprocessing: "I am so frustrated today. Nothing seems to be going right, and everything I do just feels like it's falling apart."
- After Preprocessing: "I am so frustrated today."

This step ensures a more uniform and efficient dataset, contributing to better model performance.

2.2 Description of the frameworks

This paper presents and compares five frameworks that are based on the structure of CNN, LSTM, and BERT. Hence, a brief description of these ML frameworks is presented and then the proposed frameworks in this study are explained.

1. Convolutional Neural Network (CNN)

CNN is a specialized type of artificial neural network designed primarily for analyzing and processing data that has a grid-like structure, such as images. The architecture of CNNs is inspired by the visual processing mechanisms of the human brain, particularly the organization of the animal visual cortex. A CNN consists of several layers, each serving a unique goal in the network's function. The fundamental concept of CNNs is their ability to automatically and adaptively learn spatial feature hierarchies, ranging from low-level edges and textures to high-level object elements and complete objects, directly from raw data.

The fundamental building block of a CNN is the convolutional layer, in which a set of learnable filters (or kernels) slides over the input data to create feature maps. Each filter is designed to recognize a certain type of feature (e.g., an edge, corner, or texture) within the input data. This process of convolution helps the network capture local patterns, and because these filters are applied uniformly across the entire input, CNNs are highly effective in maintaining spatial hierarchies and translation invariance, making them particularly well-suited for image recognition tasks.

Following the convolutional layers, CNNs usually contain pooling layers, which perform a down-sampling operation to diminish the spatial dimensions of the feature maps and thus control overfitting by decreasing the number of factors and computations in the network. Pooling layers, commonly using max-pooling, take the maximum value from a small neighborhood in the feature map, preserving the most prominent features while discarding less important information. This reduction process not only makes the computation more efficient but also provides a form of spatial invariance to small translations in the input image.

Following multiple convolutional and pooling layers, CNNs normally incorporate one or more fully connected layers, similar to those in CNNs. These layers interpret high-order features of the convolutional layers and predict results based on those features, such as class scores for classification tasks. These CNNs combine convolutional layers with pooling and fully connected layers to develop a hierarchical representation of the input data, making them super effective in image and video recognition, image classification, medical image processing, and even NLP.

Among various powerful and versatile architectures, CNNs deal with different types of data characterized by a grid topology. Their capability of learning from the data itself and adaptively obtaining representations of spatial hierarchies of features has revolutionized several fields of AI, with special reference to those dealing with visual data. Their structured yet flexible architecture makes them a cornerstone in developing more innovative and specialized DL frameworks [16], [17].

2. Long Short-Term Memory (LSTM)

LSTM is a special kind of RNN architecture that was specifically designed to avoid the problems that traditionally faced most RNNs in learning the features of long-term dependencies. This is generally the case because RNNs are suitable when one is dealing with data in sequences due to the hidden state that carries information about past inputs in a sequence. However, they usually have a problem with long-term dependencies because of some notorious issues, like gradient values exploding or vanishing during training. These are overcome in LSTMs, which include a memory cell with a set of gates that control information flow.

The core of an LSTM is the cell state, a kind of conveyor belt running through the whole sequence to carry information with a gradient that does not get wiped out. This helps the cell to remember long-range information without being overwritten at each time step. In addition, there will be an input gate, a forget gate, and an output gate into the cell state. Three gates control the flow of information in and out of the cell state. While the input gate determines how much new information is added from the current input to the cell state, the forget gate decides the proportion of information in the cell state to retain or drop. The output gate regulates how much information from the cell state needs to be output onto the final hidden state.

The forget gate is very important in LSTMs because it makes the network selective about which part to forget and which to remember. And that becomes an important ability if one wants to learn tasks such as context remembering, which has very, very long sequences-for example, language modeling or time series prediction. The forget gate takes a look based on what this new input is

saying and what was there in the previous hidden state to identify parts of the cell state that need to be kept. The selective memory here enables the LSTM to memorize valuable information and discard unimportant data. This helps resolve the problem of vanishing gradients, hence letting the network learn dependencies on very long-time intervals.

As a result, LSTMs have seen wide applications in many areas where data is sequential. In NLP, they do very well for machine translation, speech recognition, and tweet generation-things that are examples of tasks requiring long-range context. For time series forecasting, LSTM networks can learn to identify temporal dependencies and trends that would otherwise be obscure. Furthermore, they are also applied in anomaly detection, video analysis, and any other area where sequence data are common. By overcoming various shortcomings of traditional RNNs, LSTMs have become among the key building blocks of many state-of-the-art deep learning frameworks; this enabled a significant performance gain in various kinds of tasks involving sequential data [18], [19].

The present paper utilizes two frameworks, namely ALBERT and Distill BERT, which have been developed based on the BERT transformer model. The next section briefly introduces these three frameworks.

3. *BERT model*

BERT is an acronym that stands for Bidirectional Encoder Representations from Transformers, a revolutionary new language representation framework from Google to improve the understanding of natural languages. Unlike other older frameworks, which would process the tweet in a unidirectional manner, BERT uses a transformer architecture to read tweets bi-directionally, considering context from both the left and the right sides of a word simultaneously. In the process, BERT adopts a bidirectional approach, which makes it much stronger for the capture of deeper and richer language understanding.

The core novelty of BERT is its training methodology, which consists of two major phases: pretraining and fine-tuning. During pre-training, it learns the representations of languages by being trained on large corpora of tweets through two unsupervised tasks: masked language modeling and next-sentence prediction. In masked language modeling, some words in a sentence are randomly masked, and BERT is asked to predict these masked words from the surrounding context. This is designed to have the framework learn much about the relationship between words in a sentence. In contrast, the next-sentence prediction involves identifying whether a given sentence contextually follows from a previous sentence. This helps BERT understand the logical relations between sentences at the sentence level.

After pretraining, BERT can be fine-tuned on specific NLP tasks with relatively small datasets. This involves re-training pre-trained models on specific tasks such as question answering, sentiment analysis, and named entity recognition, among others. Fine-tuning requires minimal alteration of the model architecture; hence, it is highly flexible and efficient. This is the case when BERT can be fine-tuned for different tasks, remarkably improving the performance of natural language processing applications by drawing from its broad knowledge in the use of language and doing well on particular tasks in need with only limited additional training.

It caused a revolution in the world of NLP and set new standards for many tasks. Its usage in academia and industry is very popular. Its architecture and training method have influenced many more frameworks, some of which are its variants: ALBERT light version of BERT, and Distill BERT-smaller, faster, and cheaper than BERT. The success of BERT wowed the need for large tweet corpora pre-training and fine-tuning on a task, which helped settle the eventual direction of research in language frameworks [20], [21].

4. *ALBERT (A Lite BERT) model*

ALBERT represents a new variant of BERT, developed in the direction of improving efficiency and scalability while preserving strong performance on NLP tasks. ALBERT was developed by researchers at Google Research and aimed to overcome some of the main limitations of BERT, including its large size and high computational cost, so that it could be more applicable for resource-constrained environments.

The novelties in ALBERT are all related to the methods of reducing parameters without breaking the performance of the model. The major reduction techniques used in this model include factorized embedding parameterization and cross-layer parameter sharing. In factorized embedding parameterization, the big vocabulary embedding matrix is split into two smaller matrices. The number of parameters gets reduced in the embedding layer, which reduces the memory consumption by the model. On the contrary, cross-layer parameter sharing means that the parameters are shared between different layers of the network, instead of having unique parameters for each layer. This decreases not only the total count of parameters but also brings the framework to learn more robust and generalizable features.

Another major improvement in ALBERT is the better training mechanisms, which result in faster convergence and much better performance. ALBERT replaces the NSP task utilized in BERT with a sentence-order prediction (SOP) task. The SOP task involves a model identifying the correct order out of two given sentences, helping him understand sentence-level coherence and relations better. This change remedies certain weaknesses in the NSP task and helps in improving performance for downstream NLP tasks.

These are combined in such a way that ALBERT, with far fewer parameters than BERT, achieves amazing results. Put differently, the largest ALBERT model has roughly 18 times fewer parameters compared to the largest BERT model. It also yields results similar to or even outperforming that of BERT on several benchmark datasets. This makes ALBERT more practical in applications where computational resources and memory are limited [22], [23].

5. Distil BERT (Distilled BERT)

Distil BERT, created by Hugging Face, is lighter than BERT while striving for performances very close to BERT but much lighter and faster. The first objective of Distil BERT is to make BERT cheaper computationally and lighter in memory to adapt it to resource-limited applications. This is achieved by an operation called knowledge distillation, in which a small model, the student, learns from a bigger one, the teacher.

The process peculiar to Distil BERT makes the original BERT the teacher model and its student model a compressed version with fewer parameters. More precisely, Distil BERT keeps about 60% of the parameters of BERT and runs twice as fast while retaining 97% of BERT's performance on various NLP tasks. Accomplished herein is an efficiency in the model, reduced from twelve layers in the transformer architecture to six, but with the dimensions of hidden layers and the embedding for inputs.

Knowledge distillation involves training a smaller model to replicate the outputs of a larger model. Instead of learning directly from the raw data, the student model learns from the softened output probabilities produced by the teacher model. This method allows the student model to capture much of the essential knowledge and performance characteristics of the teacher model. It uses a three-part loss function during training: language modeling loss, distillation loss, and cosine distance loss. It makes sure the model learns the structure of the language through language modeling loss, while distillation loss aligns the outputs of the student with those of the teacher. Cosine distance loss serves to align the hidden states between student and teacher frameworks. Where Distil BERT seems to create a difference, though, is more in the practicality of real-world usage, where computational resources and response time are so critical. In such cases, large-scale NLP frameworks will hardly be able to be delivered on mobile applications, edge devices, or environments with limited infrastructure due to the immense consumption of resources. Given its reduced size and speedier inference time, Distil BERT meets these challenges effectively, opening innovative NLP capabilities to be available widely and with more deployments [24], [25].

6. Glove algorithm

Glove, or Global Vectors for Word Representation, is an unsupervised learning framework developed by investigators of Stanford University to obtain the vector representation of words. Unlike other traditional word embedding frameworks relying on local context windows, such as Word2Vec, Glove uses the global statistical information of a corpus in creating dense vector representations. These word representations capture the semantic links between words by observing a word co-occurrence matrix in a corpus, enabling words that are semantically similar to have similar vectors.

The key insight with Glove is that the meaning of a word can be induced from its co-occurrence with other words across the entire corpus. By constructing a huge co-occurrence matrix in which each entry contains the frequency with which two words appear together, the probability of word pairs is framed through the Glove framework. Then, the algorithm will factorize this co-occurrence matrix to find a set of word vectors most likely to predict the likelihood of the occurrence of such word pairs. The result is a set of word vectors where the dot product of two-word vectors corresponds to the logarithm of the probability of their co-occurrence.

One of the key advantages of Glove over other strategies for creating word embeddings is that it captures both local and global contexts. While the local context provides an insight into the immediate semantic relationship among words, it's the global context that helps a word to be understood within the wider context of the whole corpus. These dual considerations enable Glove to give more robust and meaningful embeddings, which is very useful in applications requiring a deep understanding of the language, such as semantic similarity, solving analogies, and machine translation.

The Glove training process works in a way that it tries to find the cost function to be minimized so that the variance between the predicted and actual word co-occurrence probabilities is reduced. This is done through stochastic gradient descent: iteratively, the word vectors are changed such that the co-occurrence statistics can be best fitted [26], [27], [28].

All models employed the subsequent hyperparameters: a learning rate of 0.001, a batch size of 32, a dropout rate of 0.3, and embedding dimensions of 128 for CNN and LSTM models, and 300 for GloVe embeddings. Models underwent training for 30 epochs utilizing the Adam optimizer ($\beta1 = 0.9$, $\beta2 = 0.999$, $\varepsilon = 10^{-7}$) and employed categorical cross-entropy for multi-class classification. These settings were consistently applied to all models (CNN, 3CNN-3LSTM, 3CNN-3LSTM-GloVe, DistilBERT, and ALBERT) to guarantee comparability. Experiments were performed utilizing TensorFlow on a system equipped with an Intel i7 processor, 32GB of RAM, and an NVIDIA GTX 1080 GPU.

A learning rate of 0.001 was chosen for its equilibrium between convergence and stability. The Adam optimizer was employed due to its flexible learning rate, which is optimal for deep learning applications. A batch size of 32 achieved a balance between efficiency and performance, while a dropout rate of 0.3 reduced overfitting. The models underwent training for 30 epochs to guarantee effective learning while preventing overfitting.

A flowchart depicting the architecture of each model is included below for clarity. This graphic clearly illustrates the data flow across the models and the interactions among various layers.

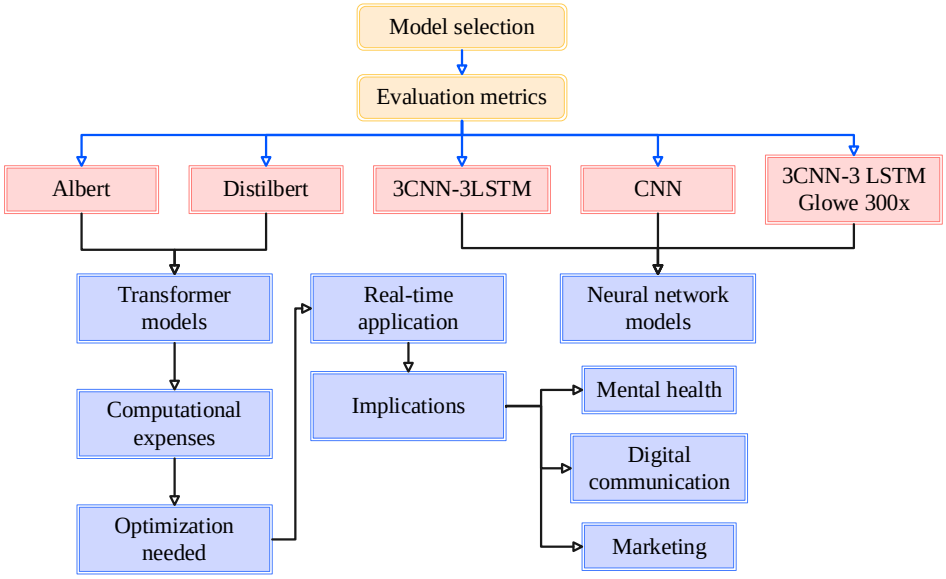

Figure 11: Flowchart Illustrating the Architecture of the Emotion Detection Models

This study employs a hybrid model that integrates a 3CNN-3LSTM architecture with pre-trained Glove embeddings to improve emotion detection efficacy. Glove embeddings were selected for their ability to encapsulate intricate semantic links between words by utilizing global co-occurrence statistics from an extensive corpus, rendering them ideal for comprehending nuanced emotions in tweets. The pre-trained Glove embeddings, possessing a dimension of 300, offer a strong representation of word meanings, hence enhancing the model's ability to interpret the emotional context of tweets. The embeddings were refined during training by permitting the neural network to make minor adjustments to the embedding weights. This fine-tuning allows the model to customize the embeddings for the unique job of emotion recognition, enhancing its capacity to learn from the dataset. Adjusting the embeddings instead of employing them as static representations enables the model to more effectively grasp task-specific subtleties and enhance performance in emotion categorization tasks.

2.3 Method of study

Preprocessing of tweet data, as described in the previous section, yielded a dataset that consisted of 65% training data, 20% test data, and 15% validation data. This further went into the training of both the neural network-based frameworks: CNN, 3CNN-3LSTM, 3CNN-3LSTM-GloVe 300x, and transformer-based frameworks: Distil BERT and ALBERT.

The CNN architecture consists of three convolutional layers with 128, 64, and 32 filters, respectively. These are followed by a global max pooling layer, then a dense layer with 64 neurons for the classification of sentences into one of six classes. Therein, embedding dimensionality, sentence length, and vocabulary size were set to 128, 44, and 3000 words, respectively.

The 3CNN-3LSTM model has implemented three CNN layers of 128, 64, and 32 filters, each followed by three bidirectional LSTM layers containing 64 units in each. The final layer contains a dense layer with 64 neurons meant for classification and further an output layer that contains 6 neurons, corresponding to the six classes. The embedding dimension in this model is kept at 128, the sentence length at 44, and the vocabulary size at 3000.

The architecture of the 3CNN-3LSTM-GloVe 300x model is the same as in the 3CNN-3LSTM model, with three CNN layers of 128, 64, and 32 filters. Then, this model uses three bidirectional LSTM layers consisting of 64 units each. It includes one dense layer with 64 neurons for classification and another layer with 6 neurons to match the number of classes. This model uses a pre-trained Glove 300x weight matrix for the embedding matrix, which results in an embedding dimensionality of 300. Sentence length and vocabulary size are the same as in the other frameworks, set to 44 and 3000 words, respectively.

For the transformer frameworks, Distil BERT and ALBERT, the tokenizers provided with the frameworks were used, meaning the vocabulary size is based on the frameworks' data and is not restricted to this study's data. The default weights for the embedding matrix and embedding dimensions were used. Each transformer model includes a dense layer with 64 neurons for classification and a layer with 6 neurons. Only these two layers were fine-tuned, while the rest of the pre-trained layers were kept frozen to decrease training costs.

The frameworks are assessed utilizing metrics including accuracy, precision, recall, and F1-score. Additionally, the time consumption of the framework's during evaluation is compared. The results of the evaluation are presented in the next chapter.

3 Evaluation results

As mentioned before, the productivity of the frameworks on data classification is assessed through the accuracy,

precision, recall, and F1 score. Fig. 12 compares the productivity of the frameworks according to these metrics.

The assessment of model performance encompasses the frequently utilized metrics of accuracy, precision, recall, and F1-score. These metrics are adequate for multiple reasons: accuracy delivers a comprehensive evaluation of the model's classification capability; precision and recall elucidate the model's efficacy regarding specific emotion categories, especially in cases of data imbalance; and the F1-score acts as a balanced metric for both precision and recall, yielding a more refined comprehension of the model's performance. To address potential class imbalances where some emotions (e.g., "happiness" or "sadness") may be more dominant than others the Matthews correlation coefficient (MCC) is also utilized as a supplementary evaluation tool. The MCC offers a more dependable assessment of performance, particularly for imbalanced datasets, as it takes into account both true and erroneous positives and negatives. MCC scores span from -1 (complete disagreement) to +1 (complete agreement), with 0 signifying no improvement over random classification. The use of MCC facilitates a more comprehensive assessment of the models, especially in datasets where specific classes may predominate.

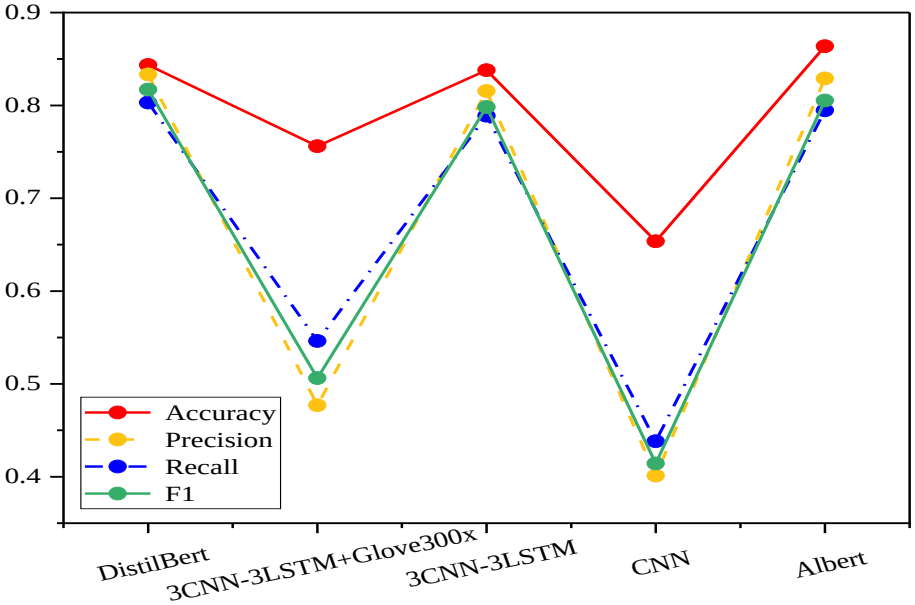

Figure 7: The comparison of the frameworks' performance

According to Fig. 12, the ALBERT displays the highest accuracy, and Distil BERT achieved the highest recall and F1 scores. Besides, the 3CNN-3LSTM shows promising results with high values according to the evaluation metrics. On the other hand, frameworks such as 3CNN-3LSTM-GloVe 300x and CNN show disappointing performance on the classification task. One reason the 3CNN-3LSTM-GloVe 300x model performs poorly is that some words from this study's data may not be included in the matrix, leading to the exclusion of these words and their values during classification.

The hybrid model with Glove embeddings demonstrated significant potential owing to its capacity to incorporate pre-trained semantic knowledge. Nevertheless, it exhibited subpar performance in the trial, chiefly because of the Glove embeddings' inability to adequately encapsulate the subtleties of the emotion-specific language inside the dataset. The embeddings, while proficient in broad language comprehension, failed to accommodate the particular emotional nuances present in tweets, resulting in inadequate performance. This issue underscores the necessity for additional refinement of pre-trained embeddings to more accurately align with the emotional context of the dataset.

Fig. 12 contrasts model performance across four metrics: accuracy (red), precision (yellow), recall (blue), and F1-score (green). The data illustrates the performance variability across different parameters, with accuracy exhibiting the greatest fluctuation. Recall and precision exhibit analogous trends; however, the F1-score reconciles these variations, offering a comprehensive performance assessment.

Table 2 shows the obtained metric values of each model:

Table 2: The obtained values of the frameworks' evaluation

frameworks	**accuracy**	**precision**	**recall**	**f1**
Distil Bert	0.843517	0.833462	0.803066	0.816986
3CNN-3LSTM+Glove300x	0.756063	0.477026	0.546286	0.506241
3CNN-3LSTM	0.83792	0.815483	0.788826	0.798185
CNN	0.653685	0.401201	0.438198	0.414339
Albert	0.863806	0.828999	0.794562	0.805243

The numerical results in Table 2 show the outperformance of the Distil BERT on classifying the textual data, compared to the other studied frameworks. Table 2 presents the evaluation measures (Accuracy, Precision, Recall, and F1-score) for each framework. To enhance the visual clarity and interpretation of these data, the metric labels (e.g., Accuracy, Precision, etc.) have been distinctly highlighted. Moreover, although the numerical values are provided, the incorporation of confidence ranges or error margins is essential for a comprehensive assessment of the data. An error margin of approximately 0.01 is deemed necessary to enhance the comprehension of the uncertainty and dependability of the provided data. This table displays the performance metrics for each model. The notable disparities in precision and recall, particularly for the suboptimal models (CNN and 3CNN-3LSTM-GloVe 300x), arise from their challenges in accurately classifying certain emotions. CNN exhibits low recollection as it fails to recognize a significant number of pertinent events. The 3CNN-3LSTM-GloVe 300x model exhibits deficiencies in precision and recall, attributable to the constraints of the pre-trained GloVe embeddings in encapsulating the complete spectrum of emotion-specific semantic attributes.

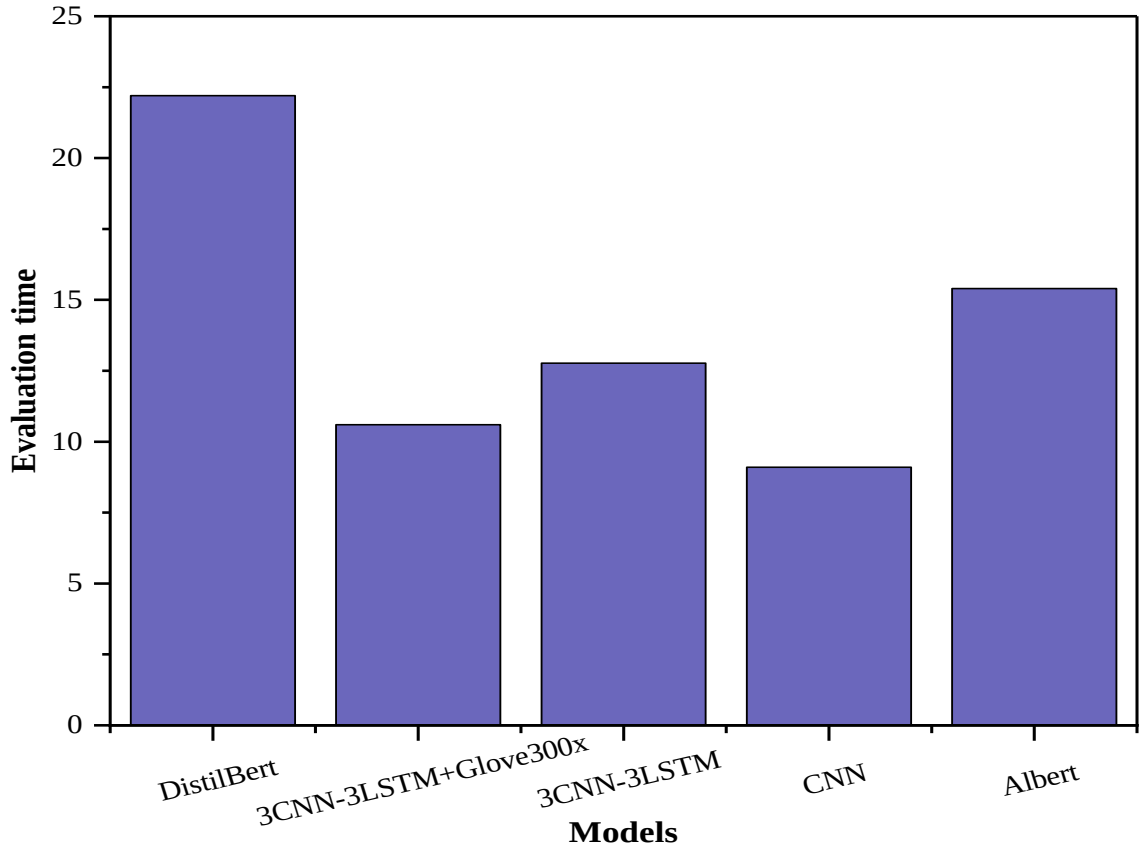

Figure 8: Comparing the time consumption of the frameworks during the evaluation

The frameworks are evaluated 10 times, and the meantime consumed by each framework is presented in Fig. 13. It shows that the transformer model tends to require more time due to its large number of parameters. The 3CNN-3LSTM-GloVe 300x model takes less time than the 3CNN-3LSTM model because the Glove 300x embedding matrix fixes some of the parameters and reduces the overall complexity.

Fig. 13 juxtaposes time expenditure, standardized by accuracy, to elucidate the trade-off between computational expense and performance. The bar chart illustrates the performance of several models, with the tallest bar indicating the model with the highest performance. Transformer models, like as DistilBERT and ALBERT, exhibit slower processing speeds but enhanced accuracy, whereas simpler models like CNN and 3CNN-3LSTM-Glove demonstrate faster performance at the expense of accuracy. This comparison highlights the necessity for optimization to reconcile efficiency and performance, especially for real-time applications.

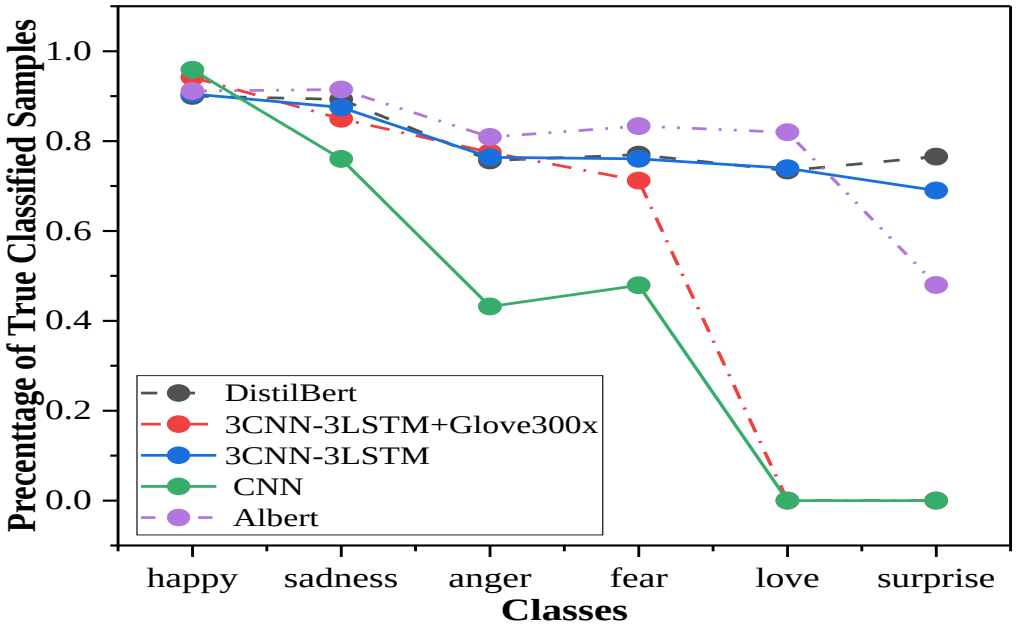

Figure 9: Performance comparison of different frameworks in emotion classification

The provided chart in Fig. 14 displays the productivity of different frameworks in correctly classifying samples across various emotion classes. In general, all frameworks perform well in the "happy" and "sadness" classes, with percentages close to 1.0. Performance tends to drop for more complex emotions such as "fear" and "love." Fig. 14 illustrates the contrasts in the efficacy of five emotion detection methods. DistilBERT and ALBERT demonstrate superior and more consistent performance, whereas the 3CNN-3LSTM + Glove300x and 3CNN-3LSTM models exhibit variability. The CNN model demonstrates the poorest performance, offering a distinct comparison of the models' efficacy.

Distil BERT shows consistently high performance across most classes. ALBERT and 3CNN-3LSTM show relatively high performance in all of the classes, except for the ALBERT model, which is weak in classifying the "surprise" class. Although the frameworks such as CNN and 3CNN-3LSTM-GloVe 300x show acceptable performance in the classes of "happy" and "sadness," they are unable to classify any sample in the classes of "love" and "surprise."

The "happy" and "sad" classes exhibited the highest accuracy among all emotion categories. This is likely due to the imbalance in class distribution, as these two emotions are more prominent in the dataset than others such as "fear" and "surprise." The increased prevalence of "happy" and "sad" tweets enables the algorithms to assimilate these categories, resulting in enhanced categorization performance more effectively. The disparity in class distribution may have influenced the elevated accuracy rates noted for certain classes, and this factor should be taken into account when analyzing the data.

Fig. 15 in the following compares the frameworks based on the ROC curves. The Receiver Operating Characteristic (ROC) curve is a visual tool utilized to appraise the effectiveness of a binary classification framework. It depicts the balance between the true positive rate and the false positive rate across different threshold values.

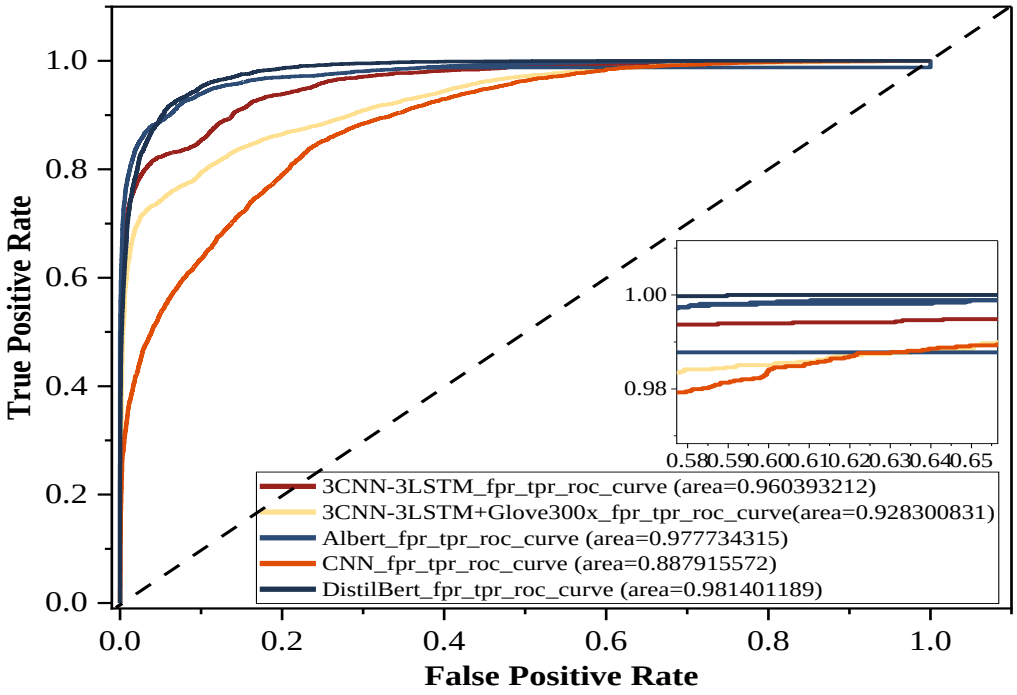

Figure 10: ROC curve analysis of different frameworks

Fig. 15 depicts the ROC curve of the frameworks and the area under the curve (AUC). A higher AUC indicates that the curve is closer to the top-right corner, signifying that the model was able to predict most samples correctly (high true positive rate). As shown, the two transformer frameworks, Distil BERT and ALBERT, exhibit the highest AUC values and outperform the other frameworks studied.

Figure 15 presents the ROC curve analysis, illustrating the trade-off between true positive rate (TPR) and false positive rate (FPR). DistilBERT achieves the highest AUC of 0.9814, followed by ALBERT with an AUC of 0.9777, indicating superior performance in emotion detection. The 3CNN-3LSTM model achieves an AUC of 0.9604, while the 3CNN-3LSTM + GloVe300x and CNN models have lower AUCs of 0.9283 and 0.8879, respectively. These results demonstrate the enhanced effectiveness of transformer-based models. The high AUC values indicate that DistilBERT and ALBERT are well-suited for real-world applications, such as social media emotion detection, marketing, and mental health monitoring, where precise emotion classification is crucial for informed decision-making and effective interventions.

The choice to freeze the majority of the transformer layers and exclusively fine-tune the classification layers was implemented to save computational expenses during training. This method, albeit efficient, probably constrained the model's performance relative to comprehensive fine-tuning, wherein all layers are modified during training. Restricting the transformer layers from modification inhibits the model's capacity to adjust to task-specific characteristics in the lower layers, hence limiting its ability to discern nuanced patterns in the data. Comprehensive fine-tuning would enable the model to adjust both the lower and upper layers, potentially enhancing performance, albeit at a somewhat greater computational expense. The trade-off between computational efficiency and model performance is essential, and a more detailed examination of its effects on findings would enhance the study's conclusions. Future research may investigate the equilibrium between these two elements, examining methodologies such as layer-wise fine-tuning or dynamic freezing tactics to enhance both efficiency and performance.

The Wilcoxon signed-rank test was conducted, as shown in Table 3, to compare the performance of five classification frameworks (ALBERT, DistilBERT, 3CNN-3LSTM, 3CNN-3LSTM+Glove300x, and CNN). Results indicate that ALBERT outperformed both 3CNN-3LSTM+Glove300x and CNN significantly ($p < 0.01$), while its performance difference with 3CNN-3LSTM approached significance ($p = 0.062$). DistilBERT also demonstrated significantly higher performance than CNN ($p < 0.05$) and 3CNN-3LSTM+Glove300x ($p < 0.05$), but did not differ significantly from ALBERT ($p > 0.05$). Overall, ALBERT and DistilBERT showed statistically comparable performance, both consistently outperforming the CNN-based architectures.

Table 3: Statistical analyses result based on Wilcoxon

Comparison	p-value
ALBERT vs DistilBERT	0.687
ALBERT vs 3CNN-3LSTM	0.062
ALBERT vs 3CNN-3LSTM+Glove300x	0.031
ALBERT vs CNN	0.031
DistilBERT vs 3CNN-3LSTM	0.125
DistilBERT vs 3CNN-3LSTM+Glove300x	0.062
DistilBERT vs CNN	0.031
3CNN-3LSTM vs 3CNN-3LSTM+Glove300x	0.25
3CNN-3LSTM vs CNN	0.062
3CNN-3LSTM+Glove300x vs CNN	0.25

3.1 Discussion

Emotion detection in textual data has garnered significant attention in recent years owing to its potential applications across several domains, including mental health, marketing, and social media analysis. Social media platforms, especially Twitter, have emerged as a prolific source of user-generated content that conveys emotions instantaneously. This study evaluates the efficacy of five distinct emotion detection frameworks, encompassing transformer-based models such as DistilBERT and ALBERT, with neural network architectures including CNN and 3CNN-3LSTM. The objective is to evaluate the capacity of these models to categorize emotions, including anger, fear, happiness, love, sadness, and surprise, within Twitter data.

The research employs diverse machine learning frameworks, encompassing pre-trained transformer models (DistilBERT and ALBERT), neural networks (CNN and 3CNN-3LSTM), and a hybrid model that integrates neural networks with pre-trained word embeddings (3CNN-3LSTM-GloVe 300x). The transformer-based models, ALBERT and DistilBERT, leverage pre-trained weights and utilize the WordPiece tokenizer for input data processing. The neural network models, CNN and 3CNN-3LSTM, utilize convolutional layers followed by LSTM layers to extract both spatial and temporal characteristics. The hybrid model incorporates GloVe embeddings to improve feature representation. The preprocessing procedures were the elimination of stop words, punctuation, and non-English characters, followed by the division of the dataset into training, validation, and test sets.

Table 2 demonstrates that the transformer-based models, DistilBERT and ALBERT, surpass the neural network models regarding accuracy, precision, recall, and F1-score. DistilBERT attained the highest recall and F1 scores, whereas ALBERT demonstrated the highest accuracy. The CNN and 3CNN-3LSTM models exhibited commendable performance, especially in identifying the "happy" and "sad" categories, although had difficulties with more intricate emotions like as "fear" and "love." The hybrid 3CNN-3LSTM-GloVe 300x model exhibited inferior performance relative to the other models, presumably due to the constraints of the GloVe embeddings in encapsulating subtle emotional nuances. These findings underscore the benefits of transformer-based models, especially in their ability to discern nuanced emotions in tweet, while also indicating the necessity for additional optimization to mitigate their computing demands. The imbalance in class distribution, with "happy" and "sad" classes being more predominant, presumably influenced the elevated accuracy in those categories.

Although transformer-based models exhibit exceptional performance, their processing requirements pose a considerable constraint for real-time applications, especially in resource-limited settings. The requirement for extensive pre-training on substantial datasets renders these models less accessible for real-time emotional analysis on social media sites. The performance of the hybrid model utilizing Glove embeddings was suboptimal, indicating that fine-tuning pre-trained embeddings is essential for enhancing the model's efficacy. The dataset of the study demonstrates an imbalance in class distribution, potentially affecting the models' performance. The findings indicate that additional study is required to enhance transformer-based models for effective deployment, and the incorporation of hybrid models with fine-tuned embeddings may present a potential avenue for future exploration. This study offers critical insights into the advantages and disadvantages of diverse emotion recognition methods, with substantial implications for applications in sentiment analysis, mental health assessment, and social media analytics.

The results indicate strengths and weaknesses among various emotion categories; however, a more thorough qualitative error analysis is necessary to comprehend the fundamental causes of misclassification. Analyzing misclassified tweets could reveal whether inaccuracies arise from linguistic ambiguity, such as polysemous terms or phrases with multiple meanings, or from conflated emotional categories, where emotions like "anger" and "fear" may exhibit similar expressions. Moreover, model limitations, including challenges in capturing small contextual details, may also lead to inaccuracies. Tweets conveying ambivalent emotions or sarcasm may provide difficulties for the models, as they depend significantly on context to accurately identify the appropriate emotional classification. An exhaustive analysis of these misclassifications would yield significant insights into the domains where the models can be enhanced, aiding in the refinement of the model's capacity to discern emotions in more intricate, real-world social media data.

Transformer-based models in this study exhibit robust performance; nevertheless, their substantial processing requirements impede real-time implementation. To tackle this issue, optimization strategies such as model pruning, which diminishes size by eliminating less significant weights, quantization to reduce memory consumption, and knowledge distillation, wherein a smaller model is trained to emulate a bigger one, can be investigated. In addition to DistilBERT, subsequent progress in knowledge distillation may produce more efficient models without compromising performance. These techniques will enhance the computational efficiency of transformer models, rendering them more appropriate for implementation in resource-limited settings such as mobile applications and edge devices.

4 Conclusion

This work performed a holistic assessment of five varied frameworks for Emotion detection in tweets through categorization into six emotions, such as anger, fear, happiness, love, sadness, and surprise. The different metrics used to assess the performance of the frameworks were accuracy, precision, recall, and F1 score, among others, with a view to understanding the trade-offs inherent in the application of the frameworks.

These two pre-trained transformer-based frameworks, Distil BERT and ALBERT, turned out to be the best in all

the metrics of accuracy, precision, recall, and F1-score. The efficiency of these two models regarding the right classification of tweets into the six categories of emotions is due to large-scale pre-training on big tweet corpora, which helps to learn subtle features and contextual information with the help of their advanced attention mechanisms. Whereas all other frameworks went quite fast, Distil BERT and ALBERT took some more time to evaluate, as transformer architectures have too many parameters and computational complexity.

Among all the non-transformer frameworks, the most effective architecture was that of the 3CNN-3LSTM model, which presented acceptable results in the light of accuracy and F1 score, though not comparable with the actual results of the transformer frameworks. On the other hand, while being simpler and less resource-intensive, the CNN model yielded relatively low performance. A hybrid model, 3CNN-3LSTM-GloVe 300x, which uses pre-trained GloVe embeddings, has also been implemented; however, it did not give better performance in the experiment as expected. While pre-trained word embeddings had some promising advantages, they failed to achieve the high accuracy and F1 scores of the transformer frameworks and those of the 3CNN-3LSTM model. This suggests that using pre-trained embeddings in neural network frameworks often requires extensive tuning and may not always outperform pre-trained transformer frameworks.

This review tends to focus on how transformer-based models stay ahead in ED tasks due to their accuracy and robustness. It suggests that future research efforts lie in further fine-tuning this kind of framework toward size and time efficiency, considering hybrid frameworks that balance performance and efficiency, and fine-tune them on the task and dataset level.

In conclusion, this study offers valuable insights into the strengths and weaknesses of different frameworks for ED in tweets. Transformer-based frameworks demonstrate significant potential due to their high performance, but their high computational demands underscore the need for ongoing innovation in model optimization and integration. This study illustrates the prospective applications of emotion detection algorithms in marketing and mental health. Emotion detection in marketing allows organizations to customize advertising campaigns that align with consumers' emotional states, hence improving engagement and customer happiness. Emotion detection in mental health can function as a mechanism for assessing emotional well-being, enabling mental health professionals to recognize fluctuations in emotions, such as heightened grief or rage, and to intervene promptly. These findings connect research with practical applications, offering actionable information that might enhance individualized marketing efforts and facilitate mental health monitoring.

Abbreviation

Abbreviation	Description	Abbreviation	Description
CNN	Convolutional Neural Network	TF	term frequency
BERT	Bidirectional Encoder Representations from Transformers	IDF	Inverse document frequency
ALBERT	A Lite BERT	NSP	Next Sentence Prediction
Distil BERT	Distilled BERT	SOP	Sentence-Order Prediction
LSTM	Long Short-Term Memory	AUC	Area Under the Curve
Glove	Global Vectors for Word Representation	ROC	Receiver Operating Characteristic
NLP	Natural Language Processing	TBED	Text-Based Emotion Detection
ED	Emotion Detection	GPEL	General-Purpose Emotion Lexicons
UMM	unigram mixture model	RNN	Recurrent Neural Network
SLDA	supervised Latent Dirichlet Allocation	DLSTA	DL Assisted Semantic Text Analysis

Authorship contribution statement

Weiguo Huang: Writing-Original draft preparation, Conceptualization, Supervision, Project administration.
Chen Zhang: Methodology, Software
Bin Zhang: Validation

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author statement

All the authors have read and approved the manuscript. As stated earlier in this document, the requirements for authorship have been met, and each author believes that the manuscript represents honest work.

Ethical approval

All authors have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

References

[1] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, “Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches.” [Online]. Available:

https://www.researchgate.net/publication/348740926

[2] L. Mathew and V. R. Bindu, "Efficient transformer-based sentiment classification models," *Informatica*, vol. 46, no. 8, 2022.

[3] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," Jul. 01, 2020, *John Wiley and Sons Inc*. doi: 10.1002/eng2.12189.

[4] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "Emotion Detection of Textual Data: An Interdisciplinary Survey," in *2021 IEEE World AI IoT Congress, AIIoT 2021*, Institute of Electrical and Electronics Engineers Inc., May 2021, pp. 255–261. doi: 10.1109/AIIoT52608.2021.9454192.

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

[6] D. Seal, U. K. Roy, and R. Basak, "Sentence-Level Emotion Detection from Text Based on Semantic Rules," in *Advances in Intelligent Systems and Computing*, vol. 933, Springer Verlag, 2020, pp. 423–430. doi: 10.1007/978-981-13-7166-0_42.

[7] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi, and S. K. Shahzad, "Emotion Detection of Contextual Text using Deep learning," in *4th International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2020 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ISMSIT50672.2020.9255279.

[8] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics (Switzerland)*, vol. 9, no. 3, Mar. 2020, doi: 10.3390/electronics9030483.

[9] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," Jul. 01, 2023, *Elsevier B.V.* doi: 10.1016/j.inffus.2023.02.028.

[10] A. R. Murthy and K. M. Anil Kumar, "A Review of Different Approaches for Detecting Emotion from Text," *IOP Conf Ser Mater Sci Eng*, vol. 1110, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1757-899x/1110/1/012009.

[11] S. Madhuri and S. V. Lakshmi, "Detecting emotion from natural language text using hybrid and NLP pre-trained models," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 10, pp. 4095–4103, 2021.

[12] H. F. Shahzad, A. A. Saleem, A. Ahmed, K. S. H. Ur, and R. Siddiqui, "A review on physiological signal-based emotion detection," *Annals of Emerging Technologies in Computing (AETiC)*, vol. 5, no. 3, 2021.

[13] N. Khan, A. Singh, and R. Agrawal, "Enhancing feature extraction technique through spatial deep learning model for facial emotion detection," *Annals of Emerging Technologies in Computing (AETiC)*, vol. 7, no. 2, pp. 9–22, 2023.

[14] A. Koufakou, E. Nieves, and J. Peller, "Towards a new benchmark for emotion detection in NLP: A unifying framework of recent corpora," in *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, 2024, pp. 196–206.

[15] F. M. Plaza-del-Arco, A. Curry, A. C. Curry, and D. Hovy, "Emotion analysis in NLP: Trends, gaps and roadmap for future directions," *arXiv preprint arXiv:2403.01222*, 2024.

[16] Z. Li, W. Yang, S. Peng, and F. Liu, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," 2004.

[17] J. Gu *et al.*, "Recent Advances in Convolutional Neural Networks," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.07108

[18] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.09586

[19] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," Jul. 01, 2019, *MIT Press Journals*. doi: 10.1162/neco_a_01199.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[21] Koroteev MV, "BERT: A Review of Applications in Natural Language Processing and Understanding."

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.11942

[23] M. Ryu, "[RE] ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." [Online]. Available: https://huggingface.co/

[24] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[25] P. Bambroo and A. Awasthi, "LegalDB: Long distilbert for legal document classification," in *Proceedings of the 2021 1st International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies, ICAECT 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAECT49130.2021.9392558.

[26] L. Dipietro, A. M. Sabatini, and P. Dario, "A survey of glove-based systems and their applications," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*,

vol. 38, no. 4, pp. 461–482, Jul. 2008, doi: 10.1109/TSMCC.2008.923862.

[27] D. J. Sturman and D. Zeltzer, “A Survey of glove-based input,” 1994.

[28] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation.” [Online]. Available: http://nlp.