PERFORMANCE EVALUATION OF MULTILABEL EMOTION CLASSIFICATION USING DATA AUGMENTATION TECHNIQUES

Zahra Ahanin, Maizatul Akmar Ismail*, and Tutut Herawan

Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya
50603 Kuala Lumpur, Malaysia

Emails: zahrahnnx@gmail.com[1], maizatul@um.edu.my[2]*(corresponding author), tutut@um.edu.my[3]

ABSTRACT

One of the challenges of emotion classification is the existence of low annotated datasets, that makes the task more complex. Certain existing datasets often suffer from imbalanced data for the emotion classes. Several data augmentation approaches can help to overcome the challenges regarding imbalanced datasets. However, the existing data augmentation techniques in emotion classification lack consideration for the contextual nuances of emotions and this area is still relatively underexplored. In this work, we study the impact of data augmentation on classification performance of three machine learning models including Logistic Regression, BiLSTM and BERT and compare frequently used methods to address the issue. Specifically, we assessed Easy Data Augmentation (EDA) and contextual Embedding-based data augmentation (BERT) on two datasets. Based on the experimental results, we combined two BERT-based augmentation techniques including insert and substitute, to generate data for minority emotion classes. Furthermore, we proposed a data augmentation method using ChatGPT. Compared to the baseline models, incorporating the BERT augmentation techniques with BERT model resulted in improvements of +4.34% and +5.56% in Macro F1 score on the SemEval-2018 and GoEmotions datasets, respectively. Moreover, the proposed augmentation technique utilizing ChatGPT yielded improvements of +3.55% and +4.83% on the same datasets.

Keywords: Text classification; Deep learning; Class imbalance; NLP; Data augmentation; ChatGPT.

1.0 INTRODUCTION

Recurrent Neural Networks (RNN) such as Long Short-Term Memory (LSTM), and transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) are becoming increasingly important across various machine learning applications. Their rising prominence is largely attributed to their remarkable performance in the field of natural language processing (NLP), which includes tasks such as sentiment analysis, emotion classification, and machine translation. Emotion classification refers to the task of detecting emotions such as happiness, sadness, anger, etc., in text, which has gained interest due to its wide range of applications from customer feedback analysis [1] to mental health monitoring [2]. The recent advances in Machine Learning and Deep Learning enabled more nuanced and context-aware emotion classification in text.

One of the challenges related to emotion classification is the availability of annotated data. There are different taxonomies such as Ekman [3] or Plutchik [4] for categorizing emotions, which are widely used by annotators. Due to the complexity of human language, which often involves informal writing styles with many grammatical and spelling errors, specifically on social media platforms, it is not a trivial task to detect emotions. It is further complicated where annotators assign one or more labels to a text message implying a message may express more than one emotion, which often results in imbalanced dataset. Data imbalance means that the number of instances in some categories is much higher or lower than in others [5]. In the case of imbalanced dataset, the classifier learns very little about the minority class (performance of the classifiers is biased toward the majority class) resulting in a higher error rate for the minority class [6]. Considering deep learning-based models exhibit enhanced performance when provided with extensive labeled datasets for training [7], [8], data augmentation can be employed to compensate the minority class, and generate additional, synthetic data using the training data.

Data Augmentation (DA) refers to the process of constructing synthetic data and expanding dataset by creating additional data instances based on the existing training data [9]. DA can provide several benefits such as serving as a form of regularization in deep learning-based algorithms, addressing class imbalance, and enabling small companies to access large volumes of data to train more effective models despite having limited access to data

[10]. Various techniques have been proposed to augment text in NLP applications, such as replacing words, inserting new words, deleting or swapping words (EDA) [11], inserting punctuation marks in the sentence [12], using reverse translation to translate text to another language and then back to the original language [13], and paraphrasing using generative language models such as ChatGPT [14].

According to the recent review [15], despite remarkable success of DA systems in computer vision tasks, their impact on NLP research has been comparatively limited, particularly in emotion and sentiment analysis. General NLP tasks and challenges often encounter limited success when employing DA techniques. One of the primary challenges lies in establishing universal rules for automatically transforming textual data while maintaining the quality of the labelling, which is especially sensitive in domains such as sentiment analysis [10].

Emotion analysis and classification is crucial in various applications such as mental health [16], sarcasm detection [17], customer feedback monitoring [18], and examining patient bahaviour and treatment quality [19]. However, existing models often overlook the contextual nuances of emotions. The EDA method primarily operates on individual words; however, emotions can be conveyed through text even without the explicit use of emotional terms. BERT augmentation techniques are not tailored specifically for emotion text and may insert words that lack emotional context. By exploring innovative data augmentation techniques tailored specifically for emotion text, we aim to address these challenges and improve the performance of emotion classification models.

In this study, we aim to investigate the impact of data augmentation on the classification of emotions within limited and imbalanced datasets. Firstly, we study and compare several augmentation techniques using EDA and BERT methods across three machine learning models. Then, we explain their performance and comparative advantages. Secondly, we combined two BERT augmentation techniques (insert, substitute) to assess its influence on performance of classification. Lastly, we proposed an augmentation method that integrates ChatGPT and compared the results, in terms of Micro F1 and Macro F1 scores, with the state-of-the-art research in the field of emotion classification. Micro F1 score on imbalanced datasets often prioritize predicting the majority class, while Macro F1weights the performance of each class equally.

The remainder of the paper is structured as follows: A literature review is provided in Section 2. Outline of the methodology and data augmentation methods are presented in Section 3. In Section 4, we described experimental setups, dataset details, and performance metrics. Then, the results and the discussion about the performance of the various data augmentation methods are presented in Section 5. Finally, we provide conclusions and state our future research direction in Section 6.

2.0 RELATED WORKS

Annotating data for machine learning tasks can be costly and time-consuming, leading researchers to explore data augmentation techniques as an alternative (Table 1). In Computer Vision, data augmentation has been widely employed to improve image classification models by generating synthetic datasets through operations like flipping and rotating training images [20]. Despite the success of data augmentation in Computer Vision tasks, its benefits have not been as pronounced in the field of NLP, largely due to the complexity of human language. In NLP, data augmentation techniques typically involve augmenting text at the word level within sequential data. This can be achieved by modifying input sequences using various techniques such as word deletion, swapping, insertion, or synonym replacement [11]. Alternatively, neural networks can guide word replacement in a sentence based on semantic closeness or contextual word representations [21]. In the former approach, lexical databases like WordNet are used to randomly replace words with synonyms, while the latter approach employs machine learning techniques such as BERT to generate augmented data based on surrounding context. BERT embeddings capture semantic similarity between words, enabling more effective insertion or replacement of words.

Oversampling is one of the basic approaches in data augmentation, which repeats the documents in the minority class. Synthetic Minority Over-sampling Technique (SMOTE) [22] generates synthetic examples of the underrepresented class until a balanced distribution between the minority and majority classes is achieved. This approach forces the model to give higher weight to classes with fewer samples, which may lead to overfitting [23]. To overcome the limitation of oversampling approach, other approaches such as combining oversampling with under-sampling [24], and removing word and inserting synonyms can be used. The findings by Olusegun et al., [24] indicate SMOTE improved the accuracy of model, however, SMOTE combined with random under-sampling does not make noticeable improvement to the model's performance. According to research by Madabushi et al., [25] synonym insertion and oversampling show similar results without improving the baseline model (BERT), while randomly dropping words decreased the performance of the classifier and produced lower scores.

Wei and Zou [11] proposed an Easy Data Augmentation (EDA), which uses four techniques such as replacing random words with their synonym, deleting random words, inserting random words from a dictionary such as WordNet, and swapping the positions of random words in the sentence. According to the results, random insertion

of words yielded high performance gains, and the combination of four techniques improved the performance of the classifier. Rather than using dictionary, Handoyo et al., [26] employed GloVe [27] word embeddings in the context of sarcasm detection. Data augmentation using word embeddings generates more relevant data by replacing words with similar meanings in the word embedding space. They evaluated the performance of data augmentation on various balanced and imbalanced datasets, and the best performance was achieved on small and very imbalanced datasets, with an improvement of 2.1% in terms of F1 score.

Hu et al. [28] examined the performance of various augmentation methods such as oversampling, BERT, Word2Vec, and WordNet in an imbalanced dataset. They generated new text by replacing the words. According to the results, all approaches improved the results, while data augmentation using BERT achieved the best performance in detecting minority class and outperformed other models. They further stated that BERT augmentation achieved the best result when combined with pre-trained BERT model. Therefore, recent research shows that data augmentation can benefit NLP applications where obtaining sufficient training data is challenging or costly. The recent study by [14] leveraged ChatGPT for data augmentation by paraphrasing the document for the task of sentiment analysis (positive, negative, neutral). ChatGPT is a variant of the GPT (Generative Pre-trained Transformer) language model that generates high-quality synthetic data. However, ChatGPT is effective for conversational applications, and it might not sufficiently diversify the dataset. Therefore, the effectiveness of augmentation varies depending on the type and the dataset in question [14].

Table 1. A summary of comparison among existing data augmentation methods

Authors (Year)	Model + Applications	Advantages	Disadvantages
Olusegun et al., (2023) [24]	Combined SMOTE with random under-sampling for the task of emotion classification	Added synthetic minority class examples to the training dataset until obtaining a balanced class distribution. Compared the method using multiple deep learning architectures including LSTM, C-LSTM and Bi-LSTM.	Using oversampling may leads to overfitting while under-sampling lead to the loss of important patterns.
Woźniak and Kocoń (2023) [14]	Utilized ChatGPT to paraphrase document and enhance the training data for sentiment analysis tasks.	Used ChatGPT's generative ability to generate new documents while keeping the sentiment. Compared multiple ChatGPT methods including a) paraphrase, b) paraphrase and keep the label sentiment	This method generates a completely new sentence, but this may not adequately diversify the dataset, which leads to limiting the model's enhanced generalization
Hu et al., (2022) [28]	Generate new document by replacing 90% of the words using BERT.	Utilized Contextualized Word Embeddings to generate words based on their surrounding context, thereby minimizing noise and preserving the sentence's intended meaning. Compared the method using multiple ML models such as LR, LSTM and BERT.	Not tailored specifically for emotion text, thus yielding low corpus-level variability in this context.
Wei and Zou (2019) [11]	Replacing, removing, inserting, and swapping words for the task of Text Classification, with the use of lexical databases.	Added word features that may not exist in training set. Reduced overfitting by adding noise in the dataset. Compared the proposed method using LSTM and Convolutional neural networks (CNNs) models.	Lacks the capability to capture implicit emotions expressed in text. Ignores position of word within the sentence, which can significantly change the meaning of the sentence.

Some studies used attention models to improve the model performance by putting more weight on relevant words in the context of emotion classification. Baziotis et al., [29] proposed a multi-layer self-attention mechanism with Bidirectional Long Short-Term Memory (Bi-LSTM) to improve the model performance by amplifying the contribution of important words. Jabreel and Moreno [30] proposed a form of attention mechanism and multi-label classification using Binary Relevance (BR) [31] method. The BR method breaks down the multi-label problem into multiple binary classification problems, where each label is treated as a separate binary classification problem. Ahanin and Ismail [32] leveraged Twitter features such as Emoji to improve the performance of Bi-LSTM and

attention mechanism (Bi-LSTM + Att). Ameer et al., [33] employed RoBERTa and multiple-attention mechanism which outperformed the other state-of-the-art models. These studies employed machine learning and deep learning models, which have been shown to obtain better performance when trained on a greater amount of labeled dataset.

As it was mentioned, manually labeling or annotating data is an expensive and time-consuming task, and the available annotated emotion dataset is small, and often imbalanced. To compensate for this drawback, data augmentation techniques can be used to generate more sample data. In this research, to generate new documents for minority classes, two approaches are used: 1) EDA, which involves inserting, deleting, swapping, or randomly replacing a word in a sentence with its synonym, using WordNet dictionary, 2) BERT, as a Contextualized Word Embeddings, involves creating training data[1] by replacing a word or inserting words based on the surrounding context. Furthermore, we proposed a data augmentation method using ChatGPT to generate new sentences by considering emotion label of the text.

3.0 METHODOLOGY

This section presents our proposed data augmentation method for multilabel emotional classification. This method, so called *ChatGPT_PartLab*, uses ChatGPT to generate new words by considering the emotion *label* of text. Moreover, we assessed EDA and BERT-based augmentation methods for the task of emotion classification.
utilized a BERT-based data augmentation method (BERT_Augs) which combines two augmentation techniques (insert, substitute), to generate new documents.

3.1 The proposed data augmentation method (*ChatGPT_PartLab)*

Our method focuses on creating new instances while preserving the emotional context of the sentence. Rather than paraphrasing the entire sentence, we retain the first half of the words (n) in the sentence and utilize ChatGPT (OpenAI's GPT-3.5) to generate a minimum of m and a maximum of n words related to the emotion *label*.
The input of the proposed method consists of a set of labeled data containing short messages in English along with their corresponding labels. Initially, data pre-processing is performed, which includes text cleaning and tokenization. Details of data pre-processing steps are provided in Section 3.4. After pre-procesing, we retain the first n words of each sentence and discard the remaining words.
The input for the Machine Learning models is the text with [PAD] token to ensure all the sentences have the same length (sequence length). Due to words limitation in social media platforms and nature of text messages shared in such platforms, the sequence length is set to 40 words (Table 3). The value of n (where n=20) is selected based on the sequence length. This method keeps the first n words from the original text message and append n words inspired by the original text by ChatGPT. To minimize the number of [PAD] tokens, the value of m is set to 10. The aim is to diversify the dataset to capture a broader range of scenarios or situations, thereby enhancing the overall augmentation process.

This process is conducted separately for each emotion label, which may increase the risk of introducing some noise into the data. The modified text messages are then merged to the original dataset to form the training set. The generated training set is then used as input to the BERT model for the task of emotion classification. (Figure 1).

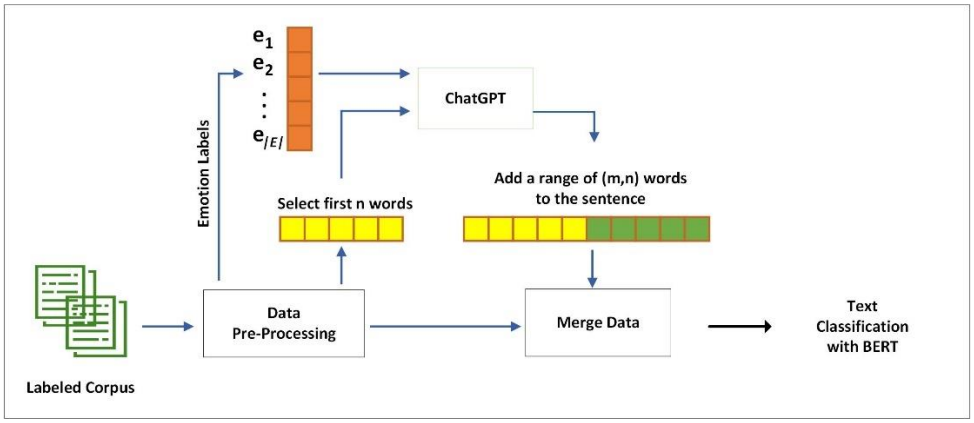

Fig. 1: The proposed data augmentation method using ChatGPT (*ChatGPT_PartLab)*

3.1.1 Prompt generation

[1] https://github.com/makcedward/nlpaug

Prompt generation consists of forcing ChatGPT to add a sentence to the original sentence while maintaining the emotion of text based on the emotion *label*. The following prompt is used in a conversation with the model:

- Based on each Tweet, generate another text with minimum 10 to maximum 20 words to show emotion of '*label'* and append at the end of the original Tweet. Do not use the word '*label'* in the sentence. Remove quotes or double quote signs. Do not add Emoji to the Tweets.

3.2 The BERT-based data augmentation method (*BERT_Augs)*

Additionally, we explored the potential of generating novel sentences by leveraging the contextualized word embeddings provided by BERT. BERT augmentation is utilized to generate two instances for each document by a) substituting existing words in a sentence with similar ones, b) inserting new words based on the surrounding context (Figure 2). The degree of alterations in the text hinges on the α value. To determine the α value, we conducted a series of tests, explained in section 3.3.

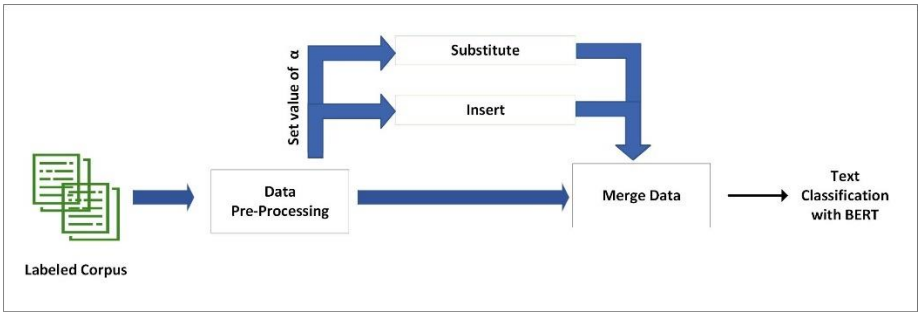

Fig. 2: The data augmentation method using combination of BERT augmentations (substitute, insert)

3.3 Performance evaluation of EDA and BERT augmentation methods

The demonstration of the process for performance evaluation of EDA and BERT augmentation methods is given in Figure 3. We performed multiple data augmentation methods with various α values. This research utilized two publicly available multi-label annotated dataset by Mohammad et al., [34] and Demszky et al., [35]. These datasets include the text messages shared on social media platforms and the corresponding emotion labels.

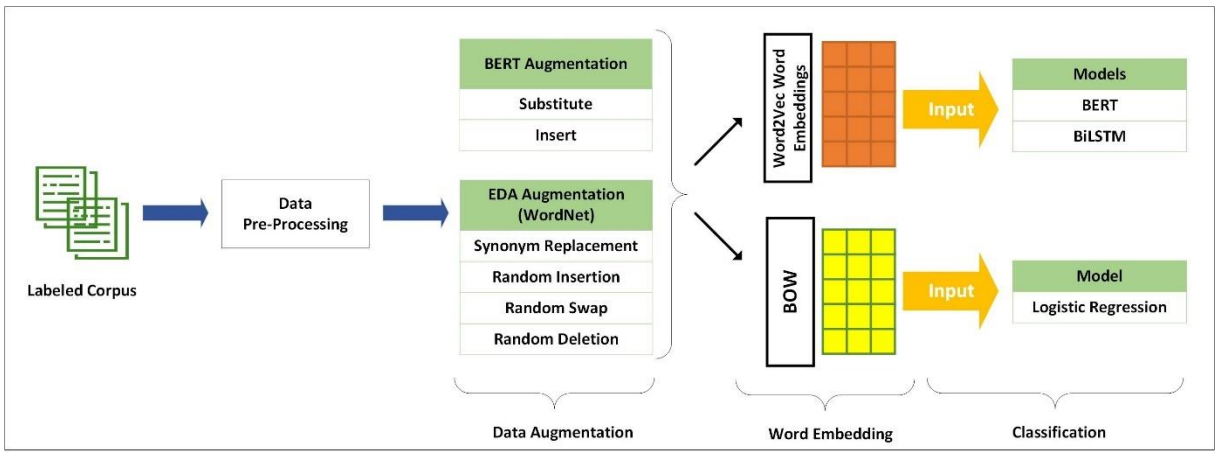

Fig. 3: The process for performance evaluation of two data augmentation methods

3.4 Data pre-processing

Data preprocessing begins with tokenization, a process where each text message is segmented into individual words or tokens. Then, all the words are converted to lowercase, and extra white spaces and line breaks are removed. Given that social media platforms such as Twitter often contain misspellings and abbreviations, we address misspellings using the Python Natural Language Toolkit (NLTK). We conducted text normalization to normalize terms including URL, number, email, money, timestamp, and date. A custom dictionary is used to replace abbreviations with their description. Furthermore, words with repeated letters are modified and any letter occurring more than two times consecutively is replaced with two occurrences. For example, the word "heyyyy"

will be changed into “heyy”. Then, all the numbers as well as punctuations, except exclamation mark (!), are removed.

3.5 Data augmentation

We compare two data augmentation methods in this study: EDA and BERT augmentations.

Easy Data Augmentation (EDA) using WordNet dictionary: Data augmentation is carried out by performing the following techniques [11]:

- Synonym Replacement (SR): Replacing *n* words in the sentence with a randomly selected synonym using WordNet dictionary.
- Random Insertion (RI): Inserting a synonym for *n* random words in a random position in the sentence using WordNet dictionary.
- Random Swap (RS): Swapping the position of random words in the sentence, for *n* times.
- Random Deletion (RD): Removing *n* random words in the sentence.

Word Embedding using BERT: This approach employs contextual word embeddings, instead of a dictionary, to insert or substitute words in a sentence. BERT generates words based on the words around it, with the aim of inserting a suitable word for augmentation. This method may replace words in the sentence randomly. In this method, data augmentation is carried out by performing the following techniques:

- Substitute: Replacing *n* words in the sentence with words predicted by the BERT model
- Insert: Inserting *n* words in the sentence with words predicted by the BERT model

Considering the dataset includes long and short text messages and the length of text messages varies, the value of n is determined based on the sentence length. Therefore, the parameter α is used, which refers to the percent of words in sentence. This parameter is multiplied by length of sentence to calculate the value of n.

Table 2 shows an example of an original sentence in training dataset and its corresponding augmented sentences using RD and RI techniques employing the EDA method. The generated text varies based on value of α, where the lower α leads to minor alterations in the sentence, whereas higher α results in more significant changes, affecting both grammar and meaning of the sentence.

Table 2: An example of an original sentence and EDA using RD and RI techniques for each value of α

Augmentation		Text
Original text		@user @user High school or not, it's still shocking. Just because you love Xbox. Good luck tomorrow!
Random Deletion	α = 0.05	high school, not it is shocking just because you love xbox. Good luck tomorrow
	α = 0.1	school is not still shocking just because you love xbox. Good luck tomorrow
	α = 0.2	high school is shocking just because I love xbox. Good luck tomorrow
	α = 0.3	school not it still shocking just because you love xbox good luck tomorrow
	α = 0.4	high school, it is still shocking just because you love xbox. Good luck
	α = 0.5	high school, not it shocking you love xbox tomorrow
Random Insertion	α = 0.05	high school not, it is still schoolhouse shocking just because you love xbox. Good luck tomorrow
	α = 0.1	high school not it is still shocking just because you love xbox good high gear luck tomorrow
	α = 0.2	dear high school, not schoolhouse, it is still shocking schoolhouse just because you love xbox. Good luck tomorrow
	α = 0.3	high schoolhouse schoolhouse school not it is still shocking hush up just because you love xbox good lie with luck tomorrow
	α = 0.4	dear dear high school not schoolhouse it is still shocking just because you love xbox dearly appall good dearly luck tomorrow
	α = 0.5	assign high thoroughly school or not it is still shocking just because portion you just now assign love xbox be intimate good be embody luck tomorrow

3.6 Classification models

We compare three classification models: Logistic Regression, BERT, and Bi-LSTM.

Logistic Regression (LR + BR): The Logistic Regression (LR) classification is designed based on our previous paper [32], and the input of this model is BOW vector representation using unigram features. At top of Logistic Regression model a Binary Relevance (BR) method is used to treat the multi-label problem.

Bi-LSTM (Bi-LSTM + Att): This model includes Bi-LSTM with an Attention model [29], [32] to give more value to essential words. The input of Bi-LSTM is a 300-dimensional word embedding developed by Baziotis et al., [29] which is trained on 550 million Twitter messages using word2vec algorithm.

BERT: The BERT [36] model BERT utilizes a multi-layer architecture based on the Transformer encoder. BERT employs a bidirectional self-attention mechanism, meaning it considers both the preceding and succeeding words when encoding each word in the input sequence at once. BERT has a vocabulary of 30k tokens, and each token is embedded into a high-dimensional vector space with 768 features. This study fine-tunes the pre-trained uncased based BERT model using Python Library "Transformers".

4.0 EXPERIMENTAL SETUPS

To evaluate the performance of various text augmentation methods, we run machine learning models for each augmentation method and compare the results with non-augmentated data (original data). Each model is trained based on different augmentation methods and different percent of words in sentence α. The "EDA" Python library [11], was used to implement WordNet augmentation methods. The BERT augmentation was implemented using "nlpaug" Python library to generate training data[2].
The deep learning models were performed on the Google Colaboratory[3] platform on a 16-GB GPU.

4.1 Hyperparameter settings

To tune the hyperparameters, we used a development dataset. In the BiLSTM-based model (Table 3) the batch size is 32, dropout value is 0.3, and Adam optimizer is utilized with learning rate of 0.001. For the transformer model (BERT), the hyperparameters are selected based on the original transformer paper [36]. For the dataset of SemEval-2018, batch size of BERT is equal to 32, and for dataset GoEmotions, batch size is equal to 16. The learning rate is set to $3e-5$ for both experiments.

Table 3: Bi-LSTM model Hyperparameter values

Hyper parameter	Values
Embedding Dimension	300
LSTM Hidden Layer Size	64
Dense Layer	20
Dropout-lstm	0.3
Dropout	0.3
Learning Rate	0.001
Batch Size	32
Epoch	100
Sequence length	40

4.2 Datasets

Table 4 and Table 5 present details of datasets utilized to evaluate the performance of the data augmentation methods. Table 4 displays explicitly the distribution of instances across different emotion labels within the SemEval-2018 and GoEmotions datasets. Distinct training, development, and test sets are provided for each dataset.

Table 4: Count of records and details of datasets

[2] https://github.com/makcedward/nlpaug
[3] https://colab.research.google.com/notebooks/welcome.ipynb

Dataset	Domain	URL	Count of Records
SemEval-2018 Task 1: E-C [34]	Twitter	https://competitions.codalab.org/competitions/17751	Train: 6,838 Development: 886 Test: 3,259
GoEmotions [35]	Reddit	https://github.com/google-research/google-research/tree/master/goemotions	Train: 43,410 Development: 5,426 Test: 5,427

SemEval-2018 [34]: It comprises English tweets, with each tweet annotated with one or more of the following eleven emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust.

GoEmotions [35]: It comprises 58,000 English Reddit comments, manually labeled as either neutral or one or more of 27 emotion categories. The annotators used a broader set of 27 emotions to encompass the nuances of human emotion.

Table 5: Data Statistics on SemEval-2018 Train Dataset and GoEmotions Train Dataset

SemEval-2018		GoEmotions					
Emotion	**Train**	**Emotion**	**Train**	**Emotion**	**Train**	**Emotion**	**Train**
anger	2544	admiration	4130	disgust	793	realization	1110
anticipation	978	amusement	2328	embarrassment	303	relief	153
disgust	2602	anger	1567	excitement	853	remorse	545
fear	1242	annoyance	2470	fear	596	sadness	1326
joy	2477	approval	2939	gratitude	2662	surprise	1060
love	700	caring	1087	grief	77	neutral	14219
optimism	1984	confusion	1368	joy	1452		
pessimism	795	curiosity	2191	love	2086		
sadness	2008	desire	641	nervousness	164		
surprise	361	disappointment	1269	optimism	1581		
trust	357	disapproval	2022	pride	111		

Data augmentation is implemented for emotion categories belonging to the minority class, characterized by the lowest number of instances. Thus, data augmentation is performed for anticipation, fear, love, pessimism, surprise, and trust emotions on SemEval-2018 dataset, while it is performed for embarrassment, grief, nervousness, relief, and pride emotions on GoEmotions dataset.

4.3 Evaluation metrics

Aligning with previous studies [7], [30], [33], Micro F1 score, and Macro F1 score are used as the performance metrics (Equation 1 to 4), where TP is True Positive, and TP_e refers to the True Positive for emotion class e. Similarly, FP refers to False Positive, and FN refers to False Negative.

$$Precision_e = \frac{TP_e}{TP_e + FP_e} \tag{1}$$

$$Recall_e = \frac{TP_e}{TP_e + FN_e} \tag{2}$$

$$F1 - \text{Micro} = \frac{2 \times Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{3}$$

$$F1 - Macro = \frac{1}{|E|} \sum_{e \in E} \frac{2 \times Precision_e \times Recall_e}{Precision_e + Recall_e} \tag{4}$$

Considering the models are training on imbalance dataset, the Macro F1 score is a more suitable metric as it weights the performance on each class equally and reflects the value of minority class. The results are based on the best model's performance over five repetitions.

The gain metric (Equation 5) is a measure used to quantify the improvement achieved by a personalized or proposed model compared to a non-personalized or baseline model [14], [37].

$$Gain = \frac{100\% \,.\, (P - B)}{(100\% - B)} \tag{5}$$

Where *P* refers to the Macro F1 score provided by the proposed method, and *B* refers to the Macro F1 score of the baseline model. The baseline model uses the original dataset without any data augmentation.

5.0 EXPERIMENTAL RESULTS AND DISCUSSION

Tables 6, 7, and 8 present the results of different data augmentation methods with Logistic Regression, Bi-LSTM, and BERT models, respectively. Table 9 indicates the performance metrics of the three aforementioned classification models on the original dataset. A comparative analysis of Tables 6, 7, and 8 against Table 10 reveals every data augmentation method across all the models enhanced the F1 scores. LR with EDA augmentation techniques achieved similar results throughout various values of α parameter (Table 6).

LR achieved Macro F1 score of 44.67 on original SemEval-2018 dataset, while employing BERT augmentation to insert new words into sentences yields improved Macro F1 scores of 46.54 ($\alpha = 0.05$), 46.65 ($\alpha = 0.1$), 46.64 ($\alpha = 0.2$), 46.14 ($\alpha = 0.3$), and 46.53 ($\alpha = 0.4$). The improvement of F1 score is consistent across all data augmentation methods using BiLSTM+Att model (Table 7).

BERT augmentation achieved higher results when it is combined with BERT classification model (Table 8). This finding is consistent with the findings of [28], which shows the effectiveness of pre-trained BERT in various natural language processing tasks. Notably, the BERT model demonstrated the most significant enhancement, reaching a score of 56.82($\alpha = 0.1$).

We observed the value of α can lead to different results across three classification models. Generally, α did not affect the result of LR classification, possibly due to several reasons. Firstly, LR is trained based on BOW features, which word order is ignored. This implies that two documents could possess identical or highly similar representations if they contain the same words, regardless of their differing meanings, which explains the consistent results of RS augmentation throughout different values of α. Additionally, in BOW, the number of occurrences of a word (term frequency) indicates the importance or relevance of word. Therefore, using the augmentation technique that can capture contextual nuances and insert words relevant to semantics of the dataset is important, since the LR classifier uses word frequencies as features to make predictions. Thus, EDA techniques such as Synonym Replacement and Random Insertion have not shown to be effective compared to similar techniques in BERT augmentation using LR classifier. In contrast, the EDA method provided comparative results with BERT augmentation when using Bi-LSTM+Att model (Table 7). This model uses pre-trained word embeddings, which are trained on massive corpora. Unlike traditional methods that represent words as numbers (word counts or weights), word embeddings use dense vector representations, which capture linguistic information and semantics of the word. Therefore, Bi-LSTM models can effectively interpret the meaning of words and sentences, which contributes to their improved performance. Despite the ability of Bi-LSTM model to capture the complex semantic relationships, its performance is notably influenced by the value of α when using EDA method. For instance, employing the Synonym Replacement technique yielded a decrease in F1 score from 56.02 ($\alpha = 0.05$) to 53.08 ($\alpha = 0.5$). Similarly, a decrease in F1 score was observed from 55.28 ($\alpha = 0.05$) to 52.87 ($\alpha = 0.5$) when employing Random Deletion technique, which is aligned with the recent study by Madabushi [25]. In addition to Bi-LSTM, the performance of BERT (Table 8) is also impacted by the by value of α, where higher α values lead to a decrease in the F1 score of the BERT model. When α is set to 0.5, half of the sentence undergoes alteration, potentially altering the original sentence's meaning and consequently leading to a decline in the model's performance. In the GoEmotions dataset, we observed a similar pattern, BERT augmentation outperforms other augmentation methods (Table 9).

Table 6: Macro F1 scores of augmentation methods using LR on SemEval-2018 dataset for each value of α

Augmentation method	**Macro F1 (%)**

		0.0	0.05	0.1	0.2	0.3	0.4	0.5
BERT	Substitute	-	46.24	46.49	46.33	46.23	46.10	-
	Insert	-	46.54	46.65	46.64	46.14	46.53	-
EDA WordNet	Synonym Replacement	-	46.15	46.32	46.03	46.00	45.59	45.59
	Random Insertion	-	45.69	45.60	46.07	45.58	45.69	45.64
	Random Swap	-	45.81	45.81	45.81	45.84	45.84	45.84
	Random Deletion	-	46.01	45.90	45.29	45.53	45.97	45.31
Original dataset (non-augmentation)		44.67	-	-	-	-	-	-

Table 7: Macro F1 scores of augmentation methods using Bi-LSTM+Att on SemEval-2018 dataset for each value of α

Augmentation method		Macro F1 (%)						
		0.0	0.05	0.1	0.2	0.3	0.4	0.5
BERT	Substitute	-	54.99	55.67	54.22	54.91	54.61	54.16
	Insert	-	55.13	55.63	54.55	56.14	54.43	54.81
EDA WordNet	Synonym Replacement	-	56.02	55.33	55.51	55.47	55.11	53.08
	Random Insertion	-	55.80	55.53	56.13	56.08	55.04	53.91
	Random Swap	-	55.59	54.95	55.72	55.35	55.58	54.39
	Random Deletion	-	55.28	55.43	55.74	55.33	54.45	52.87
Original dataset (non-augmentation)		52.63	-	-	-	-	-	-

Table 8: Macro F1 scores of augmentation methods using BERT on SemEval-2018 dataset for each value of α

Augmentation method		Macro F1 (%)						
		0.0	0.05	0.1	0.2	0.3	0.4	0.5
BERT	Substitute	-	54.03	56.78	54.73	55.65	53.31	50.63
	Insert	-	56.56	56.82	56.39	56.39	56.08	53.69
EDA WordNet	Synonym Replacement	-	54.83	55.41	53.54	53.66	54.87	53.82
	Random Insertion	-	54.93	55.44	54.92	54.71	54.73	53.97
	Random Swap	-	53.52	54.97	54.19	53.47	53.73	53.61
	Random Deletion	-	55.49	54.61	54.26	54.26	53.94	53.90
Original dataset (non-augmentation)		52.72	-	-	-	-	-	-

Table 9: Macro F1 scores of augmentation methods using LR on GoEmotions dataset for each value of α

Augmentation method		Macro F1 (%)						
		0.0	0.05	0.1	0.2	0.3	0.4	0.5
BERT	Substitute	-	35.55	35.86	35.24	35.55	35.51	-
	Insert	-	36.01	35.61	35.46	36.22	35.51	-
EDA WordNet	Synonym Replacement	-	35.55	36.92	35.44	35.50	34.96	34.78
	Random Insertion	-	34.85	35.66	35.22	35.43	35.10	34.75
	Random Swap	-	34.78	34.77	34.78	34.82	34.82	34.76
	Random Deletion	-	34.99	34.94	35.00	35.25	35.31	35.36
Original dataset (non-augmentation)		34.14	-	-	-	-	-	

To further investigate the effectiveness of data augmentation in multilabel classification, we combined two BERT (BERT_Augs) augmentation methods ($\alpha = 0.1$), and compared the Macro F1 score and Micro F1 score with the state-of-the-art research. The results are presented in Tables 10 and 11 for SemEval-2018 and GoEmotions

datasets, respectively. The BERT model with BERT_Augs yielded competitive results (57.06) when compared to the BERT model by Alhuzali and Ananiadou [38] (57.8). Moreover, it outperformed GRU model by Jabreel and Moreno [39] and Bi-LSTM by Baziotis et al., [29] in terms of Macro F1 score, both of which incorporated attention mechanisms to enhance classification accuracy. Similarly, in GoEmotions dataset (Table 11) integrating BERT augmentation techniques (BERT_Augs) leads to an improvement of +4.56% in Macro F1 score, and outperformed the other models. According to the results, the proposed augmentation technique with ChatGPT (ChatGPT_partLab) demonstrates enhancements in Micro F1, suggesting the potential for higher quality data augmentation. We believe in the ChatGPT_partLab method there is a possibility that the generated content may deviate from the original text. This means that some of the generated examples might not accurately convey the intended emotion or align perfectly with the expected emotional tone of the original text.

Table 10: Results of emotion classification on different models for SemEval-2018 dataset

Model	**Macro F1**	**Micro F1**
BERT + Proposed augmentation (ChatGPT_partLab)	56.27	70.21
BERT + BERT_Augs	57.06	69.75
BERT	52.72	69.16
Bi-LSTM based on [29]	52.63	69.04
LR + BR	44.67	60.12
RoBERTa+MA [33]	60.3	74.2
Bi-LSTM + Att [32]	56.4	71.1
BERT [38]	57.8	71.3
RF+BR [40]	55.9	57.3
GRU [39]	56.4	69.2
Bi-LSTM + Att [29]	52.8	70.1

In both datasets ChatGPT_partLab method achieved higher Micro F1 compared to BERT_Augs, albeit with slightly lower Macro F1 scores. The lower Macro F1 could be attributed to the ChatGPT_partLab method's approach of performing augmentation for each emotion label separately, possibly indicating a limitation in understanding the complex interplay of multiple emotions within a text. This could lead to higher Macro F1 scores for BERT augmentation.

An observation we made is that ChatGPT tends to repetitively use the same words to express a particular emotion. For instance, while generating new sentences for the emotion "*embarrassment*," the term "*discomfort*" was recurrently utilized. This repetition might suggest a limited vocabulary diversity in ChatGPT-generated sentences. In this case, BERT may have a more sophisticated understanding of language semantics and syntax compared to ChatGPT, particularly in the context of emotion classification. This could result in better performance in terms of Macro F1, which evaluates the model's ability to capture the nuances of individual emotion classes.

Table 11: Results of emotion classification on different models for GoEmotions dataset

Model	**Macro F1**	**Micro F1**
BERT + Proposed augmentation (ChatGPT_partLab)	50.83	57.22
BERT + BERT_Augs	51.56	56.94
LR + BR	34.14	47.85
BERT [35]	46.	-
Bi-LSTM + Att [32]	41.	-
BERT + Hybrid features [7]	49.	-

Table 12: Comparison of augmented text using different augmentation method, the value of α is equal to 0.1 for BERT and EDA methods

Augmentation Method	**Text**

	Original text	@user @user High school or not, it's still shocking. Just because you love Xbox. Good luck tomorrow! 💪
BERT	Substitute	high school not, it is still shocking . just know you love them . good luck tomorrow !
	Insert	high school not, it is still shocking . just because you still love xbox . so good luck tomorrow !
EDA WordNet	SR	high school, not it is still appalling just because you love xbox. Good luck tomorrow
	RI	high school not it is still shocking just because you love xbox good high gear luck tomorrow
	RS	high tomorrow not it is still shocking just because you love xbox good luck school
	RD	school is not still shocking just because you love xbox. Good luck tomorrow
Integrated ChatGPT	Insert	high school or not it is still shocking . just because you love xbox . good luck tomorrow ! 💪 the empathy expressed left me supportive .

Table 12 illustrates an instance of an original sentence in training dataset and its corresponding augmented sentences generated through BERT and EDA augmentation methods ($\alpha = 0.1$). Compared to RI (EDA method), inserting random words in BERT method provided a more understandable, and grammatically correct sentence. We repeated word substitution using BERT and the generated sentence is "*like school not it is still shocking . just because you love xbox . good bye tomorrow!*". In this sentence, "*good luck*" is replaced with "*good bye*", which is commonly used to end a conversation. Using BERT augmentation techniques yields higher performance accuracy; however, the optimal choice of data augmentation method may vary depending on the classification model.

The gain metric (Equation 5) is computed to evaluate the effectiveness of the data augmentation methods. Figure 4 shows the gains per augmentation method on SemEval-2018 dataset. The improvements are measured based on Macro F1 score for BERT and Bi-LSTM+Att models with and without data augmentation (baseline). All the augmentation techniques achieved positive gains and the BERT augmentation techniques incorporated with BERT model displayed substantial improvement over the baseline.

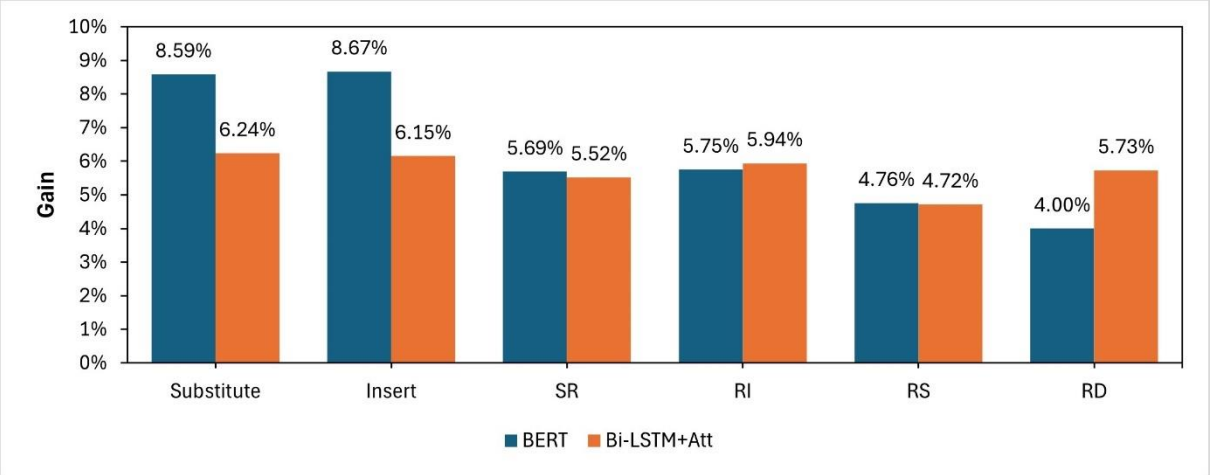

Fig. 4: Gain metric for each augmentation technique ($\alpha = 0.1$) on SemEval-2018 dataset in Macro F1 metric

Furthermore, we conducted a comparison of the gain metric between the baseline, which used the BERT model without data augmentation, and our proposed ChatGPT_partLab and BERT_Augs methods. Both augmentation techniques exhibited positive gains over the baseline, with BERT_Augs showing a more substantial impact compared to the ChatGPT_partLab method (Figure 5) across both datasets. As it was discussed, it may reflect the importance of enriching text with more context-aware words on classification performance.

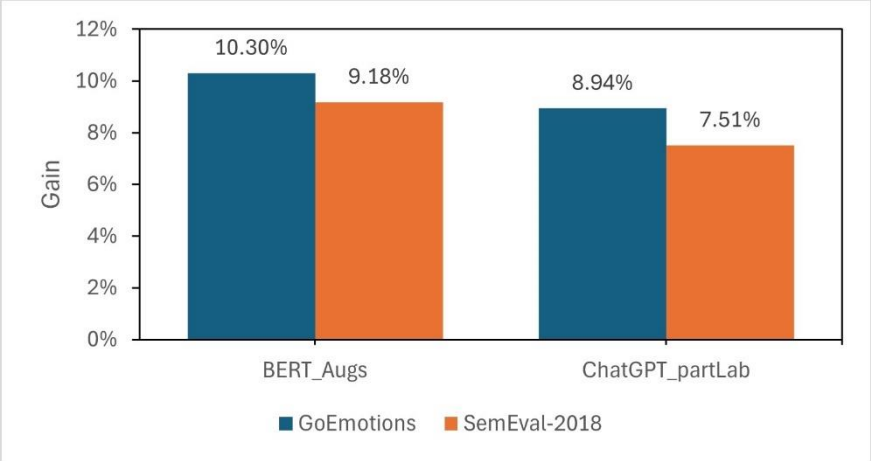

Fig. 5: Gain metric for proposed augmentation method per emotion dataset in Macro F1 metric

7.0 CONCLUSION

This paper proposed a data augmentation method, called ChatGPT_partLab, which uses ChatGPT to generate new instances while focusing on preserving the emotional context of sentences. We first provided a comparison of two data augmentation methods, EDA (using WordNet dictionary) and BERT, on two datasets originating from social media platforms to address the challenges related to the availability of annotated data, specifically on multilabel emotion classification where machine learning classifier learns very little about the minority classes. The accuracy of the model is determined based on Macro F1 score, which treats all the classes equally. We reported the performance of six augmentation techniques across three classifiers. According to the findings, BERT augmentation provided higher performance accuracy than EDA (WordNet) method in both datasets. We examined how the size of alterations in sentence (α), influences the effectiveness of the augmentation method. Experimental results show that lower value of ($\alpha < 0.4$) can potentially improve the model accuracy, whereas, modifying half of the sentence ($\alpha = 0.5$) leads to lower performance accuracy of Bi-LSTM and BERT models, where input sequences is important. Furthermore, we combined two BERT augmentation techniques, insert and substitute, to add more word features. According to the results, combining the techniques improved the F1 score in both datasets. In both datasets, the ChatGPT_partLab method outperformed BERT-based augmentations in terms of Micro F1 score, with slightly lower Macro F1 scores. One potential avenue for enhancing the F1 score could involve generating new sentences while considering multiple labels simultaneously, rather than focusing on individual labels. We believe that our study compared and assessed the potential of different data augmentation methods in multilabel emotion classification. The findings hold promise in light of the widespread adoption of large-scale neural networks that need substantial training data to learn effectively.

While this study investigated datasets in the context of sentiments and emotions, it did not specifically address emotion cues within social media content. Future research in this domain could explore the effect of augmentation techniques that retain the emotional content of the text and focus on emotion-induced features such as Emojis, GIFs, and even punctuation marks. Moreover, further investigation is needed to explore the effectiveness of integrating multiple data augmentation methods.

FUNDING

This research is funded by the Universiti Malaya International Collaboration Grant (ST005-2023).

REFERENCES

[1] G. Ren and T. Hong, "Investigating Online Destination Images Using a Topic-Based Sentiment Analysis Approach," *Sustainability*, vol. 9, no. 10, p. 1765, Sep. 2017, doi: 10.3390/SU9101765.

[2] X. Zhang *et al.*, "Community Governance Based on Sentiment Analysis: Towards Sustainable Management and Development," *Sustainability*, vol. 15, no. 3, p. 2684, Feb. 2023, doi: 10.3390/SU15032684.

[3] P. Ekman, "An argument for basic emotions," *Cogn Emot*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[4] R. Plutchik and H. Kellerman, *Emotion, theory, research, and experience*. Academic press, 1980.

[5] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[6] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, “Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review,” *Algorithms*, vol. 16, no. 5, p. 236, Apr. 2023, doi: 10.3390/a16050236.

[7] Z. Ahanin, M. A. Ismail, N. S. S. Singh, and A. AL-Ashmori, “Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages,” *Sustainability*, vol. 15, no. 16, p. 12539, Aug. 2023, doi: 10.3390/su151612539.

[8] C. Zhao, X. Sun, and R. Feng, “Multi-strategy text data augmentation for enhanced aspect-based sentiment analysis in resource-limited scenarios,” *J Supercomput*, vol. 80, no. 8, pp. 11129–11148, May 2024, doi: 10.1007/s11227-023-05864-2.

[9] C. Shorten, T. M. Khoshgoftaar, and B. Furht, “Text Data Augmentation for Deep Learning,” *J Big Data*, vol. 8, no. 1, p. 101, Dec. 2021, doi: 10.1186/s40537-021-00492-0.

[10] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, “Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers,” *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 1, pp. 135–150, Jan. 2023, doi: 10.1007/s13042-022-01553-3.

[11] J. Wei and K. Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 6382–6388.

[12] A. Karimi, L. Rossi, and A. Prati, “AEDA: An Easier Data Augmentation Technique for Text Classification,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 2748–2754. doi: 10.18653/v1/2021.findings-emnlp.234.

[13] X. Liu, J. He, M. Liu, Z. Yin, L. Yin, and W. Zheng, “A Scenario-Generic Neural Machine Translation Data Augmentation Method,” *Electronics (Basel)*, vol. 12, no. 10, p. 2320, May 2023, doi: 10.3390/electronics12102320.

[14] S. Woźniak and J. Kocoń, “From Big to Small Without Losing It All: Text Augmentation with ChatGPT for Efficient Sentiment Analysis,” in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, Dec. 2023, pp. 799–808. doi: 10.1109/ICDMW60847.2023.00108.

[15] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, “Data augmentation techniques in natural language processing,” *Appl Soft Comput*, vol. 132, p. 109803, Jan. 2023, doi: 10.1016/j.asoc.2022.109803.

[16] C. S. Abdul Razak, S. H. Ab Hamid, H. Meon, H. A/P Subramaniam, and N. B. Anuar, “TWO-STEP MODEL FOR EMOTION DETECTION ON TWITTER USERS: A COVID-19 CASE STUDY IN MALAYSIA,” *Malaysian Journal of Computer Science*, vol. 34, no. 4, pp. 374–388, Oct. 2021, doi: 10.22452/mjcs.vol34no4.4.

[17] V. Govindan and V. Balakrishnan, “INVESTIGATING THE IMPORTANCE OF HYPERBOLES TO DETECT SARCASM USING MACHINE LEARNING TECHNIQUES,” *Malaysian Journal of Computer Science* , vol. 37, no. 1, 2024.

[18] M. R. A. Rashid, K. F. Hasan, R. Hasan, A. Das, M. Sultana, and M. Hasan, “A comprehensive dataset for sentiment and emotion classification from Bangladesh e-commerce reviews,” *Data Brief*, vol. 53, p. 110052, Apr. 2024, doi: 10.1016/j.dib.2024.110052.

[19] S. A. Waheeb, N. Ahmed Khan, and X. Shang, “AN EFFICIENT SENTIMENT ANALYSIS BASED DEEP LEARNING CLASSIFICATION MODEL TO EVALUATE TREATMENT QUALITY,” *Malaysian Journal of Computer Science*, vol. 35, no. 1, pp. 1–20, Jan. 2022, doi: 10.22452/mjcs.vol35no1.1.

[20] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/S40537-019-0197-0.

[21] S. Kobayashi, “Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations,” in *The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics (ACL), 2018, pp. 452–457. doi: 10.18653/V1/N18-2072.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[23] U. R. Salunkhe and S. N. Mali, "Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach," *Procedia Comput Sci*, vol. 85, pp. 725–732, 2016, doi: 10.1016/j.procs.2016.05.259.

[24] R. Olusegun, T. Oladunni, H. Audu, Y. Houkpati, and S. Bengesi, "Text Mining and Emotion Classification on Monkeypox Twitter Dataset: A Deep Learning-Natural Language Processing (NLP) Approach," *IEEE Access*, vol. 11, pp. 49882–49894, 2023, doi: 10.1109/ACCESS.2023.3277868.

[25] H. T. Madabushi, E. Kochkina, and M. Castelle, "Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data," *ArXiv*, 2020.

[26] A. T. Handoyo, H. rahman, C. J. Setiadi, and D. Suhartono, "Sarcasm Detection in Twitter - Performance Impact While Using Data Augmentation: Word Embeddings," *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS*, vol. 22, no. 4, pp. 401–413, Dec. 2022, doi: 10.5391/IJFIS.2022.22.4.401.

[27] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[28] L. Hu, C. Li, W. Wang, B. Pang, and Y. Shang, "Performance Evaluation of Text Augmentation Methods with BERT on Small-sized, Imbalanced Datasets," in *2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*, IEEE, Dec. 2022, pp. 125–133. doi: 10.1109/CogMI56440.2022.00027.

[29] C. Baziotis *et al.*, "Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning," *ArXiv*, 2018, doi: 10.48550/arXiv.1804.06658.

[30] M. Jabreel and A. Moreno, "A deep learning-based approach for multi-label emotion classification in Tweets," *Applied Sciences (Switzerland)*, vol. 9, no. 6, 2019, doi: 10.3390/app9061123.

[31] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[32] Z. Ahanin and M. A. Ismail, "A multi-label emoji classification method using balanced pointwise mutual information-based feature selection," *Comput Speech Lang*, vol. 73, p. 101330, May 2022, doi: 10.1016/J.CSL.2021.101330.

[33] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification in texts using transfer learning," *Expert Syst Appl*, vol. 213, p. 118534, Mar. 2023, doi: 10.1016/J.ESWA.2022.118534.

[34] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 1–17.

[35] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, 2018, doi: 10.48550/arXiv.1810.04805.

[37] J. Kocoń *et al.*, "ChatGPT: Jack of all trades, master of none," *Information Fusion*, vol. 99, p. 101861, Nov. 2023, doi: 10.1016/j.inffus.2023.101861.

[38] H. Alhuzali and S. Ananiadou, "SpanEmo: Casting Multi-label Emotion Classification as Span-prediction," in *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 1573–1584.

[39] M. Jabreel and A. Moreno, "A deep learning-based approach for multi-label emotion classification in Tweets," *Applied Sciences (Switzerland)*, vol. 9, no. 6, 2019, doi: 10.3390/app9061123.

[40] I. Ameer, N. Ashraf, G. Sidorov, and H. G. Adorno, "Multi-label emotion classification using content-based features in twitter," *Computacion y Sistemas*, vol. 24, no. 3, pp. 1159–1164, 2020, doi: 10.13053/CYS-24-3-3476.