

A Comprehensive Evaluation of Word Embedding Methods with Deep Learning Models for Multi-Label Emotion Classification

Zeynep KARACA^{1*} , İlhan AYDIN¹ 

¹ Fırat University, *Engineering* Faculty, Computer Engineering Department, Elazığ, Türkiye

*Corresponding author: zkaraca@firat.edu.tr

Abstract

This study investigates the effects of different word embedding methods on the performance of deep learning models in the field of emotion analysis. Using the GoEmotions dataset developed by Google, emotion classification was performed on multi-label texts. In the experimental procedure, three distinct word embedding methods ‘GloVe, FastText, and Word2Vec’ were applied, and the resulting vector representations were fed into various deep learning architectures (CNN, RNN, LSTM, GRU). In addition, an attention mechanism was integrated into these architectures to analyze the impact of attention layers on model performance. The findings indicate that word embedding methods produce statistically meaningful differences in the success of emotion classification models. In particular, the highest precision was achieved (71.60%) in the scenario combining the CNN+Attention architecture with GloVe embeddings. On the other hand, the most balanced and effective overall performance was observed for the BiGRU+Attention architecture paired with GloVe, yielding an F1 score of 56.11%. A literature review revealed no prior study that comparatively evaluates three different embedding methods across multiple deep learning architectures on a multi-label dataset such as GoEmotions. These results underscore the need for careful selection of appropriate word embedding techniques in emotion analysis applications and provide original contributions to the integration of word representation techniques with deep learning architectures.

Keywords

Natural language processing,
Text classification,
multi-label
emotion analysis,
Word embedding,
Deep learning

Çok Etiketli Duygu Sınıflandırmasında Kelime Gömme Yöntemlerinin Derin Öğrenme Modelleri ile Kapsamlı Bir Değerlendirmesi

Zeynep KARACA^{1*} , İlhan AYDIN¹ ,

¹ Fırat Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Elazığ, Türkiye

*Sorumlu yazar: zkaraca@firat.edu.tr

Özet

Bu çalışmada, duygu analizi alanında kullanılan farklı kelime gömme yöntemlerinin derin öğrenme modellerinin performansı üzerindeki etkileri incelenmiştir. Google tarafından geliştirilen GoEmotions veri kümesi kullanılarak çok etiketli metinler üzerinde duygu sınıflandırması gerçekleştirilmiştir. Deneysel süreçte GloVe, FastText ve Word2Vec olmak üzere üç farklı kelime gömme yöntemi uygulanmış ve elde edilen vektör temsilleri çeşitli derin öğrenme mimarilerine (CNN, RNN, LSTM, GRU) girdi olarak verilmiştir. Ayrıca, bu mimarilere dikkat (attention) mekanizması entegre edilerek, dikkat katmanının model performansı üzerindeki etkisi analiz edilmiştir. Elde edilen bulgular, kelime gömme yöntemlerinin duygu sınıflandırma modellerinin başarımı üzerinde anlamlı farklılıklar yarattığını göstermektedir. Özellikle, CNN+Attention mimarisi ile GloVe gömme yönteminin kullanıldığı senaryoda en yüksek kesinlik değeri (%71.60) elde edilmiştir. Öte yandan, en dengeli ve etkili performans, BiGRU+Attention mimarisi ile GloVe yönteminin birlikte kullanıldığı durumda %56.11'lik F1 skoru ile gözlemlenmiştir. Literatür incelemesi sonucunda, GoEmotions veri kümesi gibi çok sınıflı bir yapıya sahip veri seti üzerinde üç farklı gömme yönteminin farklı derin öğrenme mimarileri ile karşılaştırmalı olarak değerlendirildiği benzer bir çalışmaya rastlanmamıştır. Bu sonuçlar, duygu analizi uygulamalarında uygun kelime gömme yönteminin seçiminde dikkatli olunması gerektiğini ortaya koymakta ve kelime temsil teknikleri ile derin öğrenme mimarilerinin entegrasyonuna yönelik özgün katkılar sunmaktadır.

Anahtar kelimeler

Doğal dil işleme,
Metin sınıflandırma,
Çok etiketli duygu
analizi,
Kelime gömme,
Derin öğrenme

1. INTRODUCTION

Emotion analysis is a significant subfield of natural language processing (NLP) that seeks to identify individuals' emotional states from written expressions. The widespread adoption of social media and online platforms has made text-based emotion classification critically important across many domains, including psychology, education, and social media analytics. Word representations used in this area—namely, word embedding methods—have a decisive influence on classification performance because they enable models to learn semantic and contextual relationships among words more accurately.

The GoEmotions dataset, created by Google, comprises over 58,000 English Reddit comments and features a multi-label structure with 27 distinct emotion categories plus one neutral class [1]. A review of the literature indicates that investigations into the interaction between deep learning architectures and word embedding methods have frequently been confined to binary positive–negative classifications, while comprehensive analyses covering a large number of emotion classes remain limited. Owing to its multi-label nature, GoEmotions offers an extensive testbed for neural network–based emotion classification models. Kumar et al. [2] performed a comparative evaluation of CNN, RNN and BERT models on the GoEmotions dataset. In that study, the CNN model reported a precision of 41% and an F1 score of 43%, whereas the RNN model achieved a precision of 40% and an F1 score of 41%. Although the study presented overall accuracy metrics, it did not include a detailed class-level analysis comparing model predictions with true labels. In this context, the present study systematically compares various deep learning models and word embedding techniques to address these gaps and evaluates model performance using multiple metrics. Birannavar et al. [3] conducted emotion analysis on the GoEmotions dataset using Support Vector Machines (SVM), Naïve Bayes, and BERT models. The precision and F1-score obtained with the Naïve Bayes model were reported as 60% and 44%, respectively. The study did not provide class-level accuracy analyses or detailed comparisons of model predictions, which limits the assessment of model performance across different classes. Younas et al. [4] performed multi-label emotion classification on the GoEmotions dataset using classical deep learning architectures. Their proposed Stacked LSTM architecture was augmented with encoder–decoder structures designed to capture contextual dependencies such as prior emotional states and social context. In comparative experiments, the BiLSTM model achieved an F1 score of 53%, while the CNN model reached an F1 score of 58%. Although the proposed model offered a strong approach to modelling contextual structures, the results indicate that the BiLSTM and CNN baselines attained higher F1 scores in terms of overall performance. Gonzalez et al. [5] evaluated the performance of DNN Pool and BiLSTM models using FastText word embeddings on the GoEmotions dataset. Zhou et al. [6] introduced the Emotion Quantization Network (EQN), an approach that maps label values to energy-density levels for multi-emotion detection. Yusifov et al. [7] focused on classical machine learning algorithms rather than neural-network-based models, comparing the performance of various algorithms on GoEmotions. Their experiments were

conducted on a subset of 10,621 samples and 7 emotion classes; annotations with low consistency were removed during preprocessing. The most frequent annotations were retained, punctuation was stripped, and texts were stemmed. In the experimental analysis, an SVM-based classification method achieved one of the best results with 62% precision, 61% recall, and a 60% F1 score. The authors concluded that examples in the GoEmotions dataset exhibit complex intra-class structures in feature space, which classical machine learning methods fail to adequately partition. These findings suggest limited generalization capacity for traditional approaches in the context of multi-label emotion classification. Ni et al. [8] carried out a comparative experimental study by training LSTM and BiLSTM models on GoEmotions to evaluate the effectiveness of classical deep learning methods for multi-label emotion classification. Their results indicated that the LSTM-based model produced successful outcomes, with the highest performance attained by the BiLSTM model. Olah et al. [9] assessed classical machine learning approaches on the GoEmotions dataset and reported that the XGBoost algorithm yielded more balanced results. These observations further underline the limitations of context-independent word representations used by traditional methods for emotion classification tasks. Bhardwaj et al. [10] addressed the multi-class emotion classification problem on GoEmotions using conventional machine learning algorithms. Their study focused on six selected core emotion classes drawn from the dataset. In comparisons among Naïve Bayes (NB), Support Vector Machines (SVM), and Gradient Boosting (GB), the GB algorithm achieved the highest score. The results indicate that while traditional methods can perform at an average level on a limited set of emotion classes, they generally underperform relative to deep-learning-based approaches due to their inability to capture contextual and semantic representations. Ramakrishnan et al. [11] evaluated the impact of training epochs on multi-label emotion classification by selecting minority classes from the 28 labels and utilizing the RoBERTa-Large model. Following four training epochs, the study reported an F1-score of 63.4% and an Average Precision of 66.2%. Monicadevi et al. [12] introduced a dual-label classification approach for the GoEmotions dataset based on Plutchik's theory. Evaluating BERT, XLNet, and RoBERTa models with auxiliary sentiment features, they found that BERT achieved the highest overall accuracy 63.81%, while XLNet and RoBERTa performed better on fine-grained emotions. The study concluded that while distinct emotions like admiration are easily identified, complex states such as grief remain challenging to classify. Peng et al. [13] introduced the MER-MPCK model, which enhances multi-label emotion recognition by fusing textual data with multiple perspectives of commonsense knowledge via multi-head attention. Using the COMET tool to extract five distinct insights—such as speaker intent and reactions the model integrates these with RoBERTa-based semantic features. Experimental results on the GoEmotions dataset demonstrated that MER-MPCK outperformed baseline models with a MacF1 score of 52.79%, surpassing XLNet 48.50%, BERT 50.10%, and RoBERTa 51.07%. The study concluded that while XLNet's autoregressive approach limited its global context capture, RoBERTa's extensive pre-training proved superior for capturing complex linguistic features in multi-label tasks.

Table 1 presents the F1-score and precision results of the emotion analysis studies surveyed in the literature.

Table 1. Results of literature studies on the GoEmotions dataset

Reference	Model	Precision	F1 Score	Number of Emotion Tags
Kumar et al. [2]	CNN	41	43	28
	RNN	40	41	28
	BERT	40	46	28
Birannavar et al. [3]	Naive Bayes	60	44	28
Younas et al. [4]	Stacked LSTM	55	50	7
	BiLSTM	-	53	7
	CNN	-	58	7
Gonzales et al. [5]	FastText +DNN Pool	42	45	28
	FastText +BiLSTM	44	51	28
Zhou et al. [6]	EQN	56	52	28
Yusifov et al. [7]	SGD	59	60	7
	SVC	62	60	7
	KNN	52	43	7
	CatBoost	62	60	7
Ni et al. [8]	LSTM	51.53	53.12	7
	BiLSTM	57.12	57.79	7
Olah et al. [9]	XGBoost	47	44	28
Bhardwaj et al. [10]	SVM	55.6	54.5	7
	GB	57.3	56.1	7
	NB	43.3	40.0	7

Previous studies on the GoEmotions dataset indicate that different modeling approaches substantially affect emotion-classification performance. Research employing classical machine-learning methods has typically reported low and limited F1 scores for algorithms such as SVM, Naïve Bayes, and XGBoost; however, some methods (e.g., CatBoost) have been reported to reach F1 scores of up to 60%. These findings suggest that traditional approaches—lacking contextual information—offer only limited success in classifying emotional content. Deep learning-based methods generally demonstrate superior performance: studies applying sequential and convolutional models such as LSTM, BiLSTM, and CNN have reported F1 scores exceeding 50%, while the BCDF model has emerged with a notable F1 score of 69%. Collectively, these results underscore the clear advantage of contextual representations for emotion-classification tasks.

To address this gap, the present study systematically compares three word-embedding methods—GloVe, FastText, and Word2Vec—across various deep-learning architectures (CNN, LSTM, GRU) and their attention-augmented variants (BiLSTM+Attention, BiGRU+Attention, CNN+Attention) on the GoEmotions dataset. The experimental results demonstrate that embedding choice plays a critical role in model performance. Specifically, the CNN+Attention architecture combined with GloVe embeddings achieved

the highest precision (71.60%), while the highest F1 score (56.11%) was observed for the BiGRU+Attention architecture, also using GloVe embeddings.

An examination of existing literature reveals that a significant portion of research conducted on the GoEmotions dataset has narrowed its scope by reducing the 28 emotion labels to 7 core categories. This trend leads to an incomplete analysis of the extensive emotion space provided by the dataset and tends to overlook the inherent complexity of its multi-label structure. The primary motivation of this study is to transcend these limitations by preserving the original 28-class structure and providing a comprehensive analysis of the interplay between various word-embedding techniques and deep learning architectures.

In this context, the original contribution of this work lies in offering a systematic comparative evaluation of three distinct word-embedding techniques across multiple deep-learning models. Furthermore, by integrating these models with attention mechanisms, the study quantifies the individual and combined contributions of both embedding methods and attention layers to classification performance. The lack of such large-scale comparative studies in the literature establishes this work as a significant reference point and a benchmark for emotion analysis research. Our findings demonstrate that the careful selection of word-embedding methods, optimized with attention-based architectures, materially enhances performance in high-dimensional emotion classification tasks.

The remainder of this paper is organized as follows: Section 2, **Material and Method**, describes the proposed system architecture, word-embedding methods, and evaluation criteria. Section 3, **Results**, details the utilized parameters, loss curves, and comparative analyses with the literature, alongside heatmaps and prediction outcomes. Finally, Section 4, **Discussion and Conclusion**, provides a general assessment based on the obtained data.

2. MATERIAL AND METHOD

In this section, the system architecture employed for emotion analysis and the specific models and word-embedding methods used in the study are described in detail. The primary objective of the study is to compare the effects of different word-embedding techniques and deep-learning architectures on emotion-classification performance. To this end, three distinct word-embedding methods—GloVe, FastText, and Word2Vec—were employed. Each embedding method provided fixed-dimensional vector representations that converted textual input into numerical form. The embedding representations were integrated with and evaluated across the following model architectures: CNN, LSTM, GRU, BiGRU+Attention, BiLSTM+Attention, and CNN+Attention. Each model was trained and tested separately with each of the three embedding methods. During model training, precision, recall, and F1 score were used as the primary evaluation metrics, and model performance was analyzed on the held-out test set.

Hyperparameter settings were controlled throughout the training process: key parameters such as the number of epochs, learning rate, and batch size were optimized and then held constant across all models to ensure a fair comparative evaluation. The proposed architecture is illustrated in Figure 1.

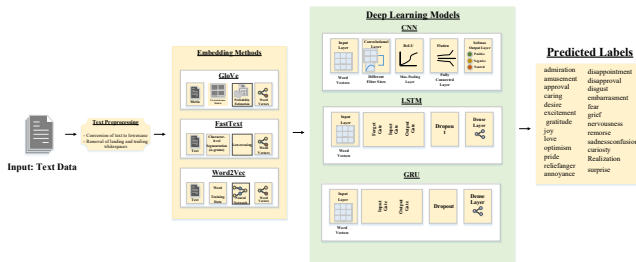


Figure 1. Architecture of the Proposed System

Figure 1 presents the study’s methodological framework in a stepwise manner. In the first stage, input text is preprocessed and converted into tokenized numerical vector representations. Different word-embedding methods (GloVe, FastText, and Word2Vec) are applied in this phase to transform the textual data into fixed-dimensional numerical vectors. In the second stage, the resulting embedding vectors are fed as inputs to various deep-learning models (CNN, LSTM, GRU and their attention-augmented variants). Each model’s architecture comprises an embedding layer, one or more hidden layers, and an output layer. In the final stage, model predictions are produced in a format appropriate for the multi-label classification task, and model performance is evaluated using metrics such as precision, recall, and F1 score.

2.1. Feature Extraction with Word-Embedding Methods

Representing textual data in numerical form is a prerequisite for machine-learning and NLP applications, since it renders text suitable as model input. Word-embedding methods were developed for this purpose: they represent words as high-dimensional vectors that capture semantic similarities in a numeric format. In the present study, three distinct word-embedding methods were evaluated.

Global Vectors for Word Representation (GloVe), unlike Word2Vec, is a statistics-driven approach. Introduced by Pennington et al. [14] in 2014, GloVe learns word vectors by exploiting a word co-occurrence matrix. The primary objective is to model relationships between words in the vector space based on their co-occurrence probabilities. The core optimization objective of the model is given by Equation (1):

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \quad (1)$$

Equation (1) denotes the **optimization function** that governs the learning of word vectors in the GloVe model. Here, X_{ij} denotes the co-occurrence frequency of word i with word j . The vectors w_i and $\tilde{w}_j$ represent the target-word and context-word embeddings, respectively, while b_i and $\tilde{b}_j$ denote bias terms. The function $f(X_{ij})$ is a weighting function that prevents low-frequency word pairs from exerting undue influence on the model and assigns greater importance to frequently observed pairs. Consequently, the model learns vector representations that more accurately capture the statistical relationships among words in the vector space.

Word2Vec, developed by Mikolov et al. [15] in 2013, is a technique that represents words as semantically meaningful dense vectors. The model is typically trained using either the CBOW or the Skip-Gram architecture [15]. The CBOW architecture predicts a target word from its surrounding context words, whereas the Skip-Gram architecture predicts surrounding context words given a target word. In both approaches, words are represented as one-hot vectors within a vocabulary of size V , and d -dimensional embedding vectors are learned by the end of training. The Skip-Gram model is formalized in Equation (2).

$$\max \prod_{t=1}^T \prod_{-c \leq j \leq c, j \neq 0} P(w_{t+j}|w_t) \quad (2)$$

w_f : Hedef Kelime

c : Bağlam Pencere Boyutu

$P(w_{t+j}|w_t)$: Hedef kelimenin komşu kelimeleri tahmin etme olasılığını gösterir.

In the Word2Vec model, w_t denotes the target word and c denotes the size of the context window. The objective of the model is to predict the neighboring words surrounding the target word w_t . In this context, $P(w_{t+j}|w_t)$ denotes the probability of observing a context word at distance j given the target word. Thus, the model aims to learn the conditional probabilities between target and context words so as to represent their semantic and contextual relationships as dense vectors.

FastText, developed by Facebook AI, is an extension of Word2Vec that processes words not only as atomic units but as compositions of subword units. In this model, each word is represented as the sum of its character n-grams; the resulting word vector is the aggregate of these subvector components. This approach is particularly effective for morphologically rich languages and provides improved generalization for rare or previously unseen words [16]. The word vector formulation is given in Equation (3).

$$\vec{w} = \sum_{g \in G_w} \vec{z}_g \quad (3)$$

In the FastText model, each word is represented not only as a single unit but also through its constituent subunits. In this approach, a word w is expressed as a set G_w of its character n -gram substrings. Each subunit $g \in G_w$ is

associated with a vector representation ($\vec{z}_g$), and the final word vector is computed as the sum of these sub-vectors. This enables the model to generate richer and more generalizable representations, particularly for rare or unseen words, while also capturing affix and root structures more effectively in morphologically rich languages.

2.2. Proposed Models

In NLP tasks such as sentiment analysis, while classical machine learning methods have achieved certain levels of success, the increasing volume and diversity of data have highlighted the advantages of deep learning architectures. These architectures stand out due to their capacity to learn more complex structures and their ability to more effectively represent contextual information and semantic relationships. In this study, the proposed models were applied within the framework of multi-label emotion classification on the GoEmotions dataset, following a unified methodological approach for each model. During the preprocessing stage, texts were lowercased, tokenized at the word level, and a vocabulary was constructed from words that appeared at least five times in the training data. This threshold was selected to reduce noise and sparse representations in the training corpus, thereby eliminating infrequent tokens, which are often attributed to typographical errors or singular instances. Consequently, the embedding matrix size and computational cost of the models were reduced, while the potential adverse impact of rare tokens on model performance was minimized. For labeling, the multi-label nature of the dataset was accommodated by encoding each instance into a multi-hot (binary) vector representation. To ensure comparability across experiments, pretrained GloVe, FastText, and Word2Vec embedding vectors were separately loaded for each experiment, with embedding layers fixed as non-trainable. The embedding matrix dimensions were aligned with the dimensionality of the corresponding pretrained vectors. All models were trained using the binary cross-entropy loss function, with the Adam optimizer employed for parameter updates. The proposed hybrid attention architecture is illustrated in **Figure 2**.

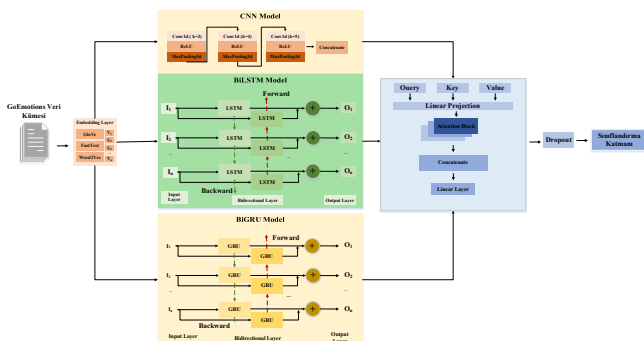


Figure 2. Integration of Embedding Methods and Deep Learning Models with the Attention Mechanism

The models proposed in this study were selected to provide a comparative assessment of the strengths and weaknesses of different deep-learning architectures for

emotion-classification tasks. Convolutional Neural Network (CNN)-based architectures can effectively capture local patterns (analogous to n-grams) in text and thereby identify short phrases or clusters of words that carry emotional content. In this work, three distinct CNN-based models were developed for each embedding method. Fixed-length vector sequences produced by the embedding layer were supplied to one-dimensional convolutional (Conv1D) layers. Parallel convolutional filters with different kernel sizes were applied, each convolutional output was reduced via MaxPooling1D, and the pooled outputs were concatenated to form a single feature vector. The resulting feature representation was passed sequentially through Dropout and Dense layers, and the final layer consisted of as many neurons as there are labels, with sigmoid activations to perform multi-label classification. Models were trained using the binary cross-entropy loss function and the Adam optimizer. Long Short-Term Memory (LSTM) networks are widely used in NLP tasks owing to their capacity to capture long-term dependencies in sequential data [17,18]. In this study, three distinct LSTM-based models integrated with different embedding methods were evaluated. Each model employed a bidirectional LSTM layer following the embedding layer; the LSTM outputs were regularized with Dropout and subsequently passed through fully connected layers for classification. In attention-augmented variants, the time-step outputs of the LSTM were supplied to a multi-head attention mechanism, and the resulting attention-weighted summary vectors were fed to the classification layer. The Gated Recurrent Unit (GRU) features a simpler architecture compared with LSTM and provides a comparably effective means of modelling long-range dependencies while offering greater computational efficiency [19]. In this study, GRU-based models were implemented as two-layer GRU architectures: the embedded word sequences were fed into the stacked (two-layer) GRU, after which the final hidden state of the network was regularized with Dropout and passed to a sigmoid-activated fully connected layer. Models were trained using the binary cross-entropy loss function and the Adam optimizer.

Bidirectional architectures—Bidirectional LSTM (BiLSTM) and Bidirectional GRU (BiGRU)—process inputs in both forward and backward directions, thereby learning sentence context from both sides and yielding richer representations [20–22]. In the present work, the output sequences produced by the BiLSTM were supplied to a multi-head attention layer; the attention-weighted summary vectors were then regularized with Dropout and propagated through dense layers to produce independent predictions for each emotion class. The number of attention heads in the multi-head attention mechanism was set to 4. The attention mechanism enables the model to focus on emotion-salient regions of the sentence rather than treating all tokens with equal importance, and it has been shown to improve performance particularly in multi-label classification problems [22, 23]. Accordingly, the proposed CNN+Attention, BiLSTM+Attention, and BiGRU+Attention architectures combine the strengths of convolutional models for detecting important short token groups and recurrent architectures for capturing long-

range contextual dependencies. As an original contribution, and in contrast to much of the existing literature that is typically limited to a single model or a single embedding method, this study systematically compares six deep-learning architectures with three different pretrained word-embedding methods (GloVe, FastText, and Word2Vec). By doing so, it provides a detailed analysis of the interaction between embedding choice and model architecture and presents novel findings specific to multi-label emotion-classification tasks. The systematic implementation steps of the proposed multi-label emotion classification framework are illustrated in Algorithm 1.

Algorithm 1. Multi-label Emotion Classification via Attention-Augmented Hybrid Architectures

```

Procedure Preprocess_and_Vectorize(D):
    Convert all text to lowercase and apply word-level
    tokenization.
    Build vocabulary V: Include only tokens with frequency
    f >= 5 reduce sparsity.
    Encode labels as multi-hot binary vectors  $Y \in \{0,1\}^{28}$ .
    Transform text into fixed-length sequences S.
    Return (S,Y)

Procedure Build_Hybrid_Model(E, M):
    Initialize Embedding Layer: Load weights from E and set
    as non-trainable.
    If M is CNN-based:
        Apply parallel Conv1D filters (different
        kernels) → MaxPooling1D.
    Else if M is Recurrent-based (LSTM/GRU):
        Initialize Bidirectional or Stacked layers.
    If M includes Attention:
        Apply Multi-head Attention (Number of heads
        = 4).
        Generate weighted context vector C.
    Final Layers: Dropout Dense → Dense(28 units)
    → Sigmoid Activation.
    Compile with Binary Cross-Entropy loss and Adam
    optimizer (lr = 1e-3).
    Return Model

Main Training and Evaluation:
    Split data: 80% Training, 20% Test.
    For each combination of (E, M):
         $Model_{curr} \leftarrow Build\_Hybrid\_Model(E, M)$ 
        Train  $Model_{curr}$  for 10 epochs with batch size
        64.
        Generate predictions P on test set.
        Calculate Precision, Recall, and F1-score
        using Equations (4),(5), and (6).

End For
End Procedure

```

Algorithm 1 outlines the end-to-end workflow, beginning with a frequency-based vocabulary filter ($f \geq 5$) to mitigate data sparsity within the 28-class GoEmotions dataset. The integration of a 4-head attention mechanism within the hybrid architecture allows the models to selectively weigh emotional cues across recurrent or convolutional hidden states. Finally, the implementation of a sigmoid activation function at the output layer treats each class as an independent Bernoulli distribution, facilitating the precise detection of overlapping emotional states inherent in the multi-label framework.

2.3. Evaluation Metrics

To evaluate model performance, this study employed three primary classification metrics commonly used in multi-label classification problems: **precision**, **recall**, and

F1 score. These measures are important for assessing the model's performance from different perspectives. **Precision** is defined as the ratio of correctly classified positive instances to the total number of instances predicted as positive. This metric is particularly effective for assessing model accuracy in contexts where false positive classifications carry significant consequences [24, 25]. The computation of precision is given in Equation (4):

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

In Equation (4), **TP (True Positives)** denotes instances that the model labeled as positive and that are actually positive, whereas **FP (False Positives)** denotes instances that the model labeled as positive but that are actually negative [26]. **Recall** indicates the proportion of actual positive instances that the model correctly identifies; it reflects the model's ability to retrieve all positive examples that should be classified as such. The recall metric is computed as shown in Equation (5):

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

Equation (5) expresses recall mathematically as the ratio of true positives (TP) to all actual positives (TP + FN). In other words, it indicates the percentage of positive instances that are correctly identified. The F1 score is the harmonic mean of precision and recall and is commonly used as a performance metric for imbalanced datasets. The F1 score metric is computed as shown in Equation (6):

$$F1\ score = 2x \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

These evaluation measures provide a comprehensive assessment of a model's overall performance by balancing its ability to make correct predictions (precision) with its capacity to capture all relevant true labels (recall). Because the label distribution in the emotion-classification dataset used in this study is imbalanced, the F1 score plays a critical role in evaluating the model's overall performance. This metric furnishes a balanced evaluation by taking both precision and recall into account.

3. RESULTS

The GoEmotions dataset used in this study is a comprehensive, human-annotated dataset consisting of 58,000 Reddit comments written in English, where each comment is labeled as neutral or assigned to one of 27 different emotion categories. Compared with other datasets, GoEmotions stands out for encompassing not only basic emotions but also more fine-grained and complex ones. These emotion categories are not limited to commonly studied emotions such as anger, sadness, and happiness, but also include nuanced, context-sensitive emotions such as admiration, amusement, gratitude, and love. In this respect, the GoEmotions dataset provides a

rich, multi-label, and detailed resource for emotion classification studies [2]. Furthermore, with its diversity of emotions and extensive vocabulary, it represents a highly varied dataset [24]. Figure 3 illustrates the classification of the emotions in the dataset into three primary affective polarities: positive, negative, and neutral.

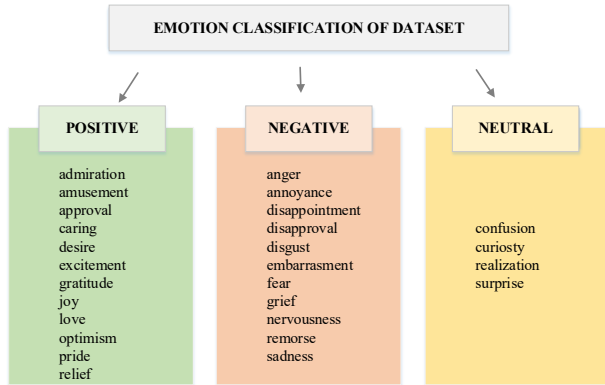


Figure 3. Classification of Emotions by Polarity

Table 2 presents sample sentences selected from the emotion classification dataset used in this study, along with their manually assigned emotion labels prior to modeling. These examples reflect the diversity of emotions and the multi-label data structure, highlighting the semantic variability encountered during the model training process.

Table 2. Sample Sentences from the GoEmotions Dataset and Their Corresponding Emotion Labels

Text	Labels
Shit, I guess I accidentally bought a Pay-Per-View again.	annoyance, embarrassment
Maybe that's what happened to the great white shark.	confusion, realization
We need more boards and to create a bit more space.	desire, optimism
Slowing things down now.	neutral
Super, thanks.	gratitude
I miss them being alive.	grief, sadness
I never thought it was at the same moment, but I guess it was.	confusion, disappointment, neutral
Thank you friend.	gratitude
To make her feel threatened.	fear

The performance of the deep learning models employed in this study was evaluated using three different word embedding methods. The key hyperparameters utilized in the experiments are presented in Table 3.

Table 3. Utilized Parameters

Parameter	Value
Number of Epochs	10
Learning Rate	1e-3
Batch Size	64
Optimization Algorithm	Adam

In order to evaluate the performance of deep learning architectures with different word embedding methods, the variations in training and validation losses with respect to the number of epochs were examined. Within this scope, CNN, GRU, LSTM, BiGRU+Attention, CNN+Attention, and BiLSTM+Attention models were individually trained with GloVe, FastText, and Word2Vec embeddings, and the trends of the loss functions throughout the learning process were compared. Figure 4 illustrates the changes in loss values per epoch for each model and embedding combination.

An examination of Figure 4 reveals that, across all models, the loss values consistently decrease as the number of epochs increases, indicating a stable improvement during the learning process. CNN- and GRU-based architectures exhibit a rapid reduction in loss, particularly in the initial epochs, whereas LSTM-based models demonstrate a more gradual learning trajectory. The BiGRU+Attention and BiLSTM+Attention models, enhanced with attention mechanisms, achieve lower loss values compared to the other models, thereby standing out in terms of overall performance. Furthermore, it is observed that the GloVe embedding method yields more stable and lower loss values in most models, while FastText and Word2Vec display more fluctuating learning trends in certain cases. These findings highlight the decisive influence of embedding methods on model performance and demonstrate that the integration of attention mechanisms significantly enhances learning in multi-label classification tasks. Figure 5 presents the F1-score and precision values for all models.

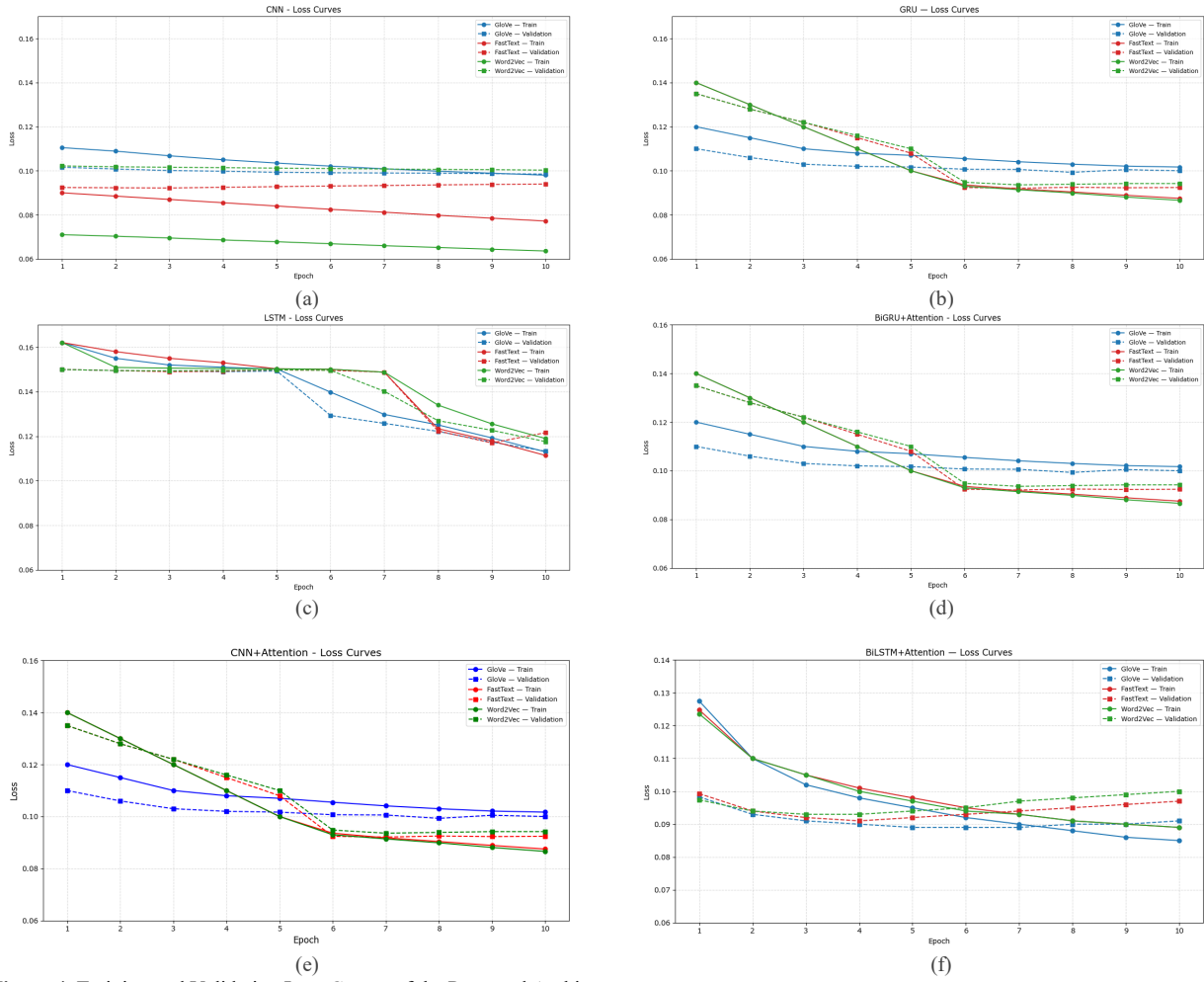


Figure 4. Training and Validation Loss Curves of the Proposed Architectures

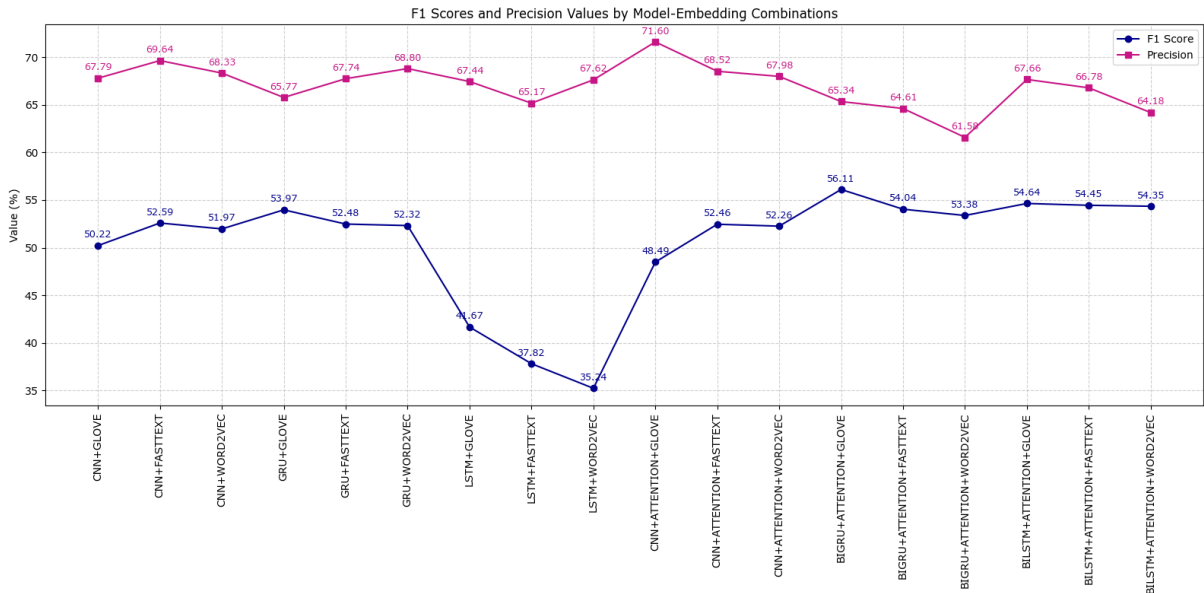


Figure 5. Evaluation Results of Models with Embedding Methods

An examination of Figure 5 shows that the highest precision value, 71.60%, is achieved when the CNN+Attention model is combined with the GloVe embedding method. However, when considering precision alone, the model demonstrates high accuracy in emotion classification. Nonetheless, the relatively lower F1-score indicates that the recall performance remains

limited. In terms of F1-score, the most successful model is the BiGRU+Attention architecture combined with the GloVe embedding method. This combination achieved an F1-score of 56.11%, demonstrating the most balanced and effective performance. Particularly in multi-label datasets with imbalanced distributions, the F1-score provides a more meaningful evaluation since it accounts for both

recall and precision. Similarly, the BiLSTM+Attention architecture, when combined with the GloVe embedding method, achieved an F1-score of 54.64%, also indicating strong performance. On the other hand, the LSTM model generally yielded lower F1-scores compared to the other models. Specifically, when used with the Word2Vec embedding method, the model appeared unable to effectively learn the multi-label structure. For the GRU+GloVe model, the precision value was calculated as 65.77%, indicating that approximately two-thirds of the predicted labels were correct. This result particularly indicates that false positives are relatively low and that the

model performs labeling with high accuracy. However, the F1-score remained at 53.97%. This score demonstrates that the model did not achieve complete success, primarily due to the imbalance in the dataset and its relatively low recall performance in certain classes. Table 4 presents a comparative analysis of the quantitative metrics, such as precision and F1-score, between the results of the proposed model and existing literature studies based on the 28 emotion classes in the GoEmotions dataset.

Table 4. Comparative performance analysis of the proposed models and existing studies on the 28-class GoEmotions dataset

Reference	Datasets	Precision	F1 score	Number of Emotion Tags
Kumar et al. [2]	CNN	%41	%43	28
	RNN	%40	%41	28
	BERT	%40	%46	28
Birannavar et al. [3]	Naive Bayes	60	44	28
Gonzales et al. [5]	FastText+DNN Pool	42	45	28
	FastText+BiLSTM	44	51	28
Zhou et al. [6]	EQN	56	52	28
Olah et al. [9]	XGBoost	47	44	28
Proposed Study	CNN+ GloVe	67.79	50.22	28
	CNN+FastText	69.64	52.59	28
	CNN+ Word2Vec	68.33	51.97	28
	GRU+ GloVe	65.77	53.97	28
	GRU+FastText	67.74	52.48	28
	GRU+Word2Vec	68.80	52.32	28
	LSTM+ GloVe	67.44	41.67	28
	LSTM+FastText	65.17	37.82	28
	LSTM+Word2Vec	67.62	35.24	28
	CNN + ATTENTION + GloVe	71.60	48.49	28
	CNN + ATTENTION + FastText	68.52	52.46	28
	CNN + ATTENTION + Word2Vec	67.98	52.26	28
	BiGRU + ATTENTION + GloVe	65.34	56.11	28
	BiGRU + ATTENTION + FastText	64.61	54.04	28
	BiGRU + ATTENTION + Word2Vec	61.58	53.38	28
	BiLSTM + ATTENTION + GloVe	67.66	54.64	28
	BiLSTM + ATTENTION + FastText	66.78	54.45	28
	BiLSTM + ATTENTION + Word2Vec	64.18	54.35	28

Based on the results detailed in Table 4, a significant improvement is observed in the performance of the proposed models compared to the existing literature, particularly in terms of precision and F1-score metrics. In multi-label datasets with imbalanced distributions like GoEmotions, increasing the precision rate is critical for minimizing false positives. While the literature reports precision values of 41% by Kumar et al. [2] and 60% by Birannavar et al. [3], the CNN+Attention+GloVe combination introduced in this study achieved a precision rate of 71.60%. This increase of over 10% compared to the closest result in the literature demonstrates that the attention mechanism effectively focuses on emotion-laden keywords within the text.

In terms of F1-score, the 52% success rate reported by Zhou et al. [6] using the EQN model was improved to 56.11% with the BiGRU+Attention+GloVe architecture in this study. The bidirectional information flow of the BiGRU architecture captures contextual depth, enabling more stable classification within the complex 28-class emotion space. Furthermore, unlike the singular

approaches in the literature, three different embedding methods were systematically tested in this work. The

findings reveal that the synergy between the GloVe embedding method and bidirectional structures is superior to Word2Vec and FastText in learning multi-label structures. In particular, the performance decline observed when the LSTM model is used with Word2Vec confirms that the compatibility of static embedding methods with the model architecture is a decisive hyperparameter in emotion analysis tasks. Table 5 presents the emotion labels predicted by the model.

Table 5. Comparative Analysis Based on Model Prediction Outcomes

Text	True Labels	Model + Embedding	Predicted Labels
“100%! Congrats on your job too!”	Gratitude	BiGRU+Attention+Glove	Admiration, gratitude
		CNN+Glove	<None>
“It’s great that you’re a recovering addict.”	Admiration, curiosity	BiLSTM+Attention+Glove	Admiration, curiosity
		CNN+Word2Vec	Admiration
“If the pain doesn’t go away after 4 hours.”	Sadness, neutral	BiGRU+Attention+Word2vec	Sadness
		CNN+FastText	<None>

Examining Table 5, it is observed that different embedding methods and deep learning architectures exhibit varying levels of success in multi-label emotion classification. For example, for the expression “100% ! congrats on your job too!”, the BiGRU+Attention+GloVe model can produce the predictions “admiration, gratitude” consistent with the ground-truth label, whereas the CNN+GloVe model fails to produce any label prediction. Similarly, in the example “It’s great that you’re a recovering addict,” the BiLSTM+Attention+GloVe model captured both correct labels, while CNN+Word2Vec predicted only the “admiration” label. In the third example, for the expression “If the pain doesn’t go away after 4 hours,” the BiGRU+Attention+Word2Vec model correctly predicted only the “sadness” label and missed the “neutral” label; CNN+FastText produced no prediction. These results demonstrate that BiGRU- and BiLSTM-based architectures enriched with attention mechanisms, in combination with embedding methods, exhibit stronger performance and provide a marked advantage over CNN-based models in capturing correct labels within a multi-

label framework. When the results were examined, it was observed that some models failed to predict any emotion labels for certain examples. This phenomenon was particularly common in basic CNN- and LSTM-based architectures, whereas advanced models incorporating attention mechanisms were notably better able to capture the multi-label structure. For example, the BiLSTM+Attention architecture produced results closest to the ground-truth labels in most examples, which demonstrates the contribution of contextual learning to emotion classification performance. Furthermore, the GloVe embedding method was found to provide more stable predictions across most models. To present the label-wise performance of the models used in the study in greater detail, the successes achieved for different emotion labels were comparatively analyzed for each model. In this context, the CNN, BiGRU+Attention, and BiLSTM+Attention architectures were evaluated together with different word embedding methods, and the results were visualized. Figure 6 shows a heatmap of the models’ F1 score and precision performance across 28 labels (gratitude, admiration, sadness, neutral, etc.).

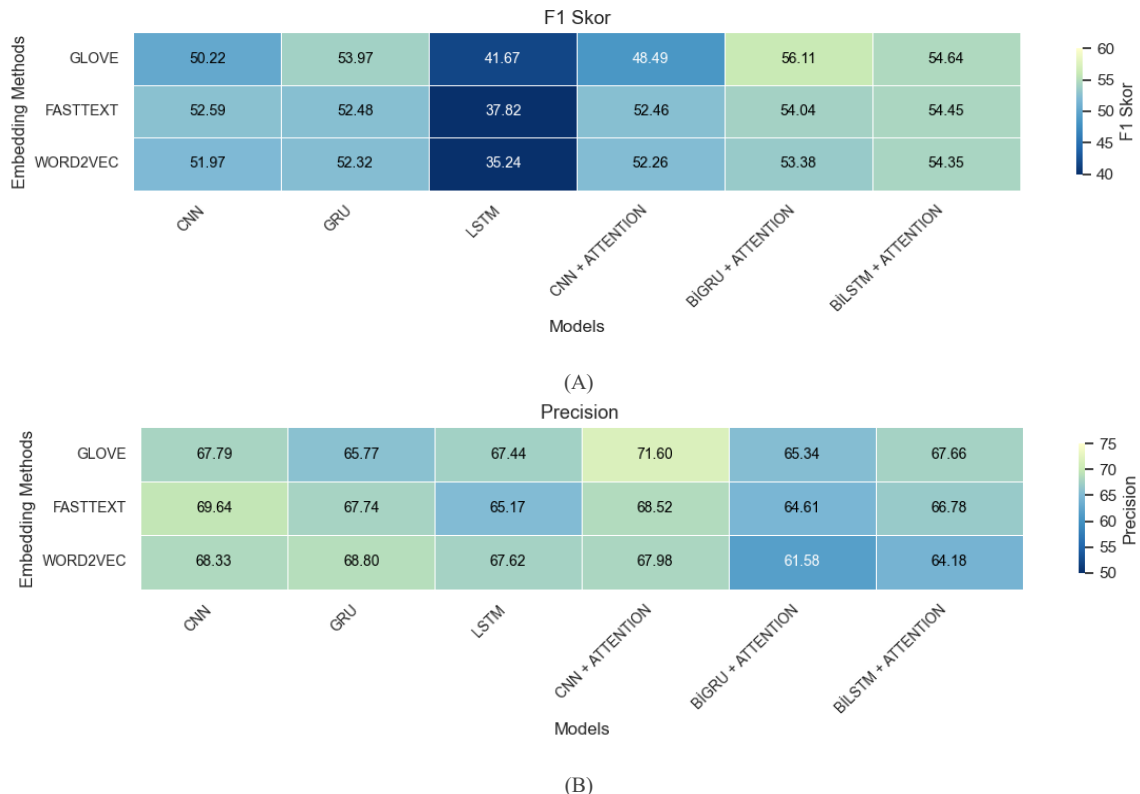
**Figure 6.** Heatmap showing F1-score and precision for each model across 28 emotion labels.

Figure 6 shows that **BiGRU- and BiLSTM-based models enriched with attention mechanisms** achieve markedly higher performance across all labels compared with CNN-based models. In particular, for the **gratitude** and **admiration** labels, the BiGRU+Attention+GloVe model attained the highest performance at 72% and 69%, respectively. CNN-based models exhibited relatively lower performance; for example, CNN+GloVe recorded the lowest score of 38% on the sadness label. Overall, the heatmap results indicate that the combined use of embedding methods and attention mechanisms substantially improves label-wise performance in the multi-label emotion classification task.

The results of this study demonstrate that high model performance can be maintained across the full spectrum of 28 emotion classes without the need to reduce the label count to 7, which is a common limitation in existing literature. Specifically, the qualitative analysis in Table 5 reveals a critical finding: while baseline CNN and LSTM architectures often fail to generate any labels for complex expressions (resulting in 'None' predictions), the attention-augmented models successfully uncover 'hidden emotional layers' within the text. This proves that the proposed **BiGRU+Attention+GloVe** architecture does not merely offer a statistical increase in F1-scores but also provides a more robust semantic understanding required for multi-label classification. Furthermore, the observed synergy between bidirectional structures and GloVe embeddings serves as a methodological contribution, confirming that hyperparameter compatibility is as decisive as model depth in high-dimensional emotion analysis. These findings establish a new benchmark for researchers seeking to classify fine-grained emotions in imbalanced and sparse datasets.

4. DISCUSSION AND CONCLUSION

Within the scope of this study, various deep learning architectures and word embedding methods were comprehensively compared to address the multi-label emotion classification problem. Experimental analyses revealed that the architectures exhibit differing sensitivities to contextual emotional expressions. In particular, BiGRU- and BiLSTM-based models augmented with attention mechanisms achieved higher F1 scores and demonstrated more balanced performance in multi-label classification compared to baseline CNN, LSTM, and GRU architectures. This finding indicates that the attention mechanism's ability to focus on emotionally salient words makes a meaningful contribution to classification performance. In future work, we plan to investigate the effects of Transformer-based architectures (e.g., BERT, RoBERTa, GPT), which excel in advanced contextual modeling, on multi-label emotion analysis. Additionally, to mitigate the class imbalance present in the dataset, strategies such as reordering or removing labels with low sample counts—particularly the neutral class—will be considered. These directions are expected to further enhance the accuracy and generalizability of multi-label emotion classification. This study was conducted solely on the GoEmotions dataset, and the pronounced class imbalance in the dataset) particularly

the low sample counts for certain labels (may have affected the model's learning and the evaluation metrics. To address these limitations in future work, we recommend validating the models on different data sources, applying resampling or class-weighting techniques, or reorganizing labels to ensure more balanced training.

REFERENCES

- [1] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, ve S. Ravi, "GoEmotions: A Dataset of Fine Grained Emotions", 03 Haziran 2020, *arXiv*: arXiv:2005.00547. doi: [10.48550/arXiv.2005.00547](https://doi.org/10.48550/arXiv.2005.00547).
- [2] V. Kumar, A. Bansal, ve U. Gupta, "Comparative Analysis of Text based Emotion Detection on GoEmotions Dataset", içinde *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Ara. 2023, ss. 364-369. doi: [10.1109/ICAC3N60023.2023.10541692](https://doi.org/10.1109/ICAC3N60023.2023.10541692).
- [3] N. Birannavar, P. Prasad, G. Dandgal, M. Ganiger, P. Redekar, ve V. Badiger, "Performance Evaluation of Sentiment Analysis on Reddit Comments: Insights and Improvement Opportunities for Naive Bayes, SVM, and BERT Models", içinde *2025 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, Şub. 2025, ss. 1-5. doi: [10.1109/ICCECE61355.2025.10940395](https://doi.org/10.1109/ICCECE61355.2025.10940395).
- [4] A. Younas, S. Riaz, S. Ali, R. Khan, M. Ullah, ve D. Kwak, "Stacked LSTM Model for Contextual Correlation Detection Among Multiple Emotions", *IEEE Access*, ss. 1-1, 2025, doi: [10.1109/ACCESS.2025.3582764](https://doi.org/10.1109/ACCESS.2025.3582764).
- [5] N. Alvarez-Gonzalez, A. Kaltenbrunner, ve V. Gómez, "Uncovering the Limits of Text-based Emotion Detection", 2021. doi: [10.18653/v1/2021.findings-emnlp.219](https://doi.org/10.18653/v1/2021.findings-emnlp.219).
- [6] J. Zhou, S. Luo, ve H. Chen, "Expansion Quantization Network: An Efficient Micro-emotion Annotation and Detection Framework", 30 Mayıs 2025, *arXiv*: arXiv:2411.06160. doi: [10.48550/arXiv.2411.06160](https://doi.org/10.48550/arXiv.2411.06160).
- [7] E. S. Yusifov ve I. S. Sineva, "An Intelligent System for Assessing the Emotional Connotation of Textual Statements", içinde *2022 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*, May. 2022, ss. 1-8. doi: [10.1109/WECONF55058.2022.9803516](https://doi.org/10.1109/WECONF55058.2022.9803516).
- [8] Y. Ni ve W. Ni, "A multi-label text sentiment analysis model based on sentiment correlation modeling", *Front. Psychol.*, c. 15, Ara. 2024, doi: [10.3389/fpsyg.2024.1490796](https://doi.org/10.3389/fpsyg.2024.1490796).
- [9] J. Olah, S. Baruah, D. Bose, and S. Narayanan, "Cross Domain Emotion Recognition using Few Shot Knowledge Transfer," Oct. 11, 2021, *arXiv*: arXiv:2110.05021. doi: [10.48550/arXiv.2110.05021](https://doi.org/10.48550/arXiv.2110.05021).

- [10] A. Bhardwaj, N. A. Wasi, and M. Abulaish, "Unlocking Emotions in Text: A Fusion of Word Embedding and Lexical Knowledge for Emotion Classification".
- [11] Y. Sazaki, A. F. Yusof, A. R. C. Hussin, and C. W. Howe, "Effect of Training Epochs on Multi-Label Emotion Classification Using RoBERTa with Focal Loss," in *2025 19th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Oct. 2025, pp. 1–4. doi: [10.1109/TSSA68467.2025.11303830](https://doi.org/10.1109/TSSA68467.2025.11303830).
- [12] M. M and S. Gokila, "Beyond Single Emotions: A Dual-Label Classification System for Enhanced Emotion Understanding in NLP Using Transformer-Based Models," in *2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2025, pp. 1491–1497. doi: [10.1109/ICOEI65986.2025.11013237](https://doi.org/10.1109/ICOEI65986.2025.11013237).
- [13] Q. Peng, Y. Wan, and X. Zeng, "Multi-Label Emotion Recognition Model Integrating Multi-Perspective Common Sense Knowledge," in *2025 7th International Conference on Natural Language Processing (ICNLP)*, Mar. 2025, pp. 421–425. doi: [10.1109/ICNLP65360.2025.11108467](https://doi.org/10.1109/ICNLP65360.2025.11108467).
- [14] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: [10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162).
- [15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017, doi: [10.1162/tacl_a_00051](https://doi.org/10.1162/tacl_a_00051).
- [16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sept. 07, 2013, *arXiv*: [arXiv:1301.3781](https://arxiv.org/abs/1301.3781). doi: [10.48550/arXiv.1301.3781](https://doi.org/10.48550/arXiv.1301.3781).
- [17] S. K. Putri, A. Amalia, and T. F. Abidin, "Sentiment Analysis Multi-Label of Toxic Comments using BERT-BiLSTM Methods," in *2024 International Conference on Electrical Engineering and Informatics (ICELTICs)*, Sept. 2024, pp. 120–124. doi: [10.1109/ICELTICs62730.2024.10776338](https://doi.org/10.1109/ICELTICs62730.2024.10776338).
- [18] D. Alnahas, F. Aşık, A. Kanturvardar, and A. M. Ülkün, "Opinion Mining Using LSTM Networks Ensemble for Multi-class Sentiment Analysis in E-commerce," in *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, Dec. 2022, pp. 1–6. doi: [10.1109/IISEC56263.2022.9998264](https://doi.org/10.1109/IISEC56263.2022.9998264).
- [19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," Dec. 11, 2014, *arXiv*: [arXiv:1412.3555](https://arxiv.org/abs/1412.3555). doi: [10.48550/arXiv.1412.3555](https://doi.org/10.48550/arXiv.1412.3555).
- [20] A. N. Ma'Aly, D. Pramesti, and H. Fakhurroja, "Comparative Analysis of Deep Learning Models for Multi-label Sentiment Classification of 2024 Presidential Election Comments," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, July 2024, pp. 502–507. doi: [10.1109/ICICoS62600.2024.10636889](https://doi.org/10.1109/ICICoS62600.2024.10636889).
- [21] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, July 2005, doi: [10.1016/j.neunet.2005.06.042](https://doi.org/10.1016/j.neunet.2005.06.042).
- [22] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*: [arXiv:1706.03762](https://arxiv.org/abs/1706.03762). doi: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762).
- [23] N. T. M. Trang and N. N. Hung, "BiGRU-Attention Sentiment Analysis for Enhancing Task-Oriented Chatbots," in *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, Jan. 2024, pp. 476–483. doi: [10.1109/ICMCSI61536.2024.00075](https://doi.org/10.1109/ICMCSI61536.2024.00075).
- [24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, July 2009, doi: [10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002).
- [25] M. Coskun, "Enhancing Reliability in Deep Learning Diagnosis of Brain Tumors Using Grad-CAM," *TDFD*, vol. 15, no. 2, pp. 39–48, Jul. 2026, doi: [10.46810/tdfd.1749282](https://doi.org/10.46810/tdfd.1749282).
- [26] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," Aug. 13, 2020, *arXiv*: [arXiv:2008.05756](https://arxiv.org/abs/2008.05756). doi: [10.48550/arXiv.2008.05756](https://doi.org/10.48550/arXiv.2008.05756).