Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 8, 2026: 3031-3045
ISSN:2302-8149
e-ISSN:2540-9719

LGI: Label-Graph Inference Multi-Benchmark Emotion Detection Transformer Framework for Intelligent Chatbots

[1]Karam Abdullah*
[1]Department of Computer Science, College of Education for Pure Science, University of Mosul, 41002, Mosul, Nineveh, Iraq
*e-mail: karamalnuaymi@uomosul.edu.iq

(***received***: 28 June 2026, ***revised***: 27 July 2028, ***accepted***: 28 July 2026)

Abstract

The fundamental ability of intelligent chatbots is to detecting emotions evident in intelligent recommendation systems, learning agent systems and mental health systems. Emotions are a key element in facilitating interaction and communication between humans. Researchers face challenges in studying the transmission of emotions through text. Despite significant advances in language transformation models, emotion recognition remains a major challenge due to the high semantic overlap between emotion categories and unbalanced distributions in datasets. This paper suggests an improved framework for enhanced transformer model based on a Label-Graph Interface (LGI) for ratings to robustly detect emotions across texts using various criteria specific to social media and conversations between humans and human-machine. It combines a pretrained transformer encoder with lightweight label relation inference layer that uses a row normalized graph prior and sample adaptive gate to propagate evidence among related emotion labels. In single label case, cross entropy loss is used to optimize the model, but in multi label detection a binary cross entropy is used with logits, class aware positive weighting and threshold selection based on validation sets. Six datasets were used in order to assess the framework: DAIR Emotion, achieved an macro-F1 to 89.10%. TweetEval Emotion is the second dataset reached an macro-F1 is 81.56%. DailyDialog, the macro-F1 arrived to 58.62%. GoEmotions, got micro-F1 about 60.03%. Empathetic Context dataset reached to 85.80% Top-3 accuracy and macro-F1 is 85.69% on the 32-class task. Lastly EmoWOZ is the most important dataset because it contains real conversation between human-machine has an accuracy about 92.86%, macro-F1 to 59.95% and AUC to 94.97%. The results demonstrate that the significant value and usefulness of relational inference at the classification level for emotion-based chatbot systems.

Keywords: *intelligent chatbot, emotion detection, label-graph inference, transformer encoder, natural language processing.*

1 Introduction

The significant development of the Internet with artificial intelligence techniques has contributed to increased risks arising from the erroneous use of social networking applications and chat robots. We refer to it as having a positive or negative impact on the individual or society. Hence the need to interpret and analyze the feelings of the texts in order to ensure a clear and correct understanding of the interaction between human beings and machines. Emotions are vital in facilitating interaction and communication among individuals. Researchers face an obstacle in studying the transmission of emotions through textual texts. Emotions from the text will be used to determine the interaction between humans and the computer that controls communication and other different aspects. The processing of natural languages (NLP) combines language and computer techniques to enable computers to understand human languages and improve human-computer interaction [1]. Recently, pre-trained transformer models have been heavily relied upon in natural language processing, especially in BERT-based structures, to capture contextual information and improve text-based emotion detection [2]. Feelings analysis is important in neuro-linguistic programming that extracts information from different types of text, such as automatic chat programs, web sites and reviews and classifies it based on its polarity as positive, negative or neutral. Emotion analysis is an area specialized in the exploration of an opinion that focuses on extracting feelings and views from textual statements organized, semi-organized or unorganized in relation to a particular subject matter [3].

ISSN:2302-8149
e-ISSN:2540-9719

Textual data are sequenced in nature, which means that the order and interdependence of words within the phrase, known as context, it is essential to understand the full meaning of the text. Conventional unsupervised machine learning approaches not only disregard the order or arrangement of words in documents, but they are also constrained by a small and very restricted input size. Therefore, the justification for using comprehensive computational techniques to examine textual material.

It is important to note that datasets classifications vary from one to another depending on the criteria that used. Some datasets consist of four or six classes and others more like in GoEmotions there are (27-emotions plus a neutral category) in a multi- classification environment, while in Empathetic Context, there are 32 emotions. Therefore, the model design must handle all general and specific classifications.

It is worth noting that pretrained transformer-based language model such as RoBERT and BERT have improved natural language processing understanding in general by learning contextual representation in large datasets [4],[5]. More specifically, model in particular domain offer performance improvements, such as BERTweet on Twitter-like dataset, by adapting pretraining to social communication languages [6]. One of the most significant challenges facing classifiers is the strong overlap between emotions such as annoyance, anger and disgust in the text. Ignoring these relationships renders the final classification useless and reflects the overall weakness of the systems. However, there are three main gaps require further research and resolution. First, previous studies have focused on working with and optimizing a single dataset. Second, many studies implicitly leave label relation in the classifier header, despite emotion classifications having a semantic structure. Lastly, there is a lack of variation in the metrics used to evaluate datasets. Therefore, this research addresses these gaps by proposing a unified framework and evaluating it using consistent metric packages. To solve this problem, this paper presents an improved framework for transformer model based on LGI for classifications as shown in Figure 1. From this standpoint, lightweight layer was added to connect the classifications above the head of the transformer classifier. This layer uses a classification diagram that collect prior semantic knowledge. The basic logarithms are modified through a special gate that controls the extent of the influence of each sample across the graph. This design maintains the advantages of pre-trained transformers without the need for a complex neural network to process conversation data.

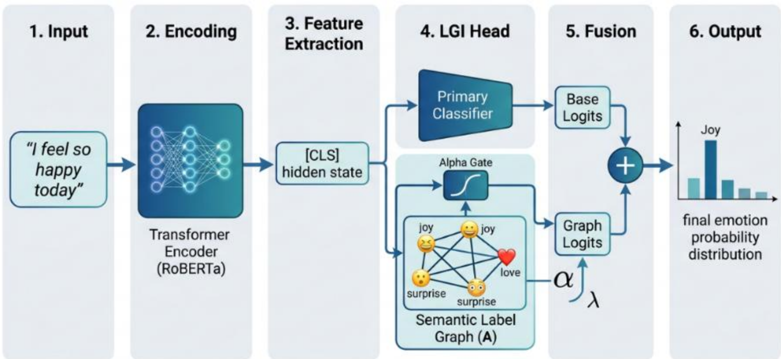

Figure 1 Framework for transformer model based on label-graph inference

The framework was evaluated using six difficult and different levels of datasets. Single-label classifications such as DAIR Emotion and TweetEval Emotion were used for conversations and tweets. DailyDialog was used for conversational speech classification, with a strong predominance of the neutral emotion category. For multi-label emotions, Go Emotions was used. Empathetic Context also played significant role in the evaluation, comprising 32 emotion categories. The EmoWOZ introduced the final and the most important decision in the success of the work, due to its inclusion of real human-machine dialogues through seven important emotional categories. This multi-criteria evaluation of the six datasets is important because a model that performs well on single dataset may fail to generalize across emotion classifications and data patterns, therefore, several datasets were adopted.

The novelty of this study in its lightweight LGI layer, where transformer models are enhanced through a learnable label-graph and contextual gate adapted to the sample. This enables the modeling

 ISSN:2302-8149
e-ISSN:2540-9719

of emotion label relations within the context without modifying the pre-trained transformer structure or incurring significant computational cost.

The main contributions of this study:

- A lightweight LGI layer that model's relation between emotions classes.
- Adaptive contextual gate mechanism to control the contribution of label graph.
- A unified framework that supports single-label and multi-label emotions detection tasks.
- Six benchmark datasets that comprehensive evaluating the framework.

2 Literature Review

Many words may change in meaning over time, and this increases the difficulty of determining a unified model for identifying emotions from texts or chatbots [7]. For this reason, time must be taken into account in the process of training different networks, the historical progress of the emergence of sentiment analysis and emotion recognition algorithms must also be studied.

Evolution of text emotion detection: The development of text emotion detection, alternatively referred to as "sentiment analysis", has witnessed notable advancements throughout the decades. The first attempts to discover emotions from text began in the fifties and sixties of the last century by adopting certain rules, such as searching for a specific word or phrase that indicates to specific emotion. Emotion-based dictionaries were used in the 1990s that include phrases and words associated with specific emotion, lists of words with degrees of emotion were created to identify emotions from the text [8]. Sentiment analysis has developed significantly with the beginning of machine learning in the current century, where sentiments have been classified into negative, positive, and neutral by using SVM and Naïve Bayes. Deep learning, convolutional artificial neural networks, recurrent neural networks, and pre-trained models have played the greatest role in developing sentiment analysis to the form we currently see of accuracy, superiority and realism. Many researchers have published articles on the development of sentiment analysis. Mäntylä presented a research review of more than 6,996 review papers from Scopus. It was found that studies on sentiment analysis that the beginning was about analyzing public opinion at the beginning of the twentieth century. He added that the subjective analysis of the text occurred in the nineties. He pointed out that the emergence of computer-based sentiment analysis occurred with the emergence of subjective texts on the Internet. Therefore, after the year 2004 was considered an important date for the publication of more than 99% of research papers in this field stated that in recent years. They illustrate the exponential expansion of modern sentiment analysis, which has increased by a factor of 50 over a span of ten years from 2005 to 2016. The quantity of citations has increased in tandem with the number of papers [9].

Emotion detection: It is considered one of the most important factors that help a person make right decisions, understand the emotions of others, and provide feedback in different situations [10]. This field has diverse uses in fields such as medical treatment, interaction between humans and computers, marketing, and entertainment. There are different types of emotion analysis, the most important of which is facial emotion analysis, this system has the capability to perceive and analyze face attributes, including eyebrow movements, eye direction, and mouth configuration, in order to recognize emotions such as joy, sorrow, rage, and astonishment [11]. Another type of sentiment analysis is voice and speech analysis, emotions can be identified by analyzing vocal characteristics, including tone of voice, pitch, speed, and strength. Sound recognition technology is frequently employed to analyze these vocal cues and deduce the speaker's emotional status [12].

Emotion detection from the text: It is important in dealing with the natural language, which requires finding and extracting the feelings conveyed in the written text. This is potentially useful in a variety of areas, including emotional analysis, feedback studies and mental health surveillance. Blogs and social media sites such as Facebook, Tweeter and Instagram provide huge amounts of data that reflect user feelings in different aspects of their daily lives, and thus the task of capturing these feelings will be a very difficult one. In [13] Gafoor et al., presented an advanced potential model called the "textual emotional recognition in multidimensional space" (TERMS). This model accurately represents the individual limits of emotions and corresponding data contained in the content, allowing for accurate and reliable identification of emotions. In [14], Yee et al. Suggested a way to identify feelings in the text using the ALBERT-Bi-LSTM network and SVM-NB classifier.

ISSN:2302-8149
e-ISSN:2540-9719

Initially, the data is processed using "Albert pre-training module". Thereafter, Bi-LSTM network is used to obtain vector-related data, which is then converted into an appropriate distinctive coordination of classification. The aim of this process is to enhance the accuracy of the identification of emotions. Ultimately, the label is used to obtain emotional electrodes of the text, which include positive and negative classifications. Negative feelings can be classified into three subcategories: anger, sadness and disgust. In [15], Rot et al. Use each of the non-controlled methodologies and supervise the various data sets. An unsupervised method is used to determine the feelings of tweeting that are automatically obtained from Twitter, which is generally accessible. Different automated learning methods were used, including Multinomial Naive Bayes (MNB), Support Vector Machines, and Maximum Entropy, to identify tweeting feelings and assess the effectiveness of different sets of features. In [16], Song et al. The web crawler technique, the six basic emotional models of Ekman, and the well-defined deep learning model BERT were used to conduct emotional analysis on Weibo, the largest social networking site in China. The use of BERT for in-depth learning has led to a rate of sensitivity of 88.50 percent when applied to a large set of social media texts. In [17], Li et al. presented a user-sensitive model aimed at analyzing the emotional aspects of public opinion events in small blogs. The compilation and preparation of the text of the commentary on public opinion events on small blogs leads to three distinct categories of incentive texts: "Fear", "Rage" and "Sadness". Subsequently, a new algorithm was developed to extract emotional distinctive words through the use of the “linear discriminant analysis” model (LDA), the emotional dictionary, and manual explanatory commentary. The inspired text that was picked up is converted into a word conveyor using Word2Vec. To complete the classification of emotions, the important characteristics of the text are extracted through the use of bidirectional long-term memories (Bi-LSTM) and telepathic convolutional neural networks (CNN) to collect long-range delusional data. Test results indicate that the value of F1 for six automated learning models increased on average by 3.66 percent, while the value of F1 for seven deep learning models on average increased by 1.84 percent.

Emotion detection from intelligent chatbots: In [18], [19] and [20], authors describe a brief history and infrastructure for the development of chatbots. The main objective of the chatbot is to make requests used in the form of text or voice entry. Chatbots are used as an alternative for persons in many civilian tasks, such as airports, police stations, hospitals and supermarkets, to provide advice, technical support and answer various questions. Taking advantage of the emotional situation in this area will increase the accuracy of the Chatbot. Emotional intelligence is a growing direction in the development of automated chatbots, with a focus on creating sympathetic agents and emotional intelligence capable of discovering user feelings and generating appropriate responses. Emotions greatly affect conversation, and chatbots that simulate human behavior can increase relationship, motivation and participation. Researchers are exploring ways to improve these capabilities to better engage users [21]. In [22], Ehteshma et al. Researched the process of generating emotions during talks that serve a wide range of purposes, using the theory of cognitive valuation. Authors present a methodology for calculating useful factors and identifying six basic feelings. Emo-Bot, an emotional chatbot, examines sound and text inputs, produces appropriate emotions, and provides corresponding responses. Empirical assessments prove that Emo-bot produces more accurate and reliable responses. In [23], Kozowski et al. introduced a project that aims to enhance the identification of mood and emotion in polish script by using a driven data dilation technique and larger NL models. The purpose of this technique is to develop a therapeutic dialogue system that examines the mood and feeling expressed in speech The language model is enhanced by substituting BERT model with "RoBERTa", and the emotion in the text bulk is magnified using a distinctive method inspired by semi-supervised learning approaches. Tera-bot, a conversation system with sympathetic capabilities, utilizes cognitive-behavioural therapy techniques to assist mental patients. Lee et al.'s study [24] found that emotional expression of chatbots significantly influences individuals' intentions to engage in climate change mitigation behavior. The study involved 577 American individuals and found that emotional expression led to greater intentions than factual information and reduced social presence.

3 Research Method

This work aims to build an intelligent system capable of inferring emotions from English-language conversational texts using deep learning models, as show in Figure 2, to support chatbots

ISSN:2302-8149
e-ISSN:2540-9719

with adaptive emotional responses. The methodology follows a series of integrated steps, beginning with text preprocessing and tokenization, transformer-based contextual encoding, a standard classification head that produces base logits, and a lightweight LGI layer that refines the logits according to emotion label relations. The intelligent system for detecting emotions in chat messages is developed using pre-trained deep learning models, with the goal of proving automatic responses consistent with the detected emotion.

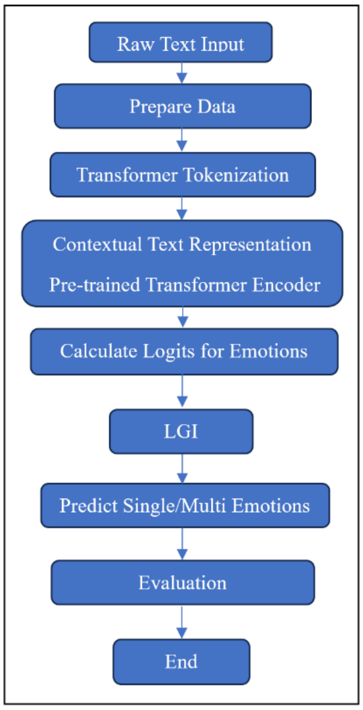

Figure 2 General workflow of the proposed LGI emotion detection system

3.1 Data preparation

A dataset of short text phrases classified according to the n emotions was prepared $L = \{y_1, y_2, \ldots, y_n\}$. Each text sample was represented as a pair (x, y), where x is the text and y $\in L$ is the correct sentiment classification. This data was used to evaluate the model's performance. The data was divided into three parts: train, validation and test. The data preparation phase begins by loading the dataset and then examining the label field to determine whether the task is multi-label or single-label. In the case of multi-labeling, the list of labels for each sample is converted into a fixed-length multi-vector that covers all classes. In the single case, the class indexes and names are taken directly from the data schema. From the training data, two supervisory aids are extracted: the label matrix "*A-init*" after it has been normalized and its main diagonal zeroed by initializing the learnable graph, and the positive weight vector 'pos-weight' calculated from the positive frequencies to address the problem of imbalance between classes.

3.2 Text processing

Text preprocessing is done by a fast-Hugging Face tokenizer associated with the selected transformer. Every input utterance is tokenized into "*input_ids*" and "*attention_mask*", then clipped to a fixed maximum length of 128 tokens to avoid overflow and padded such that all sequences in one batch have the same length. These fields are returned as "*PyTorch*" tensors; everything else is discarded, so the model consistently sees just a couple of tensors for every example. The model has been set up to return hidden states, so after a forward pass, the representation of the first token "*CLS*" from the last layer is available; that vector is fed into a lightweight linear head, which produces a single scalar α conditioned on the context, for the current input. Using this α, the final logits are offset inside the learnable label-graph inference (LGI) step, and the same tokenization settings are used at inference time, so the inputs of chats are encoded the same as during training.

ISSN:2302-8149
e-ISSN:2540-9719

3.3 Model Architecture

The proposed framework presents the single / multi-label emotion detection, that augments a pre-trained transformer by incorporating a learnable contextual gate model for LGI mechanisms. Given an input sentence and an inventory of L emotion labels, a transformer encoder RoBERT produces last-layer hidden states and a pooled representation $h_{cls} \in R^H$. A linear head maps h_{cls} to initial logits $z \in R^L$ via the Equation 1, as illustrated in Figure 3.

$$z = Wh_{cls} + b \tag{1}$$

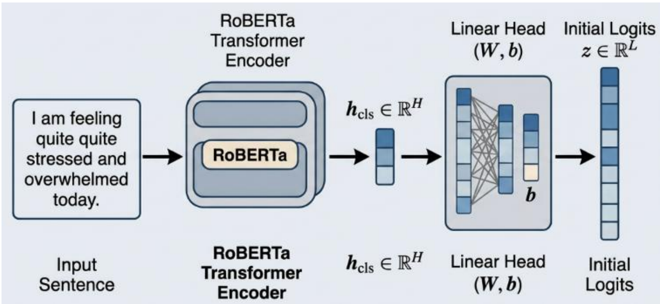

Figure 3. Transformer encoder and initial classification

RoBERTa provide a strong contextual representation and can be fine-tuned across human-machine, short social text and conversational utterances. Using RoBERTa in all datasets also ensure performance difference were associated with LGI not related to change in the backbone model.

In single-label datasets, logits are normalized by softmax during evaluation, also in multi-label dataset, each logit is passed through a sigmoid function and then compared it with specific label thresholds on the validation set. The approach relaxes that assumption by injecting an explicit learnable structure over labels directly at the logit level in a manner that is both end-to-end trainable and sample-adaptive.

The core addition is the LGI layer which is a square matrix $A \in R^{L \times L}$ that models' affinities between labels. A initialized from training co-occurrence statistics to provide an informative prior: letting $Y \in \{0,1\}^{N \times L}$ denote the multi-hot label matrix, we define $C = Y^{\top}Y$, zero its diagonal, and compute a degree-normalized adjacency by the Equation 2.

$$A_0 = D^{-1/2}\, CD^{-1/2} \tag{2}$$

where D is the sum of matrix diagonal.

The parameter A is then optimized jointly with the encoder. For stability and interpretability at each forward pass, we symmetrize and clip to obtain the effective graph as in Equation 3.

$$A\mathrm{eff} = clip((A + A^{\mathrm{T}})/2, 0, 1) \tag{3}$$

and we zero the diagonal at inference to prevent self-reinforcement. This yields a clear semantics in which large entries of Aeff capture learned co-activation relations after controlling for textual content, as illustrated in Figure 4.

ISSN:2302-8149
e-ISSN:2540-9719

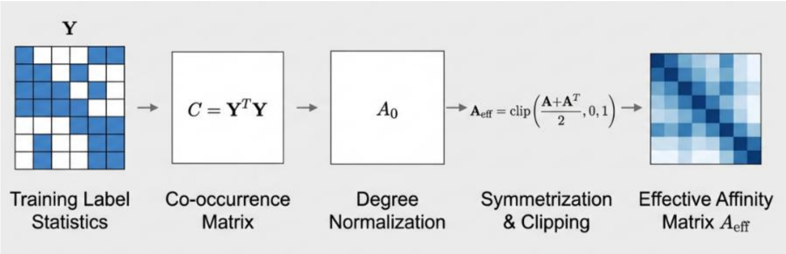

Figure 4 Label graph interaction layer

To avoid a uniform, context-agnostic propagation, we compute a per-example scalar gate α(x)∈ (0,1) from h_{cls} using a one-layer head and a sigmoid. The final logits are defined by a residual update in Equation 4, see Figure 5.

$$\acute{z} = z + \alpha(x)(z\, A_{eff}) \tag{4}$$

Which retains the evidence from the encoder directly but adds a content dependent correction that incorporates label dependencies. When the sentence context is in accordance with the learned correlations, $\alpha(x)$ increases and the model enhances complementary emotions; when context tends to convey unusual pairs, the gate suppresses the pass and dampens spurious co-activations.

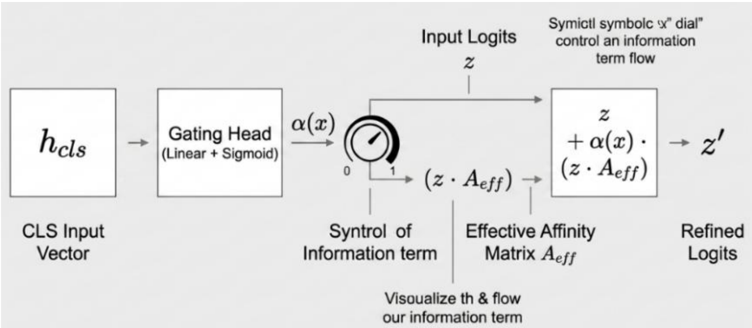

Figure 5 Contextual gating mechanism

Training minimizes an asymmetric binary cross-entropy on z′ that is tailored to class imbalance and error asymmetry in multi-label settings. With π= σ (z′) and the ground truth y ∈ {0,1} L, we reweight positives by β inverse to label frequency and apply an asymmetric focal factor so that negatives contribute less. All parameters (encoder, classifier, gate, and graph) are optimized end-to-end using "*AdamW*" with early stopping on validation F1. After training, probabilities are π=σ (z′), and because labels differ in base rates and separability, we calibrate per-label thresholds on the validation split by grid search. The learned artifacts (graph, gate, thresholds) are saved for reproducible deployment.

The LGI-step involves a single matrix-vector multiplication z*A*eff with a complexity of O(L2) per example, which is negligible for typical emotion inventories, and sparsity reduces this cost further in practice. Importantly, operating at the logit level avoids mixing the encoder with graph layers, keeping the representation numerical stable and preserving the inductive bias of the pre-trained model as well. Empirically, this residual, context-gated propagation leads to improved recall for underrepresented emotions without sacrificing precision for dominant ones, suggesting that predictive structure is captured, rather than dataset artifacts.

In all, the approach equips a powerful pre-trained encoder with a principled, learnable mechanism for modeling label dependencies that is adaptively controlled by sentence context. The key novelty-sparse, end-to-end-learned label graph gated per example and applied as a residual at the logit level-consistently improves multi-label emotion recognition without adding much extra overhead to keep the system lightweight and interpretable, and easy to deploy.

ISSN:2302-8149
e-ISSN:2540-9719

4 Results and Analysis

The experimental evaluation of the proposed LGI is discussed in this section. This methodology assesses through extensive experiments conducted on six challenging datasets that covering multi-label, single-label, social- media, human-machine conversation and fine-grained emotion recognition settings.

4.1 Dataset

The huge and relevant datasets contribute to raising the level of intelligence of the chatbot. It plays a crucial role in shaping the chatbot, as well as raising the level of understanding and response to users and continuously developing its capabilities through natural language processing algorithms, machine learning, and deep learning.

Detecting emotions in chatbot texts is extremely important, however there is only few datasets like "DAIR Emotion dataset" [25], it is a single label emotion classification dataset that contain Twitter messages with six emotion categories like joy, fear, anger, sadness, surprise and love. It's widely used and provide relatively balance benchmark for emotion classification in social text. The second dataset is TweetEval Emotion [26], it is subset derived from SemEval-2018 Task 1 [27]. In this dataset four emotion categories were used: joy, anger, sadness and optimism. it is also a single label emotion classification dataset that used to evaluate the strength of the proposed approach on noisy, short and informal social text. The third dataset is DailyDialog introduced in [28], it human-written conversation and wide scope of daily life topics that reflect everyday communication. The dataset was annotated manually with emotion information and dialogue action. It has seven emption classes: anger, fear, sadness, no emotion, disgust, surprise and happiness. It evaluated the model under conversational conditions. The fourth dataset is GoEmotions Dataset [29], it's a fine grained, large range, multi label emotion. It has (27-emotions plus a neutral category) contain English Reddit comments. This dataset enables association one text to multiple emotion labels, that handle multi label, label imbalance and semantic overlap emotion. Therefor it very suitable and challenging dataset for evaluating the proposed method. The fifth dataset is Empathetic Contexts Dataset [30], it's an emotional circumstance version return to the EmpatheticDialogues dataset that have a 32-emotion label, a context- based version was used to evaluate the proposed framework using a single and highly accurate classification. The last is EmoWOZ [31], it is a large dataset designed for emotion detection with service-oriented procedure, such as hospital and hotel booking systems, or restaurant and shopping searchers. The most important feature of EmoWOZ dataset is its inclusion of real human-machine dialogues. It is an open source that contain more than 83,000 user phrases divided in to 7-emotional categories: Neutral, satisfied, fearful, exited, dissatisfied, abusive and apologetic. This study evaluates the proposed approach through the following six datasets in details as in Table 1.

Table 1 Show the important details and differences between the six datasets

Dataset	Size	Task	Lab.	Text domain	Challenge	Metrics
DAIR Emotion [25]	Train 16k Val 2k Test 2k	Single-label	6	Twitter-style English messages	General emotion classification	Accuracy, macro-F1, AUC
TweetEval Emotion [26]	Train 3.26k Val 374 Test 1.42k	Single-label	4	Tweets	Short/noisy social-media text	macro-F1
DailyDialog [28]	Train 87.1k Val 8k Test 7.7k	Single-label	7	Conversation utterances	Neutral/no-emotion dominance	macro-F1, micro-F1
GoEmotions [29]	Train 43.4k Val 5.43k Test 5.43k	Multi-label	28	Reddit comments	Multi-label rare emotions	macro-F1
Empathetic Context [30]	Train 76.6k Val 12k Test 10.9k	Single-label	32	Emotional situation/dialogue contexts	Fine-grained semantic overlap	Accuracy, Top-3

ISSN:2302-8149
e-ISSN:2540-9719

EmoWOZ [31]	Train 66.4k Val 8.5k Test 8.6k	Single-label	7	Contain chatbot conversation	Imbalanced	macro-F1, micro-F1

4.2 Experimental Setup

All experiments conducted on ASUS laptop TUF F15 with an Intel core i7 -11370H processor and 40GB of RAM. The platform was Windows 11 and python 3.10 with spyder IDE. The GPU was Nvidia RTX 3070 with 8GB DDR6. Tokenization was performed using the fast-Hugging Face tokenizer with maximum sequence length was set to 128 tokens to ensure a consistent input and reduce memory consumption. The main hyperparameters were kept consistent across datasets, the batch size was set to 4,8 for training and evaluation respectively, with 100 epochs and early stop patience equal 2, random seed is 42, learning rate was 1×10-5, with AdamW optimizer. six different datasets were used to ensure broad data representation and to conduct an objective and efficient evaluation of the framework.

4.3 Evaluation

The proposed framework yielded promising results through comparisons with the state-of-the-art and its practical application to the six aforementioned datasets, which gave the proposed methodology a comprehensive and credible character through real, consistent and approved measurements in this section, it should be noted that the standard metrics used in single-label classification are accuracy that calculated by matching samples that have correct predictions to the ground-truth class, F1 micro, F1 macro and AUC-macro that consider more informative in imbalance classes. For multi-label classification, F1-micro, F1-macro are used because label-wise accuracy can be misleading when many labels are absent. The label-wise accuracy in multi-label classification calculated by representing each text with a binary vector consisting of 28 labels of emotions in GoEmotions dataset after that independently converting the model logits in to probabilities using sigmoid function, and then converting each probability into a binary prediction using specific threshold for validation.

4.3.1 Quantitative Results

After conducting practical experiments on the six datasets as shown in Table 2, it was found that in the single-label task the accuracy matches the F1-micro always due to all false positive come upon false negative cross the all samples. It is also noticeable that the gap between F1-micro and F1-macro scales is due to an imbalance in the datasets. Overall, the proposed framework achieved success on the DAIR dataset where the model obtains an accuracy of 93.20%, a F1-micro of 93.20%, a F1 macro of 89.10% and macro-AUC of 99.48%, these results clearly indicate that the proposed framework achieved high and balanced performance when the classification categories were, well-defined, and fairly distributed. The framework effectiveness on social media is confirmed by its analysis of short, noisy texts in the TweetEval Emotion dataset, achieving an accuracy of 84.24% and F1 macro of 81.56%. The proposed model in DailyDialog dataset achieved an accuracy of 85.08%, and the F1 macro decreased to 58.62% due the dominance of the “No emotion” category and to the low number of samples for rare categories. GoEmotions achieved a label-wise accuracy of 96.68%, F1 micro of 60.03%, F1 macro of 50.29%, and AUC-macro of 92.86%. A decrease in F1 micro and F1 macro is noted, and this attributed to the nature of the dataset itself not to the proposed framework, which is inherently challenging, comprises (27 categories plus a neutral class). Unlike the other single category datasets, a single text my express several overlapping emotions simultaneously, significantly complicating prediction. Furthermore, negatively impact performance. Despite these difficulties, the proposed framework achieved a high AUC, indicating strong recognition capabilities. The proposed framework for observing quantitative results performed well on the Empathetic Contexts dataset, achieving accuracy of 95.24%, F1 macro of 85.69%, AUC of 96.52% and Top-3 accuracy of 85.80%, this indicates that the correct emotions are consistently classified within the probabilistic candidate categories, thus maintaining a high level of recognition performance. Finally, the LGI model achieved on the EmoWOZ dataset overall accuracy and F1-micro of 92.86% with perfect AUC about 94.97%, this approve that the methodology was able to detect the emotions accurately and the model is considered very effective due to its classification of prevalent categories such as neutral and satisfied

 ISSN:2302-8149
e-ISSN:2540-9719

emotions, but the imbalance dataset itself led to a decrease in the F1-macro scale and difficulty in classifying extremely rare categories.

Table 2 Illustrate comparative results of the proposed LGI emotion detection system

Dataset	Accuracy	F1-micro	F1-macro	AUC	Note
DAIR Emotion	93.20%	93.20%	89.10%	99.48%	Best overall performance
TweetEval Emotion	84.24%	84.24%	81.56%	95.07%	Strong performance on short text
DailyDialog	85.08%	85.08%	58.62%	89.80%	Affected by dominant no-emotion class
GoEmotions	96.68%	60.03%	50.29%	92.86%	label-wise accuracy due to the dataset is multi-label class.
Empathetic Context	95.24%	95.24%	85.69%	96.52%	Fine grained 32 class highest Top-3 = 85.80
EmoWOZ	92.86%	92.86%	59.95%	94.97%	Chatbot

The proposed model demonstrated superior performance compared to the previous work across the six datasets as illustrate in Table 3. Notably, the proposed framework achieved highest performance in DAIR dataset across the three metrics: Accuracy, F1-macro and AUC. Similarly, in TweetEval Emotion dataset, LGI achieved higher F1-macro, neglecting the accuracy and AUC because they were not mentioned in published work [26]. In DailyDialog dataset, the F1-macro result closely matches the published work [32], and significantly outperforms F1-micro, neglecting the accuracy and AUC. In the multi-label GoEmotion dataset the F1-macro better than the published work about 5%. The Empathetic Contexts dataset achieved more higher than the published work in Accuracy by a difference of 12 %, neglecting the two scales F1 and AUC, according to the published resource [33]. Finally, in comparison with the previous work [31], the publisher of EmoWOZ dataset, it is noted that the proposed LGI methodology is superior in both F1-micro and F1-macro scales. This is clear evidence of the superiority of the proposed methodology in performing emption detection and achieving the highest standards in this field.

Table 3 Comparative and ablation results of the proposed LGI, RoBERTa-baseline and The published works

Dataset	Published work	Published work Result	RoBERTa-baseline Result	Proposed LGI result
DAIR Emotion	Yan et al., 2025 [34]	Acc=88.50% F1= 89.0% AUC= 93.0%	Acc=76.15% F1-macro=78.30% AUC=87.60	Acc=93.20% F1-macro=89.10% AUC=99.48
TweetEval Emotion	Barbieri et al., 2020 [26]	Acc= - F1= 79.8% AUC= -	Acc=73.34% F1-macro= 75.18% AUC= 91.12%	Acc=84.24% F1-macro= 81.56% AUC= 95.07%
DailyDialog	Ghosal et al., 2020 [32]	Acc= - F1-macro= 51.05% AUC= -	Acc= 80.38% F1-macro= 50.32% AUC= 84.70%	Acc= 85.08% F1-macro= 58.62% AUC= 89.80%
GoEmotions	Demszky et al., 2020 [29]	Acc= - F1= 46.0% AUC= -	Acc=90.65% F1-macro= 44.19% AUC= 90.71%	Acc=96.68% F1-macro= 50.29% AUC= 92.86%
Empathetic Context	Binh et al., 2025 [33]	Acc= 83.0% F1= - AUC= -	Acc=82.16% F1-macro= 80.70% AUC= 90.10%	Acc=95.24% F1-macro= 85.69% AUC= 92.86%
EmoWOZ	Feng et al., 2022 [31]	Acc= - F1-macro= 54.43% AUC= -	Acc= 90.67% F1-macro= 52.70% AUC= 92.78	Acc= 92.86% F1-macro= 59.95% AUC= 94.97

ISSN:2302-8149
e-ISSN:2540-9719

4.3.2 Qualitative Results

Regardless of the numerical performance, and for broader understanding of the proposed framework's behavior, qualitative analysis reveals that the model performs exceptionally well on the DAIR (as shown in Table 4) and TweetEval (as in Table 5) emotion datasets, in these datasets, emotions are more pronounced, enabling the transformer encoder to effectively capture distinct emotional texts. Conversely, the model encountered greater challenges on the DailyDialog dataset (as in Table 6), primarily due to the dominance of a no-emotion or nature emotion category. Consequently, performance declined on less common emotions such as fear, sadness, disgust and anger. The model exhibits strong differentiation on GoEmotion dataset (as illustrated in Table 7), as well as on the AUC scale. However, the qualitative behavior reflects difficulty in identifying multi-label emotions, since many emotion categories are semantically interconnected, and a single text can express more than one emotional state. Therefore, errors arise from partial overlaps in emotions rather than from incorrect overall classification. The most accurate single- label classification environment in Empathetic Contexts dataset (as shown in Table 8), which consists of 32 closely related emotional categories. The accuracy of the Top-3 result indicates that the model's strength often determines the correct emotional range. Some emotions, such as anxiety, fear, terror and apprehension are semantically similar to contentment, joy hope, and satisfaction; therefore, the primary cause of some errors may be subtle semantic ambiguity rather than a complete failure to understand the emotional meaning in texts. However, at EmoWOZ dataset, which is the closest to a realistic representation of chatbots, as shown in Table 9, the model correctly predicted all responses due to its high accuracy and superior ability to detect emotions. Finally, the qualitative results confirm that the proposed framework is effective in capturing general emotions and recognition emotional categories.

Table 4 Qualitative results for real texts in single-label DAIR emotion dataset

Text	True label	Predicted label	Result
"I feel a little mellow today"	Joy	Joy	Correct
"I don't feel particularly agitated"	Fear	Fear	Correct
"I was feeling a little vain when I did this one"	Sadness	Sadness	Correct
"I am feeling outraged it shows everybody"	Anger	Anger	Correct
"I want each of you to feel my gentle embrace"	Love	Love	Correct
"I feel shocked and sad at the fact that there are so many sick people"	Surprise	Surprise	Correct

Table 5 Qualitative results for real texts in single-label tweeteval emotion dataset

Text	True label	Predicted label	Result
"@user is the man"	Joy	Joy	Correct
"In need of a change! #restless"	Sadness	Sadness	Correct
"I am revolting"	Anger	Anger	Correct
"Sleep is and will always be one of the best remedies for a tired and weary soul"	Optimism	Optimism	Correct

ISSN:2302-8149
e-ISSN:2540-9719

Table 6 Qualitative results for real texts in single-label dailydialog dataset

Text	True label	Predicted label	Result
"Hey man, you wanna buy some weed?"	No emotion	No emotion	Correct
"I am not cheating. When you pass go, you collect $200, Everyone knows that!"	Anger	Anger	Correct
"Can you take he away from me?"	Disgust	No emotion	Wrong
"But there are only two more days until the bidding closes!"	Fear	Fear	Correct
"Great. We can chat while enjoying a cup there."	Happiness	Happiness	Correct
"I'm sorry I'm so late. I had a really bad day"	Sadness	Sadness	Correct
"Really? What did you get one for?"	surprise	surprise	Correct

Table 7 Qualitative results for seven emotions in real texts in single-label empathetic contexts dataset

Text	True label	Predicted label	Result
"I received concert tickets for Christmas"	Excited	Excited	Correct
"My baby is sleeping"	Content	Content	Correct
"My parents keep taking my money"	Angry	Angry	Correct
"I have a dog that lives near us and his owner never picks up his poop. It makes me feel sick!"	disgusted	disgusted	Correct
"Today when I went outside to check my banana plant I noticed the leaf was wilted"	Sad	Sad	Correct
"I was so happy the day I learned I was pregnant with our first child!"	Joyful	Surprised	Wrong
"When I was 10 years old I watched the film "The Grudge" and was so scared of dark that I couldn't navigate around my own house"	afraid	terrified	Wrong

Table 8 Qualitative results for seven emotions in real texts in multi-label goemotion dataset

Text	True label	Predicted label	Result
"it's great that you're a recovering addict, that's cool. Have you ever tried DMT?"	Curiosity, gratitude	Curiosity, gratitude	Correct
"Lol! But I love your last name. XD"	Amusement , love	amusement, love	Correct
"A surprise turn of events! I'm so glad you heard such great thinks about yourself !"	Pride, surprise	Pride, surprise	Correct
"I know people around here may mock from my optimism but what if we get actual desings for the crusaders in heroes"	Confusion, optimism	Confusion, surprise	Wrong
"Lol looks delicious"	Admiration, amusement	Admiration, amusement	Correct
"Eff your video – love Canada CA stupid geolock"	Anger,	Anger,	Correct

ISSN:2302-8149
e-ISSN:2540-9719

	annoyance	annoyance	
"I'm so proud of you for taking care of you and your babies, mama! We love you and stay strong."	Admiration, caring, love	Admiration, caring, love	Correct

Table 9 Qualitative results for real texts in chatbot EmoWOZ dataset

Text	True label	Predicted label	Result
"Once you find a restaurant, make sure you get phone number, address."	Neutral	Neutral	Correct
"I am excited about seeing local tourist attraction. The attraction should be in the same area as the restaurant."	Excited	Excited	Correct
"I am also looking for a place to stay"	Dissatisfied	Dissatisfied	Correct
"Ok thank you can you suggest a hotel as well with the above criteria"	Satisfied	Satisfied	Correct
"Your horrible and helpful at all"	Abusive	Abusive	Correct
"I I'm not ok"	Fearful	Fearful	Correct
"Sorry please book for Monday for 3 night and for 4 people"	Apologetic	Apologetic	Correct

5 Conclusion

The study introduced the proposed LGI emotion detection framework that based on transformer contextual representation with label relation modeling to enhance the emotion recognition efficiency over six diverse benchmark datasets. The included datasets that used to evaluate the LGI was DAIR Emotion, TweetEval Emotion, DailyDialog, GoEmotions, Empathetic Contexts, and EmoWOZ, they were covering the single / multi -label, fine grained and human-machine conversation emotion classification tasks. The experimental results shows that the framework accomplish powerful and consistent performance despite the complexity of the datasets. It is worth mentioning that the combination of the label relation data enabled the LGI to capture dependencies between emotion classes and increase recognition ability. The high AUC metric among all datasets indicates that the LGI outperform strong recognition capability and reliable probability approximation. The qualitative results appear that some recognition errors occurred very similar emotions class not between widely separated emotions, thus increasing the performance of the emotions representation.

Finally, the proposed LGI perform a strong and robust solution for emotion classification in text data, thus it can be used in intelligent chatbots, affective computing applications and sentiment aware systems.

There are some limitations to the study. The proposed framework remains sensitive to class imbalance and semantic overlap between very similar emotions. Also, the experiments are limited to English-language datasets, which may affect its generalizability to other world languages. Additionally, the input length of 128 words restricts generalization to longer conversations.

Future researchers should explore new approaches to address the problem of semantic imbalance and overlap between related emotions. Also, may be expand the framework to include other world languages and utilize higher input length, requiring more advanced computer specifications and explore meta learning mechanisms in this field.

References

[1] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. Pappas, "*A Systematic Review of Applications of Natural Language Processing and Future Challenges with Special Emphasis in Text-based Emotion Detection*", Artificial Intelligence Review, Vol. 56, No. 12, pp. 15129-15215, 2023, DOI: 10.1007/s10462-023-10509-0.

[2] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "*Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches*". Artificial Intelligence Review, Vol. 54, No. 8, pp. 5789-5829, 2021, DOI: 10.1007/s10462-021-09958-2.

ISSN:2302-8149
e-ISSN:2540-9719

[3] P. Thakur, D.R. Shrivastava, and A. DR, "*A Review on Text based Emotion Recognition System*", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 7, No. 5, 2018.[Online]. Available: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse01752018.pdf.

[4] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "*Pre-Trained Models for Natural Language Processing: A Survey*", Science China Technological Sciences, Vol. 63, No. 10, pp. 1872-1897, 2020, DOI: 10.1007/s11431-020-1647-3.

[5] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "*BERT Applications in Natural Language Processing: A Review*", Artificial Intelligence Review, Vol. 58 No. 6, Art. No. 166, 2025, DOI: 10.1007/s10462-025-11162-5.

[6] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, and S. Azam, "*A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges*", IEEE Access, Vol. 12, pp. 26839-26874, 2024, DOI: 10.1109/ACCESS.2024.3365742.

[7] A. Al Maruf, F. Khanam, M. M. Haque, Z. M. Jiyad, M. F. Mridha, and Z. Aung, "*Challenges and Opportunities of Text-based Emotion Detection: A Survey*", IEEE Access, Vol. 12, pp. 18416-18450, 2024, DOI: 10.1109/ACCESS.2024.3356357.

[8] U. Frevert, M. Scheer, and A. Schmidt, "*Emotional Lexicons: Continuity and Change in the Vocabulary of Feeling 1700-2000*", Oxford Unversity Prees, 2014. [Online]. Avalabile: https://reviews.history.ac.uk/review/1682.

[9] M. V. Mäntylä, D. Graziotin, and M. Kuutila,"*The Evolution of Sentiment Analysis—A Review of Research Topics, Venues, and Top Cited Papers*", Computer Science Review, Vol. 27, pp 16-32, 2018, DOI: 10.1016/j.cosrev.2017.10.002.

[10] S. Jaiswal, and G. C. Nandi, "*Robust Real-Time Emotion Detection System using CNN Architecture*", Neural Computing and Applications, Vol. 32, No. 15, pp. 11253-11262, 2020, DOI: 10.1007/s00521-019-04564-4.

[11] A. Akula, , G. Budha, G. Bingi, U. Chanda, A. R. Borra, D. B. Yadav, and M. Saravanan, "*Emotion Recognition from Facial Expressions using CNNs*", International Journal of Engineering & Extended Technologies Research (IJEETR), Vol. 8, No. 1, pp. 120-125, 2026, DOI: doi.org/10.15662/IJEETR.2026.0801013.

[12] C. Yerukonda, "*Multimodal Context-Aware Emotion Recognition using Transformer-based Fusion of Facial and Speech Cues*", In 7th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 1782-179, IEEE, 2026, DOI: 10.1109/ICICV68925.2026.11554829.

[13] Y. Ghafoor, S. Jinping, F. H. Calderon, Y. H. Huang, K. T. Chen, and Y. S. Chen, "*TERMS: Textual Emotion Recognition in Multidimensional Space*", Applied Intelligence, Vol. 53, No. 3, pp. 2673-2693, 2023, DOI: 10.1007/s10489-022-03567-4.

[14] Z. Ye, T. Zuo, W. Chen, Y. Li, and Z. Lu, "*Textual Emotion Recognition Method based on ALBERT-BiLSTM Model and SVM-NB Classification*", Soft Computing, Vol. 27,No. 8, pp. 5063-5075, 2023, DOI: 10.1016/j.specom.2022.03.002.

[15] J. K. Rout, K. K. R. Choo, A. K. Dash, S. Bakshi, S. K. Jena, and K. L. Williams, "*A Model for Sentiment and Emotion Analysis of Unstructured Social Media Text*", Electronic Commerce Research, Vol. 18, pp. 181-199, 2018, DOI: 10.1007/s10660-017-9257-8.

[16] J. Song, M. Wang, and Y. Li, "*An Exploratory Study of Deep Learning-based Sentiment Analysis among Weibo Users in China*", Current Psychology, pp. 1-14, 2023, DOI: 10.1007/s12144-023-05493-1.

[17] C. Li, and F. Li, "*Emotion Recognition of Social Media Users based on Deep Learning*", PeerJ Computer Science, Vol. 9, Art. No. e1414, 2023, DOI: 10.7717/peerj-cs.1414.

[18] J. Cahn, "*CHATBOT: Architecture, Design, & Development*", Ph.D. Thesis, University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science, 2017.

[19] E. Adamopoulou, and L. Moussiades, "*An Overview of Chatbot Technology*", In IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 373-383, Springer, Cham, 2020, DOI: 10.1007/978-3-030-49186-4_31.

[20] M. Nißen, D. Selimi, A. Janssen, D. R. Cardona, M. H. Breitner, T. Kowatsch, and F. von Wangenheim, "*See You Soon Again, Chatbot? A Design Taxonomy to Characterize User-*

 ISSN:2302-8149
e-ISSN:2540-9719

Chatbot Relationships with Different Time Horizons", Computers in Human Behavior, Vol. 127, Art. No. 107043, 2022, DOI: 10.1016/j.chb.2021.107043.

[21] G. Bilquise, S. Ibrahim, and K. Shaalan, "*Emotionally Intelligent Chatbots: A Systematic Literature Review"*, Human Behavior and Emerging Technologies, Vol. 2022, No. 1, 2022, pp. 9601630, DOI: 10.1155/2022/9601630.

[22] M. Ehtesham-Ul-Haque, J. D'Rozario, R. Adnin, F. T. Utshaw, F. Tasneem, I. J. Shefa, and A. A. Al Islam, "*EmoBot: Artificial Emotion Generation through an Emotional Chatbot during General-Purpose Conversations*", Cognitive Systems Research, Vol. 83, Art. No. 101168, 2024, DOI: 10.1016/j.cogsys.2023.101168.

[23] M. Kozłowski, K. Gabor-Siatkowska, I. Stefaniak, M. Sowański, and A. Janicki, "*Enhanced Emotion and Sentiment Recognition for Empathetic Dialogue System using Big Data and Deep Learning Methods*", In International Conference on Computational Science, Cham: Springer Nature Switzerland, pp. 465-480, 2023, DOI: 10.1007/978-3-031-35995-8_33.

[24] S. Lee, Y. Park, and G. Park, "*Using AI Chatbots in Climate Change Mitigation: A Moderated Serial Mediation Model*", Behaviour & Information Technology, Vol. 43, pp.1-17, 2024, DOI: 10.1080/0144929X.2023.2298305.

[25] E. Saravia, H. C. T. Liu, Y. H. Huang, J. Wu, and Y. S. Chen, "*CARER: Contextualized Affect Representations for Emotion Recognition*", In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3687-3697, 2018, DOI: 10.18653/v1/D18-1404.

[26] F. Barbieri, J. Camacho-Collados, L. E. Anke, and L. Neves, "*TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification*", In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1644-1650, 2020, DOI: 10.18653/v1/2020.findings-emnlp.148.

[27] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "*Semeval-2018 Task 1: Affect in Tweets*", In Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 1-17, 2018, DOI: 10.18653/v1/S18-1001.

[28] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "*Dailydialog: A Manually Labelled Multi-Turn Dialogue Dataset*", In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Vol. 1, pp. 986-995, 2017, [Online]. Available: https://aclanthology.org/I17-1099/

[29] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "*GoEmotions: A Dataset of Fine-Grained Emotions*", In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4040-4054, 2020, DOI: 10.18653/v1/2020.acl-main.372.

[30] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, "*Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset*", In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5370-5381, 2019, DOI: 10.18653/v1/P19-1534.

[31] S. Feng, N. Lubis, C. Geishauser, H. C. Lin, M. Heck, C. van Niekerk, and M. Gasic, "*Emowoz: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems*", In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 4096-4113, 2022, [Online]. Available: https://aclanthology.org/2022.lrec-1.436/

[32] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "*Cosmic: Commonsense Knowledge for Emotion Identification in Conversations*", In Findings of the association for computational linguistics: EMNLP 2020, pp. 2470-2481, 2020, DOI: 10.18653/v1/2020.findings-emnlp.224.

[33] B. Vu, N. Keshri, S. Chandna, M. Jalali, S. Mehraeen, "*A Systematic Approach to Fine-Tuning Transformers for Emotion Detection on the Empathetic Dialogues Benchmark*", International Journal of Information Technology, Vol. 17, No. 7, pp.3895-3912, 2025 DOI: https://doi.org/10.1007/s41870-025-02645-3.

[34] J. Yan, P. Pu, and L. Jiang, "*Emotion-RGC Net: A Novel Approach for Emotion Recognition in Social Media using RoBERTa and Graph Neural Networks*", Plos one, Vol. 20, No. 3, Art. No. e0318524, 2025, DOI: 10.1371/journal.pone.0318524.