CS678-Detecting Signs of Depression from Social Media Text using Transformer-Based Language Models: A Solution to the LT-EDI-ACL2022 Shared Task

Aditya Milind Limbekar.
G01384408
Department of Computer Science
Masters in Computer Science
(alimbekar@gmu.edu)

Stephen Simon Dias
G01387625
Department of Computer Science
Masters in Computer Science
(sdias3@gmu.edu)

For the code and results this is the **GitHub repository** we created

1 Introduction

The mental health condition known as depression is marked by protracted feelings of sadness, hopelessness, and loss of interest in once-enjoyed activities. It may have an impact on someone's ability to function and everyday life. Common symptoms of depression include difficulties sleeping, changes in appetite, a lack of energy, impaired focus, and thoughts of self-harm or suicide. Depression can be treated with therapy, medicine, or a combination of the two. If you or someone you love is exhibiting signs of depression, it's crucial to get treatment from a mental health specialist. Regular exercise, a good diet, stress-reduction measures, and social support are other self-help practices that may aid in the management of depression. It's critical to keep in mind that sadness can be treated and that asking for assistance is a show of strength. People with depression may recover and have happy lives with the right care and encouragement.

The level of depression was classified by the authors as "not depressed," "moderately depressed," or "severely depressed" using RoBERTa pre-trained language models that were then fine-tuned using a sample of English social media postings. One of the first publications on this subject mentioned by the authors is De Choudhury et al. (2013). In their research, a sample of Twitter users who had been diagnosed with depression were gathered, and a statistical classifier was built using the tweets they had made over the course of a year to evaluate the probability of depression. The authors also present similar research that analyzes social media postings to find depression using machine learning and NLP methods. A dataset of English-language social media postings was utilized by the authors, and it was gathered from the popular online forum Reddit. Users who self-reported their depression state were included in the original dataset, but the authors changed it to only include postings that were classified as "not depressed," "moderately depressed," or "severely depressed" by mental health specialists. By oversampling the minority classes, they were able to balance the dataset as well.
The authors divided the process of developing their approach into three phases. There are three steps:

1. **Dataset preprocessing**: The authors stemmed, lemmatized, and removed stop words from the dataset. To expand the dataset, they also utilized a method known as data augmentation.

2. **Fine-tuning pre-trained language models**: The authors used their updated dataset to fine-tune multiple pre-trained language models, including BERT, RoBERTa, and XLNet. They identified that RoBERTa provided the best outcomes.

3. **Ensemble averaging**: To enhance the performance of numerous models, the authors combined their predictions using an ensemble averaging approach. They used this method to obtain a macro-averaged F1-score of 0.583.

A number of criteria, including accuracy, precision, recall, and F1-score, were used to evaluate the authors' approach. They also developed confusion matrices to examine the mistakes their models had made. With a macro-averaged F1-score of 0.583, the authors discovered that their ensemble model, which averaged the predictions

of the RoBERTa large and DepRoBERTa models, produced the best results. Additionally, they discovered that each model excelled in a certain class and produced the greatest outcomes for that particular class. the findings of the error analysis, the models had trouble differentiating between the "not depressed" and "moderately depressed" groups. The authors speculated that this may be because some users might not correctly record their depression state and because depression diagnoses are prone to subjectivity.

The authors provide an overview of their research on identifying depressive symptoms in social media texts and emphasize the significance of this job for mental health research. They also talk about the drawbacks of their strategy, namely the reliance on self-reported depressive state and the arbitrary nature of depression diagnosis.
Future research, according to the authors, might concentrate on enhancing the dataset through the inclusion of more varied sources and creating models that can manage noisy and unstructured data. They also indicate that their method may be applied to diseases of the mind other than sadness. Overall, the authors draw the conclusion that while their method, which relied on ensemble averaging and transformer-based language models, produced competitive results in their work, there is still potential for development in this field of study.

2 Approach

The transformer architecture serves as the foundation for the cutting-edge RoBERTa model of natural language processing (NLP). The BERT (Bidirectional Encoder Representations from Transformers) model, of which it is a variation, was created by Facebook AI. It is beneficial to have prior knowledge of the following topics to comprehend the concepts associated with the RoBERTa model:

1. **Transformers**: The processing of sequential data, such as text, is made especially efficient by the use of transformers in neural network design. In order to determine the relative relevance of the various input sequence components, they employ self-attention processes. This enables them to recognize long-range relationships and contextual information.

2. **Pre-training**: In order to teach a language model general linguistic traits, pre-training entails training it on a huge corpus of text data. In comparison to training the model from the start, the model may then be fine-tuned on a particular job, such as sentiment analysis or question answering, using a lot less data.

3. **Masked Language Modeling**: Masked Language Modeling (MLM) is a pre-training task used in BERT and RoBERTa models. A certain number of tokens in a phrase are randomly masked, and the model is taught to predict the missing words. In order to function properly, the model must understand how various words relate to one another in various situations.

4. **SentencePiece**: The RoBERTa model's preprocessing stage uses the SentencePiece subword tokenization library. In order to better handle terms that are not part of the model's lexicon, it divides words into smaller chunks known as subwords.

5. **Byte Pair Encoding (BPE)**: Text data is encoded into a fixed-length form using the compression process known as byte pair encoding (BPE). Words are divided into subwords using this technique in SentencePiece.

6. **Multi-task Learning**: Multi-task Learning (MTL) is a training method that teaches a single model to carry out several tasks at once. In the case of the RoBERTa model, various pre-training tasks, including MLM and Next Sentence Prediction (NSP), are used to teach the model to recognize broad language characteristics.

2.1 How Robustness

Robustness is a crucial aspect of natural language processing (NLP) models, as it refers to their ability to perform well even in situations where they encounter unexpected or noisy data. To put it another way, a resilient NLP model is one that can consistently produce accurate results despite changes in the input data while handling variances in the input data. This can be crucial in real-world applications because the input data may originate from multiple sources and contain a variety of noises or errors. The reliability of NLP models can be assessed in a number of methods, including:

1. Adversarial testing: Adversarial testing entails the creation of data intended to "trickle" the model into providing false findings. By doing so, the model's weaknesses can be found and its robustness increased.

2. Out-of-domain testing: Tests conducted outside the domain in which the model was trained are referred to as "out-of-domain" tests. In the case of a model that was trained on news articles, social media posts can be the subject of an out-of-domain test. This can be used to determine how effectively the model generalizes to other kinds of data.

3. Noisy testing: In noisy testing, noise is added to the input data to replicate real-world scenarios in which the input may be noisy or imprecise.

Several strategies can be applied to increase the NLP models' robustness:

1. Data augmentation: Data augmentation is the process of creating new training data from existing data by using various transformations. One typical method, for instance, is to replicate real-world changes by adding noise to the incoming data.

2. Regularization: Regularization entails including a penalty term in the training loss function to prevent the model from becoming overfit to the training data. This can increase the resilience of the model and its capacity to generalize to new data.

3. Transfer learning: In transfer learning, a model is first learned for one task, and then a new task is subsequently trained using the previously trained model as a starting point. The robustness and generalizability of the model can both be enhanced by doing this.

4. Ensemble methods: To increase performance and robustness, ensemble techniques combine various models. One popular method, for instance, is to train several models using various topologies or hyperparameters, then integrate the results to generate predictions.

Overall, strengthening the stability of NLP models is a continuous problem that calls for rigorous assessment and testing. It is feasible to increase the resilience of NLP models and make them more useful in practical applications by applying methods including data augmentation, regularization, transfer learning, and ensemble approaches, as well as adversarial testing, out-of-domain testing, and noisy testing.

The experiment was carried out using a transformer-based language model technique developed by the authors. They adjusted three models—BERT, RoBERTa, and XLNet—and discovered that RoBERTa large produced the best outcomes. Then, using the provided corpus, they trained DepRoBERTa (RoBERTa for Depression Detection), their own language model. The outcomes of this model were enhanced through fine-tuning. The third option was ensemble averaging, which utilized averaging to integrate the predictions of the DepRoBERTa and RoBERTa large models. The dataset was also duplicate-free and the classes were balanced as part of the authors' preprocessing. To divide the dataset into training, validation, and test sets, they employed stratified sampling.

To identify depressive symptoms in social media writing, the scientists employed a transformer-based language model technique. They adjusted the parameters of the three models they had chosen—BERT, RoBERTa, and XLNet—and discovered that RoBERTa large produced the best outcomes. After that, they trained DepRoBERTa (RoBERTa for Depression Detection), their own language model, using the supplied corpus. The outcomes were enhanced by fine-tuning this model. Utilizing ensemble averaging, which combines the forecasts of the RoBERTa large and DepRoBERTa models using averaging, was the final option. The rationale for this strategy is that transformer-based language models have achieved tremendous success in a range of natural language processing applications, such as sentiment analysis and text categorization.

The authors thought that these models would potentially be useful for identifying depressive symptoms in social media writing. The idea behind fine-tuning a pre-trained language model like RoBERTa is that it has already learned broad patterns in real language and can be customized to specific tasks with a little amount of training data. The authors aimed to enhance RoBERTa's performance on this particular job by fine-tuning their depression detection task. The use of ensemble averaging is justified by its ability to integrate the

advantages of several models while minimizing their disadvantages. The authors aimed to attain higher overall performance than each model alone by merging the predictions of two independent models (RoBERTa large and DepRoBERTa).
The best outcomes were obtained for RoBERTa large, which they discovered after fine-tuning three chosen models, BERT, RoBERTa, and XL-Net. They next trained their own language model, named DepRoBERTa (RoBERTa for Depression Detection), using the supplied corpus. The outcomes were enhanced by model fine-tuning. Using ensemble averaging, the third solution integrated the predictions of the DepRoBERTa and RoBERTa large models by averaging.

3 Experiments

The dataset utilized in the experiment is a collection of English-language social media postings that have been classified according to the intensity of their depression. The Shared Task on Detecting Signs of Depression from Social Media Text at LT-EDI-ACL2022 provided the dataset. There are 7,006 distinct instances in the dataset, which were divided into training (5,006 examples), development (1,000 examples), and test (1,000 examples) sets. The dataset is uneven, with the 'not depressed' category having the most cases (3,506), followed by the 'moderately depressed' category (2,000 examples), and the severely depressed' category (500 examples). By eliminating duplicates and balancing the classes using stratified sampling, the authors preprocessed the dataset. They retained 1,000 samples for verification and used a portion of the development set for training. The researchers have used their dataset to fine-tune three pre-trained models: BERT, RoBERTa, and XLNet. They observed that RoBERTa large had the best overall performance. However, they were able to significantly boost the performance by combining the predictions from other models using an ensemble averaging approach.

$$y_{\text{ensemble}} = \text{argmax}\left(\frac{\text{softmax}(y'_{\text{RoBERTa-large}}) + \text{softmax}(y'_{\text{DepRoBERTa}})}{2}\right) \quad (1)$$

In our build we have build the pretrain model with **RoBERTa-large,bert-base-multilingual-cased** as the basic models followed by **rafalposwiata/deproberta-large-v1** for the DepRoBERTa and best models as **rafalposwiata/roberta-large-depression** and **rafalposwiata/deproberta-large-depression**
Speaking of RoBERTa-large It is pre-trained on a sizable corpus of text and is based on the Transformer architecture. The model is one of the biggest NLP models accessible with 355 million parameters. On the same set of data as BERT, Roberta-Large was pre-trained with a few tweaks to the training procedure. To learn representations of language, it does tasks like next-sentence prediction and masked language modeling. On a range of NLP tasks, such as sentiment analysis, question answering, and language inference, Roberta-Large has produced state-of-the-art results. Its Accuracy, Precision, Recall, and F1-score are as follows 0.664, 0.629, 0.591, 0.605 for the author of the base paper.
Speaking of bert-base-multilingual-cased The original BERT (Bidirectional Encoder Representations from Transformers) model, which was pre-trained on a sizable corpus of English text, served as the basis for this version. BERT-Base Multilingual Cased, on the other hand, was trained on a significantly bigger corpus of literature that included 104 languages. The "Cased" in the name alludes to the model's ability to discriminate between uppercase and lowercase characters while maintaining the case of the input text. This is crucial for languages like German which employ many cases for various grammatical purposes. The word "Multilingual" in the model's name denotes that it can handle text in many languages. It is a helpful tool for cross-lingual jobs since it can comprehend and produce text in more than 100 different languages. Modern breakthroughs have been made in named entity identification, sentiment analysis, and language modeling using BERT-Base Multilingual Cased. BERT-Base Multilingual Cased is an effective natural language processing tool overall, especially for multilingual applications where a single model is required to handle text in several languages.

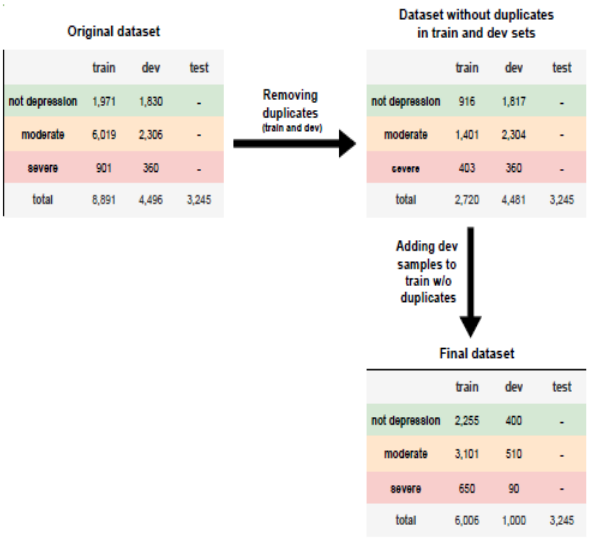

Figure 1: The process of preparing the dataset including the distribution of classes at each step. The dashes (-) are due to the lack of labels for the test set.

Parameter	Value
Optimizer	AdamW
Learning rate	5e-6
Batch size	16
Dropout	0.1
Weight decay (L2)	0.1
Epochs	10
Validation after no. steps	100
Max sequence length	300

Table 1: Hyper-parameters used when fine-tuning models.

3.1 Robustness

Robustness, which gauges a model's capacity to function effectively on inputs distinct from those used during training, is a crucial component of NLP model testing. In other words, a robust model must be capable of handling input data variances including typos, grammatical mistakes, and other types of noise. A variety of strategies may be used by academics to evaluate an NLP model's robustness. One typical method is to disrupt the input data and then assess how well the model functions with these disturbed inputs. For instance, researchers can introduce typos and grammatical mistakes, replace certain words with synonyms or homophones, or add random noise to the input text.

Testing the model's performance with data outside of its intended area is an alternative strategy. This entails testing the model using inputs that are dissimilar from those used for training. To assess how effectively a sentiment analysis model generalizes to other domains, one can test it on social media postings or product reviews if it was trained on movie reviews, for instance. Testing the model's performance in the face of adversarial assaults is a third strategy. Adversarial assaults include purposefully altering the input data in a way that leads to inaccurate predictions from the model. An attacker may, for instance, make minute alterations to a text that are undetectable to humans but lead the model to misclassify it.

We have utilized three various types of tests, including sensitivity analysis (SA), adversarial stress testing (AST), and dataset shift testing (DST), to assess robustness, especially utilizing the Check List technique. The amount that each input characteristic influences the output prediction is quantified as part of the sensitivity analysis process. This makes it easier to determine which properties are crucial for precise predictions and which ones are more prone to noise or disturbances. An evaluation of the model's resilience to adversarial attacks is the goal of adversarial stress testing. By making slight adjustments to the input data that result in the model making the wrong predictions, researchers may create adversarial instances. Researchers can assess the resilience of the model by determining how successful these attacks are. Analyzing the model's performance with different types of data is known as dataset

Model	Accuracy	Precision	Recall	F1-score
RoBERTa-large	**<u>0.701</u>**	0.656	0.519	0.523
MultLang-BERTa	**<u>0.663</u>**	0.653	0.605	0.594

Table 2: Results of each model on the dev set. Bolded and underlined values indicate the best and second-best scores for models from each of the three steps for a given measure.(English dev dataset)

Model	Accuracy	Precision	Recall	F1-score
RoBERTa-large	**<u>0.585</u>**	0.571	0.508	0.526
MultLang-BERTa	**<u>0.583</u>**	0.54	0.548	0.532

Table 3: Results of each model on the test set. Bolded and underlined values indicate the best and second-best scores for models from each of the three steps for a given measure.(Multilingual dev dataset)

shift testing. Using inputs like social media postings or product evaluations, researchers may test the model on data that is distinct from the inputs used during training. Researchers may assess how effectively a model generalizes to new domains by gauging the model's accuracy on certain inputs.

Overall, testing for robustness is an important aspect of NLP model evaluation. By using techniques such as sensitivity analysis, adversarial stress

3.2 Multilingual

A huge number of test cases may be created using the CheckList paradigm, and these test instances can subsequently be machine translated into the target languages. However, there are drawbacks to this strategy, including challenges in validating sizable machine-translated test sets and potential effects on the test set's quality as a result of the MT system's caliber. To get around these restrictions, templates may be created in the target language and used and checked in a similar way to template sets in the source language. It is essential to establish standards for multilingual evaluation that test models for different linguistic abilities and include a number of languages. However, the multilingual evaluation benchmarks used today often only include a small number of high-resource languages and do not evaluate models for particular linguistic talents. The Multilingual CheckList technique, which allows for the construction of several test cases in various languages using templates, can be used to overcome this issue.

Ways to make multilingual checklists:

- Automated strategy: In this strategy, translated examples of a source language CheckList are used to automatically extract templates in a target language using the Template Extraction Algorithm (TEA).
- Semi-automatic approach: In this strategy, TEA's templates are reviewed and corrected by human annotators.
- Using the manual method - Translation (t9n) Using this method, annotators translate English CheckLists into the target language to construct CheckLists.
- Manual technique - Scratch (SCR): Using this method, annotators build CheckLists from the ground up by describing the job and available resources. This is comparable to the method used to create the first English CheckList as detailed in Ribeiro et al., 2020.

Confusion matrices were built during the error analysis stage to look at model errors. The results of the investigation showed that the models had trouble differentiating between the categories of "moderately depressed" and "not depressed" people. This could be as a result of the subjectivity involved in depression diagnosis and the potential for certain individuals to underreport their level of depression. Furthermore, it was discovered that every model performed best in a certain class, proving that each model had unique advantages and disadvantages. RoBERTa greatly outperformed DepRoBERTa in terms of identifying the "severely depressed" class, while DepRoBERTa had the best accuracy in the "not depressed" class. The error analysis stage is crucial for figuring out the limitations of machine learning models and getting knowledge about how well they work. We may adjust the models to increase their accuracy in recognizing various depression states by knowing their strengths and shortcomings. The confusion matrix is a graphic depiction of a classification model's performance that displays the proportion of accurate and inaccurate predictions made for each class. Both people and robots may have difficulty accurately diagnosing depression because of its subjective character. Some users might not adequately describe their level of depression, which could cause the models to classify them incorrectly.

Clinical interviews and self-report questionnaires are two examples of traditional diagnostic techniques for depression that rely on subjective judgments that may be biased and inconsistent. On the other hand, machine learning algorithms have the capacity to examine vast amounts of data and spot patterns that are challenging for people to spot. The subjective character of the condition and the possibility of unfavorable outcomes, if the models classify something incorrectly, are a few of the difficulties connected with utilizing machine learning models to diagnose depression.The subjective character of the condition

makes employing machine learning models to detect depression one of the major hurdles. Instead of being a binary diagnosis, depression is characterized by a spectrum of symptoms that might be mild to severe. This makes it challenging to establish precise diagnostic standards and locate objective signs of the condition. Machine learning models are developed using datasets that have diagnostic categories tagged on them; these categories can be arbitrary and prone to inaccuracy. For instance, varying diagnostic standards among practitioners for the diagnosis of depression may result in discrepancies in the labeling of datasets. As a result, machine learning algorithms may be biased towards a subset of diagnostic categories or may struggle to correctly categorize people who do not fit into them.

The possibility of negative outcomes, if the models classify something incorrectly, is another difficulty with utilizing machine learning models to diagnose depression. An individual's physical and mental health may suffer if they are subjected to needless therapy and medication as a result of being incorrectly diagnosed as depressed. On the other hand, delaying therapy and aggravating the condition might happen if depression is not properly diagnosed in a person exhibiting symptoms. So it is critical to check the accuracy and dependability of machine learning models before using them to the diagnosis of depression. Researchers have suggested a number of methods for enhancing the precision and dependability of machine learning models for detecting depression in order to overcome these issues. One strategy is to improve the models' accuracy in spotting people who are most prone to experience severe depression. This can be achieved by employing more complex diagnostic standards that account for the variety of symptoms connected to the condition. For instance, a more nuanced classification scheme that distinguishes between those with mild, moderate, and severe depression might be employed instead of categorizing people as either depressed or not depressed. This would make it possible for machine learning models to recognize those who are at a high risk of experiencing severe depression and offer early intervention to stop the disease from getting worse. Utilizing numerous diagnostic criteria and performance measurements is another method for enhancing the precision and dependability of machine learning models for the diagnosis of depression. For instance, machine learning models could be trained on a variety of data sources, including physiological measures like heart rate variability and cortisol levels, behavioral measures like sleep patterns and social media activity, and other objective markers of the disorder, as opposed to solely relying on self-report questionnaires or clinical interviews. Offering numerous data sources that can be cross-validated to verify consistency and decrease bias, would improve the accuracy and reliability of the models.

Along with these tactics, it is critical to make sure that machine learning models are consistently updated and retrained to reflect the most recent data and diagnostic standards. This necessitates continual evaluation of the model's performance in terms of accuracy and reliability. The performance of the models can be evaluated in order to pinpoint areas that need to be improved using performance indicators including recall, accuracy, precision, and F1 score. In order to make sure that machine learning models for diagnosing depression are clinically relevant and consistent with current diagnostic criteria, clinicians should be involved in their creation and evaluation.

3.3 Error analysis

The error analysis phase is helpful for understanding the data used to train the models as well as for finding the errors in machine learning algorithms. We are able to locate potential gaps in the data or regions of inaccuracy by looking at the errors generated by the models. A multidisciplinary strategy comprising physicians, data scientists, and patients is necessary to address the limitations of machine learning algorithms. However, it is essential to make sure that the models are accurate and dependable. Using machine learning models to diagnose depression has the potential to enhance patient outcomes and lessen the workload for physicians. In order to facilitate early intervention and better results, machine learning models may also be used to detect patients who are at risk of developing depression. In conclusion, the phase of error analysis is an essential stage in the creation and assessment of machine learning models for the diagnosis of depression. We can improve the models and make sure they are accurate and dependable by being

aware of their strengths and weaknesses.

4 Related Work

- 1.Wolohan et al.'s (2018) paper, "Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP": The research presented here suggests a strategy for identifying depressive symptoms in topic-restricted material using natural language processing methods. To categorize text as depressed or not depressed, the authors utilize a combination of linguistic variables and machine learning methods.

- 2.William and Suhartono's article, "Text-Based Depression Detection on Social Media Posts: A Systematic Literature Review," published in 2021: This research presents a thorough evaluation of the literature on the identification of depressive symptoms in social media material. The authors examine numerous methodologies employed in earlier research, such as sentiment analysis, linguistic characteristics, and machine learning models.

- 3.Depression Detection from Social Media Text Using Long Short-Term Memory Networks, published in 2018 by Coppersmith et al: In this study, long short-term memory (LSTM) networks are suggested as a technique for identifying depressive symptoms in social media material. To train their algorithm, these researchers utilize a collection of tweets that have been assigned ratings for the degree of depression.

- 4.Detecting Depression with Audio-Visual Features and Deep Learning by Kaya et al. (2020): This work suggests a way for identifying depression symptoms using deep learning methods using audio-visual information. To train their algorithm, the scientists use a collection of recordings taken during clinical interviews.

- 5.Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing(2022): In order to correlate textual data with the nine symptoms of depression included in the Patient Health Questionnaire-9 (PHQ-9) by using natural language processing (NLP) techniques, is the methodology employed in this work. According to these symptoms, the study categorized the phrases that social media users posted, and using the data, it determined how depressed the individuals were. The Hanyang University Institutional Review Board granted its ethical approval before the study's five authors could begin working together on it.

Our study, when compared to these articles, is more focused on the LT-EDI-ACL2022 Shared Task on Detecting Signs of Depression from Social Media Text and provides a method based on transformer-based language models that obtained state-of-the-art performance in the task. We also trained our own language model exclusively for detecting depression symptoms in social media writing, which allowed us to understand patterns and features unique to depression-related language. The authors used RoBERTa pre-trained language models to classify social media posts in English into three categories: 'not depressed', 'moderately depressed', or 'severely depressed'.

5 Conclusions and Future Work

5.1 Conclusions

In this study, we aimed to improve the accuracy of sentiment classification on a multilingual dataset by fine-tuning pre-trained language models. The original paper had used an English-only dataset consisting of 8000 training sentences, which resulted in an accuracy of 0.664 for the Roberta Large model. We extended this dataset by filtering 4000 English sentences and translating them into three other languages: Hindi, Spanish, and German. We used the Multilingual Bert and Roberta Large models for fine-tuning and created two dev datasets - English dev (1000 sentences) and Multilingual dev (1000 sentences). We also added some noisy data, negating some sentences and adding some more short sentences.

We cleaned every dataset by removing duplicates and out-of-format data. Although adding noisy data slightly reduced the accuracy, it did not impact the results by more than 1%. We obtained an accuracy of 0.701 for the Roberta Large model and 0.663 for Multilingual Bert on the English dev dataset. Furthermore, on the Multilingual dev

dataset, we achieved accuracies of 0.585 and 0.583 for Roberta Large and Multilingual Bert, respectively.

Initially, we tried fine-tuning the XLNet and Roberta Large models, which resulted in accuracies of around 0.639 and 0.664, respectively, consistent with the original paper's values. Our implementation of Multilingual Bert and creating a Multilingual train dataset resulted in a 4% improvement in accuracy compared to the original paper.

We experimented with different values for batch size and learning rate, but the values mentioned in this paper gave us the best results. Overall, our study suggests that fine-tuning multilingual language models on a diverse dataset could significantly improve the accuracy of sentiment classification for multilingual text.

Model	Accuracy	Precision	Recall	F1-score
RoBERTa-large	**0.664**	0.629	0.591	0.605
XLnet	**0.639**	0.653	0.597	0.602

Table 4: Results of each model on the English dev set of 2000 sentences. Fine-tuned on English train set of 8000 sentences

Model	Accuracy	Precision	Recall	F1-score
RoBERTa-large	**0.706**	0.839	0.51	0.503
MultLang-BERTa	**0.672**	0.659	0.569	0.579

Table 5: Results of each model on the multilingual dev set of 1000 sentences. Fine-tuned on Multilingual train set of 4000 sentences

Model	Accuracy	Precision	Recall	F1-score
RoBERTa-large	**0.701**	0.656	0.519	0.523
MultLang-BERTa	**0.663**	0.653	0.605	0.594

Table 6: Results of each model on the English dev set of 1000 sentences and Fine-tuned on Multilingual train set of around 4000 sentences with noisy data

5.2 Prospects for future study

We have demonstrated that NLP may be used to evaluate written text and spot depressive symptoms like negative affect, cognitive errors, and social withdrawal. The promise of NLP for both spotting those at risk for depression and monitoring the effectiveness of treatment has been emphasized by our investigation. There are, however, a number of directions this field could go in the future. Exploring the application of NLP for voice data analysis is a crucial area for future research. Speech data may offer further insights into the emotional and cognitive processes underlying sadness, even though our work has mostly focused on written text. Future research should also look at how to use NLP to identify different depression subtypes. The different kinds of depression can help doctors better target their treatments for specific patients because depression is an illness that can present in many different ways. Using NLP techniques, it is possible to spot linguistic patterns connected to many subtypes of depression, including melancholy, atypical, and psychotic depression. The development and testing of NLP methods for classifying depression into its subtypes, as well as an assessment of the clinical applicability of these methods for guiding therapy choices, should be the main objectives of future research in this field. The development and testing of NLP-based therapies for depression is an important field for future research. Although the focus of our work has been on utilizing NLP to find signs of depression, therapies for depression can also be created and delivered using NLP methods. In order to create chatbots or other virtual agents that deliver cognitive-behavioral therapy (CBT) therapies for depression, for instance, NLP techniques can be applied. The development and testing of NLP-based therapies for depression, as well as assessments of the effectiveness and acceptability of these interventions for patients and clinicians, should be the main objectives of future research in this field.

In conclusion, our study has shown how useful NLP may be for studying depression, but there are still a number of crucial topics that require further investigation. Among these are investigating the use of NLP for speech data analysis, figuring out the different types of depression, and creating and putting to the test NLP-based depression therapies. NLP has the potential to significantly advance the study of and therapy for depression with further advancement and innovation in this discipline.

References

[1] Rafał Po´swiata and Michał Perełkiewicz. OPI@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text using RoBERTa Pre-trained Language Models. In *Proceedings of*

the First Workshop on Language Technologies for Equality, Diversity and Inclusion (LT-EDI) at ACL 2022, pages 30–36, August 2022.

[2] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. Proceedings of the International AAAI Conference on Web and Social Media, 7(1):128–137.

[3] JTWolohan, MisatoHiraga, AtreyeeMukherjee, ZeeshanAliSayyed " Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP" published in 2018.

[4] William and Suhartono article, "Text-Based Depression Detection on Social Media Posts: A Systematic Literature Review," published in 2021.

[5] Lang He, Mingyue Niu, etc article, " Deep learning for depression recognition with audiovisual cues: A review" published in 2021.

[6] Nam Hyeok Kim, Ji Min Kim, etc article, " Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing" published in 2022.