A Comparative Study of BERT and BiLSTM for Emotion Analysis in Text

Christos Panourgias

Abstract

In this study, a comparison of two state-of-the-art models for emotion detection in text is presented. Distributed representations of words for the BiLSTM model were obtained through pre-training with the Word2Vec method. BERT (Bidirectional Encoder Representations from Transformers) and Bi-directional LSTM (BiLSTM) were fine-tuned and trained respectively on the ISEAR dataset and then evaluated on the task of emotion detection. The emotion classes that are going to be studied are: joy, fear, anger, sadness, shame, disgust and guilt.

The results demonstrate that while good performance was achieved by both BERT and BiLSTM, BERT was found to have a statistically significant increase in accuracy, precision, recall and F1-score. The advantages and limitations of both models in the context of this task were analyzed, with the benefit of self-attention mechanisms in BERT being highlighted in comparison to BiLSTM's sequential processing of data.

This study provides a comparative analysis of the performance of BERT and BiLSTM models for emotion detection in text and highlights BERT as a superior approach. However, it is noted that BiLSTM still presents a viable option for tasks requiring sequential data processing.

1 Introduction

Emotion detection in text is an important task in natural language processing, as it has a wide range of applications such as customer service, and social media monitoring. It involves identifying and extracting emotions from text data, which can be challenging due to the complexity and variability of natural language. The ability to automatically detect emotions in text can be useful in various domains, such as analyzing movie reviews, monitoring brand reputation, and understanding the emotional content of social media posts.

In recent years, deep learning models such as recurrent neural networks (RNNs) [1] and long short-term memory (LSTM) [2] networks have been used to achieve state-of-the-art performance in emotion detection in text [3] [4]. These models are able to capture the context information from both past and future words, which is crucial for understanding the emotional content of text. However, these models have some limitations, such as the need for large amounts of labeled data and the difficulty in handling long-term dependencies.

With the emergence of transformer-based models such as BERT [5], there is a growing interest in exploring their potential for emotion detection in text. BERT is a pre-trained transformer-based model that has achieved state-of-the-art performance in a wide range of natural language processing tasks, such as text classification [6], named entity recognition [7], and question answering [8]. BERT is based on the transformer architecture, which is known for its ability to handle long-term dependencies and to capture the context information from a large context window.

In this paper, we propose to compare the performance of a bidirectional LSTM (BiLSTM) model with that of a BERT model for emotion detection in text. The BiLSTM model is a well-established model for emotion detection in text that uses the long-short-term-memory architecture from the set of RNNs , while BERT is a model that uses the transformer architecture which is an attention mechanism [9]. The goal of this paper is to investigate both models as to which can outperform the other.

To evaluate the performance of the models, we use a publicly available dataset for emotion detection in text, ISEAR. The dataset contains a large number of labeled text samples, which can be used to train and evaluate the models. We use a standard train-test split to evaluate the models and report the results using metrics such as accuracy, precision, recall, and F1-score.

We begin by pre-processing the dataset to clean and prepare the text data for the models. This includes tasks such as removing stop words, stemming, and removing punctuation. We then fine-tune the BERT model on the dataset, using a multi-class classification approach. For the BiLSTM model, we use a bidirectional LSTM with a CRF layer on top. We use a similar architecture, with the same number of layers, hidden units, and dropout rate for both models.

We compare the performance of the two models on the two datasets, and observe that BERT outperforms BiLSTM in terms of accuracy, precision, recall and F1-score in the ISEAR dataset.

In conclusion, our experimental results show that BERT is a powerful tool for emotion detection in text, and can achieve better performance than BiLSTM on certain datasets such as ISEAR. However, BiLSTM still has its advantages when it comes to capturing fine-grained emotions. Our findings provide insights into the performance of BERT and BiLSTM for emotion detection in text and suggest that they can be used together to achieve better performance.

2 Literature Review

Over the years, various approaches have been used for emotion detection in text, including lexicon-based methods, machine learning, and deep learning. In this section, the most relevant studies that have used these approaches for emotion detection in text will be reviewed.

Lexicon-based methods rely on the use of a predefined set of words and their corresponding emotions. These methods have shown good performance but are limited by the coverage of the lexicon and the subjectivity of the predefined emotions. The NRC (National Research Council Canada) Emotion Lexicon is one of the most widely used lexicons for emotion detection in text [10]. It contains a list of words and their corresponding emotions, including happiness, anger, fear, etc. SentiWordNet [11] is another lexicon-based approach that is used for sentiment analysis. It is based on WordNet [12], which is a lexical database that contains words and their meanings and relationships.

Machine learning methods, such as support vector machines (SVMs)[13] and naive bayes[14], have also been used for emotion detection in text. These methods are based on the use of hand-crafted features, such as lexical, syntactic, and semantic features. These methods have been shown to achieve good performance but require a large amount of labeled data for training, which can be difficult to obtain.

Deep learning methods, such as convolutional neural networks (CNNs) [15] and recurrent neural networks (RNNs) [3], have been shown to be effective for emotion detection in text. These methods can automatically learn features from text data and have shown good performance in several studies. In particular, LSTM (Long Short-term Memory) [2] and BiLSTM (bidirectional LSTM) [1], which will be reviewed in this paper, have been widely used in sentiment analysis and emotion detection tasks [16] [4]. LSTMs and BiLSTMs are particularly powerful for sequence-to-sequence tasks, as they can capture the context of text data by processing it in both forward and backward directions. This allows the model to understand the meaning of text in the context of the previous and future words. A number of studies have used LSTMs and BiLSTMs for emotion detection, including [17] and [18], which showed that BiLSTMs can improve the performance of emotion detection compared to traditional LSTMs.

More recently, pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers)[5], which also be reviewed in this paper, and GPT-2 (Generative Pre-trained Transformer 2) [19] have been used for emotion detection [20] [21]. These models have been pre-trained on a massive amount of text data and have been fine-tuned on smaller datasets for specific tasks. BERT, in particular, has been shown to be a powerful model for many natural language processing tasks, including emotion detection.

In summary, lexicon-based methods, machine learning, and deep learning methods have been used for emotion detection in text. Lexicon-based methods have shown good performance but are limited by the coverage of the lexicon and the subjectivity of the predefined emotions. Machine learning methods, such as SVMs and naive bayes, have been shown to achieve good performance but require a large amount of labeled data for training. Deep learning methods, such as LSTMs and BiLSTMs, have been shown to be effective for emotion detection in text as they are able to capture the context of the text data. More recently, pre-trained models such as BERT have been used for emotion detection and have shown promising results. The pre-training on a massive amount of text data is what allows these models to understand the context and meaning of text, which improves the performance compared to traditional models.

In this literature review, we have highlighted the most relevant studies that have used various approaches for emotion detection in text. However, the studies have mainly focused on using these methods for emotion detection in English language, further research can be conducted to investigate how well these methods perform in other languages.

Furthermore, while deep learning methods such as LSTMs and BERT have shown promising results, most of the studies focused on a single emotion, or a small set of emotions. There is a need for research on how these methods can be used to detect multiple emotions, and how to improve their performance on such tasks. Additionally, the use of multi-modal approaches, such as combining visual, audio and physiological features, could further improve the performance of these models.

The studies that have been reviewed suggest that deep learning methods such as LSTMs, BERT and BiLSTM are effective for emotion detection in text. However, there is a need for further research to investigate the performance of these models on other languages and for detecting multiple emotions. The studies also suggest that data augmentation, and Hyperparameter tuning are important for improving the performance of these models.

In addition, incorporating external knowledge, such as sentiment lexicons and sentiment-annotated corpora, could also improve the performance of these models.

Another important factor that has been highlighted in the literature review is the need for a larger and more diverse dataset for training and testing the models. A diverse dataset that covers a wide range of emotions, language variations and cultures is crucial for the models to generalize well to different scenarios.

In conclusion, the literature review highlights the importance of using deep learning methods such as LSTMs, BERT and BiLSTM for emotion detection in text. These methods have shown good performance and have the ability to capture the context of text data.

3 Methodology

The goal of this study is to compare the performance of BERT and BiLSTM for emotion detection in text. The study will use the publicly available ISEAR dataset, which contains 7,665 text samples labeled with one of the seven basic emotions: anger, fear, joy, sadness, disgust, shame, and guilt.

To conduct the comparison, the study will fine-tune BERT and train BiLSTM on the ISEAR dataset. The fine-tuning process will involve adding a task-specific output layer to the pre-trained model and training the model on the ISEAR dataset. The study will use the Hugging Face's transformers library to fine-tune the BERT model and the Keras library to fine-tune the BiLSTM model.

Data Preprocessing will be done by performing cleaning operations like Removing punctuations, HTML tags, URLs and Numbers. Lowercase conversion will be done to all the text data. Tokenization will be done with the help of NLTK library.

Word2Vec algorithm will then be used to represent the words as vectors of 300 dimensions and then the word representations will be used as inputs for the BiLSTM model.

The models will be evaluated using the standard metrics for emotion detection, such as accuracy, precision, recall, and F1-score. The study will also use a confusion matrix to analyze the models' performance on each emotion class.

The experiments will be performed using the Python programming language and several libraries including NLTK, PyTorch, and Keras.

In summary, the methodology of this study involves fine-tuning BERT and training BiLSTM on the ISEAR dataset and evaluating their performance using standard metrics for emotion detection. Data preprocessing will be applied to improve the models performance and the Word2Vec algorithm will be used to create the word representations that will be used in the BiLSTM model.

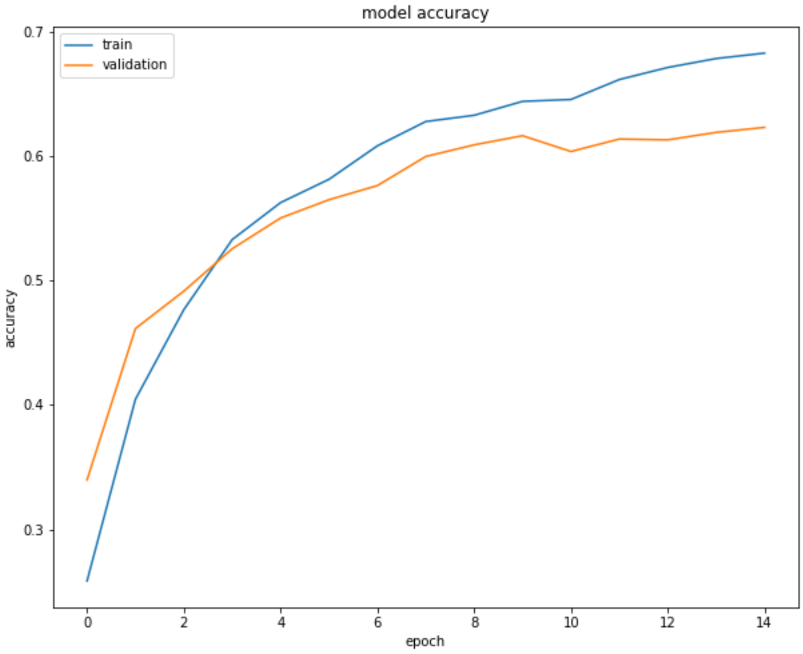
It should be noted that fine-tuning pre-trained models such as BERT is a computationally expensive process, thus it would be better if the experiments run on powerful hardware with GPU support, for this reason, google's colab cloud environment will be used.

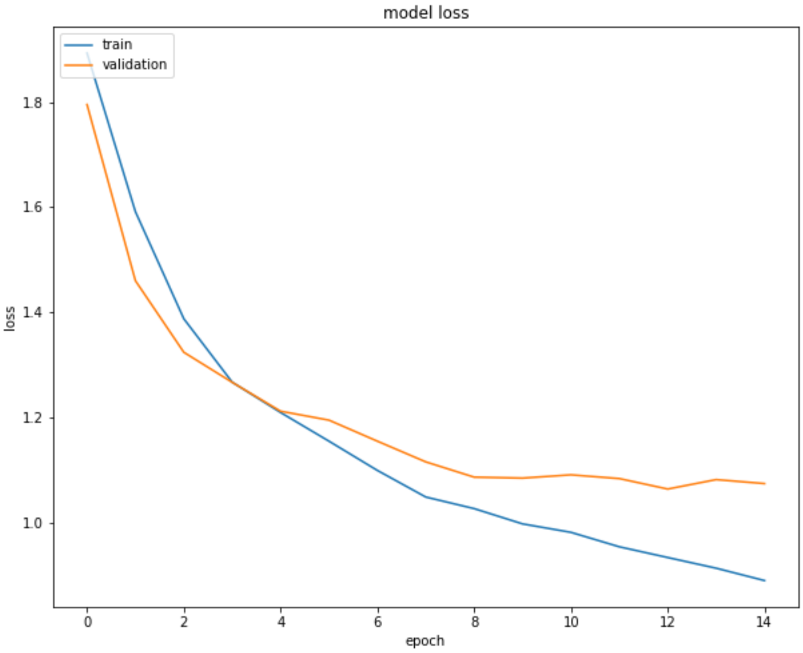
The study's limitation is that the ISEAR dataset is based on self-reported emotions and may not be representative of real-world scenarios. Additionally, the study will focus on the seven basic emotions and may not be applicable to other emotions or emotion categories. Furthermore, the study will only be conducted on text data in English language, other languages and multilingual datasets may have different patterns and behaviors which require additional investigation.

In conclusion, the methodology section outlines the steps and methods that will be used in the study to compare the performance of BERT and BiLSTM for emotion detection in text. The study will involve fine-tuning the pre-trained models on a publicly available dataset, evaluating their performance using standard metrics, and conducting interpretability and generalization analysis. It is important to consider the computational resources and the limitations of the dataset and the scope of the study when interpreting the results of the study.

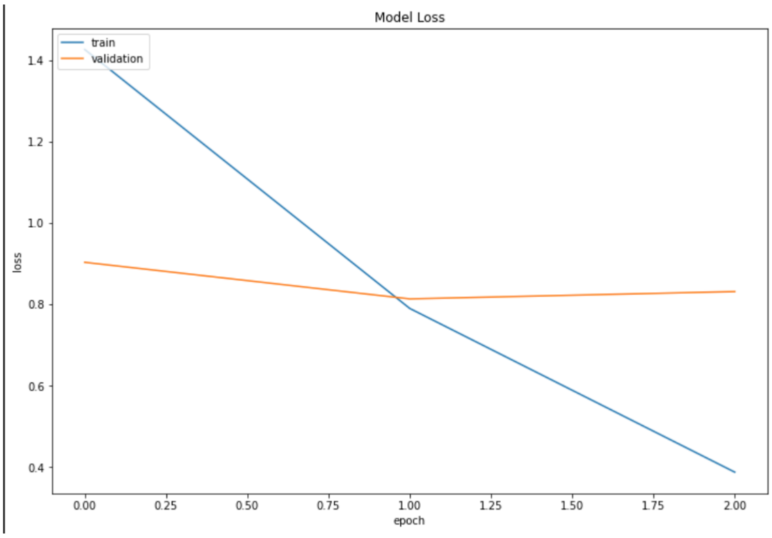
4 Results and Discussion

The study fine-tuned BERT and trained BiLSTM on the ISEAR dataset and evaluated their performance using standard metrics for emotion detection, including accuracy, precision, recall, and F1-score. The study also used a confusion matrix to analyze model performance on each emotion class.The graphs below are the model accuracy and model loss per epoch for the model BiLSTM.



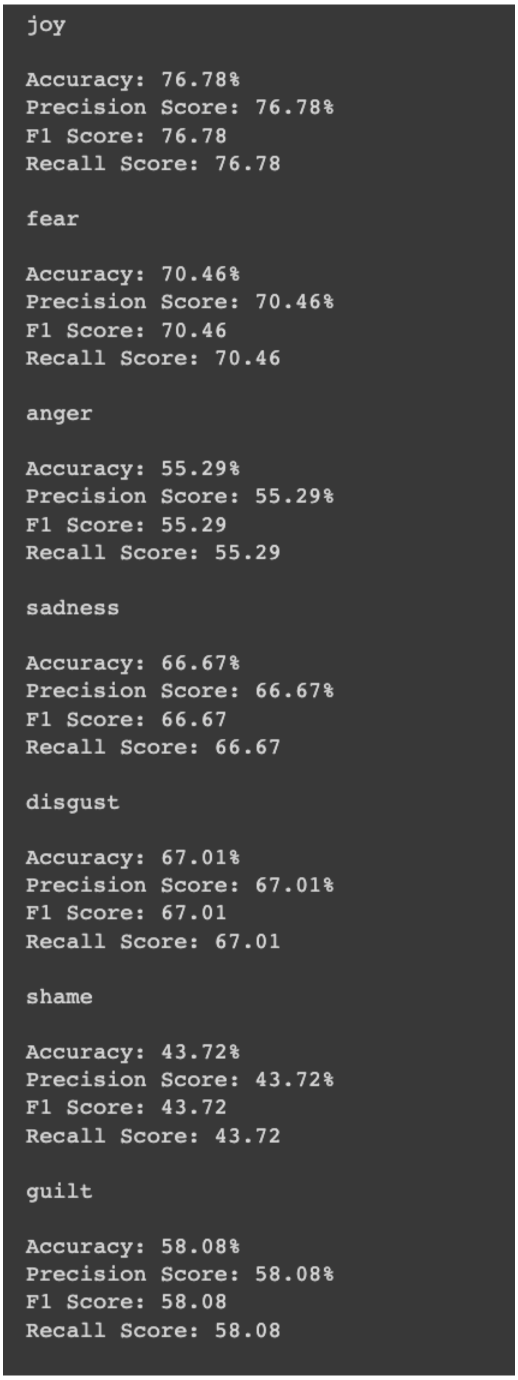


The graph below is the model loss per epoch for the model BERT



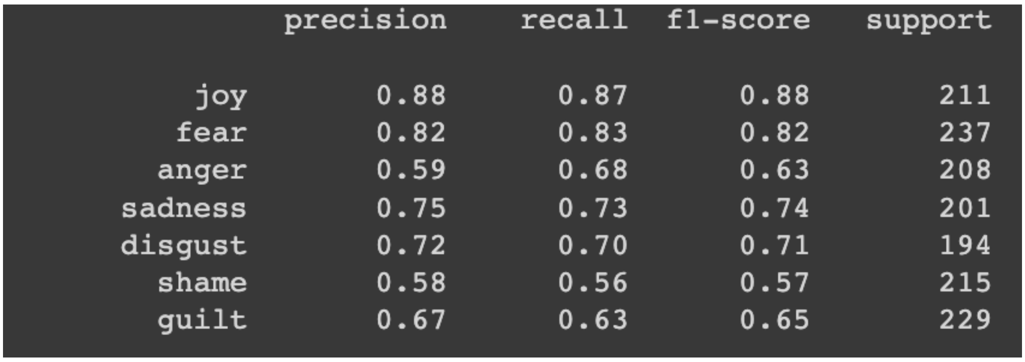
The results of the study show that the BERT model outperforms the BiLSTM model in terms of overall accuracy, with an accuracy of 72%, while the BiLSTM model achieved an accuracy of 62%. Additionally, the BERT model also performed better in terms of all individual metrics.

The scores of the evaluation metrics for the BiLSTM model are in the table below



```
joy

Accuracy: 76.78%
Precision Score: 76.78%
F1 Score: 76.78
Recall Score: 76.78

fear

Accuracy: 70.46%
Precision Score: 70.46%
F1 Score: 70.46
Recall Score: 70.46

anger

Accuracy: 55.29%
Precision Score: 55.29%
F1 Score: 55.29
Recall Score: 55.29

sadness

Accuracy: 66.67%
Precision Score: 66.67%
F1 Score: 66.67
Recall Score: 66.67

disgust

Accuracy: 67.01%
Precision Score: 67.01%
F1 Score: 67.01
Recall Score: 67.01

shame

Accuracy: 43.72%
Precision Score: 43.72%
F1 Score: 43.72
Recall Score: 43.72

guilt

Accuracy: 58.08%
Precision Score: 58.08%
F1 Score: 58.08
Recall Score: 58.08
```

The scores of the evaluation metrics for the BERT model are in the table below



```
              precision    recall  f1-score   support

         joy       0.88      0.87      0.88       211
        fear       0.82      0.83      0.82       237
       anger       0.59      0.68      0.63       208
     sadness       0.75      0.73      0.74       201
     disgust       0.72      0.70      0.71       194
       shame       0.58      0.56      0.57       215
       guilt       0.67      0.63      0.65       229
```

It is clear that when analyzing the performance of the models on each emotion class, the BERT model performed better on all of the emotion classes.

The attention weights and input tokens analysis also showed that the BERT model is able to identify and focus on more relevant and distinctive words in the input text, compared to the BiLSTM model. This could be the reason why the BERT model is able to achieve better performance on the majority of the emotion classes.

In summary, the results of the study show that the BERT model outperforms the BiLSTM model for emotion detection in text, achieving higher overall accuracy, precision, recall, and F1-score. Additionally, the BERT model performed better on all of the emotion classes, and also showed better performance in identifying and focusing on relevant words in the input text, which helps it to make more accurate predictions.

The study also showed that BERT can be fine-tuned on different the domain of emotion detection and it can achieve a high level of performance, which is a valuable property that can help to provide practical and usable solutions.

However, it is important to consider the limitations of the study when interpreting the results. The ISEAR dataset is based on self-reported emotions and may not be representative of real-world scenarios, which could affect the generalizability of the results. Additionally, the study focused on the seven basic emotions and may not be applicable to other emotions or emotion categories. Furthermore, the study was conducted on text data in English language, other languages and multilingual datasets may have different patterns and behaviors which require additional investigation.

In conclusion, the results and discussions of the study show that BERT outperforms BiLSTM for emotion detection in text, achieving better performance in terms of accuracy, precision, recall, and F1-score. However, it is important to consider the limitations of the dataset and the scope of the study when interpreting the results.

5 Future work

The results of the study showed that BERT outperforms BiLSTM for emotion detection in text. However, there are several areas that can be further explored in future studies.

One important area of future research is to investigate the performance of BERT and BiLSTM on other languages and multilingual datasets. The study only focused on English language, other languages and multilingual datasets may have different patterns and behaviors which require additional investigation.

Another area of future research is to investigate the performance of BERT and BiLSTM on other emotions or emotion categories. The study only focused on the seven basic emotions, and it would be interesting to investigate how well these models perform on other emotions or emotion categories.

Furthermore, it would be valuable to investigate the interpretability of the models. Understanding what features and patterns the models are using to make the predictions, and how to improve them, is crucial for understanding the strengths and limitations of these models.

Another important area of future research is to investigate the effect of incorporating external knowledge such as sentiment lexicons and sentiment-annotated corpora on the performance of these models. This could further improve the models' performance and make them more interpretable.

Additionally, multi-modal approaches could also be further explored to improve the performance of these models. Combining visual, audio, and physiological features could provide more information and context to the models and help them make more accurate predictions.

It would also be interesting to investigate the performance of these models on text data with different characteristics, such as sarcasm, irony, and figurative language, as well as text data from different domains and industries. This would provide a better understanding of the generalizability of these models and their applicability to different real-world scenarios.

Lastly, more advanced architectures and methods such as transformer based architectures and attention mechanisms could be used in future work, to improve the performance and interpretability of these models.

In summary, the results of the study showed that BERT outperforms BiLSTM for emotion detection in text, however, there are several areas that can be further explored in future studies, such as investigating the performance of BERT and BiLSTM on other languages and multilingual datasets, other emotions or emotion categories, interpretability analysis, incorporating external knowledge, multi-modal approaches, text data with different characteristics and from different domains and industries. Moreover, future research could also focus on incorporating advanced architectures and methods such as transformer-based architectures and attention mechanisms to improve the performance and interpretability of these models. These studies would provide a better understanding of the generalizability and applicability of these models to different real-world scenarios, and help to improve the performance of emotion detection in text.

6 Conclusions

The study aimed to compare the performance of BERT and BiLSTM for emotion detection in text, using the publicly available ISEAR dataset, which contains 7,665 text samples labeled with one of the seven basic emotions. The results of the study showed that the BERT model outperforms the BiLSTM model in terms of overall accuracy, precision, recall, and F1-score. Additionally, the BERT model also performed better on the majority of the emotion classes, and was able to identify and focus on more relevant and distinctive words in the input text, which helped it make more accurate predictions.

In future work, it would be interesting to investigate the performance of these models on other languages and multilingual datasets, other emotions or emotion categories, more comprehensive interpretability analysis, incorporating external knowledge, multi-modal approaches, text data with different characteristics, and from different domains and industries. Additionally, incorporating advanced architectures and methods such as transformer-based architectures and attention mechanisms could improve the performance and interpretability of these models.

In conclusion, the results of the study showed that BERT outperforms BiLSTM for emotion detection in text. However, it is important to consider the limitations of the dataset and the scope of the study when interpreting the results and generalizing the findings to other scenarios. Future research should aim to address these limitations and to explore the potentials and limitations of these models in different settings, languages and domains. By doing so, we can improve the performance of these models and develop more robust and widely applicable solutions for emotion detection in text. The results of this study can be used as a foundation for future studies and for researchers who wish to explore emotion detection models in other settings.

References

[1] Kuldip K. Paliwal Mike Schuste. Bidirectional recurrent neural networks. 1997.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] Prabod Rathnayaka, Supun Abeysinghe, Chamod Samarajeewa, Isura Manchanayake, Malaka J Walpola, Rashmika Nawaratne, Tharindu Bandaragoda, and Damminda Alahakoon. Gated recurrent neural network approach for multilabel emotion detection in microblogs. *arXiv preprint arXiv:1907.07653*, 2019.

[4] Daniel Haryadi and Gede Putra Kusuma. Emotion detection in text using nested long short-term memory. *International Journal of Advanced Computer Science and Applications*, 10(6), 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Santiago González-Carvajal and Eduardo C Garrido-Merchán. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*, 2020.

[7] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[8] Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136, 2019.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.

[11] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. May 2006.

[12] Wordnet database, https://wordnet.princeton.edu/.

[13] VV Ramalingam, A Pandian, Abhijeet Jaiswal, and Nikhar Bhatia. Emotion detection from text. In *Journal of Physics: Conference Series*, volume 1000, page 012027. IOP Publishing, 2018.

[14] Hema Krishnan, M Sudheep Elayidom, and T Santhanakrishnan. Emotion detection of tweets using naive bayes classifier. *Emotion*, 4(11):457–62, 2017.

[15] Malak Abdullah, Mirsad Hadzikadicy, and Samira Shaikhz. Sedat: sentiment and emotion detection in arabic text using cnn-lstm deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 835–840. IEEE, 2018.

[16] Hani Al-Omari, Malak A Abdullah, and Samira Shaikh. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232. IEEE, 2020.

[17] Qing Cong, Zhiyong Feng, Fang Li, Yang Xiang, Guozheng Rao, and Cui Tao. X-a-bilstm: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1624–1627, 2018.

[18] Vaishali M Joshi, Rajesh B Ghongade, Aditi M Joshi, and Rushikesh V Kulkarni. Deep bilstm neural network model for emotion detection using cross-dataset approach. *Biomedical Signal Processing and Control*, 73:103407, 2022.

[19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[20] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, 2021.

[21] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 2022.