Effect of Text Preprocessing Methods on the Performance of Social Media Posts Classification

Yasemin ATAYOLU[1*], Yakup KUTLU[2]
[1,2] Department of Computer Engineering, Iskenderun Technical University, Turkey
ORCID: 0009-0005-5139-0381, 0000-0002-9853-2878
E-mails: yaseminatayolu@gmail.com, yakup.kutlu@iste.edu.tr
*Corresponding author.

***Abstract*— This study investigates the classification performance of RoBERTa, BERT, XLNet, and T5 for mental health-related tasks and evaluates the impact of various text preprocessing steps. Initially, the classification results of these models were obtained by applying a minimal preprocessing step, specifically the removal of repeated expressions. Subsequent analysis assessed the effects of additional preprocessing techniques, including the removal of stop words, technical stop words, URLs, and punctuation, as well as text normalization steps such as lemmatization and conversion to lowercase. RoBERTa achieved the highest classification accuracy and F1 score, particularly excelling in the detection of depression and suicide tendencies. All preprocessing steps, apart from removing repeated expressions, reduced overall classification accuracy and performance. For the depression class, converting text to lowercase also had a positive effect, showing an inverse relationship between preprocessing intensity and performance in most cases. The results underscore the need to tailor preprocessing steps carefully to the task and dataset.**

***Keywords*— Text preprocessing, mental health classification, large language models,**

I. INTRODUCTION

Depression is the emotional expression of feeling helpless and powerless to meet high personal expectations [1]. Choudhury and co-authors demonstrated that the behavior of individuals with severe depression on social media may differ from those without these symptoms [2]. Suicide is the act of ending one's own life. Edwin S. Shneidman, the founder of contemporary suicidology, initiated efforts to prevent suicide [3]. Language, through speech patterns and word choice, can serve as a predictor for suicidal thoughts and behaviors, reflecting emotional states that may help identify individuals at risk. [4]. Anxiety is a future-oriented mood linked to anticipating negative events [5]. Wang and Bashir revealed differences in social media behavior between individuals with and without anxiety symptoms. Since social media reflects individuals' emotional states and daily lives, it could serve as a novel resource for mental health assessments [6].Social media data is available in real time, facilitating mental health monitoring and risk prediction [7].

Thought is consciously manifested through linguistic expressions that guide it. More-over, although a word may seem to carry a fixed meaning, its meaning changes in different contexts, requiring careful analysis to understand these differences [8]. The order of words in sentences and the order of sentences in paragraphs can provide important insights into cognitive functioning. Advanced syntactic analysis methods in natural language processing can also be used to evaluate linguistic complexity and cognitive health [9].

The current study systematically evaluates the impact of various preprocessing techniques on classification performance. Using datasets from Twitter, Reddit, and Instagram, this study compares four transformer-based models—RoBERTa, BERT, XLNet, and T5—and investigates their ability to detect depression, anxiety, and suicidal tendencies under different preprocessing conditions.

II. RELATED WORKS

A. Text Preprocessing in Social Media

General text preprocessing methods are suitable for most datasets, but content-focused tech-niques are required to preserve more semantic information. Advanced text preprocessing techniques are evaluated in addition to these methods [10]. For example, in a study examining the impact of stop words on classification, removing stop words improved classification performance, and further removal of technical stop words enhanced the classification success [11]. In another study, experimental work measuring the impact of different preprocessing methods on three datasets showed that using words in their lemma form and in lowercase yielded high performance in most cases, while removing punctuation marks, URLs, user mentions (@mentions), and hashtags (#) often resulted in lower performance [12]. The literature consistently highlights that implementing text preprocessing steps, including the removal of stop words, significantly enhances classification performance and overall task efficacy [13-18].

B. Large Language Models in Mental Health Classification

On the Reddit platform, a study classifies five common mental health disorders -depression, anxiety, bipolar disorder, ADHD, and PTSD-, and normal mental health using unstructured user data. This research achieves the best performance with a pre-trained RoBERTa transfer learning model, reporting an accuracy and F1 score of 0.83 [19]. Novikova and Shkaruta

introduced a set of behavioral tests (DECK) to enhance BERT-based depression detection. Their method improved F1 scores by up to 53.93% but faced challenges with markers like suicidal ideation [20].
Further research introduces MM-EMOG, a multi-label emotion graph representation tailored for classifying mental health states from social media posts, which achieves up to 78% accu-racy using the BERT model [21]. Additionally, ClinicalBERT [22], customized for health-related texts, and MentalBERT [23] tailored for mental health texts, are task-specific models fine-tuned from BERT.

III. MAETRIALS & METHODS

A. *Dataset*

Social media data, leveraging the large amount of user-generated content, constitutes a valuable resource for providing insights into emotional states and psychological conditions. A review of the literature highlights numerous studies utilizing such resources. These studies classify conditions including depression [24], anxiety [25], and suicidal tendencies [26]. This study adopts a broader and more comprehensive approach by merging similar data to create a new dataset and analyzing its performance using artificial intelligence models. In this con-text, diversity in social media platforms is maintained, with data obtained from at least two different platforms for each class. A classification study is conducted based on four classes: three mental health disorders and a normal condition.

Data from various datasets sourced from social media platforms such as Twitter (X), Insta-gram, and Reddit are analyzed to examine suicidal tendencies, depression, anxiety, and nor-mal conditions.

Instagram Anxiety Dataset: This dataset, sourced from the study titled "Mpox Narrative on Instagram: A Labeled Multilingual Dataset of Instagram Posts on Mpox for Sentiment, Hate Speech, and Anxiety Analysis" [25], consists of Instagram posts. Posts written in different languages are filtered to include only English ones. Posts labeled as “No Stress/Anxiety De-tected” or “Stress/Anxiety Detected” are categorized into respective anxiety and normal classes.

Twitter Suicidal Tendency Datasets: These datasets include data obtained from the Kaggle data science platform and GitHub version control repositories. The datasets consist of posts indicating suicidal tendencies or not. The first dataset [27] uses posts labeled as "Potential Suicide post" for the suicidal tendency class and "Not Suicide post" for the normal class. The second dataset [28] categorizes posts with the label "intention" assigned as “1” into the sui-cidal tendency class and posts labeled as “0” into the normal class. The third dataset [29] contains raw data sourced only from Twitter and tagged for suicidal tendencies. This dataset is used exclusively for the suicidal tendency label.

Twitter Depression Dataset: This combined dataset is obtained from the Harvard Dataverse platform [30] Posts are categorized into depression and normal classes based on labels of Depressed/0 and Non-depressed/1.

Reddit Mental Health Datasets: These datasets, containing mental health posts from different periods, are sourced from the study "Natural Language Processing Reveals Vulnerable Men-tal Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19" [31]. They include posts under topics such as suicide, anxiety, and depression. The datasets are grouped based on different periods (pre-COVID-19: 2018 and 2019, and mid-COVID-19). Specific samples from each period are selected for classes other than normal (anxiety, suicidal tendencies, depression).

To achieve a more balanced distribution and better generalization, samples from datasets with a high number of entries are limited. After truncating the text length (2-510 words), data distribution is adjusted for equal representation. Approximately 7,000 entries each remain for Instagram anxiety and Twitter suicidal tendency datasets, and 7,000 entries are selected from the Twitter depression dataset. For Reddit anxiety data, additional entries are selected to reach 14,000 posts for the three non-normal classes. No normal-labeled data is included from Reddit. Since the normal-labeled entries from Twitter’s suicidal tendency dataset number around 6,000, an additional 6,000 entries each are selected from Instagram anxiety and Twit-ter depression datasets to create a total of 18,000 normal-labeled posts (random state 42).

A total of 18,000 normal-labeled posts, 14,000 anxiety-labeled posts from two datasets, 14,000 depression-labeled posts, and 14,000 suicidal tendency-labeled posts from four da-tasets are used, amounting to a total of 60,000 posts. Data are split into 70% (42,000) training, 10% (6,000) validation, and 20% (12,000) test sets based on equal distribution across classes and sources.

B. *Text Preprocessing*

In natural language processing (NLP), preprocessing transforms linguistic data -in this study, texts- to a format suitable for analysis and modeling. This stage involves cleaning and structuring data to enhance the performance of machine learning and deep learning models. Preprocessing aims to reduce data noise by removing unnecessary characters, non-meaningful structures, and elements that hinder the generalization ability of language models. These steps facilitate the model's ability to learn true semantic relationships within the text.

Stop-word removal involves eliminating words that do not hold meaning or functionality for modeling purposes. The list of stop-words can be customized based on the specific application. Lemmatization ensures that words are represented in their root forms while maintaining their meanings, consolidating various inflections and forms into a single, meaningful structure. Stemming, on the other hand, removes suffixes from words but may result in roots that differ from their base forms, unlike lemmatization.

The removal of URLs cleans the dataset by eliminating links, allowing the model to focus more effectively on relevant content. Similarly, HTML tags found in web content are removed to discard non-meaningful structures, producing cleaner data. Line breaks and tab characters are also eliminated to create a more structured and analyzable text format. Markdown links are removed to eliminate distractions, resulting in a more comprehensible dataset.

Usernames and email addresses are excluded, especially in social media datasets, to improve focus on the actual content. Punctuation marks are removed to help the model concentrate on meaningful word groupings. Only characters specific to the target language are retained to enhance data consistency and limit unnecessary character variability during analysis.

Lowercasing all text addresses case-sensitive differences that could otherwise impact the analysis process. Excessive or unnecessary whitespace is removed to ensure cleaner and more structured data. Repeated expressions that do not add meaning, except for intentional repetitions such as reduplications, are eliminated to reduce processing overhead and optimize computational efficiency.

C. Models

In 2017, Vaswani and co-authors introduced the Transformer network architecture in the paper "Attention is All You Need," enabling parallelization by eliminating sequential computations and considering the full context between the input and output. The architecture consists of two main components: encoder and decoder. The attention mechanism, with scaled dot-product and multi-head components, focuses on information from different positions in parallel, helping the model learn stronger and more detailed contextual relationships. Each layer includes a feed-forward network applied separately to each position, embeddings to convert tokens into vectors, and positional encodings to understand the sequence order [32].

The term "large" in Large Language Models emphasizes the scale of the model, which contains millions or billions of parameters, enabling it to capture complex language patterns and nuances. Language models are trained to predict the probability of a language sequence, making them effective for tasks like translation, summarization, and question answering. Large language models are pre-trained models based on the transformer architecture, which allows them to effectively handle long language contexts.

BERT (Bidirectional Encoder Representations from Transformers) uses a bidirectional attention mechanism to process both directions of a language simultaneously, generating context-sensitive representations. By using a masked language model (MLM) and next sentence prediction (NSP), it better understands the context of the language [33]. RoBERTa (Robustly Optimized BERT Pretraining Approach) is a language model developed by improving BERT's training design. RoBERTa enhances BERT's performance through longer training periods, larger batches, more data, removal of the NSP objective, and dynamic masking [34]. XLNet adopts the Auto-Regressive (AR) approach, combining the advantages of Auto-Encoding (AE) methods in language modeling tasks. Built on the Transformer-XL architecture (which learns dependencies in long texts), XLNet uses a permutation-based language modeling method, shuffling the word order rather than masking tokens [35]. Raffel and co-authors proposed T5 (Text-to-Text Transfer Transformer), a unified framework that treats all text-based language tasks as a text-to-text conversion problem, enabling various language processing tasks to be performed through transfer learning [36].

IV. RESULTS AND DISCUSSION

In this study, RoBERTa, BERT, XLNet, and T5 models are used for mental disorders classification. The models are fine-tuned. The training data is prepared in tensor format. The maximum input length is set to 256, with a dropout rate of 0.1, a learning rate of 3e-5, a linear scheduler, AdamW as the optimization algorithm, and Softmax as the activation function. Training is conducted over 3 epochs using GPUs, with a batch size of 32 for RoBERTa, BERT, and XLNet, and 16 for T5. Optuna is used to optimize the F1 score.

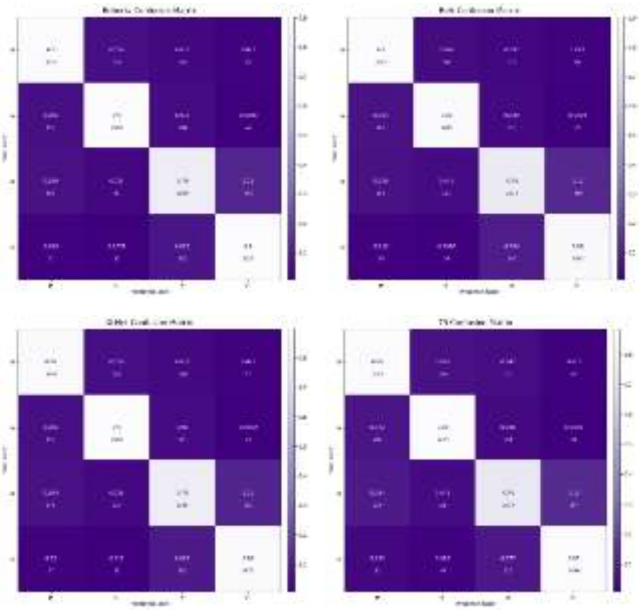

Figure 1. Classification performance of Large Language Models

The classification results of large language models are obtained by applying the preprocessing step of removing only repeated expressions as shown in Table 2. The highest performance and F1 score are achieved with the fine-tuning of RoBERTa. Additionally, for the Normal class, RoBERTa and BERT, for the Anxiety class, RoBERTa and XLNet, and for the Depression and Suicide tendency classes, RoBERTa provide the highest classification accuracy.

Table 2. Classification scores of Large Language Models

Score	**RoBERTa**	**BERT**	**XLNet**	**T5**
Acc	0.872*	0.864	0.863	0.846

F1	0.872*	0.864	0.862	0.845

In Figure 1., class-based accuracy and the number of correct predictions for each class are presented with confusion matrices for each model.

In this study, the effects of text preprocessing techniques on classification are examined step by step. In this study, it is concluded that the case of letters, different characters, links, spaces, stop words, and affixes applied to words are meaningful when classifying social media posts related to mental health.

The classification results of large language models indicate that the removal of repeated phrases, as applied in Figure 1., is the preprocessing step yielding the highest classification accuracy.

In Figure 2., the impact of text preprocessing steps and class-wise fine-tuning of the Roberta model is presented in the form of complexity matrices. A notable observation is that converting text to lowercase has a relatively negative impact on three classes, while it positively affects the depression class, increasing its accuracy to 0.80.

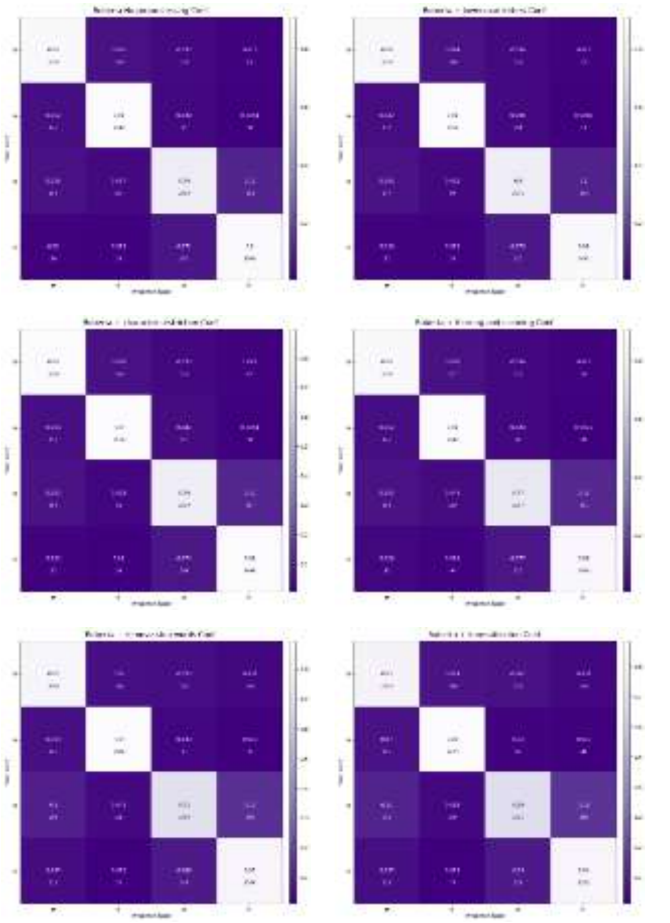

Figure 2. Classification performance of text preprocessing techniques

Removing repeated expressions renders it unnecessary to repeatedly present structures already learned by the model, potentially causing a slight decrease in classification accuracy.

In Table 3., it is observed that each text preprocessing step slightly reduces performance.

Table 3. Classification Scores of text preprocessing techniques

Preprocessing	**Acc**	**F1**
Remove repeated expressions	0.872	0.872
No preprocessing	0.872	0.871
+ Lowercase letters	0.871	0.871
+ Character restriction	0.869	0.868
+ Filtering and cleaning	0.862	0.861
+ Remove stop words	0.828	0.827
+ Lemmatization	0.810	0.809

V. CONCLUSION

This study evaluates the performance of RoBERTa, BERT, XLNet, and T5 for mental health classification tasks, with a focus on the impact of text preprocessing steps. The removal of repeated expressions improves classification accuracy and F1 scores across all classes. Converting text to lowercase enhances performance specifically for the depression class. However, additional preprocessing steps, including lemmatization, stop word removal, and character restriction, reduce overall accuracy and F1 scores.

Excessive preprocessing eliminates semantically significant features, which negatively affects classification outcomes. Steps like removing stop words and punctuation may discard critical linguistic information that is essential for detecting mental health states. These results demonstrate the necessity of preprocessing pipelines that preserve task-relevant information while avoiding unnecessary alterations. Examining the class-based results in the confusion matrices reveals that posts labeled as depression are predominantly misclassified as suicidal ideation posts. This outcome is plausible, as depression is closely associated with suicide [37].

Task-specific preprocessing strategies are essential to align the preparation of text data with classification goals. Preprocessing techniques need to account for the diverse requirements of different classes, as shown by the varying effects of lowercase conversion and other steps. Models benefit from carefully balanced preprocessing steps that minimize the loss of meaningful information.

REFERENCES

[1] Bibring, E. (1953). The mechanism of depression. In P. Greenacre (Ed.), Affective disorders; psychoanalytic contributions to their study (pp. 13–48). International Universities Press.

[2] De Choudhury, M., Gamon, M., Counts, S., &

Horvitz, E. (2013). Predicting depression via social media. In Proceedings of the international AAAI conference on web and social media (Vol. 7, No. 1, pp. 128-137).

[3] Leenaars, A. A. (2010). Edwin S. Shneidman on suicide. Suicidology online, 1(1), 5-18.

[4] Homan, S., Gabi, M., Klee, N., Bachmann, S., Moser, A. M., Michel, S., ... & Kleim, B. (2022). Linguistic features of suicidal thoughts and behaviors: A systematic review. Clinical psychology review, 95, 102161.

[5] Craske, M. G., Rauch, S. L., Ursano, R., Prenoveau, J., Pine, D. S., & Zinbarg, R. E. (2011). What is an anxiety disorder?. Focus, 9(3), 369-388.

[6] Wang, T., & Bashir, M. (2020). Does social media behaviors reflect users' anxiety. A case study of twitter activities.

[7] Gruebner, O., Sykora, M., Lowe, S. R., Shankardass, K., Galea, S., & Subramanian, S. V. (2017). Big data opportunities for social behavioral and mental health research.

[8] Jackendoff, R. (1996). How language helps us think. Pragmatics & Cognition, 4(1), 1-34.

[9] Voleti, R., Liss, J. M., & Berisha, V. (2019). A review of automated speech and language features for assessment of cognitive and thought disorders. IEEE journal of selected topics in signal processing, 14(2), 282-298.

[10] Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.

[11] Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. Plos one, 16(8), e0254937.

[12] Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. Multimedia Tools and Applications, 80, 35239-35266.

[13] Samad, M. D., Khounviengxay, N. D., & Witherow, M. A. (2020). Effect of text processing steps on twitter sentiment classification using word embedding. arXiv preprint arXiv:2007.13027.

[14] Ladani, D. J., & Desai, N. P. (2020, March). Stopword identification and removal techniques on tc and ir applications: A survey. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 466-472). IEEE.

[15] Jefriyanto, J., Ainun, N., & Al Ardha, M. A. (2023). Application of Naïve Bayes Classification to Analyze Performance Using Stopwords. Journal of Information System, Technology and Engineering, 1(2), 49-53.

[16] Rahimi, Z., & Homayounpour, M. M. (2023). The impact of preprocessing on word embedding quality: A comparative study. Language Resources and Evaluation, 57(1), 257-291.

[17] HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. PloS one, 15(5), e0232525.

[18] Jung, G., Shin, J., & Lee, S. (2023). Impact of preprocessing and word embedding on extreme multi-label patent classification tasks. Applied Intelligence, 53(4), 4047-4062.

[19] Ameer, I., Arif, M., Sidorov, G., Gòmez-Adorno, H., & Gelbukh, A. (2022). Mental illness classification on social media texts using deep learning and transfer learning. arXiv preprint arXiv:2207.01012.

[20] Novikova, J., & Shkaruta, K. (2022). DECK: Behavioral tests to improve interpretability and generalizability of BERT models detecting depression from text. arXiv preprint arXiv:2209.05286.

[21] Cabral, R. C., Han, S. C., Poon, J., & Nenadic, G. (2024). MM-EMOG: Multi-Label Emotion Graph Representation for Mental Health Classification on Social Media. Robotics, 13(3), 53.

[22] Wang, G., Liu, X., Ying, Z., Yang, G., Chen, Z., Liu, Z., ... & Chen, Y. (2023). Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. Nature Medicine, 29(10), 2633-2642.

[23] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). Mentalbert: Publicly available pre-trained language models for mental healthcare. arXiv preprint arXiv:2110.15621

[24] Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., & Hasan, K. (2023). DEPTWEET: A typology for social media texts to detect depression severities. Computers in Human Behavior, 139, 107503.

[25] Thakur, N. (2024). Mpox Narrative on Instagram: A Labeled Multilingual Dataset of Instagram Posts on Mpox for Sentiment, Hate Speech, and Anxiety Analysis. arXiv preprint arXiv:2409.05292.

[26] Ramírez-Cifuentes, D., Freire, A., Baeza-Yates, R., Puntí, J., Medina-Bravo, P., Velazquez, D. A., ... & Gonzàlez, J. (2020). Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. Journal of medical internet research, 22(7), e17758.

[27] Mahmud, S. A. (n.d.). Suicidal Tweet Detection Dataset. Retrieved October 19, 2024, from https://www.kaggle.com/datasets/aunanya875/suicidaltweet-detection-dataset

[28] Internet. (n.d.). Twitter Suicidal Data. Retrieved October 19, 2024, from https://github.com/laxmimerit/twitter-suicidal-intention-dataset/blob/master/twitter-suicidal_data.csv

[29] Yadav, A. (n.d.). Twitter Suicide Data. Retrieved October 19, 2024, from https://github.com/warriorwizard/suicidal-ideation-detection/tree/main/Dataset

[30] Helmy, A., 2024, "Depression dataset for English tweets classified binary", https://doi.org/10.7910/DVN/ISZCSA, Harvard Dataverse, V1

[31] Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., & Talkar, T. (2020). Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. Journal of medical Internet research, 22(10), e22635.

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... ve Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[33] Devlin, J., Chang, M. W., Lee, K. ve Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[34] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 364.

[35] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. ve Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

[36] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... ve Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[37] Shneidman, E. S. (1993). Commentary: Suicide as psychache. The Journal of nervous and mental disease, 181(3), 145-147.