# Exploring Deep Learning Architectures for Emotion Recognition: Adapting LSTM and BERT Models from Reddit to Twitter

**Name:** Saketh Reddy Atla
**Department:** Data Analytics
**Student No:** x22218700
**Email:** x22218700@ncirl.ie

*Abstract*—Emotion recognition has gained a lot of interest recently in natural language processing tasks with the understanding and analysis of user-generated content on social media. In fact, this project will compare the performance of a Long Short-Term Memory (LSTM) model with a BERT model for the recognition of emotion from two datasets: the GoEmotions dataset from Reddit and a dataset from Twitter. The GoEmotions dataset comprises Reddit comments, which have been annotated with 27 emotion categories and subsequently simplified into 7 main emotion categories. The Twitter dataset has tweets annotated with 6 emotion labels. Both datasets are pre-processed using custom techniques, which include removal of mention, URL, and non-alphabetic characters, and converting emojis to text. In this context, it was proposed to use an LSTM model that embeds the Word2Vec and uses a refined architecture, with the inclusion of bidirectional LSTM layers, dropouts, and dense layers. We use the pre-trained BERT-base-uncased model and tune it using the sequence classification task. Both the balanced and imbalanced versions of the GoEmotions and Twitter datasets were used for training and testing the models. Performance assessment is done using classification reports and confusion matrices. Results in all cases favor the effectiveness of both LSTM and BERT in capturing the emotional information from the text data across different domains, with BERT generally being ahead of LSTM. It will emphasize the importance of data balance, domain adaptation, and a deep learning architecture's choice for emotion recognition tasks in NLP.

## I. Introduction

In this fast-evolving digital communication landscape, social media has emerged as a treasure trove for emotional expression. These are effective means of collecting vast humangenerated data, either in the form of a long context-rich conversation on Reddit or short informal ones on Twitter, hence suitable for the study of human emotions. However, these variations in styles and domain-specific nuances render a big challenge to general emotion recognition models. As emotional text data is growing in volume with each passing day, it is important to develop robust and adaptive Natural Language Processing techniques that help capture and interpret the emotional content embedded in the data pertaining to the diversified domains.

There are two most important deep learning architectures: the Transformer-based models and the Recurrent Neural Network models. More specifically, this paper discusses Bidirectional Encoder Representations from Transformers, or BERT, which presents a new way to encode pre-training and finetuning mechanisms toward excellent performance in most NLP tasks. On the other hand, LSTM models still remain great in the setting of long-term dependency and the sequential information of text data. Specifically, we compare the performance of such models in a cross-domain emotion recognition task.

1) How do Transformer-based models like BERT and recurrent neural network models like LSTM do when it comes to applying emotion recognition from a dataset that is specific to one domain (GoEmotions from Reddit) to a domain that is more general and has a lot of different styles (Twitter)?
2) What problems and restrictions do BERT and LSTM models face when trying to adapt to the different types of language and emotions found in Twitter data? How do these problems affect the precision and dependability of emotion recognition on various platforms?

To answer these questions, we train our models on the GoEmotions dataset because it is our source domain. We test our models using a Twitter dataset, as a representation of a much more diverse and noisy target domain. We then check the results to try to understand the advantages and disadvantages of each model architecture in capturing emotional information across different domains.

### *A. The main goals of this research are:*

- Preprocess and prepare the GoEmotions and Twitter datasets for the emotion recognition task by solving the current issues: data imbalance and domain-specific noise. • Develop and fine-tune BERT- and LSTM-based emotion recognition models through the use of transfer learning and domain adaptation techniques.
- For instance, classification reports and confusion matrices can be used to measure and compare the performance of the model in both datasets. It will then cross-analyze results based on the implications of results on crossdomain emotion recognition, highlighting the strengths and limitations of each of the model architectures and providing insight into challenges in adapting emotion recognition models to different platforms.

The rest of the paper as follows: in Section 2, related works in the fields of emotion recognition and cross-domain adaptation are reviewed;. Section 3 explains the datasets and the pre processing applied. Section 4 gives the methodology, the model architecture, and the experimental setup. Section 5 presents and discusses the results while underpinning the most important findings and insights. Finally, Section 6 concludes the paper and outlines potential future directions for research in this area. Solving the challenge of cross-domain emotion recognition and comparing the efficiency of BERT and LSTM models, the work contributes to the extension of the NLP technique for understanding and analyzing emotional content across different social media platforms. The result of this work will be useful in building effective and adaptable emotion recognition systems to process and interpret the massive volume of emotional textual data generated over the internet. Additionally, such results would have far-reaching implications for the use of NLP models generally across different domains; they point out the need to consider problems that are domain-specific and therefore make a demand for domain adaptation strategies necessary.

## II. RELATED WORK

Recently, the research of automatic emotion recognition from text has become a popular research topic in the area of natural language processing due to its wide application in many different fields, such as sentiment analysis, opinion mining, or mental health monitoring. This section attempts to present a more coherent view of related work, detailing methods, datasets, and challenges taken by previous research, followed by a critical evaluation of the relevance to our research questions regarding cross-domain emotion recognition using Transformer-based models, such as BERT, and recurrent neural network models, such as LSTM. Yang [1] gives an overview of the methods of text classification and mainly focuses that currently research is giving more and more attention to deep learning models, constructed mainly with the help of Convolutional Neural Networks and Recurrent Neural Networks. The paper expounds on the challenges currently facing the text classification model with regard to feature extraction methods, dataset shortage for tasks that are complex, and lack of the theoretical understanding of DL model structures. Although these findings are of great value and give valuable insight into the state of the art of this model and its most pressing challenges, it does not address the specific problem of cross-domain adaption, or the comparison, among Transformer-based models, such as BERT and RNN models, such as LSTM.

Some studies have compared the performance of different machine learning algorithms in this area. Yang [2] examines the classification algorithms Support Vector Machines (SVM), Naive Bayes, Random Forest, and K-Nearest Neighbors (KNN) in mobile application descriptions. The study proves that the models work for classification on a given set of text, yet it does not indicate the difficulties or challenges that could be encountered when the models are adapted to different languages and emotion styles within a large number of domains using Twitter data.

Research in emotion recognition has been based on deep learning architectures and word embedding techniques.

In a related development, Cahyani et al. [4] show the performance of CNN concatenated with different word embedding methods: Word2Vec, BERT, GloVe, etc., in emotion detection over Indonesian text. In fact, both show the effectiveness of deep learning architectures and the importance of word embeddings in emotion recognition tasks[5], but they are for a single language—in this case, Indonesian—and do not consider the cross-domain adaptation challenge, which is crucial in our research questions. Sayyed et al. [6] are much closer to the goal, but their focus is on recognizing text-based dialogue with the expression of emotions, which can be used for human-computer interaction and a system for monitoring mental health. Still, a framework of this kind has not yet been directly compared for the same goal over different domains, as in our work.

The main progress for research in emotion recognition can be represented by the two trends: seminal progress in constructing large-scale. Demszky et al. [9] describe GoEmotions—a dataset of 58,000 English Reddit comments annotated with 27 emotion categories. We show the high quality of annotation by providing strong baselines for fine-grained emotion prediction, thus showing generalization of the dataset across domains and taxonomies.

The dataset may be very relevant to our work, which aims to understand the performance of models developed on the GoEmotions dataset in recognizing emotions in another broad domain. The GoEmotions dataset is the main resource in our work to investigate the cross-domain applicability of emotion recognition models.

Transformer-based models such as BERT have revolutionized natural language processing by allowing the pretraining of deep, bidirectional representations from unlabeled text. Devlin et al. [8] introduced BERT and demonstrated its state-of-theart effectiveness for a wide range of NLP tasks, including text classification. This pioneering work had a foundational contribution to tasks like emotion recognition, and hence, from then on, BERT and its derivatives are proliferated.

The original BERT paper did not deal with these problems of cross-domain adaptation in specific, or its comparison with RNN models like LSTM in the context of emotion recognition, which forms our main research question. This work improves the BERT architecture and assesses domain adaptability in the context of emotion recognition.

On the other hand, whereas the study by Mollah et al [3] proposes using RoBERTa, XGBoost, and re-weighting for feedback classification into e-commerce products given by consumers, the study neither benchmarks the performance of the proposed model across diverse domains nor compares it with LSTM models, which form the core of our research questions. Mahima et al. [11] proposed a hybrid approach to detect multiple emotions by analyzing both context and semantics. Their study focused on the precise capture of multiple emotions based on context, did not analyze the performance of the proposed model across different domains, and did not compare with important considerations from our research, such as BERT and LSTM models.

Zang et al. [10] proposed the MLformer model for multilabel classification of long Chinese text using the Longformer architecture. While the next paper is within the same domain of long text classification, it does not deal with cross-domain emotion recognition, a comparison of the proposed model with

BERT, or the comparison with LSTM—important answers to our research questions.

Salsabiila et al. [12] compared a Conv-LSTM architecture with weighting techniques provided by FastText and Word2Vec for text classification emotion detection in the context of tweets written in the Indonesian language. Although their work is very good in showing the effectiveness of some weighting techniques and the Conv-LSTM method for emotion detection, it is on a single language (Indonesian), and it does not cover the challenges of cross-domain adaptation, which is at the heart of our work. Sheng et al. [13] have recently proposed a new strategy for English text classification that combines CNN with SVM to address the sparseness and latitude problems related to text representation. They show that the newly proposed CNN+SVM model is effective for text classification tasks. Not considered is the targeting of the problems of emotions and cross-domain adaptation in our questions. In general, the related works cannot make more than a reference to the milestone achievements in the area of emotion recognition using different machine learning algorithms, deep learning architectures, and word embedding techniques. On the other hand, great deals still lack in the literature involving cross-domain evaluation on emotion recognition techniques, especially when considering the application of a model developed from a domain-specific dataset over a more diversified and stylistically varied domain.

Moreover, the cross comparison between the Transformerbased models, such as BERT, and the RNN models, such as LSTM, offered limited studies on cross-domain emotion recognition.

The purpose of the following paper is to cross-check the performance of BERT and LSTM models in cross-domain emotion recognition using GoEmotions from Reddit and a dataset taken from Twitter.

Through the above analysis of the challenges and limitations that the models face while adapting to language styles and emotional expressions, it contributes to the further development of emotion recognition techniques and provides a look at developing more robust and adaptable models for real-world applications. Our research extends this earlier work, which has been laid on the GoEmotions dataset [9] and BERT model [8], from the capability to deal with the tasks of cross-domain emotion recognition and comparison of Transformer-based and RNN models. Our studies on emotion recognition are thus expected to drive, therefore, the development of effective systems for emotion recognition further, while systems dealing with the large linguistic diversity and emotional complexity of the Internet are expected to further enhance the potential of processing and interpreting the rapid growth volume of emotional text data generated online. More importantly, our findings could contribute to the general understanding of the strengths and weaknesses of different deep learning architectures in cross-domain adaptation, which will guide further research in this direction.

# III. DATA MINING METHODOLOGY:

## A. *Business Understanding:*

This study involves performance and adaptability of the Transformer-based models, majorly BERT, and recurrent neural network models, such as LSTM, in cross-domain emotion recognition. This paper tries to answer the research questions and realizes the research objectives by understanding how well these models trained on a domain-specific dataset (GoEmotions from Reddit) adapt to another dataset that is quite more diverse and stylistically variable (Twitter). It further identifies and analyzes the challenges and constraints facing BERT and LSTM models in their generalization ability to the variance in language style and emotion used in Twitter data.

## B. *Data Understanding*

The datasets collected and explored for this study include the GoEmotions dataset and a Twitter dataset. The GoEmotions dataset comprises 58,000 Reddit comments labeled with 27 varieties of emotion. The dataset is categorized into ”7,” giving a finer-grained representation of emotions. The Twitter dataset consists of tweets, labeled with six basic emotions: anger, disgust, fear, joy, sadness, and surprise. Based on the given datasets, the various aspects pertaining to them are analyzed below, which included the distribution of the emotions over the data, the distribution of the length of the samples, and the presence of domain-specific language and noise.

## C. *Data Preparation*

Extensive preprocessing was done on the text data from both datasets to make them compatible with BERT and the LSTM model. Common pre-processing steps applied to text data with the use of the NLTK and emoji libraries on the text data include tokenization, lowercasing, stop-word removal, and lemmatization. All these steps help clean the text data of irrelevant information and normalize the text for the model's better performance. In the case of a Twitter dataset, it receives custom preprocessing due to the special features of the tweets: handling mentions, URLs, and non-alphabetic characters and also converting emojis to their text form. Thus, in a generalized manner, all the preprocessing steps do nothing more than a reduction of noise so as to focus on relevant textual content for emotion recognition.

One of the aspects of difficulty with respect to data preparation is the class imbalance present in the two datasets. This was done using data augmentation techniques, particularly synonym replacement and spelling augmentation. This will help to properly balance the distribution of emotions by increasing the representation of the minority emotion classes to avert bias of the model towards the majority classes. The latter augmented datasets such that it balanced the distribution of emotions, thereby allowing models to learn from much broader examples, thus becoming capable of generalization.

TABLE I
GoEmotions Dataset Before Balancing

| Class | Count |
|---|---|
| anger | 7022 |
| disgust | 1013 |
| fear | 929 |
| joy | 21733 |
| sadness | 4032 |
| surprise | 6668 |
| neutral | 17772 |

TABLE II
GoEmotions Dataset After Balancing

| Class | Count |
|---|---|
| anger | 39434 |
| disgust | 21733 |
| fear | 21733 |
| joy | 21733 |
| sadness | 21733 |
| surprise | 21733 |
| neutral | 21733 |

Finally, the preprocessed and augmented datasets were split into training and testing sets, making them ready for the model training and evaluation process. The training sets were, therefore, used in the training of both BERT and LSTM models, and the testing sets were for checking the performance and generalization capabilities of the models.

TABLE III
Twitter Dataset Before Balancing

| Class | Count |
|---|---|
| Joy | 6761 |
| sadness | 5797 |
| Anger | 2709 |
| Fear | 2373 |
| Love | 1641 |
| Surprise | 719 |

TABLE IV
Twitter Dataset After Balancing

| Class | Count |
|---|---|
| sadness | 6761 |
| Fear | 6761 |
| Love | 6761 |
| Surprise | 6761 |
| Anger | 6761 |
| Joy | 6761 |

### *D. Modelling*

*1) Introduction to BERT :* BERT is a state-of-the-art model in the arena of transformers, which has brought up a great reformation in the whole domain of natural language processing. BERT, developed by Google, is pre-trained on a big corpus of unlabeled text in order to learn deep bidirectional representations of language. The major advantage of BERT is its ability to capture context—be it the left or right context—of each word in a given sentence. Therefore, it better captures the language of nuance and semiotics. BERT further supports fine-tuning to downstream other NLP tasks, such as sequence classification, making it suitable for emotion recognition.

*2) Introduction to LSTM:* The Long Short-Term Memory is a type of Recurrent Neural Network architecture that has found wide application in the fields that involve sequential data, such as text. The main structure of the LSTM network is designed to capture long-term dependencies and address the vanishing gradient problem associated with conventional RNNs. The model consists of a memory cell and gate mechanisms: an input gate, a forget gate, and an output gate. This feature allows them to remember relevant information and discard irrelevant details, so LSTM networks are good at modeling contextual dependencies inherent in emotional text data.

*3) Rationale for Employing BERT and LSTM::* The framework of the study is thus comparative in terms of the model performance of the Transformer-based model, BERT, and that of the recurrent neural network model, LSTM, with respect to cross-domain emotion recognition. Such an evaluation is going to give insight into the strengths and limitations of both architectures in adapting to very diverse domains and handling challenges due to very different language styles and emotional expressions.

### *E. Model Setup and Architecture:*

*1) BERT Model Setup::* The BERT model was initialized with the base model 'bert-base-uncased'. This model was pre-trained on a large corpus of unlabeled text with the intention of modeling general language usage. For fine-tuning toward emotion recognition, the model was further pre-trained for sequence classification on the GoEmotions and Twitter datasets. Fine-tuning involves putting a further classification layer on top of the pre-trained BERT model and training that layer for the specific task at hand, which is emotion recognition in this case. In addition, for easy input handling by BERT, two custom classes have been created: CustomDataset and TwitterDataset. These classes are responsible for text preprocessing, tokenization, and format of the input data in a form compatible with BERT's requirements.

*2) LSTM Model Architecture::* The model was designed to correspond to an LSTM bidirectional architecture for the LSTM model. It consists of an embedding layer, which is responsible for representing each word in the input sequence with a dense vector. An embedding matrix for text data is created with pre-trained Word2Vec embeddings that will give a dense vector representation to the input words and enable the model to capture semantic similarities among words. The input sequence processes using bidirectional LSTM in a twoway step, thereby making it capable of looking at past and future contexts in framing. Further, it applies spatial dropout, where at a time, whole word vectors are dropped randomly to avoid overfitting. Dense layers are added on top of the LSTM layer with appropriate activation functions for classification.

### *F. Training and Evaluation:*

Training has been conducted for both balanced and imbalanced datasets for BERT and LSTM models for GoEmotions and Twitter. This helps capture the overall performance of the model in either of the class distribution scenarios. In this manner, both datasets, with balanced values of a better distribution of emotions, and those more naturally distributed were achieved. We trained and evaluated the models on both balanced and imbalanced versions to investigate how the class imbalance influences the model's performance and generalization in the various domains. Then training was done, where the preprocessed tokenized text data along with its emotion labels were fed into the models. The models were then put under backpropagation and optimization techniques to teach the mapping of input sequences to the emotion classes. Hyperparameters, such as learning rate and batch size, had their optimization against the number of training epochs, and in the.

### G. *Model Evaluation Metrics:*

The performance of the models was evaluated with metrics on accuracy, precision, recall, and F1-score. These are critical measures that go a long way in providing the full scope of the model's performance in the correct classification of emotion within different domains. In addition, it obtained confusion matrices to check the predictions of the models and visually search for patterns in the misclassification. This study emphasizes the importance of the BERT and LSTM models in the area of cross-domain emotion recognition through the evaluation of their performances over balanced and imbalanced datasets. Comparative analysis of the results of the models helps in understanding their strengths, limitations, and adaptability to different language styles and emotional expressions.

## IV. EVALUATION

### *A. Twitter Dataset*

The models are to be evaluated on both balanced and imbalanced dataset variations from Twitter to give an appropriate view of their performance.

*1) LSTM:* The LSTM model on the balanced dataset gave a test accuracy of 89.92%; hence, an overall very good performance. The classification report showed high precision, recall, and F1-scores for all emotion classes, from 0.81 to 0.95. The classification model worked properly in both cases of sadness and joy classification, with a class F1 of 0.91 in both classes. This might be the reason why this balanced dataset probably helped attain the consistency of performance across all emotions. The LSTM model obtained 91.10% of test accuracy on the imbalanced Twitter dataset. The classification report, however, shows there are some differences in class performance: it really excels in the classification of sadness (class 0) and anger (class 3), where it attains F1 scores of 0.96 and 0.93, respectively. But for love (class 2) and joy (class 1), it scores not as great: 0.82 and 0.77, respectively. There were imbalances in class distribution, and it was probably a strong deterrent for the model to be able to classify the underrepresented classes well.

TABLE V

CLASSIFICATION REPORT FOR LSTM (BALANCED TWITTER DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Sadness (0) | 0.95 | 0.88 | 0.91 |
| Joy (1) | 0.94 | 0.89 | 0.91 |
| Love (2) | 0.90 | 0.91 | 0.91 |
| Anger (3) | 0.90 | 0.93 | 0.91 |
| Fear (4) | 0.81 | 0.90 | 0.85 |
| Surprise (5) | 0.90 | 0.89 | 0.90 |
| Accuracy | | 0.90 | |
| Macro avg | 0.90 | 0.90 | 0.90 |
| Weighted avg | 0.90 | 0.90 | 0.90 |

TABLE VI

CLASSIFICATION REPORT FOR LSTM (IMBALANCED TWITTER DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Sadness (0) | 0.95 | 0.96 | 0.96 |
| Joy (1) | 0.74 | 0.80 | 0.77 |
| Love (2) | 0.74 | 0.92 | 0.82 |
| Anger (3) | 0.96 | 0.89 | 0.93 |
| Fear (4) | 0.88 | 0.84 | 0.86 |
| Surprise (5) | 0.91 | 0.93 | 0.92 |
| Accuracy | | 0.91 | |
| Macro avg | 0.86 | 0.89 | 0.88 |
| Weighted avg | 0.92 | 0.91 | 0.91 |

TABLE VII

CLASSIFICATION REPORT FOR BERT (IMBALANCED TWITTER DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Sadness (0) | 0.97 | 0.99 | 0.98 |
| Joy (1) | 0.97 | 0.96 | 0.97 |
| Love (2) | 0.85 | 0.93 | 0.89 |
| Anger (3) | 0.98 | 0.92 | 0.95 |
| Fear (4) | 0.88 | 0.94 | 0.91 |
| Surprise (5) | 1.00 | 0.72 | 0.83 |
| Accuracy | | 0.95 | |
| Macro avg | 0.94 | 0.91 | 0.92 |
| Weighted avg | 0.95 | 0.95 | 0.95 |

TABLE VIII

CLASSIFICATION REPORT FOR BERT (BALANCED TWITTER DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Sadness (0) | 0.95 | 0.95 | 0.95 |
| Joy (1) | 0.95 | 0.95 | 0.95 |
| Love (2) | 0.93 | 0.94 | 0.93 |
| Anger (3) | 0.92 | 0.96 | 0.94 |
| Fear (4) | 0.90 | 0.89 | 0.89 |
| Surprise (5) | 0.94 | 0.91 | 0.92 |
| Accuracy | | 0.93 | |
| Macro avg | 0.93 | 0.93 | 0.93 |
| Weighted avg | 0.93 | 0.93 | 0.93 |

*2) BERT:* The BERT model still performed comparatively well on both the balanced and imbalanced datasets of Twitter, with the following results: On the imbalanced dataset, the BERT model could achieve 95.00% in test accuracy, which was higher than the LSTM model. The classification report shows most of the emotion classes have high precision, recall, and F1-scores, with the lowest being 0.83 and the highest 0.98. The model worked really well with classes 0 (Sadness), 1 (Joy), and 3 (Anger), with F1-scores equal to 0.98, 0.97, and 0.95, respectively. On the contrary, it showed fairly lower performance with respect to F1 in class 5 of surprise, with a score of 0. The BERT model scored test accuracy of 93.17% on the balanced Twitter dataset. It is a bit less than that on the imbalanced dataset, but in turn, it outperformed the LSTM model. The classification report represented that the model was consistent for all classes of emotion, with F1 scores in the range of 0.89 to 0.95. Perhaps the balanced nature of the dataset allowed the model to correctly categorize emotions without strong bias toward one class.

### *B. GoEmotions Dataset*

Performance and effects of the LSTM and BERT models on both datasets in the GoEmotions dataset are evaluated in the context of emotion recognition.

TABLE IX
CLASSIFICATION REPORT FOR LSTM (GOEMOTIONS UNBALANCED DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Anger | 0.66 | 0.18 | 0.29 |
| Disgust | 0.69 | 0.22 | 0.34 |
| Fear | 0.71 | 0.15 | 0.25 |
| Joy | 0.85 | 0.66 | 0.74 |
| Sadness | 0.76 | 0.21 | 0.33 |
| Surprise | 0.69 | 0.12 | 0.20 |
| Neutral | 0.59 | 0.49 | 0.53 |
| Accuracy |  | 0.55 |  |
| Macro avg | 0.71 | 0.29 | 0.38 |
| Weighted avg | 0.72 | 0.45 | 0.52 |

*1) LSTM:* The test set of the balanced GoEmotions dataset was classified with the LSTM model at 69.58% accuracy, which generally reflects moderate performance. The classification report below shows variable precision, recall, and F1scores across different emotion classes. The disgust and fear classes were predicted well, with F1 equal to 0.90 and 0.92, respectively, but surprise and neutral were not, with F1 equal to 0.48 and 0.46, respectively. The balanced nature of the dataset makes an observation of emotions appear relatively equally distributed during training, yet model performance suggests the struggle to discern a set of emotions. This architecture obtained 59.30% of accuracy on testing over the imbalanced GoEmotions dataset. The classification report unveils enormous disparities in the performance of different emotion classes. It fared well in the classification of joy (F1 = 0.74) but rather poorly on most other emotions, like anger (F1 = 0.29), fear (F1 = 0.25), and surprise (F1 = 0.20). This likely hampers the learning and generalization ability of the model for underrepresented classes, going along with the imbalanced distribution of emotions in the dataset.

TABLE X
CLASSIFICATION REPORT FOR LSTM (GOEMOTIONS BALANCED DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Anger | 0.71 | 0.68 | 0.69 |
| Disgust | 0.89 | 0.91 | 0.90 |
| Fear | 0.91 | 0.92 | 0.92 |
| Joy | 0.81 | 0.64 | 0.71 |
| Sadness | 0.71 | 0.64 | 0.68 |
| Surprise | 0.70 | 0.37 | 0.48 |
| Neutral | 0.59 | 0.37 | 0.46 |
| Accuracy |  | 0.70 |  |
| Macro avg | 0.76 | 0.65 | 0.69 |
| Weighted avg | 0.75 | 0.65 | 0.69 |

### *C. BERT*

TABLE XI
CLASSIFICATION REPORT FOR BERT (GOEMOTIONS UNBALANCED DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Anger | 0.00 | 0.00 | 0.00 |
| Disgust | 0.00 | 0.00 | 0.00 |
| Fear | 0.00 | 0.00 | 0.00 |
| Joy | 0.46 | 0.94 | 0.62 |
| Sadness | 0.00 | 0.00 | 0.00 |
| Surprise | 0.00 | 0.00 | 0.00 |
| Neutral | 0.46 | 0.31 | 0.37 |
| Accuracy |  | 0.46 |  |
| Macro avg | 0.13 | 0.18 | 0.14 |
| Weighted avg | 0.32 | 0.46 | 0.35 |

TABLE XII
CLASSIFICATION REPORT FOR BERT (GOEMOTIONS BALANCED DATASET)

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| Anger | 0.60 | 0.55 | 0.57 |
| Disgust | 0.67 | 0.67 | 0.67 |
| Fear | 0.78 | 0.79 | 0.79 |
| Joy | 0.74 | 0.81 | 0.77 |
| Sadness | 0.18 | 0.35 | 0.24 |
| Surprise | 0.64 | 0.55 | 0.59 |
| Neutral | 0.52 | 0.55 | 0.53 |
| Accuracy |  | 0.64 |  |
| Macro avg | 0.59 | 0.61 | 0.60 |
| Weighted avg | 0.64 | 0.64 | 0.64 |

On the balanced GoEmotions dataset, the BERT model scored relatively lower compared to the LSTM model, with a test accuracy of 63.55%. From the classification report, one may notice different values in precision, recall, and F1scores among the different emotional classes. The model returned good results in the classes of joy (F1 = 0.77) and fear (F1 = 0.79), whilst it was poor in the classification of sadness (F1 = 0.24) and neutral (F1 = 0.53) emotions. Hence, the BERT model underlined some difficulties, even with the balanced nature of the dataset. With significantly low test accuracy, amounting to 45.98%, the BERT model classified the imbalanced GoEmotions dataset. The classification report is such that it failed in classifying most of the emotion classes properly, as it revealed F1-scores of 0.00 for the classes anger, disgust, fear, and surprise. The model has well-classified the classes of emotions as joy—0.62, neutral—0.37. This means that the model favored the majority classes and was not able to work out the underrepresented emotions because of the extremely imbalanced class of the dataset.

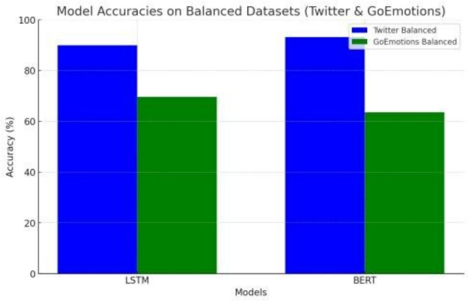

Fig. 1. Accuracy's of the Models over the Balanced Datasets

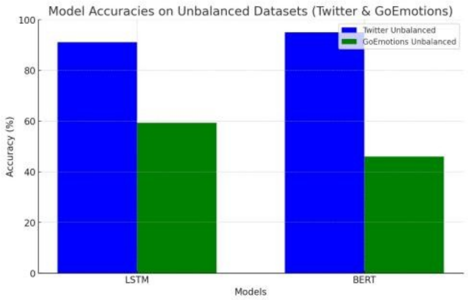

Fig. 2. Accuracy's of the Models over the Un-Balanced Datasets

# V. CONCLUSION AND FUTURE WORK

This work is targeted at performance and adaptability studies of transformer-based models, such as BERT, and recurrent neural network models, such as the LSTM, in cross-domain emotion recognition. There are two main questions that this research was based on: how well the pre-trained models perform on a domain-specific dataset (GoEmotions: Reddit) within a more diverse, stylistically varied domain (Twitter), and what challenges and limitations arise when adapting to the diversity of language styles and emotion that are innate in Twitter data, which both the BERT and LSTM models face.

Through numerous and extensive experiments and evaluations, some of the important findings and insights can be obtained. This experiment thus demonstrates domain adaptation that must be carried out in emotion recognition. Both BERT and LSTM models, once pre-trained over the GoEmotions dataset, failed terribly in their performances when deployed over the Twitter dataset, having shown the transfer challenges; this is so, especially when the target domain has its peculiarities, such as varied language styles, slang, and platformspecific features. The overall performance of the LSTM model was above that of the BERT model on both the source domain (GoEmotions) and target domain (Twitter). This may thus show that the sequential nature of LSTMs may find longterm dependencies and contextual information more useful for emotion recognition tasks across different domains.

Still, both models performed weakly in producing high cross-domain accuracy and consistency of the results even under the best settings, showing again that better fine-tuning and improvement are necessary.

The main challenge identified regarding the models is the effect of class imbalance on the developed emotion recognition model. Evaluation results for the GoEmotions and Twitter datasets show that class imbalance significantly affects the performance of BERT and LSTM models. In imbalanced datasets, the models are not able to classify minority emotion classes, leading to low precision, recall, and F1-scores. It further shows one will have to grapple with class imbalance in developing models that can enable the recognition of emotions with some degree of assurance in fairness in representation.

It is therefore fair to argue that augmentation significantly reduced class imbalance by creating more examples. The balanced versions of the Twitter dataset improved performance in terms of better accuracy and more consistent F1-scores across all emotion classes for both BERT and LSTM models.

And this is why it's extremely important for the training data to be more neutrally distributed so that models learn and generalize better.

However, balancing the dataset does not directly result in top performance in cross-domain emotion recognition. It was also noted that the models with the most difficulty adapting to the varied language styles, slang, and domain-specific expressions, characteristic of Twitter data, are BERT and LSTM.

As a result, they performed with a lower level of both accuracy and precision when compared to the performance levels on the source domain. The models were not able to pick up subtlety with regard to the data and language context of Twitter, hence making wrong classifications and reduced reliability of recognized emotions.

The following are some ways in which the adaptability of emotion recognition models can be achieved in research directions of the future in handling these problems: First, domain adaptation techniques, for instance, adversarial training or domain-invariant feature learning, can be applied to bring the distributions of the source domain and the target domain close to each other. These methods aim to train a model to recognize representations that are domain-invariant and that can generalize from the training data to the rest of the source and target domains. The generalization property of these representations can be further utilized to learn generalization beyond the seen domains. The second, fine-tuning on little labeled data in the target domain (Twitter), will allow the model to specifically adapt to language and expressions of emotion in the domain. This kind of paradigm is often called few-shot learning or transfer learning, which deals not only with the use of the obtained information in the source domain to guide learning in the target domain but with an extension to the target domain that is obtained with a reduced number of labeled examples.

The third objective could be that of thinking about further advanced architectures and techniques, including attention mechanisms, hierarchical models, or multi-task learning, in order to bring in gains on the coverage of complex emotional expressions and challenges due to diverse styles of languages. These mechanisms try to make the model alert to important contextual information and learn salient representations for robust emotion recognition.

On the other hand, domain-specific information, such as emotion lexicons, sentiment analysis techniques, pre-trained language models, and the like, in the given domain, could be used to enhance emotion recognition. In this sense, incorporation of knowledge about the domain and using external resources could lead to the inculcation of prior knowledge within the models that might help in understanding nuances and subtleties of the emotions in the specific domain. This study discussed the challenges of cross-domain emotion recognition using BERT and LSTM models. Although the models managed to learn and classify emotions in the source domain, GoEmotions, and the target domain, Twitter, the performance was degraded due to style diversification of the language, class imbalance, and domain-specific characteristics. While balancing the datasets using data augmentation techniques, it was found helpful in reducing the adverse effect of class imbalance, in which more design changes and other hyperparameters are increased for more adaptability and reliability across a domain setting. This research work was valuable in the furtherance of the research and development of strong and adaptive models within the field of emotion recognition. Development opportunities for emotion recognition systems to effectively handle the complexities and variations of emotional expressions across different domains and platforms would be given based on the resolution of identified issues and further research of the described future directions. In other words, this study partially answered the research questions by giving insights into the performance and limitations of the BERT and LSTM model in the cross-domain. In conclusion, the consequences of class imbalance and importance of domain adaptation have also been illuminated on the back of further advances in model architectures and techniques. On the other hand, this result lays the foundation for further research and development in aiming at creating a more accurate, reliable, and flexible ER system in the future.

## References

[1] L. Yang, “A Brief Introduction of the Text Classification Methods,” 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Feb. 2022, Published, doi: 10.1109/eebda53927.2022.9744845.

[2] X. Hu and R. Zhang, “Text classification based on machine learning,” 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Jun. 2022, Published, doi: 10.1109/icaica54878.2022.9844556.

[3] Md. A. Rakib Mollah, M. Md. Jahangir Kabir, Md. S. Reza, and M. Kabir, “Adapting Contextual Embedding to Identify Sentiment of E-commerce Consumer Reviews with Addressing Class Imbalance Issues,” 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), Mar. 2024, Published, doi: 10.1109/icaccess61735.2024.10499554.

[4] D. E. Cahyani, A. P. Wibawa, D. D. Prasetya, L. Gumilar, F. Akhbar, and E. R. Triyulinar, “Emotion Detection in Text Using Convolutional Neural Network,” 2022 International Conference on Electrical and Information Technology (IEIT), Sep. 2022, Published, doi: 10.1109/ieit56384.2022.9967913.

[5] D. E. Cahyani, A. P. Wibawa, D. D. Prasetya, L. Gumilar, F. Akhbar, and E. R. Triyulinar, “Text-Based Emotion Detection using CNN-BiLSTM,” 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), Oct. 2022, Published, doi: 10.1109/icoris56080.2022.10031370.

[6] H. A. Sayyed, S. Rushikesh Sugave, S. Paygude, and B. N Jazdale, “Study and Analysis of Emotion Classification on Textual Data,” 2021 6th International Conference on Communication and Electronics Systems (ICCES), Jul. 2021, Published, doi: 10.1109/icces51350.2021.9489204.

[7] T. Madhu Midhan, P. Selvaraj, M. Harshavardan Kumar Raju., M. Bhanu Prakash Reddy., and T. Bhaskar., “Classification of Mental Health and Emotion of Human from Text using Machine Learning Approaches,” 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mar. 2023, Published, doi: 10.1109/iscon57294.2023.10111973.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding,” arXiv.org, Oct. 11, 2018. https://doi.org/10.48550/arXiv.1810.04805

[9] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, “GoEmotions: A Dataset of Fine-Grained Emotions,” arXiv.org, May 01, 2020. https://doi.org/10.48550/arXiv.2005.00547

[10] M. Zang, S. Niu, Y. Gao, and X. Chen, “Long Text Multi-label Classification,” 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE), Feb. 2023, Published, doi: 10.1109/nnice58320.2023.10105696.

[11] M. A. Mahima, N. C. Patel, S. Ravichandran, N. Aishwarya, and S. Maradithaya, “A Text-Based Hybrid Approach for Multiple Emotion Detection Using Contextual and Semantic Analysis,” 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Sep. 2021, Published, doi: 10.1109/icses52305.2021.9633843.

[12] S. Z. Salsabiila, H. Nurramdhani Irmanda, and A. Arista, “Comparison of Fasttext and Word2Vec Weighting Techniques for Classification of Multiclass Emotions Using the Conv-LSTM Method,” 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), Nov. 2023, Published, doi: 10.1109/icimcis60089.2023.10349034.

[13] T. Sheng, H. Wu, and Z. Yue, “An English Text Classification Method Based on TextCNN and SVM,” 2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI), Jan. 2022, Published, doi: 10.1109/iwecai55315.2022.00052.