Combining Natural Language Processing and Machine Learning to Identify Stress from Mental Health Conditions in Subreddit Posts

Deanna Oliver, MPH, Dual M.S. Applied Information Science & Information Systems[1], Yi-Ru Pei, Applied Information Science & Information Systems[1], Yuhan Xu, M.S. Health Informatics[2], Harshvi Shah, M.S. Health Informatics[2]

[1]Cornell Tech, New York, NY, [2]Weill Cornell Medicine, New York, NY

Abstract

Accurately diagnosing mental illness poses unique challenges relative to other chronic diseases. While objective tests exist for many medical conditions, mental health diagnoses rely heavily on qualitative observations. Moreover, factors such as systematic and structural challenges, healthcare provider variabilities, communication complexities, and patient factors further enhance diagnostic challenges. This study addresses the need for more objective means of examining the application of NLP in processing mental health-related data. Utilizing a dataset of subreddit posts, we employed NLP and machine learning techniques to detect linguistic markers of stress and mental health conditions. The study compared the efficacy of various word embedding methods—ELMo, BERT, and Bag of Words (BoW)—alongside three different machine learning models—Logistic Regression, SVM, and XGBoost—to classify text data. Our results indicate that BERT embeddings, particularly when used with Logistic Regression and SVM, significantly enhance the identification of stress indicators, potentially offering a more objective lens through which mental health can be assessed. This work suggests a promising direction for augmenting traditional diagnostic methods with automated text analysis tools, aiming to reduce subjectivity in mental health diagnoses and support healthcare professionals in their clinical assessments.

1. Introduction

This study delves into the intricate challenge of achieving accurate diagnoses within the sphere of mental health disorders. The ability to precisely diagnose such conditions affects not only the trajectory and outcomes of an individual's treatment but also plays a key role in broader sociological and administrative capacities. [1] Despite the critical nature of this task, the process of accurately identifying mental health conditions is fraught with challenges, stemming from a variety of sociological, cultural, and clinical factors. The lack of objective biomarkers, the dynamic nature of diagnostic criteria, and the influence of racial and cultural biases, in addition to the subjective aspects of clinical judgment and characteristics of healthcare providers, significantly hinder the reliability of diagnoses. [2][3][4][5][6] However, in recent years, Natural Language Processing (NLP) and Machine Learning (ML) techniques have emerged as promising tools for analyzing textual data to better understand, predict, and diagnose mental health conditions. [7] In this study, we leverage NLP and ML to extract diagnoses of mental health conditions from online posts. NLP and ML offer promising tools for analyzing textual data to improve diagnostic accuracy. By analyzing online posts, NLP and ML can extract linguistic markers and themes that are indicative of specific mental health conditions, identifying subtle patterns that may be missed by human evaluators and improving diagnostic accuracy and consistency. [8]

2. Related Work

One commonality among related studies is the utilization of advanced NLP techniques such as ELMo, BERT, and Bag-of-Words, often combined with supervised learning models like Support Vector Machines (SVM) and XGBoost, to extract meaningful insights from textual data. For instance, Inamdar et al. and Bora and Kumar both leverage these techniques to detect mental stress on social media and early signs of mental illness, respectively. [9][10] Additionally, several studies emphasize the importance of large-scale data analysis in understanding counseling conversations and predicting suicidal ideation. [11][12]

Nevertheless, differences emerge in the specific applications and datasets used. For example, DeSouza et al. focus on speech and text pattern analysis to identify signs of depression, [13] while Swaminathan et al. developed a Crisis Message Detector-1 (CMD-1) to flag potential crisis messages in tele-mental health platforms. [7] Moreover, while some studies like Jackson et al. and Borah and Kumar highlight the importance of scalability and

integration with existing workflows, [8][10] others like Malhotra et al. underscore the need for specialized datasets to improve model performance. [14] Extensions of these studies include addressing the identified limitations and exploring novel applications. For instance, many studies in this domain focus on predicting a single mental health condition or the presence of stress from clinician-patient text rather than exploring the nuances of patient text across multiple conditions.

3. Material and Methods

Dataset

The dataset used in this study is the set of text posts on Reddit, containing 3,553 posts from the Reddit community between January 1, 2017 and November 19, 2018 in five different categories. This dataset, specifically curated to classify textual data into binary categories of stressed or not stressed, also encompasses aspects of the text's readability, sentiment, and lexical content, alongside social media metrics such as timestamps and karma scores. For this analysis, the dataset has been streamlined to focus on columns that are most relevant for text classification tasks involving sentiment analysis, topic modeling, or similar NLP-driven evaluations. The refined dataset includes a subset of the original features directly related to the textual content of the posts and their derived representations, including:

- Subreddit: the Reddit community where the post was published. Subreddits are specific forums dedicated to particular topics or themes, and the name of a subreddit can indicate the nature of the discussion.
- Label: A binary or categorical target variable indicating the classification label or class of the post. Depending on the task of this paper, it represents the psychological state of the post, such as positive/negative emotions or presence/absence of mental stress.
- Text: The raw text of the Reddit post. This is the unprocessed text from the data source containing the full content of the user's post within the subreddit.

Furthermore, the original dataset contained both training and test sets, but in detail we merge them into a new dereddit dataset, which is then divided into training, development, and test sets to evaluate the performance of the machine learning models in a structured way (see Methods). Table 1 reflects a summarization of the number of entries, sentences, tokens, and unique tokens in the training and test sets by NLTK.

Table 1. The number of entries, sentences, tokens, and unique tokens in the training and test sets by NLTK

	Entries	**Sentences**	**Tokens**	**Unique tokens**
Training set	2838	13345	280586	13889
Testing set	715	3378	70405	6465

Methods

We utilized a combination of NLP and ML techniques proposed by DeSouza et al. to help accurately detect mental stress in Reddit posts. [13] Different combinations of text representations and traditional ML algorithms have different performances in NLP. After evaluating and comparing them, the combination that performs optimally on this dataset is screened and the final goal is to perform binary text classification on reddit posts: marking text as stressful and non-stressful.

With respect to NLP tools, ELMo, BERT and Bag of words [1] were chosen, each of them offers unique advantages and captures the essence of language. ELMo uses character level embedding strategy to create vectors for text sequences. This approach breaks down words into characters, which can be modeled by machine learning to generate embeddings that capture word-level and sub-word-level information. In contrast, BERT takes a different approach to preserving contextual information in text sequences: representing words as vectors, utilizing a transformer architecture trained on large text datasets. As the most advanced embedding technique among the three, BERT can effectively capture the nuances and dependencies between words by recognizing the context in both directions. On the other hand, bag-of-words modeling adopts a simpler but effective approach to represent

sentences: the text is treated as an unordered collection of words, focusing only on the corresponding frequencies, regardless of syntax and word order. Despite the lack of contextual understanding, word frequency is still a relevant feature in document categorization and sentiment analysis. ML models (especially SVMs) are able to effectively utilize the high-dimensional sparse vectors they generate and achieve significant classification performance. In addition, three traditional machine learning models, Logistic Regression, SVM and XGBoost [2] were trained for text categorization and stress recognition[3].

Our approach has a total of four steps: preprocessing, word embedding, machine learning modeling and model evaluation and comparison.

First, Google Drive is installed to access the dataset stored in it. The test and training sets of the original dataset are merged and then checked for missing values, which ensures that there is no missing data for the feature under consideration. Since the dataset collects posts from different users on Reddit, its text data may have issues such as high noise, spelling, and grammatical errors. Meanwhile, users' different cultural backgrounds and individual differences lead to diversity in text content and language. Therefore, preprocessing is a necessity. Stop words, punctuation, links and direct calls to subreddit are removed from the text. Tokenization is done using NLTK and stemming extraction is implemented. In addition, TF-IDF technique is used to perform keyword extraction on the text data and we extract the first ten keywords of each text, which helps to analyze the high-frequency or focused words related to mental health and further identify their mental stress conditions. Then, we use ELMo, BERT and BoW for word embedding respectively.

Before training the machine learning model, the dataset needs to be split. First, 85% of the data is used for the initial split and 15% for the test set. Then, the initial 85% was split into 82.35% for the training set and 17.65% for the validation data to achieve an approximate 70-15-15 split ratio. Next, for each word embedding method, we construct and train Logistic Regression, SVM and XGBoost models in turn. Multiple measures: accuracy, recall, precision and F1-score are used for initial evaluation of performance. To improve the model performance, hyper-parameter tuning using grid search is performed and the model performance under the optimal parameters is tested by validation data. Finally, we obtain nine sets of combined models. The conclusions are drawn by applying the test set to the tuned models and evaluating the corresponding model performance to filter out the optimal combination model.

4. Experimental Settings

Data Preparation and Preprocessing

The re and nltk toolkits provide regular expression manipulation and natural language processing functions, respectively, which help to realize the various steps of text processing.

Feature Extraction

In keyword extraction, TfidfVectorizer class comes to convert the text into TF-IDF feature representation.

Word Embedding Experiments

For the ELMo model, we use the Elmo module from the AllenNLP library, a toolkit for processing deep learning models to accomplish word embedding for EMLo. The torch library, which is the core library of the PyTorch deep learning framework, is also introduced. The BertTokenizer and BertModel classes in the transformers library help with BERT embedding. Moreover, the embedding that implements Bag of Words uses the CountVectorizer class from the sklearn library.

Model Training, Tuning, and Evaluation

To divide the data into training, test and validation sets, we called the train_test_split function of the sklearn.model_selection module twice to implement data splitting.

In the training of machine learning models, LogisticRegression and SVC (Support Vector Machine) from the Scikit-learn library were used for Logistic Regression and SVM model training, and the XGBClassifier module

from the xgboost library was used to construct XGBoost models. In addition, the introduction of modules such as GridSearchCV was used to construct and tune relevant models and perform grid search.

Multiple types of evaluation metrics functions such as accuracy_score, precision_score, recall_score, f1_score, etc. are used to generate the final results.

5. Results

The primary objective of this study was to evaluate the efficacy of different NLP word embedding techniques and machine learning models in identifying stress indicators from mental health-related subreddit posts. Our methodology comprised the implementation of three word embedding experiments—ELMo, BERT, and Bag of Words (BoW)—each coupled with three machine learning models—Logistic Regression, SVM, and XGBoost. The performance of these models was measured using accuracy, recall, precision, and F1-score.

Performance of Word Embedding Techniques

- ELMo: Exhibited moderate performance due to computational intensity.
- BERT: Stood out for its depth of context and semantic understanding, with Logistic Regression and SVM showing enhanced performance.
- BoW: Effective results, particularly with SVM, suggesting its utility with limited computational resources.

Machine Learning Model Evaluations

- Logistic Regression: Consistent performer, especially when combined with BERT embeddings. Post-tuning improvements were significant.
- SVM: Displayed robustness with both BoW and BERT, efficiently handling high-dimensional data.
- XGBoost: Lagged behind the other models, especially with BERT embeddings, indicating less suitability for complex NLP tasks.

Quantitative Results

Our experimental findings include the precision results for the models with each feature set, before and after hyperparameter tuning. For instance, the Logistic Regression model using BERT embeddings improved from a pre-tuning accuracy of 67.6% to a post-tuning accuracy of 68.11%. Similarly, SVM with BoW embeddings witnessed an increase in accuracy from 61.24% to 66.23% after tuning.

Final Model Selection

Upon the evaluation of the test set, Logistic Regression and SVM with BERT embeddings emerged as the most effective combinations for the task at hand. They demonstrated enhanced precision in correctly identifying stress indicators, with SVM slightly edging out in precision metrics post-tuning. Ultimately, the quantitative results suggest that strategic combinations of NLP embeddings and machine learning models significantly enhance the detection of mental stress indicators in textual data.

6. Conclusion

This coding exercise presents a comprehensive exploration of text classification, leveraging various word embedding techniques—ELMo, BERT, and Bag of Words (BoW)—in conjunction with three machine learning models: Logistic Regression, SVM, and XGBoost. Through meticulous experimentation and evaluation, the exercise reveals insightful findings on the effectiveness of combining different text representation methods with machine learning algorithms for natural language processing tasks. Here's an overall conclusion based on the experiments conducted with different embeddings and models:

Word Embedding Techniques

ELMo provided deep contextualized word representations, significantly capturing the nuances in word meanings based on their context. Its performance, however, was somewhat limited by computational demands and the complexity of integrating these embeddings with traditional machine learning models. BERT emerged as the most advanced embedding technique among the three, offering rich, context-aware embeddings that capture a broad

range of syntactic and semantic nuances. Models utilizing BERT embeddings generally outperformed those with other types of embeddings, highlighting the superiority of transformer-based models in understanding complex textual data. BoW, while the simplest form of text representation among the three, offered surprisingly competitive results. Despite its inability to capture word order and semantic context, the high-dimensional sparse vectors generated by BoW were effectively utilized by the machine learning models, particularly SVM, to achieve noteworthy classification performance.

Machine Learning Models

Logistic Regression was versatile across embeddings, showing a strong ability to leverage the nuanced representations provided by BERT, achieving the best results when tuned properly. Its performance highlights the efficacy of linear models in high-dimensional spaces, especially when paired with powerful embeddings. SVM consistently performed well across all embedding techniques, especially with BoW and BERT embeddings. Its ability to manage high-dimensional data and find optimal hyperplanes for classification tasks was evident, making it the best performer in several scenarios, particularly with BoW embeddings. XGBoost, while a potent algorithm for various machine learning tasks, did not lead in performance for text classification in this exercise. Despite its advanced ensemble learning techniques, XGBoost trailed behind Logistic Regression and SVM, particularly with BERT embeddings. This outcome suggests that while gradient boosting machines are powerful for tabular data, they may not always be the first choice for text data, especially when complex embeddings are involved.

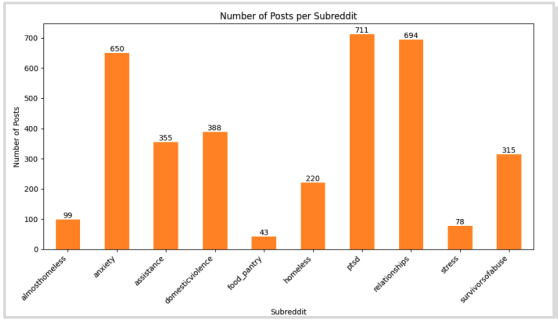
The coding exercise underscores the critical importance of choosing the right combination of word embeddings and machine learning models for text classification tasks. BERT embeddings combined with Logistic Regression and SVM models provided the most promising results, highlighting the strength of transformer-based embeddings in enhancing model performance. The success of BoW with SVM also serves as a reminder that simpler models can still be highly effective, particularly when computational resources are limited or when working with less context-dependent data.

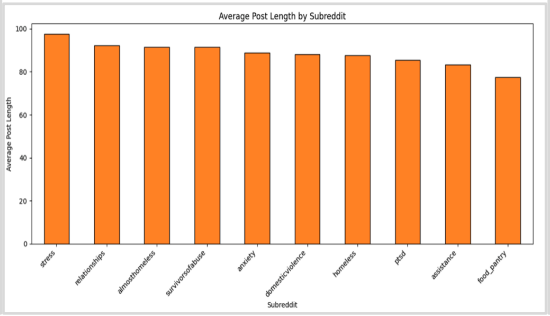
7. Discussion

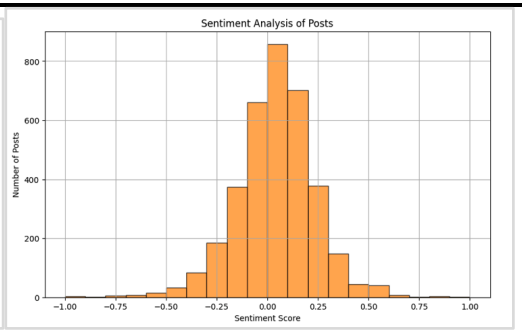
Our study underscores the complexity of mental health diagnostics and the promise of technology-enhanced solutions.Specifically, it has demonstrated the potential of combining NLP and ML techniques to improve the objectivity of mental health diagnoses by analyzing linguistic patterns in subreddit posts. Our findings reveal that word embedding techniques, particularly BERT, significantly enhance the ability of machine learning models to classify text accurately.

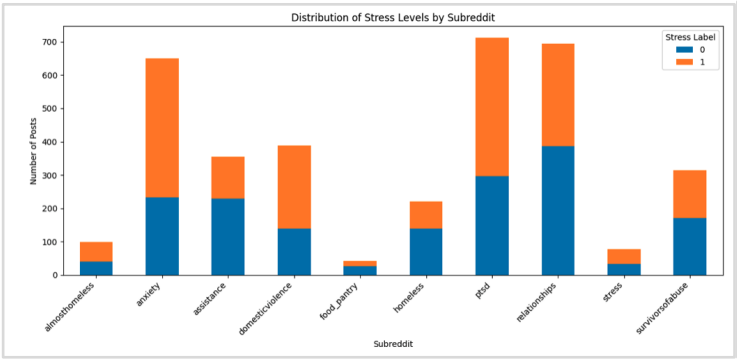
Future research could take a novel approach by investigating factors that differentiate between closely related conditions, leveraging ML and NLP to identify linguistic markers, themes, and combinations thereof that distinguish between conditions that share common symptoms but have distinct diagnostic criteria and treatment approaches. Future research may also involve investigating whether linguistic markers can differentiate between subtypes or severity levels within specific mental health conditions (e.g., linguistic features associated with major depressive disorder versus persistent depressive disorder). This type of analysis may assist clinicians in distinguishing between closely related conditions. Finally, as we advance, it will be essential to refine these methods and validate them in clinical settings, paving the way for more accurate and accessible mental health care tools.

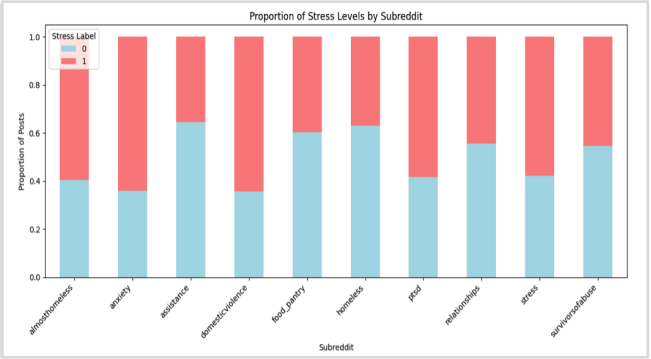
Appendix A: Summary Statistics

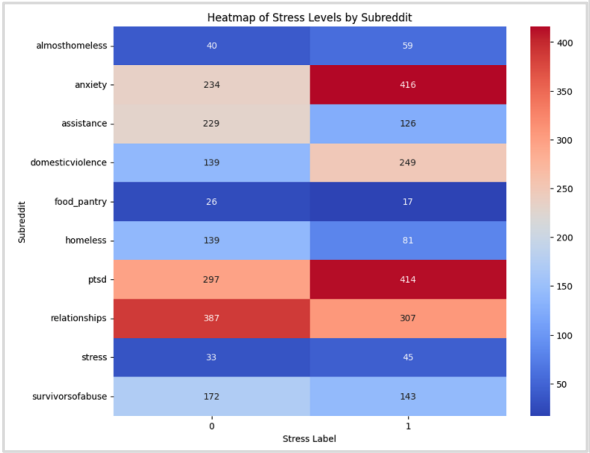


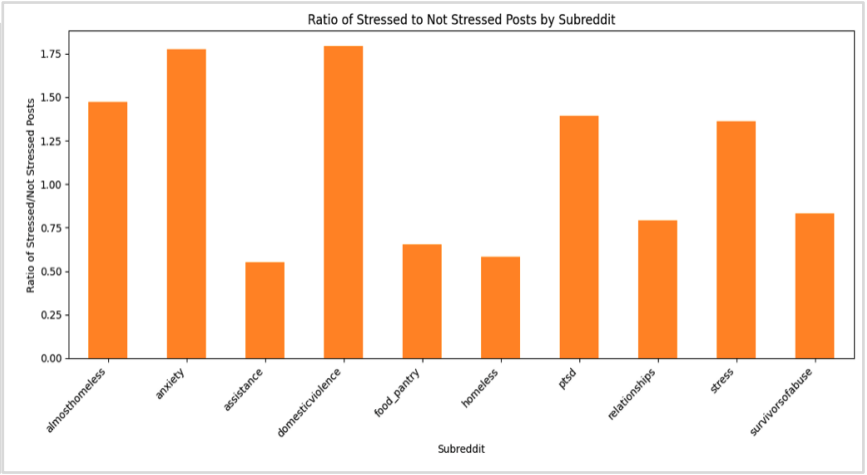












Appendix B: Precision Results by Model and Embedding Type

ELMo. The precision results for the models with ELMo feature before and after hyperparameter tuning are as follows:

Model	Metric	Pre-Tuning	Post-Tuning
Logistic Regression	Accuracy	0.6030	0.6135
	Recall (Macro-average)	0.6008	0.6082
	Recall (Micro-average)	0.6030	0.6135
	Precision (Macro-average)	0.6031	0.6099
	Precision (Micro-average)	0.6030	0.6135
	F1 Score (Macro-average)	0.5998	0.6084
	F1 Score (Micro-average)	0.6030	0.6135
	Average Precision	0.6139	0.6502
SVM	Accuracy	0.6273	0.6229
	Recall (Macro-average)	0.6238	0.6172
	Recall (Micro-average)	0.6273	0.6229
	Precision (Macro-average)	0.6315	0.6195
	Precision (Micro-average)	0.6273	0.6229
	F1 Score (Macro-average)	0.6202	0.6174
	F1 Score (Micro-average)	0.6273	0.6229
	Average Precision	0.6391	0.6438
XGBoost	Accuracy	0.5768	0.5835

	Recall (Macro-average)	0.5751	0.5786
	Recall (Micro-average)	0.5768	0.5835
	Precision (Macro-average)	0.5761	0.5795
	Precision (Micro-average)	0.5768	0.5835
	F1 Score (Macro-average)	0.5744	0.5787
	F1 Score (Micro-average)	0.5768	0.5835
	Average Precision	0.6141	0.6104

The tuned models were re-evaluated on the test set to finalize the selection based on their performance metrics. Here, precision results highlight the models' ability to identify positive instances among correctly predicted positives, with the SVM model showing a slight edge in post-tuning precision scores.

BoW. The precision results for the models with BoW before and after hyperparameter tuning, alongside other metrics, are as follows:

Model	Metric	Pre-Tuning	Post-Tuning
Logistic Regression	Accuracy	0.6180	0.6417
	Recall (Macro-average)	0.6177	0.6355
	Recall (Micro-average)	0.6180	0.6417
	Precision (Macro-average)	0.6177	0.6388
	Precision (Micro-average)	0.6180	0.6417
	F1 Score (Macro-average)	0.6177	0.6357
	F1 Score (Micro-average)	0.6180	0.6417
	Average Precision	0.6820	0.7327
SVM	Accuracy	0.6124	0.6623
	Recall (Macro-average)	0.6106	0.6598
	Recall (Micro-average)	0.6124	0.6623

	Precision (Macro-average)	0.6122	0.6600
	Precision (Micro-average)	0.6124	0.6623
	F1 Score (Macro-average)	0.6101	0.6599
	F1 Score (Micro-average)	0.6124	0.6623
	Average Precision	0.6766	0.7321
XGBoost	Accuracy	0.6180	0.5966
	Recall (Macro-average)	0.6208	0.6072
	Recall (Micro-average)	0.6180	0.5966
	Precision (Macro-average)	0.6247	0.6134
	Precision (Micro-average)	0.6180	0.5966
	F1 Score (Macro-average)	0.6158	0.5959
	F1 Score (Micro-average)	0.6180	0.5966
	Average Precision	0.6696	0.6737

Transitioning to BoW for word embedding, this iteration of the pipeline presents a conventional approach to text representation within an NLP and classification task. Despite its simplicity relative to deep learning methods like ELMo, BoW proves effective, demonstrated by the improvement in model performances, especially after hyperparameter tuning. The final evaluation on the test set reveals that the SVM model, with RBF kernel and C parameter set to 10, outperforms the others, achieving the highest accuracy and precision scores. This underscores the SVM's capability to manage high-dimensional data efficiently, a common scenario when using BoW embeddings.

BERT. The precision results for the models with BERT before and after hyperparameter tuning are as follows:

Model	Metric	Pre-Tuning	Post-Tuning
Logistic Regression	Accuracy	0.6760	0.6811
	Recall (Macro-average)	0.6748	0.6774
	Recall (Micro-average)	0.6760	0.6811

	Precision (Macro-average)	0.6762	0.6788
	Precision (Micro-average)	0.6760	0.6811
	F1 Score (Macro-average)	0.6748	0.6779
	F1 Score (Micro-average)	0.6760	0.6811
	Average Precision	0.7098	0.7741
SVM	Accuracy	0.6704	0.6792
	Recall (Macro-average)	0.6676	0.6736
	Recall (Micro-average)	0.6704	0.6792
	Precision (Macro-average)	0.6741	0.6774
	Precision (Micro-average)	0.6704	0.6792
	F1 Score (Macro-average)	0.6662	0.6742
	F1 Score (Micro-average)	0.6704	0.6792
	Average Precision	0.7080	0.7860
XGBoost	Accuracy	0.6423	0.6473
	Recall (Macro-average)	0.6407	0.6425
	Recall (Micro-average)	0.6423	0.6473
	Precision (Macro-average)	0.6425	0.6445
	Precision (Micro-average)	0.6423	0.6473
	F1 Score (Macro-average)	0.6404	0.6429
	F1 Score (Micro-average)	0.6423	0.6473
	Average Precision	0.6948	0.7557

References

[1] Jutel, Annemarie. “Sociology of diagnosis: a preliminary review.” Sociology of health & illness vol. 31,2 (2009): 278-99.

[2] Aboraya A, Rankin E, France C, El-Missiry A, John C. The Reliability of psychiatric diagnosis revisited: the clinician's guide to improve the reliability of psychiatric diagnosis. Psychiatry (Edgmont). 2006;3(1):41-50.

[3] Alarcón, Renato D. “Culture, cultural factors and psychiatric diagnosis: review and projections.” World psychiatry : official journal of the World Psychiatric Association (WPA) vol. 8,3 (2009): 131-9.

[4] Anglin, Deidre M, and Dolores Malaspina. “Ethnicity effects on clinical diagnoses compared to best-estimate research diagnoses in patients with psychosis: a retrospective medical chart review.” The Journal of clinical psychiatry vol. 69,6 (2008): 941-5.

[5] Trierweiler, S. J., Muroff, J. R., Jackson, J. S., Neighbors, H. W., & Munday, C. Clinician Race, Situational Attributions, and Diagnoses of Mood Versus Schizophrenia Disorders. Cultural Diversity and Ethnic Minority Psychology, 2005;11(4), 351–364.

[6] Trierweiler, S J et al. “Clinician attributions associated with the diagnosis of schizophrenia in African American and non-African American patients.” Journal of consulting and clinical psychology. 2000;vol. 68,1:

[7] Swaminathan A, López I, Mar RAG, Heist T, McClintock T, Caoili K, Grace M, Rubashkin M, Boggs MN, Chen JH, Gevaert O, Mou D, Nock MK. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. npj Digit Med. 2023;(6):213.

[8] Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, Roberts A, Dobson RJ, Stewart R. Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project. BMJ Open. 2017;7(1):e012012.

[9] Inamdar S, Chapekar R, Gite S, Pradhan B. Machine learning driven mental stress detection on Reddit posts using natural language processing. Hum-Cent Intell Syst. 2023;(3):80–91.

[10] Borah T, Ganesh Kumar S. Application of NLP and machine learning for mental health improvement. In: Gupta D, Khanna A, Hassanien AE, Anand S, Jaiswal A, editors. International Conference on Innovative Computing and Communications. 2022;492:219-228.

[11] D'Alfonso S. AI in mental health. Curr Opin Psychol. 2020;36:112-117.

[12] Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. Trans Assoc Comput Linguist. 2016;4:463-476.

[13] DeSouza DD, Robin J, Gumus M, Yeung A. Natural language processing as an emerging tool to detect late-life depression. Front Psychiatry. 2021;(12):719125.

[14] Malhotra G, Waheed A, Srivastava A, Akhtar MS, Chakraborty T. Speaker and time-aware joint contextual learning for dialogue-act classification in counseling conversations. 2021:*ArXiv.*

[15] Weir K. The roots of mental illness. How much of mental illness can the biology of the brain explain? Monitor on Psychology. 2012;43(6):30