Emotion Detection in Text: RoBERTa Vs Att-BiLSTM

Madhurima Dutta (madhurima.dutta@mail.utoronto.ca)
University of Toronto, 1004832577

Luke Liu (theluke.liu@mail.utoronto.ca)
University of Toronto, 1005265614

Abstract

This paper investigates the ability of natural language processing (NLP) models to recognize fine-grained emotions in diverse textual contexts, such as poems and dialogues of a script. The study compares the performance of two state-of-the-art NLP architectures, BERT (RoBERTa) and Att-BiLSTM, in recognizing six basic emotions (anger, fear, joy, love, sadness, and surprise) in three datasets: Twitter, Poem, and Reddit. The models are trained on Twitter data, with variations in the addition of external data, and tested on all three datasets to evaluate their generalization abilities. The findings reveal that RoBERTa generalizes to new forms of data better than Att-BiLSTM. Label-wise analysis of the models' performance is also provided. The results demonstrate the importance of evaluating models on diverse datasets to assess their generalization abilities accurately.

Keywords: Emotion recognition, NLP, BERT, LSTM, RoBERTa, Att-BiLSTM, generalization, sentiment analysis, machine learning

Introduction

Emotions are a ubiquitous aspect of human communication, and detecting or comprehending them is vital for effective social interaction. Although humans are adept at recognizing emotions conveyed through various forms of communication, such as speech, facial expressions, and body language, machines have long struggled with this task. However, recent advancements in natural language processing (NLP) have improved the ability of machines to detect emotions in text. NLP models trained on text messages for emotion recognition can learn to recognize emotions in new texts that they have not encountered before.

While existing studies have shown that NLP models can recognize emotions in simple text messages(Alswaidan & Menai, 2020), questions remain about their ability to recognize emotions in different forms of text data, such as Poems or dialogues of a script. Recent research(Kaur & Saini, 2018) has shown that emotion detection can vary significantly depending on the writing style. Their study examined differential emotion detection and sentiment analysis in text corpora using both formal and informal writing styles for Support Vector Machines, Decision Trees, and Naive bayes. They found that NLP models performed better in detecting emotions and sentiments in informal writing styles, such as tweets and chats, as compared to formal writing styles, such as essays and research papers. These findings raise important questions about the generalization abilities of NLP models to recognize emotions in diverse textual contexts.

In this paper, we investigate the ability of two different NLP architectures, BERT (RoBERTa) and Att-BiLSTM, to recognize fine-grained emotions such as anger, fear, and happiness in diverse textual contexts, including both formal and informal writing styles. We choose RoBERTa and Att-BiLSTM due to their popularity and reputation of good performance on NLP tasks. Our primary aim is to compare the generalization abilities of these two models to recognize emotions in different types of text data. Our results have implications for the development of more robust NLP models that can accurately generalize and recognize emotions in various types of textual data.

BERT and RoBERTa

BERT, or Bidirectional Encoder Representations from Transformers, is an NLP model made by Google that is capable of understanding contexts of words in a sentence. Using a technique called "pre-training", it is already trained on a large corpus of texts and only optionally needs to be fine-tuned for specific uses, such as emotion detection (Devlin, Chang, Lee, & Toutanova, 2019). However, researchers Yinhan Liu et al. improved upon BERT with a series of modifications to the pre-training process. These include training the model on more data, longer sequences, removing the next sentence prediction task, and dynamic changes to the masking patterns in the training (Y. Liu et al., 2019). These modifications significantly improved the performance on several NLP benchmarks, hence the name RoBERTa (Robust BERT). Amongst these benchmarks is the GLUE benchmark, which evaluates performance of NLP models on tasks such as sentiment analysis.

Att-BiLSTM

We compare the BERT based RoBERTa model to a bidirectional long short-term memory model with an attention mechanism and convolutional layers (Att-BiLSTM) (G. Liu & Guo, 2019). The attention mechanism in this model allows it to focus on specific parts of the input sequence that are most relevant to the classification task at hand. This improves the model's ability to capture relationships between different parts of the input sequence and enhances performance on more complex NLP tasks like emotion detection (G. Liu & Guo, 2019). The performance of Att-BiLSTM was compared to other state of the art models on benchmark

datasets including sentiment analysis and emotion detection, and it outperformed other models on most datasets, achieving performance comparable to BERT-based models.

Hypothesis

We hypothesize that RoBERTa will outperform Att-BiLSTM due to its pretraining on a massive text corpus. While Att-BiLSTM includes an attention mechanism and convolutional layers (G. Liu & Guo, 2019), RoBERTa has a more complex architecture with multiple layers of transformers which are designed to capture long-range dependencies in the text (Y. Liu et al., 2019). Additionally, RoBERTa's pretraining improvements include dynamic changes to masking patterns during training which give it an edge when it comes to generalizing to different types of text data.

H_1: The RoBERTa model will have significantly higher accuracy in classifying emotions in a generalized, unrecognized context.

H_0: There is no significant difference in accuracy between RoBERTa and Att-BiLSTM models in classifying emotions in a generalized, unrecognized context.

Methods

Material

Dataset For this study, we utilized three datasets: Twitter, Poem, and Reddit.

The Twitter dataset is the Emotions dataset of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise (Saravia, Liu, Huang, Wu, & Chen, 2018). It comprises 20,000 labeled text files that are divided into three data files: train, text, and val, which will serve as the primary basis for training the models.

The Poem dataset is a collection of 716 poems labeled with emotions obtained from Mendeley Data (Ponnarassery, 2017). There are a total of 717 poems categorized into the following nine emotions: love, sad, hate, anger, fear, surprise, courage, joy, and peace.

The Reddit dataset was retrieved from 58k English-speaking Reddit comments labeled for 27 emotion categories or Neutral (Demszky et al., 2020).

We chose the Reddit and Poem datasets in addition to the Twitter data to represent different emotional expression contexts. The Reddit dataset, similar to the Twitter dataset, consists of short texts that explicitly convey emotions. However, it differs slightly in that it contains unfiltered texts that include emojis and profanity. On the other hand, the Poem dataset comprises longer texts that implicitly express emotions through tone and language. Our aim was to evaluate the models' ability to recognize and generalize emotions across various textual data types and emotional contexts by testing them on these diverse datasets.

Att-BiLSTM We implemented the Att-BiLSTM model for text classification using the source code available on GitHub (Renovamen, 2021). Our implementation focuses on adapting the trainer, tester, and the Att-BiLSTM model. Both the trainer and the tester has been modified to fit to our dataset format. We also modified the tester to include additional evaluation metrics as described in the Procedure section.

Table 1: Distribution of Label count in Training Data

Dataset	anger	fear	joy	love	sad	surprise
TD1	2159	1937	5362	1304	4666	572
TD2	2928	2218	6115	2101	5320	1053

RoBERTa RoBERTa was loaded following the HuggingFace RoBERTa transformers 3.0.2 documentation (?, ?). The model was trained using a batch size of 16, and trained for 3 epochs. During training and evaluation, the batch size remained the same for both CPU and GPU. The evaluation strategy used was "steps", with evaluation occurring every 1000 steps (as specified by the logging steps parameter).

Procedure

Data Pre-processing All three datasets were filtered to have only six main emotion labels: *anger*, *fear*, *joy*, *love*, *sadness* and *surprise*. The text samples from all three dataset for both models was tokenized using a pre-trained 'bert-base-uncased' from the Hugging Face Transformers library (Devlin, Chang, Lee, & Toutanova, 2018). The labels from all three dataset was convertedf into argmax encodings, where each label is represented by an integer value starting at 0.

Training Models For our experiments, we used two types of training data: TD1, which consisted only of Twitter train dataset, and TD2, which included 10% of both Reddit and Poem data in addition to Twitter train dataset. We can refer Table 1 to learn the distribution of the samples across each emotion.

Based on the training datasets, we trained four different models: **M1**, an Attention-based BiLSTM trained on TD1; **M2**, an Attention-based BiLSTM trained on TD2; **M3**, a RoBERTa model trained on TD1; and **M4**, a RoBERTa model trained on TD2. By comparing the performance of our models trained on these two datasets, we aimed to evaluate the impact of adding a small amount of external data on model generalization.

Testing Models Since our goal is to evaluate model performance on generalizing emotion detection, we will test on all three datasets: Twitter, Poem and Reddit. Testing all four models on the Twitter dataset will provide us with a baseline to compare how well the model generalizes to the rest of the datasets.

Our main testing metric for evaluating overall performance of the models is the accuracy rate, defined as the number of correct predictions over the total number of labels. This signifies how well the model classifies.

We will also be looking further into each model by examining the accuracy rate for each emotion in the dataset. Hence,

Table 2: Accuracy Rates for each Model

Dataset	M1	M2	M3	M4
Twitter	92.7	92.8	93	93
Reddit	22.8	67.7	46.4	73
Poem	14.5	41.5	27.8	42.3

Table 3: M1 - Att-BiLSTM on TD1

Dataset		anger	fear	joy	love	sad	surprise
Twitter	Acc	95.3	88.4	96.4	76.7	96.4	63.6
	TPR	91.0	92.1	92.9	86.5	96.2	79.2
Reddit	Acc	93.9	0.7	11.8	0.2	4.8	0.6
	TPR	21.5	30	30.1	38.1	61.9	53.8
Poem	Acc	66.7	20	20.8	1.5	6.2	0
	TPR	9.1	33.3	27.5	25	33.3	0

for the label-wise analysis, we will calculate both accuracy rate and the True Positive Rate (TPR). TPR is the ratio of correct positive predictions to the total number of actual positive labels in the dataset and indicates how often the model's prediction for a particular label is correct among all the times it predicts that label. Finally, we will be plotting a confusion matrix to learn the distribution of labels for each model and dataset. This will help us learn how often is one emotion is mislabeled as another.

Results

Overall Analysis

According to Table 2, the null hypotheses can be rejected since BERT-based models (M3 and M4) demonstrated higher test accuracy across all three datasets. The study also found that the models showed better generalization to the Reddit dataset than to the Poem dataset. This is evident in Figures 2 and 3, where the Reddit dataset shows better categorization on the diagonals of the confusion matrices. Additionally, including 10% of data from both the Reddit and Poem dataset to the Twitter training set resulted in a significant improvement in the performance of both LSTM and BERT-based models, with an approximate 20% increase in accuracy.

However, it is important to note that while there was a drastic improvement in model performance, none of the models were able to achieve similar accuracy rates for the Reddit and Poem dataset as the Twitter dataset when applied to other datasets.

Label-wise Analysis

Since all 4 models predicts almost accurately for all labels with Twitter Data (as seen in Figure 1), we decided to focus mainly on the performance for Poem Data and Reddit Data.

Anger The performance of model M1 in predicting the label *anger* is excellent, with an accuracy rate of almost 90% for the Reddit dataset. However, based on the true positive rate (TPR) shown in Table 3, we can see that M1 predicts *anger* frequently, resulting in a very low true positive rate (21.5% for Reddit, 9.1% for Poem), despite the high accuracy rate (93.9% for Reddit, 66.7% for Poem). This can be observed in Figures 2 and 3, where the Att-BiLSTM model has the highest false positive rate among all labels. In the case of the Poem test set (Figure 2), all models except M1 mislabel *anger* as *sadness*, whereas all models accurately predict *anger* for the Reddit test set (Figure 3).

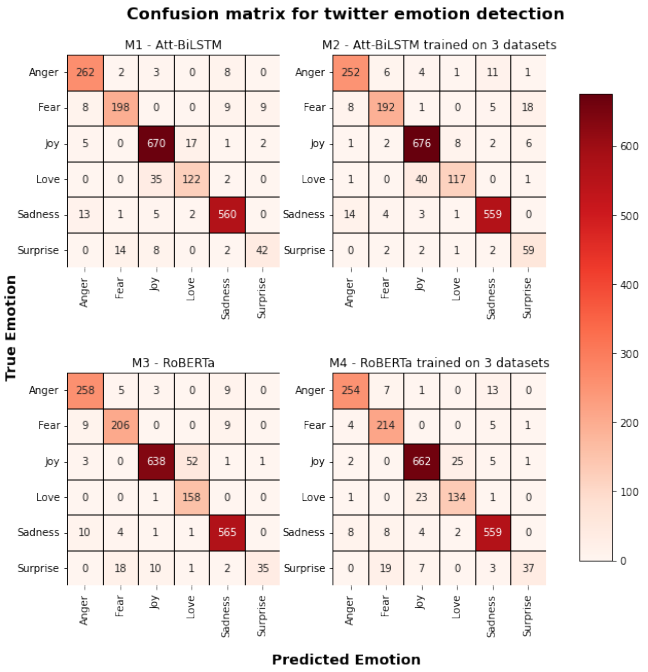

Figure 1: Confusion matrix for Twitter Dataset

Table 4: M3 - RoBERTa on TD1

Dataset		anger	fear	joy	love	sad	surprise
Twitter	Acc	93.82	91.96	91.8	99.37	97.25	53.03
	TPR	92.14	88.41	97.7	74.53	96.42	97.22
Reddit	Acc	66.45	50.90	62.09	32.05	41.43	18.54
	TPR	41.15	36.53	43.91	66.31	54.54	60.56
Poem	Acc	16.67	20.00	32.08	36.92	22.50	22.22
	TPR	15.38	4.76	29.31	39.34	36.73	40.00

Fear For the Poem dataset, Tables 3, 6, and 4 show that the performance of M1, M2, and M3 in predicting *fear* is very similar and better than that of M4. This contradicts our overall accuracy results, which indicate that M4 outperforms the rest of the models. On the other hand, for the Reddit dataset, M4 shows the best performance in predicting *fear*, with an accuracy rate of almost 60%. Models with combined datasets, i.e. M2 and M4, perform better than M1 and M3. Similar to the label *anger*, M2, M3, and M4 often confuse *fear* with *sadness* in the Poem dataset, as seen in Figure 2.

Joy With the Poem data, model M2 predicts *fear* best with almost 50% accuracy rate (Table 6). This is contrary to the fact that BERT performs better than LSTM. From Figure 2, we can see that BERT models M3 and M4 show confusion in labeling *joy*. M4 shows major confusion with *love*, while M3 also shows slight confusion with both *love* and *fear*. With the

Table 5: M4 - RoBERTa on TD2

Dataset		anger	fear	joy	love	sad	surprise
Twitter	Acc	92.36	95.54	95.25	84.28	96.21	56.06
	TPR	94.42	86.29	94.98	83.23	95.39	94.87
Reddit	Acc	66.96	63.30	75.20	79.34	78.11	71.99
	TPR	66.45	50.90	62.09	47.12	41.43	50.00
Poem	Acc	8.33	10.00	32.08	75.38	36.25	44.44
	TPR	14.29	5.88	48.57	47.12	46.03	50.00

Table 6: M2 - Att-BiLSTM on TD2

Dataset		anger	fear	joy	love	sad	surprise
Twitter	Acc	91.6	85.7	97.3	73.6	96.2	89.4
	TPR	91.3	93.2	93.1	91.4	96.5	69.4
Reddit	Acc	65.3	59.6	66.5	84.5	63	58.7
	TPR	70	58.5	37.9	42.9	46.4	66.7
Poem	Acc	16.7	20	47.2	60	32.5	44.4
	TPR	26.7	28.6	37.9	42.9	46.4	66.7

Reddit data, M4 outperforms M2 even though both models surpass 50% accuracy rate. Surprisingly, M1 shows worse performance in predicting *joy* with Reddit data than Poem data.

Love In contrast to *joy*, M4 performs better at predicting *love* for the Poem dataset, while M2 is the best performer for the Reddit dataset, with an accuracy rate of nearly 80%. Figure 2 shows that M3 is confused between *joy* and *fear* while predicting *love*. Similarly, in the confusion matrix for the Reddit dataset, 3, M3 shows confusion by mislabeling *love* as either *anger* or *joy*.

Sadness Predicting *sadness* for both the Poem and Reddit dataset follows a very similar trend across all models. From Table 5, we can see that M4 performs the best at predicting *sadness*. We can also conclude that models with combined datasets, i.e., M2 & M4, perform well in predicting *sadness*, with an accuracy rate for the Reddit dataset almost reaching 80%. In Figure 2, we can see that even though the performance of M2 is high, it mislabels *sadness* as *love* more frequently than *sadness*. Similarly, M3 frequently predicts *joy* instead of *sadness*, while M4 again predicts *love*.

Surprise We can observe from Table 3 that M1 performs poorly in labeling *surprise* for both the Poem and Reddit dataset. However, there is a significant improvement with M2 on both datasets, with an increase in accuracy of over 40%. For the Poem dataset, as seen in Figure 2, all models barely predict *surprise*. In contrast, for the Reddit dataset, as shown in Figure 3, M3 performs worst in predicting *surprise* as it frequently labels *anger* and *joy* instead.

Discussion

Our study aimed to evaluate the performance of different deep learning models for emotion detection across three datasets: Twitter, Reddit, and Poem. The overall accuracy results (Table 2) showed that BERT-based models (M3 and M4) outperformed LSTM-based models (M1 and M2) across all three datasets. Thus, we are able to reject our initial null hypthesis.

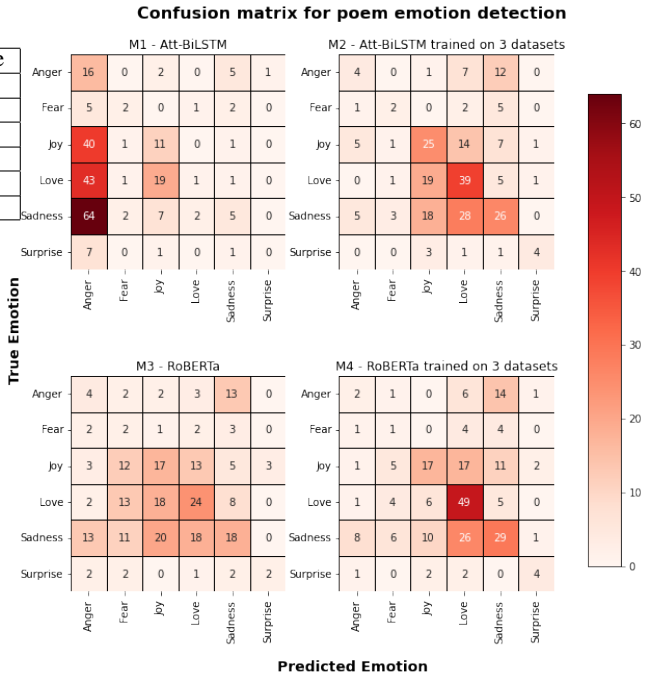

Figure 2: Confusion matrix for Poem Dataset

Every model showed better generalization to the Reddit dataset than to the Poem dataset, which could be due to the higher similarity between the Reddit and Twitter dataset. Even after the large improvement with the introduction of the 10% of both Reddit and Poem data, the models struggle to achieve 80% - a reasonable benchmark of generalizability for our task.

From Figures 1, 2 and 3, we can also see that both Att-BiLSTM and RoBERTa models generalizes better and shows higher performance just by adding 10% of both Reddit and Poem data.

A closer look at the label-wise performance of the models revealed some interesting findings. For instance, LSTM-based model M1 predicted the label *anger* with high accuracy, but with a low true positive rate (TPR), suggesting that it frequently incorrectly predicted *anger* even when it was not present. Additionally, M1 performed poorly in predicting the label *surprise* for both the Poem and Reddit dataset, which was significantly improved with the addition of 10% of data from both datasets to the Twitter training set (M2).

One interesting observation was that even though RoBERTa (M4) had the best overall accuracy across all datasets, it mislabeled Poem data with the label *joy* and *sadness* frequently. This stood out as in both TD1 and TD2, *joy* and *sadness* outnumbered the others by a vast amount which should have resulted in a bias towards predicting them better. Furthermore, M3 and M4 had relatively poor performance in predicting the label *surprise* across all datasets with M4 performing slightly better.

The models M2, M3, and M4 frequently mislabeled *anger* and *fear* samples as *sadness* in the Poem dataset. Addition-

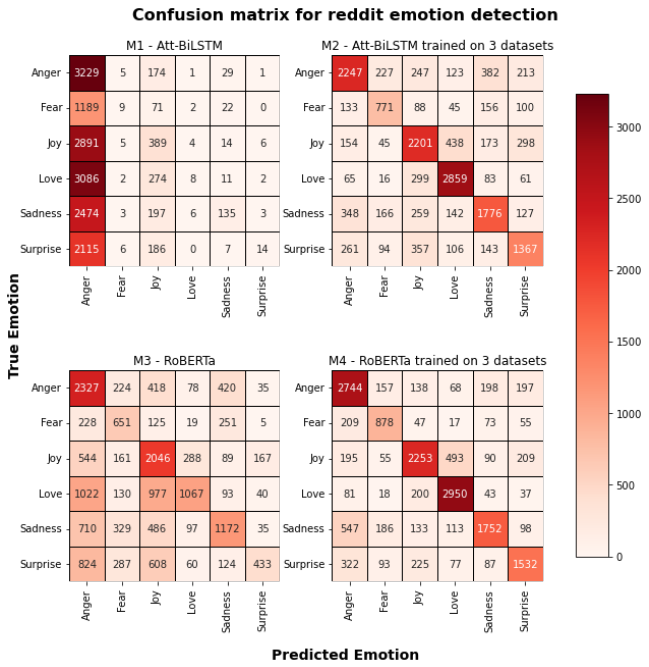

Figure 3: Confusion matrix for Reddit Dataset

ally, M3 showed confusion between *love* and *joy*.

While the Poems are labeled by professional poets and writers, there is always a degree of subjectivity and ambiguity when it comes to textual emotion, especially with larger contexts like in Poem data. Emotions may have certain key words that are indicators (Bandhakavi, Wiratunga, Massie, & Padmanabhan, 2017), and these may have overlap in similar emotions (like fear and anger, both high stress emotions (Gu, Wang, Wang, & Huang, 2016)). On top of this, commonly used features for emotion detection are not always effective enough to capture all of the nuances involved in emotion (Zad, Heidari, Jones, & Uzuner, 2021). This also explains why the architecture of the pretrained RoBERTa, which is more capable of capturing long range dependencies within the text, has significantly better performance than the Att-BiLSTM.

Limitations

The limitations of our study include the unequal distribution of samples across labels, which could affect the generalization of the models. Future studies should consider adding weights and biases to account for such imbalance. Another limitation is the lack of preprocessing of the Poem data to remove newline characters, which may have affected the model's performance. Furthermore, none of the models achieved an accuracy rate above 80%, which indicates that they may not generalize well to unseen data. This could be due to the subjective nature of emotions and the complexity of language. Future studies should continue to explore ways to improve the accuracy and generalizability of emotion detection models.

Extensions

Several directions for future research are worth exploring. First, it would be interesting to analyze the impact of different linguistic features on the classification. For example, one could investigate how much influence various features such as rhyme, alliteration, and meter have on the classification outcomes. This could provide insights into what aspects of language are most important for the classification task as well as what is causing frequent confusion and mislabelling such as between *fear* and *sadness*.

Second, the current dataset is imbalanced (as seen in Table 1), with many more samples in the *joy* category than in the other categories, and a significantly low number of samples of *surprise*. A larger, more balanced dataset would enable more robust evaluations of the models' performance across all labels.

Third, it could be valuable to train and test the models on different types of Poem data, such as haikus, sonnets, or free verse. These have varying lengths (haikus, for example, are much shorter but just as, if not more complex).

These extensions could contribute to a more comprehensive understanding of the strengths and limitations of the current models, as well as potential improvements for future research.

Conclusion

In conclusion, our study evaluated the performance of different deep learning models for emotion detection across three datasets: Twitter, Reddit, and Poem. BERT-based models (M3 and M4) outperformed LSTM-based models (M1 and M2) across all three datasets, but no model achieved high enough accuracy rates for us to conclude acceptable generalization (80%). Our label-wise analysis revealed interesting findings, such as the mislabeling of *joy* frequently with the Poem data by RoBERTa (M3), and the difficulty of correctly predicting the *surprise* label by all models.

The limitations of our study included the unequal distribution of samples across labels, the lack of preprocessing of the Poem data, and the low overall accuracy rates. Future studies should consider adding weights and biases to account for label imbalance or gather a larger, more balanced dataset for better generalization. Preprocessing the data more carefully, and exploring other types of Poem data such as haikus as well as analyzing the impact of different linguistic features on the classification could provide useful insights on the reason behind mislabeling.

Overall, emotion detection from textual data is a challenging task that requires further research and development to achieve higher accuracy rates and better generalization. Our study contributes to this area by evaluating the performance of different models across multiple datasets and providing insights into their strengths and weaknesses.

Appendix

The code can be found in this GitHub repository: madhurima236/COG403-Att-BiLSTM-and-RoBERTa.

References

Alswaidan, N., & Menai, M. E. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, *62*(8), 2937–2987. doi: 10.1007/s10115-020-01449-0

Bandhakavi, A., Wiratunga, N., Massie, S., & Padmanabhan, D. (2017). Lexicon generation for emotion detection from text. *IEEE Intelligent Systems*, *32*(1), 102-108. doi: 10 .1109/MIS.2017.22

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In *58th annual meeting of the association for computational linguistics (acl).*

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. Retrieved from http://arxiv.org/abs/1810.04805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Gu, S., Wang, W., Wang, F., & Huang, J. H. (2016). Neuromodulator and emotion biomarker for stress induced mental disorders. *Neural Plasticity*, *2016*, 1–6. doi: 10.1155/2016/2609128

Kaur, J., & Saini, J. R. (2018, Dec). *Emotion detection and sentiment analysis in text corpus: A differential study with informal and formal writing styles.* Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3291299

Liu, G., & Guo, J. (2019). Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, *337*, 325-338. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231219301067 doi: https://doi.org/10.1016/j.neucom.2019.01.078

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.*

Ponnarassery, S. (2017). *Poem Emotion Recognition Corpus (PERC).* Mendeley Data. doi: 10.17632/n9vbc8g9cx.1

Renovamen. (2021). *Text classification.* https://github.com/Renovamen/Text-Classification. GitHub.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018, October-November). CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3687–3697). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D18-1404 doi: 10.18653/v1/D18-1404

Zad, S., Heidari, M., Jones, J. H. J., & Uzuner, O. (2021). Emotion detection of textual data: An interdisciplinary survey. In *2021 ieee world ai iot congress (aiiot)* (p. 0255-0261). doi: 10.1109/AIIoT52608.2021.9454192