Text-Based Data Analysis for Mental Health Using Explainable AI and Deep Learning

Chapter *in* Studies in Computational Intelligence · April 2024

DOI: 10.1007/978-3-031-53274-0_3

5 authors, including:

Tazrin Rahman
North South University
1 PUBLICATION **1** CITATION
SEE PROFILE

Nafia Sultana
North South University
1 PUBLICATION **1** CITATION
SEE PROFILE

Text-based Data Analysis for Mental Health Using Explainable AI and Deep Learning

Tazrin Rahman[1], Rehnuma Shahrin[2], Faharia Akter Pospu[3], Nafia Sultana[4], Rashedur M. Rahman[5]

[1,2,3,4,5] Department of Electrical and Computer Engineering,
North South University,
Plot-15, Block-B, Bashundhara Residential Area, Dhaka, Bangladesh.
{ [1]tazrin.rahman, [2]rehnuma.shahrin, [3]faharia.pospu,
[4]nafia.sultana, [5]rashedur.rahman}@northsouth.edu

Abstract: This paper analyzes mental health through text-based data using explainable AI and deep learning techniques, focusing on the classification of addiction, alcoholism, anxiety, depression, and suicidal thoughts. The motivation for this research stems from the lack of suitable datasets from Twitter and the need for a more comprehensive understanding of mental health using text data. We propose a methodology that involves leveraging Reddit data, employing various text vectorization techniques (TF-IDF, Word2Vec, and GloVe), and implementing multiple classification algorithms (XGBoost, Decision Tree, SVM, Naive Bayes, Simple Gradient Descent, Stochastic Gradient Descent, K-Nearest Centroid, K-Nearest Neighbor, AdaBoost, Random Forest, and Logistic Regression) using the TF-IDF text vectorizer. The problem revolves around accurately classifying mental health-related topics based on text data. By employing a range of classification algorithms and text vectorization techniques, we aim to develop machine-learning models that effectively identify addiction, alcoholism, anxiety, depression, and suicidal thoughts within text-based discussions. This research contributes to mental health analysis using text-based data and advanced machine-learning techniques. The results highlight the potential of Reddit as a valuable resource for understanding mental health concerns and demonstrate the effectiveness of the proposed methodology. Our findings have far-reaching implications for mental health practitioners, researchers, and policymakers, facilitating the development of tailored interventions and support systems for those in need.

Keywords: Explainability, Textual analysis, LIME, Reddit, Deep learning, Text-based data.

1 Introduction

Mental health has emerged as a critical global issue, affecting individuals across all walks of life. The increasing prevalence of mental health disorders calls for comprehensive research and innovative approaches to understanding and addressing these conditions. Recently, there has been a growing interest in utilizing machine learning and AI techniques to analyze text-based data for mental health analysis. Existing studies have predominantly focused on utilizing Twitter data for mental health analysis. Twitter provides a wealth of real-time information, making it suitable for certain types of analysis, such as real-time monitoring and event tracking. However, these datasets often have limitations such as overuse, outdated information, and a lack of specificity to diverse mental health conditions. These challenges researchers seeking to develop effective treatments and interventions for a broader range of mental health concerns. The motivation for this research stems from the inadequacy of existing Twitter datasets and the need for a more comprehensive understanding of mental health using text-based data. In this study, we propose using Reddit data as an alternative source for mental health analysis. Reddit offers a unique platform characterized by longer and more detailed discussions, which can provide valuable insights into the language used by individuals when discussing their mental health concerns and the topics that are most relevant to them. This research aims to conduct a comprehensive analysis of mental health using text-based data obtained from Reddit, leveraging explainable AI and deep learning techniques. The primary goal is to develop machine learning models that can effectively classify and analyze mental health-related text data, specifically focusing on addiction, alcoholism, anxiety, depression, and suicidal thoughts.

In summary, this study is motivated by the limitations of existing Twitter datasets and the need for a more comprehensive analysis of mental health using text-based data. By utilizing Reddit data and advanced AI techniques, we aim to advance the field of mental health analysis, providing valuable insights and facilitating the development of effective interventions for a wide range of mental health conditions.

2 Related Works

Research using Reddit data for mental health analysis has attracted attention recently. Several studies have explored the potential of Reddit as a valuable resource for understanding mental health-related discussions and developing machine learning models for classification and analysis. Here are a few examples of existing research and the limitations associated with this work:

The study [13] focused on predicting suicide risk based on Twitter data. It developed a machine learning model that analyzed language patterns and linguistic cues to identify users at risk of self-harm. The research highlighted the potential of Twitter for early detection and intervention in mental health crises. The study focused specifically on suicide risk prediction, and the model's performance may vary when applied to other mental health conditions.

In [14], the authors compared Reddit and Twitter data characteristics for biomedical research, including mental health topics. The study analyzed the differences in user demographics, language patterns, and topic distributions between the two platforms. While the research provides insights into the differences between Reddit and Twitter, it does not specifically focus on mental health analysis. The findings might not directly address the specific needs of mental health researchers or the limitations associated with using Reddit data for mental health analysis.

3 Data Set Descriptions

Although there are many datasets of Twitter posts for mental health analysis, we failed to find any that suited our research requirements. There is a problem that the dataset was already used multiple times, or it needed to be updated to use for new work. So, we decided to create a new dataset for our research purpose. We used Reddit data instead of Twitter data in this case.

The decision to use Reddit data instead of available Twitter data for our research depends on various factors. Due to the platform's nature and user characteristics, Reddit data is more appropriate for certain types of analysis, such as sentiment analysis [7,12] or topic modeling [1,10]. For example, Reddit users tend to engage in longer and more detailed discussions, which may be more helpful in understanding the nuances of certain topics or communities. Additionally, Reddit data is more readily available for certain topics or communities of interest, as many subreddits are dedicated to specific topics or interests. On the other hand, Twitter data is considered more appropriate for other types of analysis, such as real-time monitoring or event tracking, due to its focus on up-to-the-minute updates and its use by a broader range of users. Reddit data has been found more suitable due to the platform's nature and the availability of data on mental health-related topics. It is also considered an inexpensive source of high-quality data. Our research required data based on different mental health conditions rather than only depression. Reddit is known for having numerous subreddits dedicated to mental health, such as r/mentalhealth, r/depression, and r/anxiety, where users often share their personal experiences, discuss treatment options, and offer support to others. The data collected from these subreddits can provide valuable insights into the language used by individuals when discussing their mental health concerns and the topics most relevant to them. Using this data, we can develop machine learning models that can classify and analyze mental health-related text-based data, which can aid in developing effective treatments and improving mental health outcomes.

But using Reddit data has some drawbacks, too. Lots of people informally do their subreddits. The data can contain incomplete words, random punctuations, and symbols, making the dataset noisy and inappropriate for our research.

4 Dataset Collection and Preparation

Although there were no available datasets that contained data from Reddit, we created our dataset. Our approach was using the web scraping method to collect data from Reddit.

Web scraping: Web scraping automatically extracts data from websites using software programs and scripts. This process allows us to easily get data from any online platform. The software programs are designed to crawl through the website's structure, extracting specific information and storing it in a structured format such as a spreadsheet or database. There are various tools available for doing web scraping. We must select the tool based on our research necessities and complex scraping tasks. Our main target is to collect as much data as possible because the more data, the better prediction. We used the 'Web Scraper - Free Web Scraping' tool, a Google Chrome extension. It is free and very easy to use. It does not require any other software installed on the device. It is an extension to the browser, so it works simply by just adding it. This scraping mainly works in the backend of the browser. We have attached the interface of the web scraper below. We created five different classes, which are called sitemaps here. Those sitemaps contain the data extracted from the website through scraping. After completing web scraping from Reddit, we exported the data as a XLSX file, which we later converted as a .CSV file.

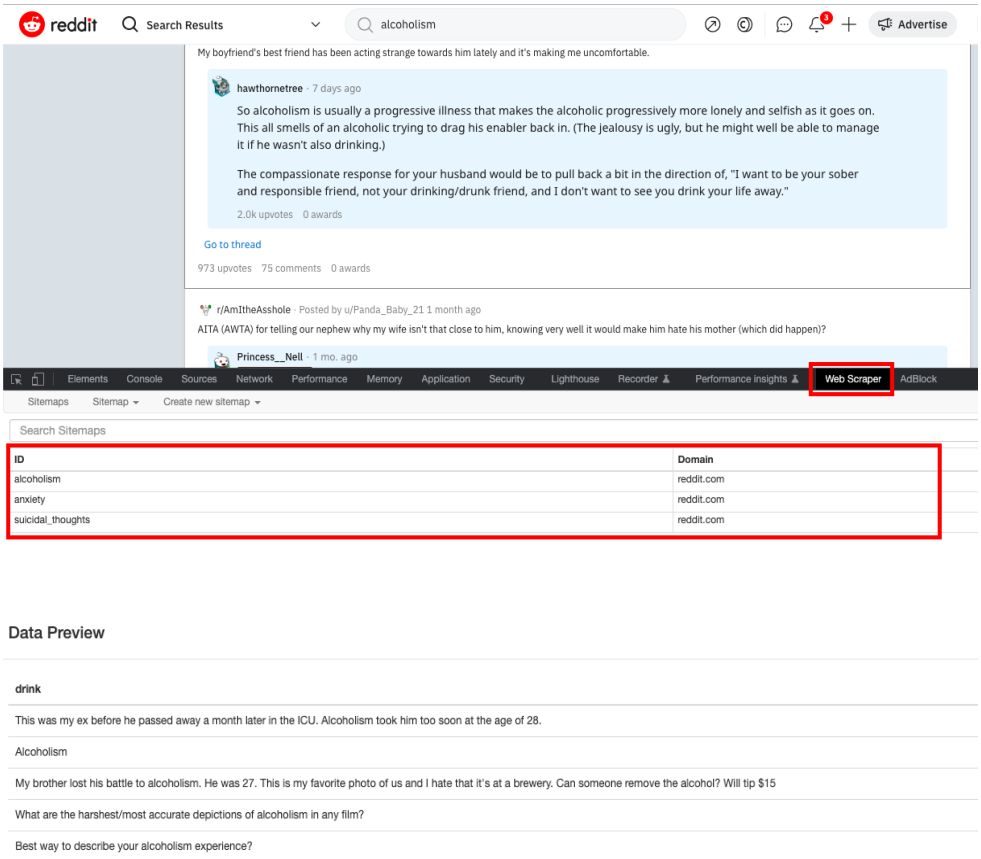

Figure 1: Web scraping tool and method

We divided our dataset into five classes and then merged them into one final dataset for our research. The final dataset name is 'mental state'. The five classes are: 'Anxiety', 'Addiction', 'Alcoholism', 'Suicidal Thought', and 'Depression'.

Here is the dataset plotting summary in the Figure 2.

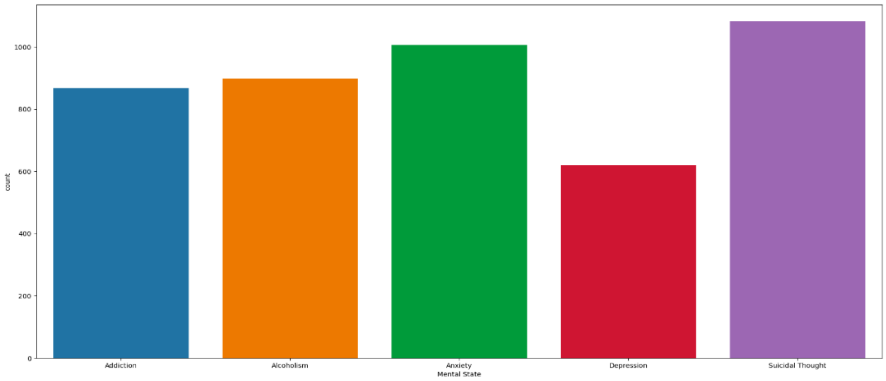

Figure 2: Dataset Summary

5 Preprocessing Techniques

Preprocessing is an important step in machine learning, where raw data is cleaned and transformed into a form that can be used for further analysis. Preprocessing aims to improve data quality and usability by removing noise and inconsistencies, making the data more consistent and easier to process. Preprocessing plays an important role in machine learning models as it can greatly impact the accuracy and effectiveness of machine learning models. Poor data preprocessing can produce inaccurate predictions and models that lack robustness and reliability. On the other hand, proper data preprocessing ensures that machine learning models can deliver accurate predictions and valuable, relevant, and meaningful insights. Figure 3 shows the data preparation process.

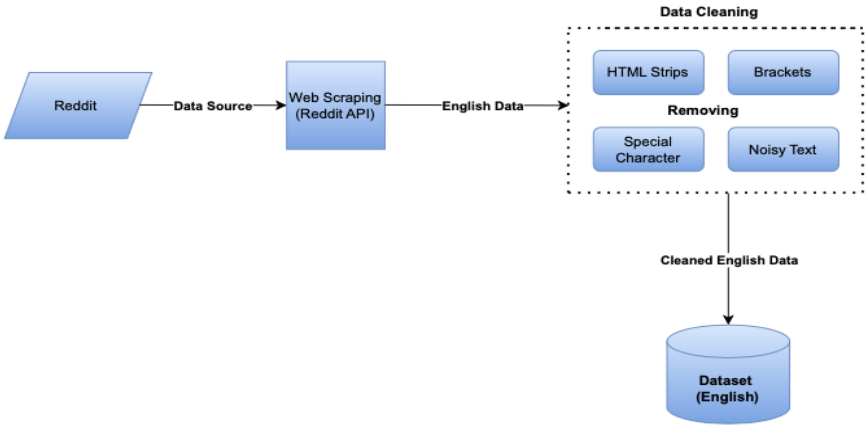

Figure 3: Dataset preparation process

After successfully handling missing values, removing noise, stemming the text, and removing stop words, we were left with the following dataset:

Full dataset: (4472,4)
Longest String length before Preprocessing: 13690
Longest String length after Preprocessing: 7323

Among all the data, we divided our dataset into two parts, which is 80% for the training set and the 20% for the test set.

6 Methodology

We started our research by collecting data from online platforms by web scraping. Almost 4500 data were collected initially from Reddit. After collecting the data, we applied different embedding matrices as feature extractors to extract features from our data. Then, those features were fed into the machine learning and deep learning models. After successful training, the models were evaluated and showed our desired output.

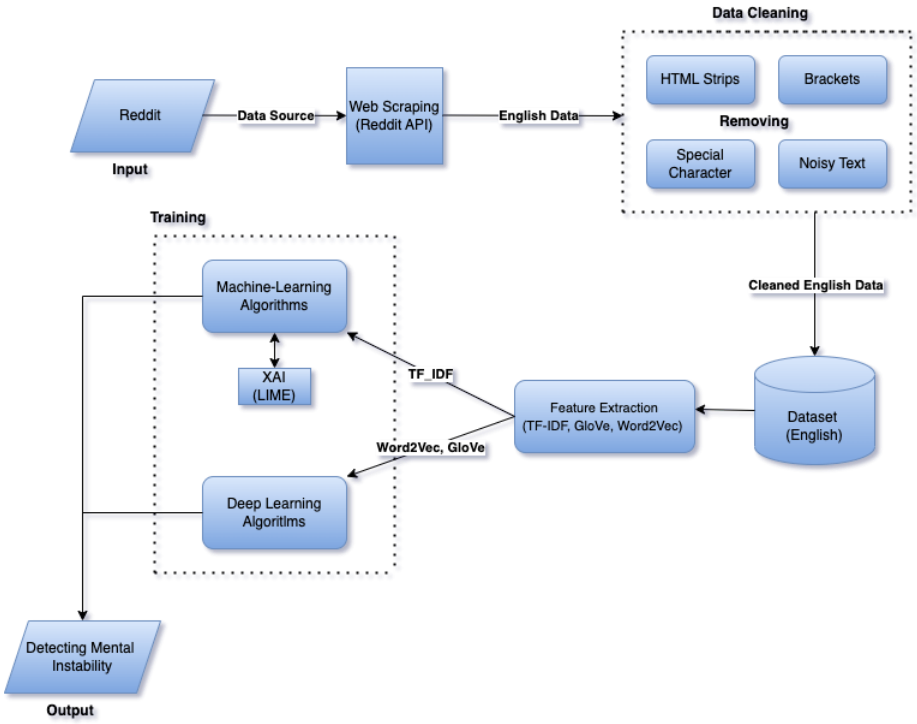

Figure 4: System diagram

After preprocessing and categorizing the data, we started with the model-building part. Model building creates a predictive model that can make predictions or classifications based on input data. Building a model typically involves choosing an appropriate

algorithm, defining model inputs and outputs, preprocessing data for modeling, training the model using a training dataset, and training the model against a test dataset.

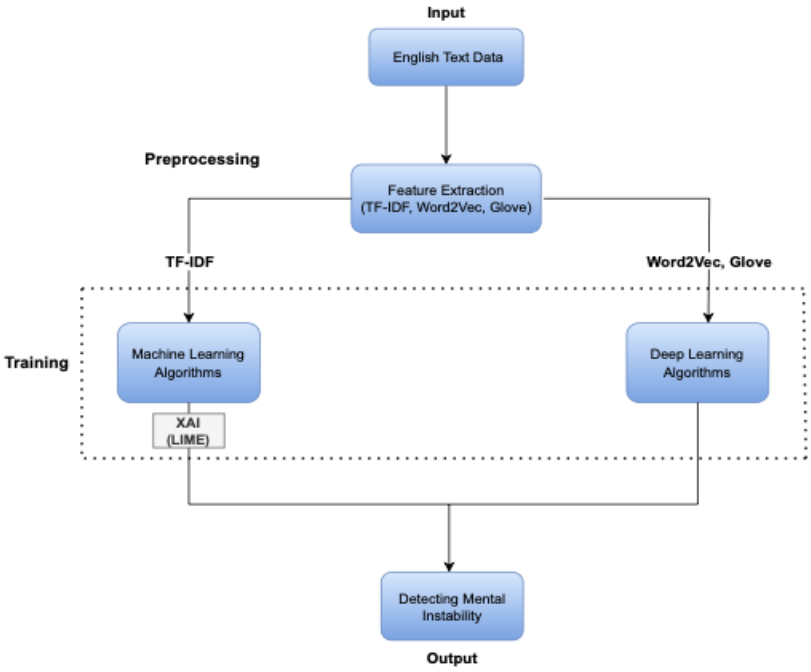

Figure 5: Model Selection Diagram

Initially, machine learning models were used to classify text data. Specifically, the models are trained to predict the class labels (e.g. 'Depression', 'Anxiety', 'Addiction', etc.) of the input text data based on patterns and relationships found in the training data. XGBoost, Decision Tree, SVM, Naive Bayes, Simple Gradient Descent, Stochastic Gradient Descent, K-Nearest Centroid, K- Nearest Neighbor, AdaBoost, Random Forest, and Logistic Regression models were used using the TF-IDF text vectorizer. All the ML models were run on the default setup with no changes in the hyperparameter. We also added a voting classifier with all the ML models for final prediction. A few variations of DNN models were used to evaluate the performance of our dataset. We used CNN (Convolutional neural network) algorithm with Word2Vec and GloVe vectors [2, 3, 9].

Hyperparameters for deep neural network methods are included in the Table 1:

Hyperparameters	Word2Vec-CNN	GloVe-CNN+BiLSTM
Pooling type	max	max
Embedding dimension	300	300
Batch size	64	64
Activation function	swish, SoftMax	swish, SoftMax
Learning rate	0.001	0.001
Optimizer	adam	adam
Epoch	35	35

Table 1: Hyperparameters for DNN models

7 Explainability Methods

Explainability methods are important for understanding the model's behavior, ensuring fairness, and building trust in its predictions [4-6,8]. Our research uses LIME as an explainable AI (XAI) method. LIME (Local Interpretable Model-Agnostic Explanations) is a model-agnostic technique used to explain the predictions of machine learning models [2]. It can be used with any type of model, especially those used for text classification tasks. There are many XAI methods, but one of the main reasons for using LIME in our research is to gain insights into how a model makes its predictions on individual text samples. LIME creates locally interpretable explanations for the predictions made by a model by identifying the most important features or words that contributed to the prediction. LIME uses a simpler model, such as a linear model or decision tree, to approximate the behavior of the complex model in the local region around a specific instance. This approach is computationally efficient and can provide meaningful insights into the contribution of different words or features to the model's prediction.

As our research is text-based, LIME mainly focuses on the words that give important and useful information in explainability. It highlights the important words for the detection and thus gives the outcome. This type of explainability is like a human prediction explanation, that is why it is widely accepted.

8. Result and Analysis

This section reviews all the models (ML and DNN models) tried on our English textual dataset to classify mental states from the text. Model performance analysis based on accuracy involves evaluating how well a model predicts the correct outcome compared to the total number of predictions. It measures how many predictions are correct from the total number of predictions. As for our ML models, we have different accuracy rates for different models. Not all models gave satisfactory results. For now, we chose accuracy to determine a model's performance. However, other matrices (precision, recall, f1) were also considered to evaluate the performance. On our dataset, XGBoost and Naïve Bayes outperformed all the models with an accuracy of 70% and 71%, respectively. Table-2 shows the accuracy for different models.

Model	Accuracy
Naive Bayes (TF-IDF)	71%
SVM (TF-IDF)	65%
XGBoost (TF-IDF)	70%
Decision Tree Classifier (TF-IDF)	59%
Stochastic Gradient Descent (TF-IDF)	68%
Logistic Regression (TF-IDF)	67%
AdaBoost (TF-IDF)	61%
K-Nearest Centroid (TF-IDF)	37%
K-Nearest Neighbor (TF-IDF)	40%

Table 2: ML Models' Accuracy Percentage

Explainable AI: What sets our research apart is the focus on explainability. Trust and transparency are crucial when dealing with sensitive topics like mental health. Therefore, we have integrated Explainable AI techniques into our model, allowing us to provide clear and understandable explanations for its predictions. XAI (Explainable Artificial Intelligence) plays a key role in the research by providing transparency, interpretability, and accountability to the classification model decision-making process. This enables us to uncover important features, patterns, and rules influencing the model's output. We separately applied the XAI method (LIME) [11] to our machine-learning models. Figure 6 and 7 provide explainability of the XAI models.

Figure 6: Explanation of Naive Bayes model using LIME

Figure 7: Explanation of XGBoost's model using LIME

Once the interpretable model is trained, LIME assigns importance weights to the features based on their contribution to the predictions. These weights explain the model's decision by highlighting the features that most influenced the prediction. As we can see, the results are shown on the prediction probabilities mentioning the five mental states. The words that are highlighted are the words predicted by the model that can refer to Alcoholism (77%) and Anxiety (23%).

Like the shown model, we applied XAI on each ML model we trained. Each of them showed results according to their model performance. However, we can conclude that our model's explanation part worked successfully.

For DNN models, we have evaluated two models with 35 epochs each. The first one is **Word2Vec-CNN,** which has an accuracy of 64%. Its training accuracy is about 98%, and its testing accuracy is 65%. This means this model needs to be trained for more and more time. By each time this model will gain better insights into the model's pattern. The second DNN model is **Glove - CNN + BiLSTM**. It has an accuracy of 68% which is better than Word2Vec. Its training accuracy is 99.9%, whereas testing accuracy is 69%. We can finally tell that among both DNN models GloVe+BiLSTM is the better model.

For the Word2Vec-CNN model, after running 35 epochs initially, the model is trained on the dataset properly for about 98.6%, and on the same dataset, it was tested with 64.32% accuracy.

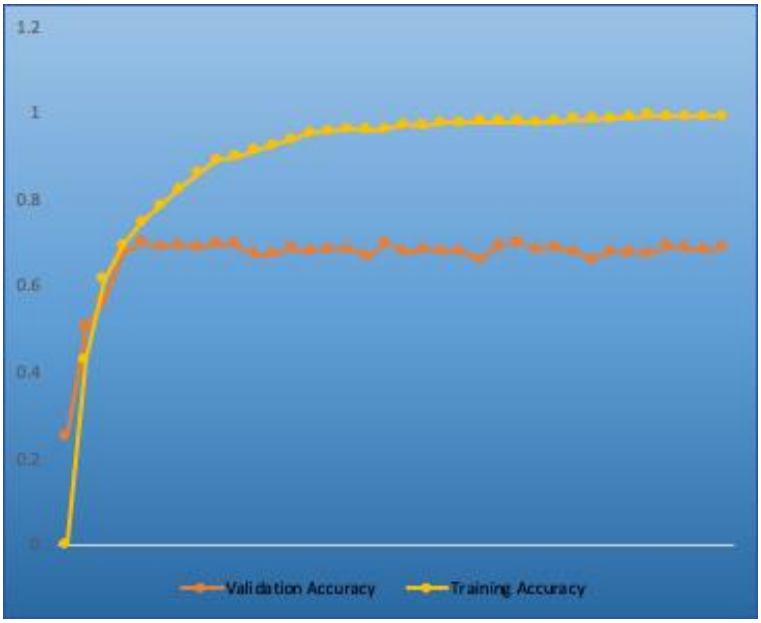

Figure 9: Accuracy graph for GloVe-CNN+BiLSTM model

Training loss is the loss the model incurs on the training data during training. It is the error that the model makes when it tries to predict the correct output for the training data. Validation loss is the model's loss on a separate validation set during training. We can see that with each epoch, the training loss started to decrease, and the accuracy started to increase. The best accuracy achieved by the model is 67.23%.

Figure 10: Loss graph for Word2Vec-CNN model

In sentiment analysis, error analysis is used for getting a closer look on the misclassified text samples. For example, suppose our model classifies user posts as mental health analysis and misclassifies depression data as anxiety data. In that case, we can examine the words or phrases that caused the misclassification to determine why the model failed. Among DNN models, GloVe-CNN+BiLSTM showed comparatively better performance. That is why we did a thorough error analysis with a confusion matrix of that particular model. Figure 11 depicts the confusion matrix.

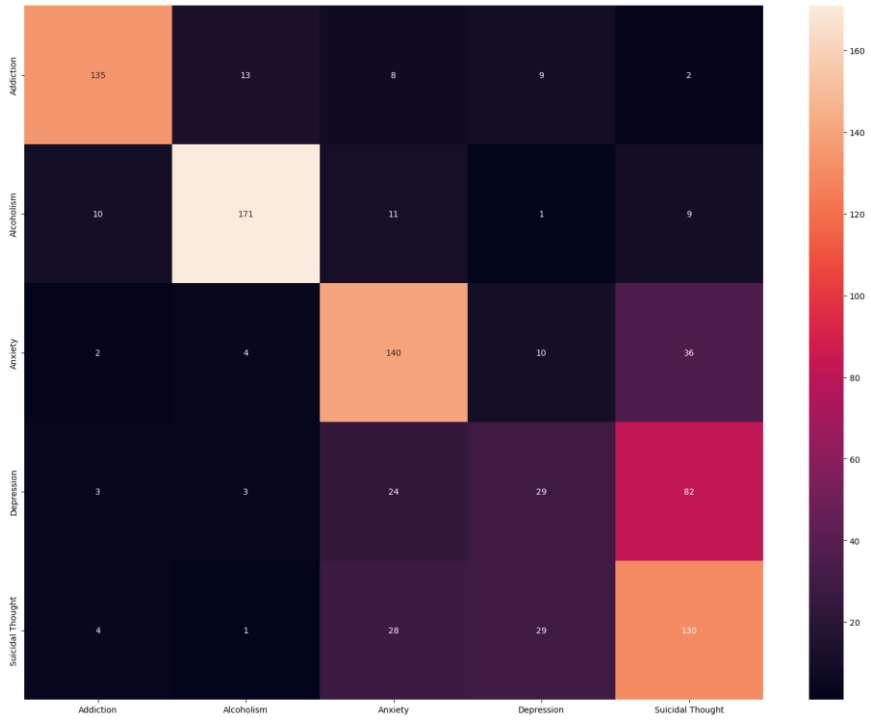

Figure 11: Confusion matrix of GloVe-CNN+BiLSTM model

The matrix depicts a class-by-class proportion of estimated labels. The matrix reveals that a small number of data points were incorrectly categorized. For example, 82 samples were predicted as 'Suicidal thoughts' in the 'Depression" class. This represents a huge error. In almost all the ML models, the 'Depression' class gave the least accuracy. It means the class 'depression' is frequently misclassified during the predictions. There are many possible reasons for inaccurate predictions, e.g., class imbalance in the dataset. So, by creating a balanced dataset, prediction errors could be reduced to a minimum.

9. Conclusion and Future Work

This research aimed to comprehensively analyze mental health using text-based data obtained from Reddit, employing explainable AI and deep learning techniques. By addressing the limitations of existing Twitter datasets and leveraging the unique

characteristics of Reddit, we sought to develop machine learning models capable of accurately classifying and analyzing mental health-related text data.

To effectively process the text data, we utilized various text vectorization techniques, including TF-IDF, Word2Vec, and GloVe, transforming the textual information into numerical representations suitable for machine learning algorithms. Implementing a range of classification algorithms, such as XGBoost, Decision Tree, SVM, Naive Bayes, Simple Gradient Descent, Stochastic Gradient Descent, K-Nearest Centroid, K-Nearest Neighbor, AdaBoost, Random Forest, and Logistic Regression, we evaluated the performance of the developed models using precision, recall, and F1 scores. These metrics provided insights into the accuracy and effectiveness of the models in correctly identifying and classifying instances of different mental health conditions.

We have taken great effort to assure the correctness and dependability of our findings throughout the whole study process. We trained our model using a large and diverse dataset covering various mental illnesses and language patterns. A rigorous testing and validation process was employed to ensure the robustness and generalizability of the results.

The findings of this research have significant implications for mental health research and practice. By accurately classifying mental health-related text data, we gain a deeper understanding of the prevalence, characteristics, and nuances associated with addiction, alcoholism, anxiety, depression, and suicidal thoughts. Furthermore, integrating explainable AI techniques in our methodology enhances the interpretability and transparency of the classification models.

In conclusion, this research contributes to mental health analysis by utilizing text-based data from Reddit and employing advanced machine-learning techniques. The outcomes of this research provide valuable insights into mental health conditions and pave the way for future advancements in personalized interventions, support systems, and mental health treatments. By harnessing the power of text-based data and innovative AI methodologies, we can continue to improve mental health outcomes and promote overall well-being in individuals facing various mental health challenges.

While our dataset was a collection of Reddit data from different regions of the world, it did not have important factors such as the age, demographic, and financial capacity of the users surveyed. Another important factor the dataset did not account for was the platform from which to collect data. Most teenagers use Instagram or Snapchat these days, while elderly people mainly use Facebook as their go-to social media platform. We could factor these important aspects into our dataset and get even more accuracy. There are also recent deep-learning models that we are currently working on. We are also planning to complete that model and apply XAI to it. Another future work is using Bengali-translated data in our model.

References

[1] M. Jamnik and D. Lane, "The Use of Reddit as an Inexpensive Source for High-Quality Data," *Practical Assessment, Research, and Evaluation*, vol. 22, no. 1, Nov. 2019, doi: https://doi.org/10.7275/swgt-rj52.

[2] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014, doi: https://doi.org/10.3115/v1/d14-1162.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv.org*, Sep. 07, 2013. https://arxiv.org/abs/1301.3781

[4] "Using Machine Learning for Sentiment Analysis: a Deep Dive," *DataRobot AI Cloud*. https://www.datarobot.com/blog/using-machine-learning-for-sentiment-analysis-a-deep-dive/

[5] M. Du, N. Liu, and X. Hu, "Techniques for Interpretable Machine Learning," *arXiv:1808.00033 [cs, stat]*, May 2019, Available: https://arxiv.org/abs/1808.00033

[6] A. H. Yazdavar *et al.*, "Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media," *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, Jul. 2017, doi: https://doi.org/10.1145/3110025.3123028.

[7] "How To Train a Neural Network for Sentiment Analysis | DigitalOcean," *www.digitalocean.com*. https://www.digitalocean.com/community/tutorials/how-to-train-a-neural-network-for-sentiment-analysis

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv.org*, Feb. 16, 2016. https://arxiv.org/abs/1602.04938

[9] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in Pre-Training Distributed Word Representations," *arXiv:1712.09405 [cs]*, Dec. 2017, Available: https://arxiv.org/abs/1712.09405

[10] I. Lage *et al.*, "An Evaluation of the Human-Interpretability of Explanation," *arXiv.org*, Aug. 28, 2019. https://arxiv.org/abs/1902.00006 (accessed Oct. 24, 2023).

[11] P. Bhatnagar, "Explainable AI(XAI) — A guide to 7 packages in Python to explain your models," *Medium*, Jun. 04, 2021. https://towardsdatascience.com/explainable-ai-xai-a-guide-to-7-packages-in-python-to-explain-your-models-932967f0634b

[12] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975, doi: https://doi.org/10.1145/361219.361220.

[13] Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, Argyle T. Tracking suicide risk factors through Twitter in the US. Crisis. 2014;35(1):51-9. doi: 10.1027/0227-5910/a000234. PMID: 24121153.

[14] Correia, R. B., Wood, I. B., Bollen, J., & Rocha, L. M. (2020). Mining Social Media Data for Biomedical Signals and Health-Related Behavior. Annual Review of Biomedical Data Science. https://doi.org/10.1146/annurev-biodatasci-030320-040844