Multi-Class Classification using BERT models for Detecting Depression Signs from Social Media Text

# Introduction

The objective of the paper is to identify indications of depression in an individual through their social media messages and posts, where people frequently express their emotions and feelings. The system is designed to classify signs of depression into three categories, namely "not depressed," "moderately depressed," and "severely depressed," based on social media postings in the English language. To achieve this objective, the paper used the pre-trained BERT model from the transformer library.

# Related Work

In order for early intervention to be viable, depression detection models need to be highly accurate and fast. (Shen et al., 2017) proposed the extraction of six feature groups, which were then used to train a multi-modal depression dictionary learning model for identifying depressed Twitter users. (Burdisso et al., 2019) introduced the SS3 text classification system, which is both easy to use and effective for early diagnosis of depression in social media streams. (Lin et al., 2020) presented SenseMood, a system that utilizes a BERT classifier and a CNN to classify social media messages and images as either depressed or not depressed.

BERT models are commonly utilized in the field of natural language processing. To gain a deeper understanding of how these models operate, (Van Aken et al., 2019) conducted a Layer-Wise Analysis of Transformer Representations. (Devlin et al., 2018) demonstrated how pre-trained models can be leveraged for natural language interpretation. (Acheampong et al., 2021) provide an overview of BERT-based models for text-based emotion recognition. (Xin et al., 2020) proposed an early stopping modification of BERT to improve inference speed.

# Experiments

## Dataset

The dataset used in this task consists of social media text in English sourced from a publicly available repository on GitHub. The dataset is composed of three columns: the pid (unique identifier), the social media text in English, and the label which is categorized into three classes: "not depressed", "moderately depressed", and "severely depressed". Both the training and test datasets contain data that belong to these three classes. The training set has 6006 entries, the test set has 3245 entries, and the development set has 1000 entries. In the training set, there are 650 "not depressed" entries, 3101 "moderately depressed" entries, and 2255 "severely depressed" entries. In the test set, there are 228 "not depressed" entries, 2169 "moderately depressed" entries, and 848 "severely depressed" entries. In the development set, there are 90 "not depressed" entries, 400 "moderately depressed" entries, and 510 "severely depressed" entries. The statistics of each set are also presented in Figure 1.

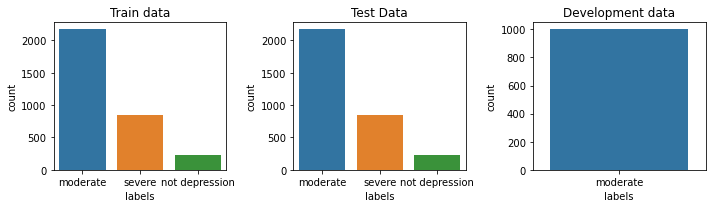


Figure 1: Statistics of the subsets in the dataset.

## Implementation

The implementation of the paper (Anantharaman et al., 2022) was performed using the GitHub repo provided by the author of the paper at this link: <https://github.com/rafalposwiata/depression-detection-lt-edi-2022/tree/main>

## Results

The use repo listed the three models for the classification of the text labelled BERT, Roberta, and DE-Roberta. The BERT showed 0.43% accuracy while Roberta showed 0.41% accuracy and the de Roberta showed a 0.43% accuracy score. The classification report of each model is shown in below Table 1.

Table 1: Classification Reports

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report of BERT** | | | | |
|  | Precision | Recall | F1 Score | Support |
| Moderate | 0.8 | 0.62 | 0.7 | 2169 |
| Not depression | 0.02 | 0.1 | 0.03 | 228 |
| Severe | 0.08 | 0.02 | 0.03 | 848 |
|  | | | | |
| accuracy |  |  | 0.43 | 3245 |
| macro avg | 0.3 | 0.25 | 0.25 | 3245 |
| weighted avg | 0.55 | 0.43 | 0.48 | 3245 |
| **Classification Report of Roberta** | | | | |
|  | Precision | Recall | F1 Score | Support |
| Moderate | 0.79 | 0.6 | 0.68 | 2169 |
| Not depression | 0.01 | 0.09 | 0.02 | 228 |
| Severe | 0.08 | 0.02 | 0.03 | 848 |
|  | | | | |
| accuracy |  |  | 0.41 | 3245 |
| macro avg | 0.29 | 0.23 | 0.24 | 3245 |
| weighted avg | 0.55 | 0.41 | 0.46 | 3245 |
| **Classification Report of DE Roberta** | | | | |
|  | Precision | Recall | F1 Score | Support |
| Moderate | 0.79 | 0.63 | 0.7 | 2169 |
| Not depression | 0.02 | 0.1 | 0.03 | 228 |
| Severe | 0.09 | 0.02 | 0.04 | 848 |
|  | | | | |
| accuracy |  |  | 0.43 | 3245 |
| macro avg | 0.3 | 0.25 | 0.25 | 3245 |
| weighted avg | 0.55 | 0.43 | 0.48 | 324 |

As the classification report showed that the model showed the highest 0.43% accuracy score for the tuned BERT model. The comparison of the reproduced results with the reported results is presented in Table 2.

Table 2: Comparison of Reported and Reproduced Results.

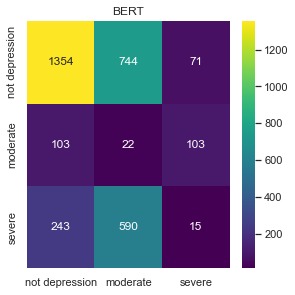
|  |  |  |
| --- | --- | --- |
| Metric | Reported | Reproduced |
| Accuracy | 0.585 | 0.43 |
| Macro F1-Score | 0.412 | 0.25 |
| Macro Recall | 0.403 | 0.25 |
| Macro Precision | 0.436 | 0.30 |
| Weighted F1-Score | 0.576 | 0.48 |
| Weighted Recall | 0.585 | 0.43 |
| Weighted Precision | 0.572 | 0.55 |

## Discussion

The code of the selected paper is available via the GitHub repository that was used here for the reproduction of the results. The BERT classifier showed a maximum 0.43% accuracy after hyperparameter tuning. But the model didn’t show the results that are reported in the published paper. Numerous factors can play a role in this situation. The reported results may be extracted from another set except for the test set and mistakenly mentioned in the paper. The model may be overfitted at the time of training and the test set also like the training set showed significant results. The seed’s value and the accuracy/loss graph of the model during training are not discussed which represents the overall behavior of the model and can assist in the reproduction of similar results.

## Error Analysis

We generated a confusion matrix of the BERT model's performance on the test set to conduct error analysis. The matrix displayed accurate sample counts and false negative/positive samples for each class. However, the authors did not provide further details regarding error analysis.



# Conclusion

By following the model parameter tuning and model training, we conclude that the results of the paper are reproducible. The results can be reproduced from the development of the code from scratch or by using the GitHub repository of the paper provided by the author. But a few configuration-like details are missing like the value of the seed which fails the researcher to produce similar results. Collectively, the methodology of the paper is implementable, and results are reproducible but similar results may require extensive attention or may require a hit and trail method with different configurations to extract similar results.

# References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829. https://doi.org/10.1007/S10462-021-09958-2/TABLES/18

Anantharaman, K., Rajalakshmi, S., Angel Deborah, S., Saritha, M., & Sakaya Milton, R. (2022). SSN\_MLRG1@LT-EDI-ACL2022: Multi-Class Classification using BERT models for Detecting Depression Signs from Social Media Text. *LTEDI 2022 - 2nd Workshop on Language Technology for Equality, Diversity and Inclusion, Proceedings of the Workshop*, 296–300. https://doi.org/10.18653/V1/2022.LTEDI-1.44

Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, *133*, 182–197. https://doi.org/10.1016/J.ESWA.2019.05.023

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://arxiv.org/abs/1810.04805v2

Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020). SenseMood: Depression detection on social media. *ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval*, *5*(20), 407–411. https://doi.org/10.1145/3372278.3391932

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T. S., & Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. *IJCAI International Joint Conference on Artificial Intelligence*, *0*, 3838–3844. https://doi.org/10.24963/IJCAI.2017/536

Van Aken, B., Löser, A., Winter, B., & Gers, F. A. (2019). How does BERT answer questions? A layer-wise analysis of transformer representations. *International Conference on Information and Knowledge Management, Proceedings*, 1823–1832. https://doi.org/10.1145/3357384.3358028

Xin, J., Tang, R., Lee, J., Yu, Y., & Lin, J. (2020). DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2246–2251. https://doi.org/10.18653/v1/2020.acl-main.204