ComTech: Computer, Mathematics and Engineering Applications, 15(1), June 2024, 65−71 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v15i1.11105 E-ISSN: 2476-907X

Psychological Stress Detection Using Transformer-Based Models

Derwin Suhartono[1]*, Irfan Fahmi Saputra[2], Andhika Rizki Pratama[3], and Gabriel Nathaniel[4]

[1-4]Computer Science Department, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia 11480
[1]dsuhartono@binus.edu; [2]irfan.saputra@binus.ac.id;
[3]andhika.pratama006@binus.ac.id; [4]gabriel.nathaniel@binus.ac.id

Received: 7th January 2024/ **Revised:** 13th June 2024/ **Accepted:** 13th June 2024

How to Cite: Suhartono, D., Saputra, I. F., Pratama, A. R., & Nathaniel, G. (2024). Psychological Stress Detection Using a Transformer-Based Model. *ComTech: Computer, Mathematics and Engineering Applications, 15*(1), 65−71. https://doi.org/10.21512/comtech.v15i1.11105

Abstract **-** Stress is a significant mental health problem that can result in a lack of concentration. It has been more widely identified through social media since people who are under stress usually post about their physical pain and tiredness. However, stress assessment through social media by professionals can be expensive and time-consuming. The research aimed to produce a stress detection system trained using a Twitter dataset to predict stress using the user's input sentence. The experiments that were done in the research used transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT (RoBERTa). The research involved data pre-processing, model training, and model evaluation to ensure high-quality train data. Since the data were imbalanced, data trimming was performed in pre-processing to select data randomly until the balance matched. This process ensured the model's effectiveness in the training and evaluation stages. The features used in these experiments were features from each pre-trained model. In evaluating the model, accuracy, loss, and F1 score were used as metrics. In the result, for BERT, accuracy reaches 0.848 with an F1 score of 0.847. Meanwhile, RoBERTa has an accuracy of 0.837 and 0.834. The results prove that BERT and RoBERTa can be used to classify stress with accuracy and an F1 score above 0.8. The experiment result shows that the BERT deep learning model can detect stress using the Twitter datasets.

Keywords**:** stress detection, transformer-based model, Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT (RoBERTa)

I. INTRODUCTION

Psychological stress is a serious condition for mental health and often leads to other severe conditions. A repeated state of stress is associated with mental health problems like self-defeating behavior (Hawk et al., 2019), concentration loss, fatigue, and loneliness (Aalbers et al., 2019). According to a study in the UK in 2018, 74% of the participants felt stressed, overwhelmed, and unable to cope with certain situations (Samele et al., 2018). Another research that was conducted in the US in 2020, Around 8 in 10 adults, or 78%, said that Coronavirus was one of the largest contributors to their mental health (stress) in their life compared to the previous year, which was mostly caused by health care, mass shootings, or climate change (American Psychological Association, 2020).

Social media provides vast free public data that can be used for studying and understanding various fields (Bhimani et al., 2019). Texts from social media often contain information about natural language and intentions, which can be learned by humans and computers. However, fitting social media into Natural Language Processing (NLP) requires adjustments to the text itself, such as stemming, lemmatization, and part of speech tagging (Hasan et al., 2019; Yogish et al., 2019). There are two approaches to sentiment analysis: dictionary-based and machine learning-based. Sentiment analysis can be divided into document-level, sentence-level, aspect-based, comparative, and sentiment lexicon acquisition (Birjali et al., 2021).

Linguistic studies have indicated that usage of social media may expose users to stress and can identify stress through their posts or tweets. In one of

*Corresponding Author

the studies, there are several indications that stressed people tend to post more about exhaustion, losing control, and physical pain, while those who are not stressed post about family time or travel (Guntuku et al., 2019). Social media also provides a benefit by raising one's awareness about their mental health (Hampton et al., 2014). Visiting professionals for mental health assessments is one way to go, even though it may be costly and time-consuming if the visitor has limited knowledge about their stress level. Throughout the years, a variety of computational approaches have been proposed to predict one's mental state by using social media (Ríssola et al., 2021).

Stress detection is a task that involves determining or assessing someone's stress level using technology and computers (Gedam & Paul, 2021). The technique includes analyzing social media posts from Reddit or Twitter (Rastogi et al., 2022). Reddit, a popular social media platform, offers full anonymity for users. It has been analyzed using the Bag of Words, BERT, and Embeddings from Language Models method for feature extraction (Inamdar et al., 2023). Moreover, another previous research has used various frameworks, including logistic regression, Naive Bayes, support vector machine (SVM), and LSTM, to compare the results (Oryngozha et al., 2024). It uses machine learning or deep learning approaches as classification models.

Based on the previous description, the research aims to study stress detection with the help of newer models (transformer-based models) than previous research, such as Winata et al. (2018). The research creates a performance comparison between Long Short-Term Memory (LSTM) as its benchmark since the model is used by the previous research mentioned. Then, Bidirectional Encoder Representations from Transformers (BERT) and a Robustly Optimized BERT (RoBERTa) are used to create a system that can detect stress from social media posts for self-diagnosis stress and early preventive action. The research is carried out using Twitter posts in English gathered from the previous research. Then, the datasets from previous research are modified accordingly. In addition, due to specific characteristics of Twitter datasets that are different from other social media platforms, the arrangement for conducting experiments to investigate similar objectives should be managed differently, too.

II. METHODS

There are two models that are the focus of the research: BERT and RoBERTa. BERT is a transformer-based machine learning technique for NLP task such as sentiment classification (Areshey & Mathkour, 2023). Meanwhile, RoBERTa is an improved version of BERT. RoBERTa replaces the next sentence prediction objective with full sentences without the next sentence prediction (Joshy & Sundar, 2022). It is trained on longer sequences (125K steps with 2K sequences) and uses a dynamic masking pattern, while BERT uses static masking. Moreover, RoBERTa has more training data (160 GB) than BERT (16 GB).

The setting of the experiment is the learning rate and batch size. The data that are utilized in the research are the datasets provided by Winata et al. (2018) about psychological stress detection from spoken language with attention-based LSTM and distant supervision. They created a list of hashtags for stress and relaxation and collected data from Twitter associated with the hashtags they defined before.

The data used in the research are the main dataset, whereas those originally utilized as pretraining data for their model are fine-tuned with another dataset. Because BERT and RoBERTa are already pre-trained with a huge amount of data, but the training is not specific to one task, the Twitter data from before are data that fine-tune the model. The data contains the whole sentence, and the label indicates whether it is associated with stress or not.

The research has several stages, including data pre-processing, model training, and model evaluation. The data pre-processing stage ensures that the train data is of good quality before it is used in the model training stage. It is also essential because pre-processing influences the model performance (Maslej-Krešňáková et al., 2020). The research applies data trimming, cleaning, tokenization, and splitting as part of data pre-processing. The dataset, containing 367,000 tweets labeled with stress or relaxation, is heavily imbalanced with a 1:5 ratio. However, oversampling may lead to overgeneralization and bias in accuracy scores (García et al., 2020). So, undersampling is done to maintain an equal ratio and avoid overgeneralization.

The initial step is data trimming. In this step, the data used are Twitter data that are already labeled since the data are not balanced. Hence, the data with the relax label are trimmed by choosing data randomly until the amount matches the amount of data with the stress label. Table 1 shows the distribution of the datasets.

Table 1 Dataset Distribution from Twitter

Stress	Relax
59,768	59,768

The next step in data pre-processing is data cleaning. In data cleaning, the data are cleaned with several methods, focusing on removing symbols, punctuation, and extra space to keep the originality of each sentence and prevent the data from being tokenized properly and having a bias when converted into an embedding vector. After the data are cleaned, it continues to tokenization. In the tokenization, the data are tokenized by adding a classification token at the beginning of the sentences and a separation token at the end. Then, the data are converted into a vector

embedding called input ID by the WordPiece tokenizer with their vocabulary. The data are also converted into an attention mask, a vector representation of each token in the input ID that indicates whether, in a particular position, there is a token representation of a word or not. Figures 1 and 2 illustrate the input and the attention mask.

The last step in the data pre-processing stage is data splitting. In this process, data are split into training and testing data using cross-validation. The cross-validation method allows each data point to take part in becoming training data and testing data. First, data are cloned ten times. Second, each dataset is split into training and testing data, noting that each clone dataset has different testing data from 9 other clone datasets. Splitting is done for the input ID and attention mask.

The next stage is model training. In the model training stage, the data from the pre-processing are used to train the deep learning model. The data that are used at this stage are training data. In this stage, the training process is done ten times to utilize the entire dataset with cross-validation. Each training process uses a new, fresh model, and the previous model that is already trained is stored in a particular way. The setting in the training process is a combination of learning rates (2e-5 and 4e-5) and batch sizes of 32 and 64. After training, this experiment conducts the testing process. The testing process is done after each iteration of the cross-validation. The data used by the model are created by creating a prediction of testing data and comparing it with ground truth or real label. The prediction results in the probability distribution of each data point or, in a simple sentence, the tendency of the sentence itself, whether it is stressed or not.

The final stage of the experiment is evaluation. Evaluation is done after the testing process is executed. The probability distribution generated in the testing process is processed into a label and compared with the ground truth or real label. With this prediction label, a confusion matrix can be generated to give a general idea of the research model when it comes to the prediction process. The confusion matrix aims to calculate True Positive, True Negative, False Negative, and False Positive values. True Positive is determined by counting how many data correctly categorize the data as stress, True Negative by counting how many data correctly categorize the data as relax, False Negative by counting how many data incorrectly categorize the data as stress, and False Positive by counting how many data incorrectly categorize the data as stress. The confusion matrix is then used for deeper evaluation with accuracy and F1 score. Furthermore, the confusion matrix value is used to calculate accuracy, and F1 score represents deeper insight into the research model.

III. RESULTS AND DISCUSSIONS

Evaluation using the Twitter dataset is done to achieve results according to methodology. All stress detection models created are going to be evaluated using accuracy, loss, and the F1 score metric. These experiments are done using cross-validation with a 10-fold. From Tables 2 to 5, metrics are shown based on multiple scenarios and through the proposed models using pre-trained models.


```
[0, 14746, 110, 920, 16, 11, 3267,
13, 5, 363, 8, 47, 120, 10, 3462,
200, 7, 2512, 137, 47, 95, 213, 7,
3267, 2, 1, 1, 1, 1, 1, 1, 1]
```

Figure 1 Input ID Sentence

```
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1 ,1, 1, 1,
1, 1, 1, 1, 1, 0, 0 ,0 ,0 ,0,
0, 0]
```

Figure 2 Attention Mask

Although data in Table 2 are proven to be better in terms of loss setup, the gap between the second-lowest loss is not quite far. For BERT, the difference between the lowest one and the second lowest lies at 0.007, which is really low. Meanwhile, the loss difference with RoBERTa lies at 0.047 between the two lowest loss values.

Table 2 Twitter Dataset's Performance Metrics Evaluation (Learning Rate of 4e-5 and Batch Size of 32)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.846	0.356	0.847
RoBERTa	0.795	0.437	0.804

Table 3 uses hyperparameters with a learning rate of 4e-5 and batch size of 64, with accuracy for BERT of 0.844 and RoBERTa of 0.841. Although in this setup, accuracy has been declared as the better one in terms of overall accuracy, it does not apply to BERT. BERT has shown the highest value compared to the other hyperparameter setups. However, the difference between the accuracy of BERT in Table 3 and the highest BERT accuracy is only on a thin margin, with the difference only at 0.004. On the other hand, RoBERTa from Table 3 proves to have the highest accuracy (0.841) in comparison to other hyperparameter setups. Table 3 shows more prominent results for accuracy than the other hyperparameter setups.

Moving to the F1 score, the model with a learning rate of 4e-5 and batch size of 64 from Table 3 gives a better overall score using BERT and RoBERTa than the others. The F1 score of BERT and Roberta has the same value of 0.840. In Table 3, RoBERTa gives out the highest score in comparison with other hyperparameter setups. The difference between the highest F1 score and the second highest is not very high in RoBERTa since the difference only lies at 0.006. Even though Table 3 gives a better F1 score compared to other tables of hyperparameter setup, in terms of BERT, it has the lowest value compared to the highest ones in Tables 2 and 4, with a 0.007 difference in score. In terms of loss, Table 4, with the setup of a learning rate of 2e-5 and batch size of 32, gives out the lowest loss compared to the other hyperparameter setup. In that setup, BERT loss has the lowest one, with 0.349. Meanwhile, Table 5 shows different perspective by mixing the setup of learning rate and batch size from Table 3 and Table 4.

The profile of this interview dataset is retrieved from the Natural Stress Emotion Corpus (Wang et al., 2012), which is an interview with 25 students (13 females) transcribed into text. Winata et al. (2018) contributed more to the dataset (3 females) with an identical setup. The total number of sentences inside the corpus is around 2,272, with 813 going to the stressed label and the remaining going to the relaxed label.

Further testing is done with the same hyperparameter setup and 10-fold cross-validation to evaluate the model better. Ten different datasets are used in Table 6. The difference between the interview and Twitter datasets is that the interview datasets are provided by previous researchers. So, this step needs to be done to check whether the model can predict new data. By using same dataset, Table 7 presents the result with different batch size setting which is 64.

Table 3 Twitter Dataset's Performance Metrics Evaluation (Learning Rate of 4e-5 and Batch Size of 64)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.844	0.435	0.840
RoBERTa	0.841	0.448	0.840

Table 4 Twitter Dataset's Performance Metrics Evaluation (Learning Rate of 2e-5 and Batch Size of 32)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.848	0.349	0.847
RoBERTa	0.837	0.370	0.834

Table 5 Twitter Dataset's Performance Metrics Evaluation (Learning Rate of 2e-5 and Batch Size of 64)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.847	0.368	0.845
RoBERTa	0.829	0.417	0.821

Table 6 Interview Dataset's Performance Metrics Evaluation (Learning Rate of 4e-5 and Batch Size of 32)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.690	0.902	0.760
RoBERTa	0.725	0.728	0.751

Table 7 Interview Dataset's Performance Metrics Evaluation (Learning Rate of 4e-5 and Batch Size of 64)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.71	1.0001	0.763
RoBERTa	0.78	0.682	0.819

For accuracy, Table 8 shows the highest accuracy value for BERT with 0.735 if the researchers compare it with other hyperparameter setups. Table 8 uses hyperparameters with a learning rate of 2e-5 and batch size of 32. The second-best performance in terms of loss is in Table 8, with a loss of 0.798 for the BERT model and 0.656 for the RoBERTa model using a hyperparameter setup with a learning rate of 2e-5 and batch size of 32. The loss difference is not that far from the best and the second-best performance: 0.046 for BERT and 0.072 for RoBERTa.

However, for RoBERTa, the best result is achieved in Table 9 using the hyperparameter with a learning rate of 2e-5 and batch size of 64 with an accuracy of 0.83. If the researchers compare the best accuracy with the second best, for BERT, the difference is really small between Tables 8 and 9, with only 0.005. Meanwhile, for RoBERTa, the difference is also small, but not as small as the BERT. Between Tables 8 and 9, the difference is 0.035.

Moving on to loss, Table 9 shows the lowest loss value compared with the loss value from other hyperparameter setups. Table 9 uses hyperparameters with a learning rate of 2e-5 and batch size of 64. With this hyperparameter, the BERT model gets a loss of 0.752, and the RoBERTa model obtains a loss of 0.584. If the researchers compare it with the others, this hyperparameter setup has the best performance for the loss value.

Now for the F1 score, the best two performances are again shown in Tables 8 and 9. The best F1 score for BERT is shown in Table 8 using a learning rate of 2e-5 and batch size of 32 with a score of 0.789. The best F1 score for RoBERTa is shown in Table 9 using a learning rate of 2e-5 and batch size of 64 with 0.853. Compared with the second-highest F1 score for BERT, the difference is only 0.007. Meanwhile, for RoBERTa, the difference is 0.028.

Even though the hyperparameter setup in Table 9 is better than Table 8, the accuracy and F1 score in the hyperparameter setup in Table 8 surpasses Table 9. As for RoBERTa, regarding loss, accuracy, and F1 score, the hyperparameter setup in Table 9 with a learning rate of 2e-5 and batch size of 64 can surpass the performance of other setups. So, this hyperparameter setup is the best for RoBERTa regarding the interview case.

Table 8 Interview Dataset's Performance Metrics Evaluation (Learning Rate of 2e-5 and Batch Size of 32)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.735	0.798	0.789
RoBERTa	0.795	0.656	0.825

Table 9 Interview Dataset's Performance Metrics Evaluation (Learning Rate of 2e-5 and Batch Size of 64)

System Baseline	Accuracy	Loss	F1 Score
BERT	0.73	0.752	0.782
RoBERTa	0.83	0.584	0.853

Before jumping to a conclusion about which hyperparameter setup to pick, a further side-by-side comparison between the Twitter and interview datasets needs to be done. The creation of the Twitter dataset for training and validation revolves around the same environment, which is without proper grammatical structure. In contrast, the interview dataset has a proper grammatical structure since it is a transcribed spoken language. It is given that the interview dataset will perform worse than the Twitter dataset prediction due to the circumstances mentioned before. With that being said, the recommended configuration for BERT and RoBERTa is a batch size of 32 and learning rate of 2e-5.

The reason behind the chosen parameter combination is to consider all the performance metrics: accuracy, loss, and F1 score. Another reason to add is that this combination is also picked due to the performance difference between Twitter and interview datasets. To put it short, performance metrics from the Twitter dataset should be better than the interview ones, whether it is accuracy, loss, or F1 score, and the reversal condition will be ruled out from consideration. All in all, among all the different scenarios for the model, a batch size of 32 and a learning rate of 2e-5 come out on top. Lastly, a comparison is made to check whether the model performs better than the previous studies or not.

Table 10 Model Performance Comparison

Experimentation	Loss	F1 Score
LSTM	0.55	0.54
BERT	0.848	0.847
RoBERTa	0.837	0.834

Table 10 compares LSTM, BERT, and RoBERTa. BERT comes out on top with the highest accuracy and F1 score, while RoBERTa is the runner-up. By determining these metrics, researchers can conclude that they can create a transformer-based model to identify stress from a Tweet or user input. It means that newer models like BERT and RoBERTa can perform better than precedent research.

With a learning rate of 2e-5 and batch size of 32, the system can achieve an accuracy of 0.848 and an F1 score of 0.847 by using BERT. Meanwhile, RoBERTa

has an accuracy of 0.837 and F1 score of 0.834. Newer models, such as BERT and RoBERTa, are able to reach higher accuracy and F1 score compared to previous studies. Both transformer-based models, like BERT and RoBERTa, are able to outperform accuracy and F1 score from the previous research by Winata et al. (2018). Both models (BERT and RoBERTa) perform greatly with accuracy and F1 score above 0.8. The deep learning of BERT outperforms the performance of RoBERTa on stress detection tasks. During experiments conducted using BERT and RoBERTa with the same dataset and hyperparameter setups, BERT can outperform RoBERTa in terms of accuracy and F1 score with a 0.011 difference for accuracy and a 0.013 difference for F1 score in this stress detection task.

IV. CONCLUSIONS

The research has run experiments on stress detection using deep learning. Stress detection using Twitter datasets is possible. Through several stages, such as data pre-processing, model training, and model evaluation, the research succeeds in showing the best performance of the BERT as a deep learning model to detect stress using the Twitter dataset. The deep learning stress detection approach can utilize social media and run automatically.

Although the experiment shows good results, the research still has limitations, including the dataset used with 119,536 tweets. Using a larger dataset will let the system obtain more insight into detecting stress through sentences. Hence, for further research, the experiment on stress detection in social media using the Twitter dataset can be improved in several aspects, such as utilizing a larger dataset for training and validation. Future research can also utilize a better pre-processing method for the dataset to create a cleaner dataset that will increase the chance that the system will recognize words. Moreover, future research can perform a better configuration for the models used in the research for better configurations to match the task and the dataset better and search for the best performance and result.

REFERENCES

Aalbers, G., McNally, R. J., Heeren, A., De Wit, S., & Fried, E. I. (2019). Social media and depression symptoms: A network perspective. *Journal of Experimental Psychology: General, 148*(8), 1454–1462. https://doi.org/10.1037/xge0000528

Areshey, A., & Mathkour, H. (2023). Transfer learning for sentiment classification using bidirectional encoder representations from transformers (BERT) model. *Sensors*, *23*(11), 1–18. https://doi.org/10.3390/s23115232

American Psychological Association. (2020, October). *Stress in America 2020: A national mental health crisis*. https://www.apa.org/news/press/releases/stress/2020/report-october

Bhimani, H., Mention, A. L., & Barlatier, P. J. (2019). Social media and innovation: A systematic literature review and future research directions. *Technological Forecasting and Social Change*, *144*, 251–269. https://doi.org/10.1016/j.techfore.2018.10.007

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*. https://doi.org/10.1016/j.knosys.2021.107134

García, V., Sánchez, J. S., Marqués, A. I., Florencia, R., & Rivera, G. (2020). Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. *Expert Systems with Applications*, *158*. https://doi.org/10.1016/j.eswa.2019.113026

Gedam, S., & Paul, S. (2021). A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access, 9*, 84045–84066. https://doi.org/10.1109/ACCESS.2021.3085502

Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019). Understanding and measuring psychological stress using social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 214–225). https://doi.org/10.1609/icwsm.v13i01.3223

Hampton, K. N., Rainie, L., Lu, W., Dwyer, M., Shin, I., & Purcell, K. (2014, August 26*). Social media and the 'spiral of silence'*. PewResearchCenter. https://www.pewresearch.org/internet/2014/08/26/social-media-and-the-spiral-of-silence/

Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019). Sentiment analysis with NLP on Twitter data. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)* (pp. 1–4). IEEE. https://doi.org/https://doi.org/10.1109/IC4ME247184.2019.9036670

Hawk, S. T., Van Den Eijnden, R. J., Van Lissa, C. J., & Ter Bogt, T. F. (2019). Narcissistic adolescents' attention-seeking following social rejection: Links with social media disclosure, problematic social media use, and smartphone stress. *Computers in Human Behavior*, *92*, 65–75. https://doi.org/10.1016/j.chb.2018.10.032

Inamdar, S., Chapekar, R., Gite, S., & Pradhan, B. (2023). Machine learning driven mental stress detection on Reddit posts using natural language processing. *Human-Centric Intelligent Systems*, *3*(2), 80–91. https://doi.org/10.1007/s44230-023-00020-8

Joshy, A., & Sundar, S. (2022). Analyzing the performance of sentiment analysis using BERT, DistilBERT, and RoBERTa. In *2022 IEEE International Power and Renewable Energy Conference (IPRECON)*

(pp. 1–6). IEEE. https://doi.org/10.1109/IPRECON55716.2022.10059542

Maslej-Krešňáková, V., Sarnovský, M., Butka, P., & Machová, K. (2020). Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. *Applied Sciences*, *10*(23), 1–26. https://doi.org/10.3390/app10238631

Oryngozha, N., Shamoi, P., & Igali, A. (2024). Detection and analysis of stress-related posts in Reddit's acamedic communities. *IEEE Access, 12,* 14932–14948. https://doi.org/10.1109/ACCESS.2024.3357662

Rastogi, A., Liu, Q., & Cambria, E. (2022). Stress detection from social media articles: New dataset benchmark and analytical study. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN55064.2022.9892889

Ríssola, E. A., Losada, D. E., & Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, *2*(2), 1–31. https://doi.org/10.1145/3437259

Samele, C., Lees-Manning, H., Zamperoni, V., Goldie, I., Thorpe, L., Wooster, E., ... & Rowland, M. (2018). *Stress: Are we coping*. Mental Health Foundation.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 587–592). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.119

Winata, G. I., Kampman, O. P., & Fung, P. (2018). Attention-based LSTM for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6204–6208). IEEE. https://doi.org/10.1109/ICASSP.2018.8461990

Yogish, D., Manjunath, T. N., & Hegadi, R. S. (2019). Review on natural language processing trends and techniques using NLTK. In *Recent Trends in Image Processing and Pattern Recognition: Second International Conference, RTIP2R 2018* (pp. 589-606). Springer Singapore. https://doi.org/10.1007/978-981-13-9187-3_53