COMP 652 Final Project
Sentiment Analysis in Mental Health

Diksha Bidikar, Katie Rizik, Laura Swafford, Danny Yoon, Lajvi Bhavsar

***Abstract*—The power of Large Language Models (LLMs) are being introduced across all sectors of life, one of which is the healthcare industry. This is an incredibly high-stakes incorporation of models due to the potential life altering outcomes of an incorrect diagnosis. This study is a comprehensive review of five different models trained on patient symptom data to predict a mental health diagnosis. Among the models tested, the best performing model with an accuracy of 80.7% and an AUROC (Area under Receiver Operating Characteristic) of 98% was BioGPT, which leverages GPT-2 architecture trained specifically on biomedical data [1].**

Domain-specific models, such as MentalBERT and BioGPT, are found to outperform general-purpose models like fastText and ALBERT, with notable improvements in their understanding of mental health terminology and colloquial expressions. In contrast, XLNet, while its performance was similar to MentalBERT, significant training time was required to produce the model. This review highlights the effectiveness of pre-trained domain-adapted models for mental health NLP tasks, focusing specifically on their performance in terms of accuracy, AUROC score, and ROC curve analysis. It also discusses the trade-offs observed between these evaluation metrics across different models.

***Index Terms*—Large Language Models, Natural Language Processing**

I. Introduction

In recent years with the rapid evolution of generative Artificial Intelligence (AI), health care workers are beginning to incorporate more Large Language Models (LLMs) into their daily practices. One common way is through AI note taking. Health care workers spend about 35% of their time documenting patient details [2], which can take time away from treating patients. A key benefit of LLMs to reducing this workload is through AI note taking to help triage patients. This can be especially beneficial for healthcare workers working in psychiatry, as it can provide objectivity and consistency when diagnosing patients that the field currently struggles with [3]. There is increasing interest in this practice as well. There has been consistent growth in the number of articles using Machine Learning (ML) and Deep Learning (DL) to help with diagnosing mental illness [4].

The models tested today are aimed at helping tackle this problem by reviewing over 50,000 patient conversations and classifying them as one of the following diagnoses [5]:

1) Anxiety
2) Normal
3) Depression
4) Suicidal
5) Stress
6) Bipolar
7) Personality Disorder

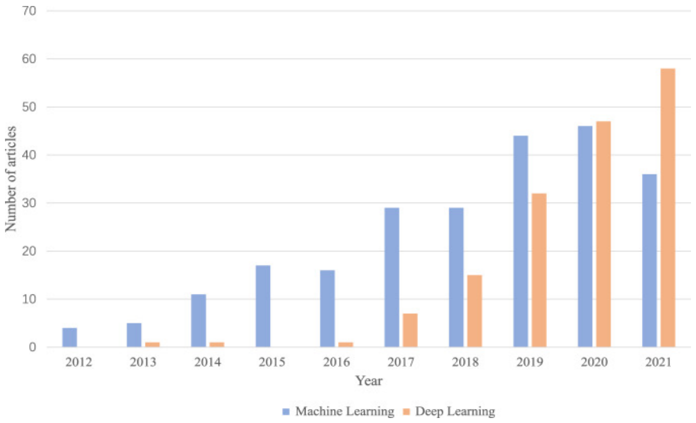
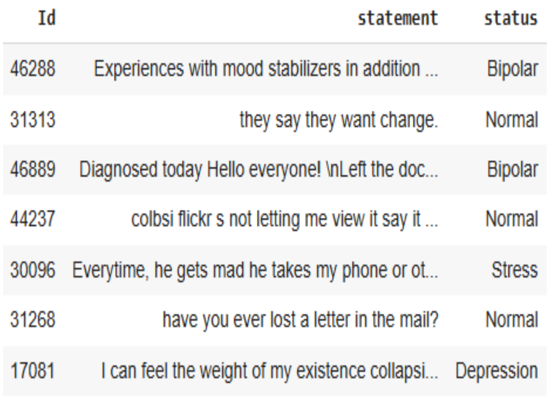


Fig. 1. NLP trends applied to mental illness detection research using machine learning and deep learning. [4]



Id	statement	status
46288	Experiences with mood stabilizers in addition ...	Bipolar
31313	they say they want change.	Normal
46889	Diagnosed today Hello everyone! \nLeft the doc...	Bipolar
44237	colbsi flickr s not letting me view it say it ...	Normal
30096	Everytime, he gets mad he takes my phone or ot...	Stress
31268	have you ever lost a letter in the mail?	Normal
17081	I can feel the weight of my existence collapsi...	Depression

Fig. 2. Example patient statements and diagnosis

This dataset contains text data from various mental health-related conversations, including social media posts and forum discussions. Figure 2 shows examples. The majority of the data belongs to ”Normal”, ”Depression” and ”Suicidal” (as shown in Figure 3), which will be good benchmarks for seeing how well the data classifies for lesser represented diagnoses like ”Stress” and ”Personality Disorder”.

This dataset will be tested across:

1) BioGPT(Generative Pre-trained Transformer for Biomedical Text Generation and Mining) [1]
2) ALBERT (A Lite BERT (Bidirectional Encoder Representations from Transformers)) [6]
3) MentalBERT (Pre-trained BERT model) [7]

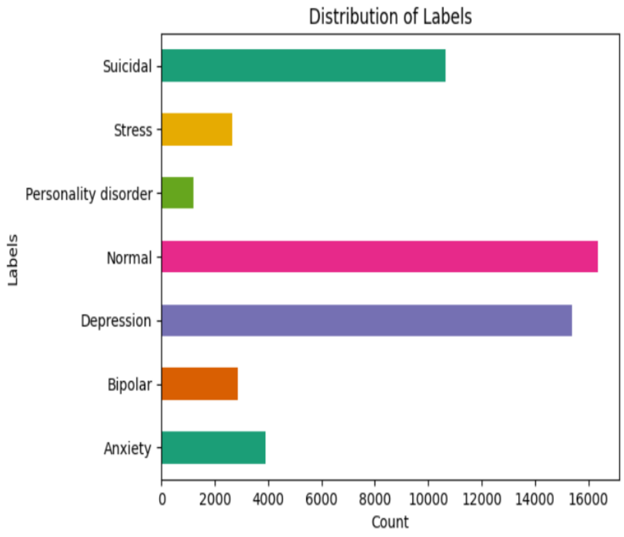


Fig. 3. Distribution of Diagnosis Labels

4) fastText [8]
5) XLNET (Generalized Autoregressive Pretraining for Language Understanding) [9]

and evaluated based on accuracy, one-to-many Area Under the Receiver Operating Characteristic (AUROC), and a confusion matrix to compare which model is best at diagnosing based on patients' symptom claims.

II. RESEARCH METHODOLOGY

A. BioGPT (Generative Pre-trained Transformer for Biomedical Text Generation and Mining)

BioGPT is a model created off of GPT-2 architecture that was created specifically for text generation using the causal language model objective. This model was chosen for this study given the clinical nature of the training data to diagnose mental illnesses.

BioGPT was trained on a corpus of fifteen million PubMed abstracts and generated a vocabulary of 42,384 words [1]. It differs from other LLMs based on biomedical text like BioBERT and PubMedBERT due to using the GPT-2 architecture and is pre-trained based on a causal language modeling. This gives it the ability to generate more coherent text related to the biomedical field. BioGPT was originally applied to 6 different natural language processing (NLP) tasks and outperformed other LLMs, one of these primary tasks being document classification [1]. For fine-tuning the model for document classification tasks, researchers generate a target sequence using the format "the type of this document is label"[1]. When testing document classification against similar models like BioBERT, PubMedBERT, and others, it out performed all with an F1 score of 85% [1]. This indicates that it will likely apply well to the diagnostic goal of this study.

For document classification, BioGptTokenizer was used to tokenize the input sentences and BioGptForSequenceClassification for the actual classification task. To stay within the memory parameters, the data was truncated to account for the length of 90% of the training data, which is 278 words per sequence. Accuracy was then used as the best model metric, and early stopping was added based on validation loss to prevent overfitting. This took 6 epochs to train with a validation accuracy of 80.4%.

In addition to accuracy, AUROC and a confusion matrix were used to measure this multi-class classification model. This model had an overall test accuracy of 80.7% and an average one-vs-rest AUROC score of 98%. BioGPT performed very well on the test data across all labels looking at the confusion matrix:

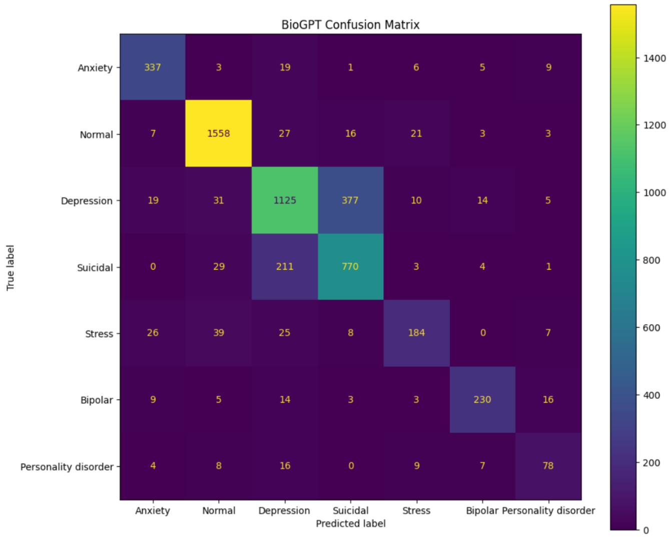


Fig. 4. BioGPT Confusion Matrix

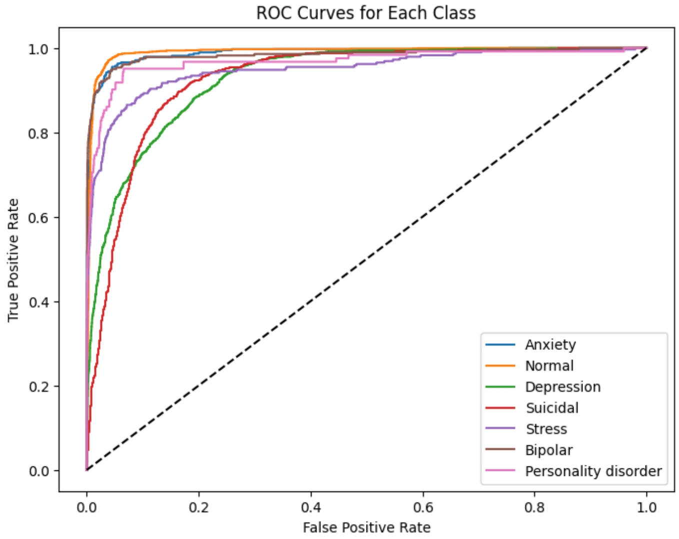


Fig. 5. BioGPT ROC CURVE

Classifying depression vs. suicidal proved to be the most difficult. There is understandable ambiguity between these two classifications coming at it from a human perspective, but this is a very high stakes classification for a model to get wrong.

This should be a key consideration when comparing a model's effectiveness for this type of dataset.

B. ALBERT (A Lite BERT)

ALBERT is a lighter, faster, and more memory-efficient version of Bidirectional Encoder Representations from Transformers (BERT). In order, to improve the training and results of BERT architecture by using different techniques, ALBERT was developed. ALBERT was proposed by researchers at Google Research in 2019. The backbone of ALBERT architecture is similar to BERT, where both have encoder layers with Gaussian Error Linear Unit (GELU) activation function. ALBERT is released in 4 different model sizes (base, large, large, xxlarge). There are three main contributions that ALBERT makes over the design choices of BERT:

1) Factorized Parameter Embedding
2) Cross-layer Parameter Sharing
3) Sentence Order Prediction

Factorization of the Embedding matrix:

In the BERT model, the input layer embeddings and hidden layer embeddings are the same size. However, in the ALBERT model, the two embedding matrices are separated. The input-level embedding (E) needs to refine only context-independent learning but hidden level embedding (H) requires context-dependent learning. This helps in reduction in parameters by 80% with a minor drop in performance when compared to BERT

Cross-layer Parameter Sharing:

To improve efficiency and reduce redundancy, the parameter sharing between layers was introduced. The three types of parameter sharing are:

- Only share Feed Forward network (FFN) parameter
- Only share attention parameters
- Share all parameters.

The default approach for ALBERT is to share all parameters across layers which leads to a 70% reduction in the overall number of parameters.

Sentence Order Prediction:

The BERT model uses Next Sentence Prediction (NSP) loss. NSP is a binary classification loss for predicting whether two segments appear consecutively in the original text, whereas ALBERT uses Sentence Order Prediction (SOP). It only looks for sentence coherence. Despite the much fewer number of parameters, ALBERT has achieved the state-of-the-art of many NLP tasks [6].

Given the sensitive and nuanced nature of mental health sentiment analysis, selecting a model with strong language understanding capabilities was critical. ALBERT offers a compelling balance between performance and computational efficiency, allowing for effective fine-tuning, even on moderately sized datasets. Its parameter-sharing mechanisms reduce the risk of overfitting, which is particularly important in mental health-related text datasets where class distributions may be imbalanced or data may be scarce. Additionally, ALBERT's ability to capture inter-sentence coherence enhances its suitability for analyzing mental health posts, which often involve complex emotional expressions and context-dependent sentiments.

Thus, ALBERT was chosen for this work due to its efficiency, robustness, and proven success across diverse NLP benchmarks.

The performance of the ALBERT model on the mental health classification task, as illustrated by its confusion matrix in Figure 6, shows significant limitations in distinguishing between several classes. Although the model achieves a test accuracy of 52.88% and an AUROC score of 84.04%, these metrics are not fully representative of its true effectiveness, particularly due to the imbalanced predictions observed. The model performs relatively well in identifying the 'Normal' and 'Depression' classes, correctly classifying a large number of instances in these categories. However, it struggles considerably with the remaining classes—Anxiety, Suicidal, Stress, Bipolar, and Personality Disorder—which are frequently misclassified as 'Depression'. Notably, none of the instances from Anxiety, Stress, Bipolar, or Personality Disorder were correctly predicted, indicating that the model is highly biased toward the more dominant or frequent classes in the dataset. This suggests that ALBERT, in its current form, lacks the nuance needed to differentiate between similar mental health categories, which could stem from class imbalance or insufficient domain-specific training. Although the AUROC score appears high, it is likely influenced by the model's good performance on the majority classes and does not reflect its poor performance on underrepresented categories.

The ROC curve analysis in Figure 7 confirms that the ALBERT model is biased toward the more frequently occurring classes, particularly 'Normal'. It is less capable of capturing the nuanced distinctions needed for minority mental health categories like Anxiety and Personality disorder.

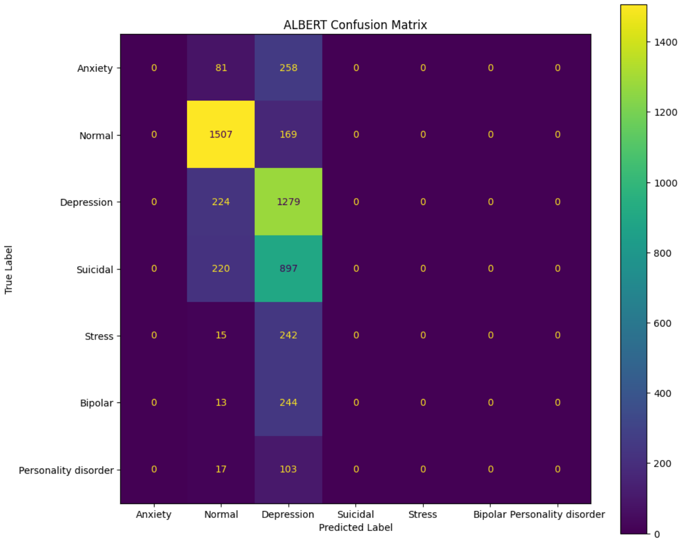


Fig. 6. ALBERT Confusion Matrix

C. XLNet (Extra Long Network)

XLNet is a deep learning transformer model that aims to test the boundaries of natural language understanding tasks. XLNet

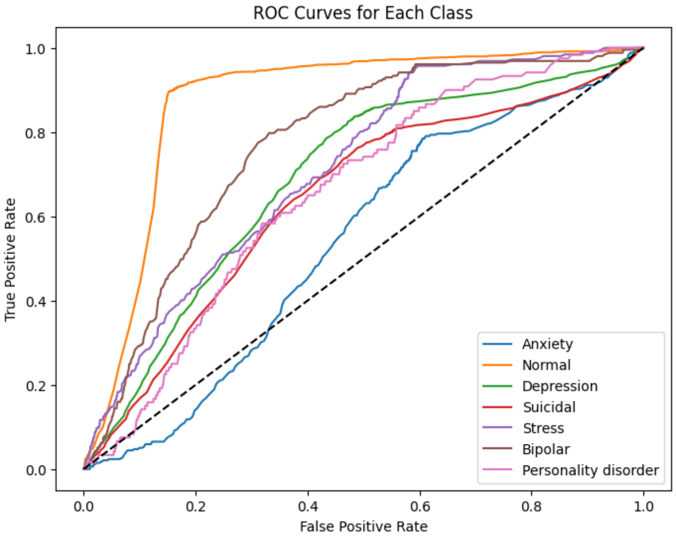


Fig. 7. Albert ROC CURVE

is based on a permutation-based training objective, where it can learn bidirectional context without predicting masked tokens. This assists XLNet in learning word dependencies more effectively while retaining the benefits of autoregressive generation. Following the Transformer-XL model, XLNet also brings recurrence mechanisms that enable it to handle longer sequences compared to earlier models like BERT. Due to this, XLNet achieved state-of-the-art performance for a wide range of natural language processing tasks [9]

XLNet is used to fine-tune for the sentiment classification model. To manage computational resources effectively while maintaining model performance, a subset of 3000 examples from the larger dataset are used. The input sequences were tokenized with a maximum sequence length of 64 tokens, and the model was trained over eight epochs using a batch size of eight and the Adam optimizer with a learning rate of 2e-5. The total training time was approximately 997.61 seconds.

During training, the model achieved a steady increase in training accuracy, starting at 66.12% after the first epoch and reaching 95.11% by the eighth epoch. Validation accuracy remained relatively stable across epochs, averaging around 71%. On the held-out test set of 600 examples, the final test accuracy was recorded at approximately 76%. The classification report indicated strong performance for common classes such as ”Normal” (precision 94%, recall 85%) and ”Depression” (precision 72%, recall 71%), while minority classes like ”Personality disorder” showed relatively lower performance due to limited support. The model achieved a weighted average F1-score of 0.76 across all classes, and the one-to-many AUROC score was calculated to be approximately 94.94%.

To better understand the model’s prediction behavior across categories, a confusion matrix was generated and is presented in Figure 8. These results demonstrate that XLNet can be effectively adapted to domain-specific tasks like medical sentiment analysis.

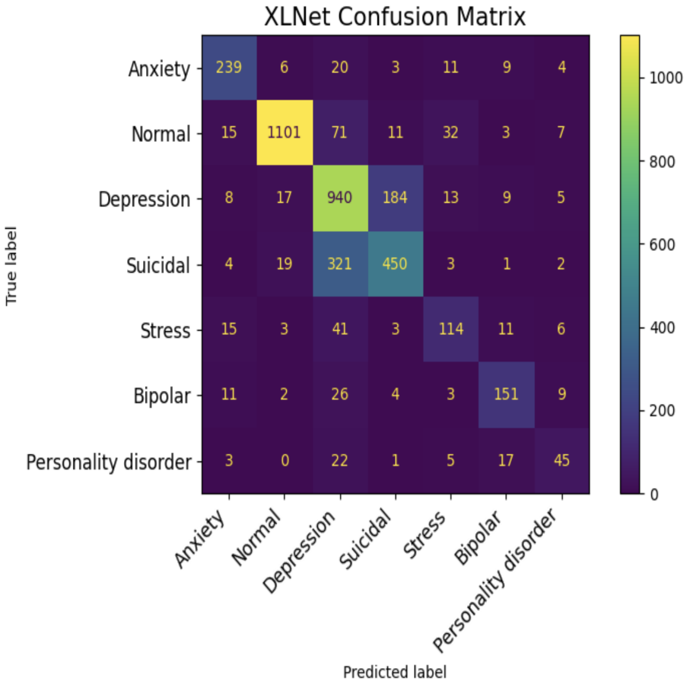


Fig. 8. XLNet Confusion Matrix

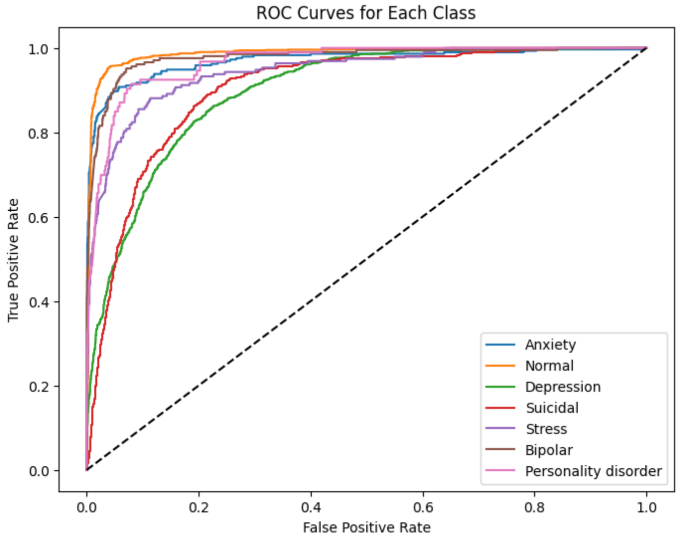


Fig. 9. XLNet ROC CURVE

D. *MentalBERT*

MentalBERT is an NLP model trained on mental health data from social forums[7]. The goal was demonstrating NLP classification on mental health disorders and creating a novel, publicly available mental health LLM, which the model creators claim was not publicly available when the model was published. They updated the model base, BERT-Base (uncased_L-12_H-768_A-12)[10], using posts on social forums from specific Reddit and Twitter communities related to mental health such as ”r/depression” [7]. Eight datasets with over 56,000 training samples, over 10,000 validation samples, and over 14,000 test samples were used to train and test the model [11], [12], [13], [14], [15], [16], [17].

MentalBERT was chosen for this study due to its unique pretraining on mental health data. For this project, the input data was tokenized using MentalBERT and truncated to the first 50 words in each entry. Data truncation was applied to improve the model fitting time. Then MentalBERT model was re-trained on 10,000 samples and validated on 7,903 samples from the patient conversations dataset[5] for for sentiment analysis categorizing text as anxiety, depression, suicidal, stress, bipolar, personality disorder, and normal. These categories have overlap with the original categories used to train MentalBERT but also include personality disorder and normal.

The MentalBERT confusion matrix (Figure 10) and the MentalBERT ROC curve (Figure 11) show the model results of the test data. Although the accuracy of the test results is 72.6%, the confusion matrix and the ROC curves show the model performs better for some categories compared to others. The normal category has the highest precision and recall, 93.2% and 90.2%, respectively. Anxiety has the second highest precision (76.4%), and bipolar has the second highest recall (82.4%). The model performed significantly worse on suicidal (precision 64.1%, recall 47.8%), depression (precision 63.6%, recall 59.0%), personality disorder (precision 56.1%, recall 56.1%), and stress (precision 50.9%, recall 42.2%) categories. As shown in 10, the model has a tendency to misidentify depression as suicidal, suicidal as depression, and personality disorder as depression.

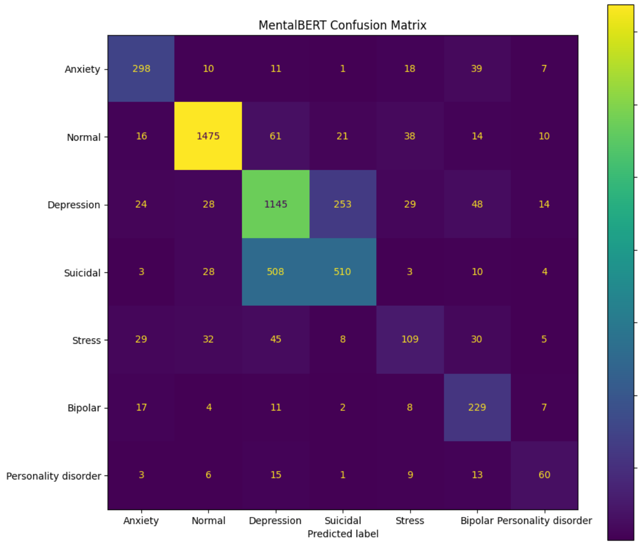


Fig. 10. MentalBERT Confusion Matrix

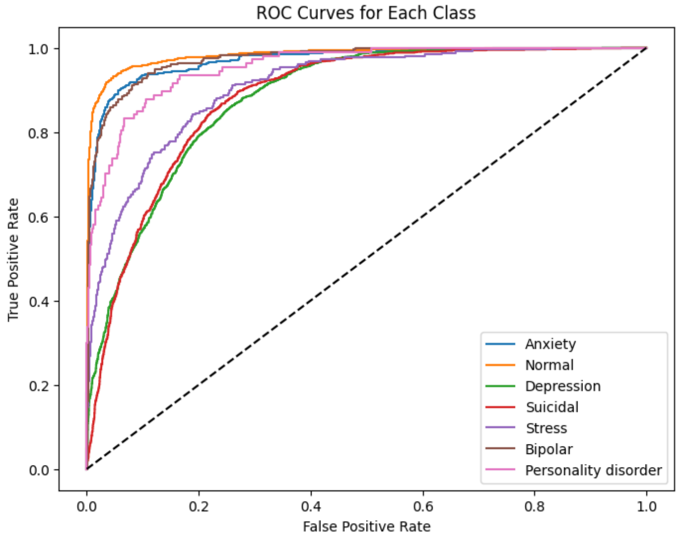


Fig. 11. MentalBERT ROC CURVE

E. fastText

fastText is a lightweight model developed by the Facebook AI team. fastText is powerful because it is as effective as deep learning classifiers, and it can be run on machines using a standard CPU. It can classify half a million sentences for up to 312K classes in less than a minute [18]. While most deep learning models are powerful and can achieve favorable results, they require a lot of time and computing resources. Simpler linear models, on the other hand, can also be very powerful if the right parameters are selected and scaled correctly; hence, the purpose behind fastText. Prior work on fastText was completed with eight datasets used from Zhang [18], where multiple methods were used for result comparisons as a baseline: N-grams, character level convolutional model (char-CNN), character based convolution recurrent network (char-CRNN), and very deep convolutional network (VDCNN) [18] [19] [20]. Results found that using bi-grams improved the performance by 1-4 percent, and the accuracy was better than char-CNN and char-CRNN and worse than VDCNN. Notably, char-CNN and VDCNN use NVIDIA Tesla K40 GPU, and the fastText model uses a simpler CPU with 20 threads. Models using convolutions are multitudes slower than the fastText model [8].

For the sentiment analysis, the preprocessed sentences of the data are passed into the fastText model as a text file; the only way data can be passed into the fastText model. In the model itself, it represents these sentences as a bag of words and trains a linear classifier with logistic regression or support vector machine [8]. A nuance to the model is the y value has to prepended to the training sentence. Since the purpose of the fastText model is to be lightweight and fast, it is unable to provide an embedding layer for training with weights. Instead, K folds is used to discretely capture results across eighteen hyper parameter configurations. Three key parameters are used in these configurations: learning rate, wordNgrams, loss function.

After training eighteen models with different configurations and capturing their accuracy scores, the validation model with the highest accuracy yielded a 99.49% accuracy using the following hyperparameter configurations: learning rate=1.0, wordNgram=3, and loss function= ”hierarchical softmax”. The hierarchical softmax version is advantageous because nodes and parent nodes’ probabilities allow the depth first search method and maximum probability to discard branches with low probabilities[21] [22]. With the test data and optimal hyperparameter configurations, fastText yielded a mediocre accuracy of 76%. The confusion matrix showed significant misclassifications between “Suicidal” and “Depression” which makes sense given these sentiments go in tandem. Otherwise, the average of the one vs rest - AUROC score shows a good score of 85.94%, proving the data generally classifies the sentiment well.

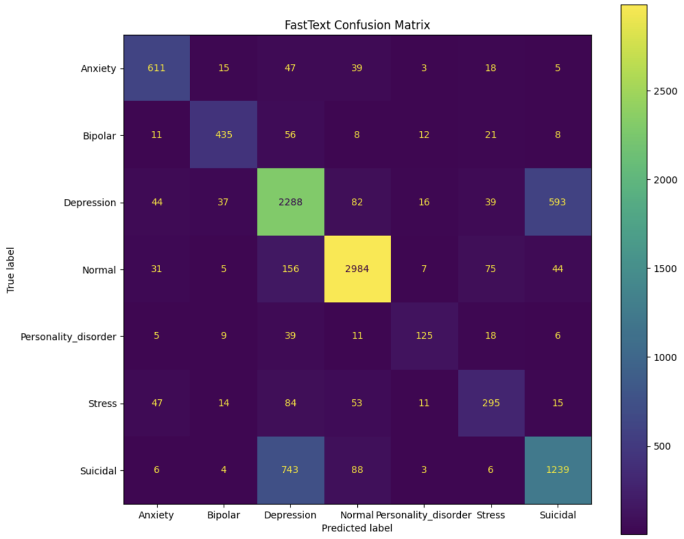


Fig. 12. fastText Confusion Matrix

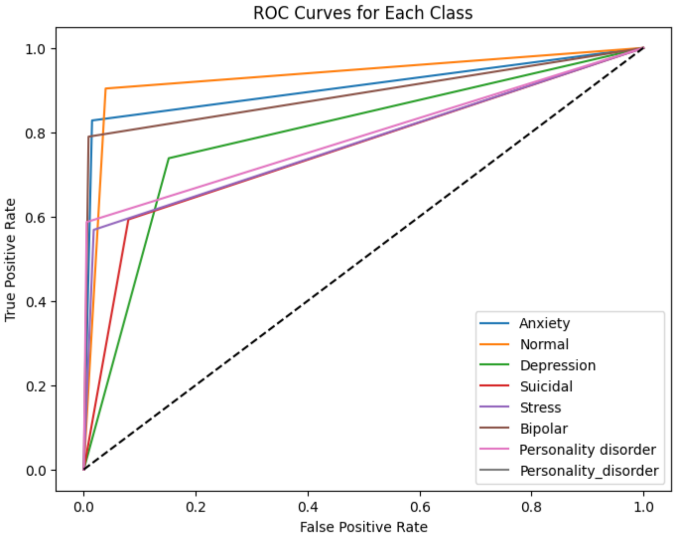


Fig. 13. fastText ROC CURVE

III. Results - Model Comparison

In order to compare the performance of the five models, three key performance metrics are used. The accuracy, a confusion matrix, and the one-to-many AUROC score. Numeric results are represented in the figure below.

TABLE I
Comparison of Model Performance

Model	Accuracy	AUROC
BioGPT	80.71%	98.0%
ALBERT	52.88%	84.0%
MentalBERT	72.61%	95.0%
fastText	75.89%	85.9%
XLNet	76.00%	94.9%

The accuracy from testing yielded a range between 52.8% and 80.7% with BioGPT yielding the highest at 80.7% and ALBERT the lowest at 52.8%. In conjunction with AUROC scores, the models yielded high numbers between 84% and 98% with BioGPT yielding the highest at 98% and ALBERT the lowest at 84%. The higher AUROC scores indicate the models generalize well and do not overfit, particularly BioGPT. It is important to consider there are seven classifications with varying distributions for each classification, as seen in Figure 2.

A model classifying sentiment with 80.7% is a good indicator of how well it works. Conversely, shortcomings commonly found were misclassifications, highlighted in the confusion matrices.

Significant misclassifications occur between ”Depression” and ”Suicidal”, which makes sense given those sentiments are similar. Additionally, the data set used for this study may be considered small on a global scale. Only 51,074 patient conversations are used, and they’re also all in English. Despite being a generally good classifier, the data distribution is skewed towards Normal, Suicidal, and Depression. Consequently, the model may not perform well if additional data is introduced in other languages or more concentrated on less common sentiments.

IV. Conclusion

The best models for sentiment analysis use transformer models, such as BERT or GPT models, as their cornerstone. While fastText and ALBERT demonstrate moderate results, they lag behind the domain-specific models due to the absence of pretraining on mental health or biomedical data. This study found that using BioGPT and MentalBERT, both models that trained on biomedical or mental health datasets, perform well for diagnostic classification. Comparing MentalBERT to ALBERT, both BERT based models, the domain specificity of MentalBERT added significant increases to accuracy and AUROC.

The ability to use multiple BERT models as a hybrid and their transformers are particularly adept at classifying sentiment [23]. Sentiment analysis can be further improved by re-training models using domain specific datasets. However, models may produce varying accuracy, precision, and recall results across different classes in a domain. When applying an LLM to a specific problem, whether in mental health or another domain, prior model results should be considered before choosing a modeling for the problem.

References

[1] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 9 2022. [Online]. Available: https://doi.org/10.1093/bib/bbac409

[2] A. Rule, S. Bedrick, M. F. Chiang, and M. R. Hribar, "Length and redundancy of outpatient progress notes across a decade at an academic medical center," *JAMA Network Open*, vol. 4, no. 7, p. e2115334, 7 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8290305/

[3] R. L. Allesøe, W. K. Thompson, J. Bybjerg-Grauholm, D. M. Hougaard, M. Nordentoft, T. Werge, S. Rasmussen, and M. E. Benros, "Deep Learning for Cross-Diagnostic Prediction of Mental Disorder Diagnosis and Prognosis using Danish Nationwide register and Genetic data," *JAMA Psychiatry*, vol. 80, no. 2, p. 146, 12 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9857190/

[4] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *npj Digital Medicine*, vol. 5, no. 1, 4 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8993841/

[5] "Sentiment Analysis for Mental Health," 7 2024. [Online]. Available: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health/data

[6] M. Ryu, Z. Lan, M. Chen, S. Goodman, R. Soricut, and P. Sharma, "Albert: A lite bert for self-supervised learning of language representations," *arXiv.1909.11942*, 1 2021.

[7] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare," 2021. [Online]. Available: https://arxiv.org/abs/2110.15621

[8] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv*, vol. 3, 8 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1607.01759

[9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: https://arxiv.org/abs/1906.08237

[10] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962v2*, 2019.

[11] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, vol. 34, no. 13, p. 10309–10319, Jun. 2021. [Online]. Available: http://dx.doi.org/10.1007/s00521-021-06208-y

[12] D. Losada and F. Crestani, "A test collection for research on depression and language use," vol. 9822, 09 2016, pp. 28–39.

[13] I. Pirina and Çöltekin, "Identifying depression on reddit: The effect of training data," 01 2018, pp. 9–12.

[14] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "Clpsych 2015 shared task: Depression and ptsd on twitter," 01 2015, pp. 31–39.

[15] E. Turcan and K. McKeown, "Dreaddit: A reddit dataset for stress analysis in social media," 2019. [Online]. Available: https://arxiv.org/abs/1911.00133

[16] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, K. Loveys, K. Niederhoffer, E. Prud'hommeaux, R. Resnik, and P. Resnik, Eds. New Orleans, LA: Association for Computational Linguistics, Jun. 2018, pp. 25–36. [Online]. Available: https://aclanthology.org/W18-0603/

[17] M. Mauriello, E. Lincoln, G. Hon, D. Simon, D. Jurafsky, and P. Paredes, "Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems," 05 2021, pp. 1–7.

[18] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *NIPS*, 2015. [Online]. Available: https://arxiv.org/abs/1502.01710

[19] Y. Xiao and K. Cho, "Efficient character-level document classification by combining convolution and recurrent layers," *arXiv*, 2016. [Online]. Available: https://arxiv.org/abs/1602.00367

[20] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Efficient character-level document classification by combining convolution and recurrent layers," *arXiv*, 2016. [Online]. Available: https://arxiv.org/abs/1606.01781

[21] J. Goodman, "Classes for fast maximum entropy training," *arXiv*, 2001. [Online]. Available: https://arxiv.org/abs/cs/0108006

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space. arxiv preprint," *arXiv*, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[23] A. S. Talaat, "Sentiment analysis classification system using hybrid bert models," *Journal of Big Data*, vol. 10, 2023. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00781-w#:~:text=In%20general%2C%20RoBERTa%20has%20higher,3G%2C%20with%2091.72%25%20accuracy.