Assessing Suicide Risk in Social Media Posts Using NLP

Louis Wu, Branndon Marion, Shruti Jain

December 2024

1 Abstract

We present a highly effective way to identify suicide risk in social media posts using transformer encoders such as BERT, MentalBERT, and DisorBERT. We observe that MentalBERT outperforms the other two models in their native form and that all three models improve significantly after fine-tuning, reaching accuracies above 90 percent. To evaluate our models, we utilize metrics such as accuracy, precision, recall, and F1 score, ensuring a comprehensive assessment of their performance. Despite having high overall accuracy, our models struggle to distinguish between suicidal intent and general depression.

2 Introduction

Suicide is a major global public health issue, with over 700,000 people dying by suicide each year according to the World Health Organization (WHO, 2024). Beyond these fatalities, countless more individuals attempt suicide, highlighting the urgent need for early detection and intervention. Despite ongoing prevention efforts, suicide remains one of the leading causes of death worldwide, particularly among young people aged 15-29. Traditional methods of assessing suicide risk, such as clinical interviews and self-reported measures, are often limited by their reliance on individuals seeking help or openly expressing their thoughts. These limitations highlight the necessity of innovative methods to identify suicide risk earlier and more effectively.

Advances in natural language processing (NLP) offer new opportunities to analyze large-scale social media data, where individuals frequently share their mental health experiences. While existing NLP models have successfully identified broader mental health concerns, such as depression or anxiety, they often fail to distinguish between these disorders and suicide ideation. This gap is significant, as linguistic patterns for suicide ideation are nuanced and can overlap with those of other mental health issues. For instance, expressions of hopelessness or isolation are common in both depression and suicidal thoughts, making accurate classification a complex task.

Our work aims to address this challenge by leveraging transformer-based models, including BERT, MentalBERT, and DisorBERT, to improve the detection of suicide risk in social media posts. Building on prior research, we employ domain-specific models and fine-tuning techniques to enhance classification accuracy. Unlike earlier studies that focus broadly on mental health, our approach specifically targets the linguistic markers of suicide ideation, ensuring a more focused and actionable outcome. By using Reddit data from both pre and post-COVID periods, we evaluate our models' ability to adapt to temporal and contextual variations in language use.

Through this work, we aim to contribute to the growing field of computational mental health by developing models that not only improve upon traditional methods but also provide actionable insights for public health interventions. Our study seeks to bridge the gap between research and real-world applications, offering a scalable, data-driven solution to an urgent global problem.

3 Background

A team of researchers at Qntfy, a mental health predictive analytics agency, first attempted to detect suicide risk in social media posts using NLP (Coppersmith et al., 2018). The authors fed labeled data through a bidirectional LSTM to detect signals around suicide attempts. Although the results were considered state-of-the-art at the time of publication, the authors did not use attention-based transformers, which have since proven to be more effective at encoding text. This paper serves as our inspiration for using transformers to detect suicide risk.

Shortly after the onset of COVID-19, a team at the Department of Brain and Cognitive Sciences at MIT attempted to use NLP to measure the impact of the pandemic on various mental health subreddits (Low et al., 2020). They did not run any experi-

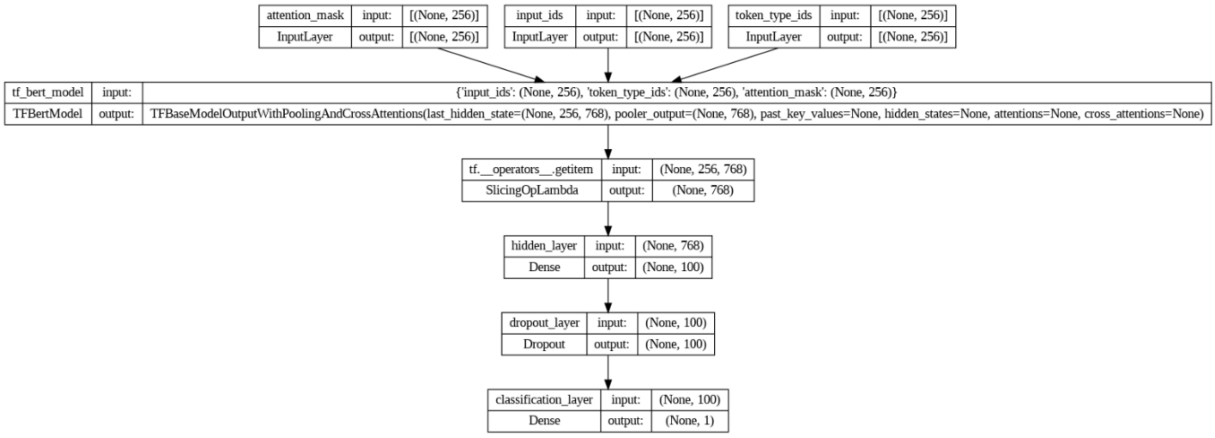

Figure 1: Architecture of the BERT-based classification model, featuring input layers, transformer layers, a hidden dense layer, dropout layer, and final classification layer with sigmoid activation.

ments focused on suicide detection, but they compiled a comprehensive labeled dataset of reddit posts which we use to train and fine-tune our models.

Researchers at Aalto University in Finland released a pre-trained masked language model, MentalBERT, to benefit machine learning research in the mental healthcare domain (Ji et al., 2022). Their model demonstrated significant improvement in the performance of mental health detection tasks. However, they did not perform any training or fine-tuning specific to suicide detection. We use their model, which is available on the Huggingface repository, to perform our task.

The following year, researchers at Universidade de Santiago de Compostela in Spain released a similar model, DisorBERT, to detect mental health disorders in social media (Aragon et al., 2023). Unlike MentalBERT, which is pre-trained, DisorBERT extends from BERT with two stages of domain adaptation through Reddit and through mental health documents, respectively. Their model is available on Huggingface and we use it for our task.

4 Methods

We use transformer models to classify Reddit posts for suicide risk. We use the labeled data provided by the MIT research team, which includes posts from 27 different subreddits: r/conspiracy, r/divorce, r/fitness, r/guns, r/jokes, r/legaladvice, r/meditation, r/parenting, r/personalfinance, r/relationships, r/teaching, r/mentalhealth, r/EDAnonymous, r/addiction, r/alcoholism, r/adhd, r/anxiety, r/autism, r/bipolarreddit, r/bpd, r/depression, r/healthanxiety, r/lonely, r/ptsd, r/schizophrenia, r/socialanxiety, and r/suicidewatch. We label posts from r/suicidewatch as positive and the remaining posts as negative. The posts in each subreddit are separated between pre-COVID and post-COVID periods, with the pre-COVID period covering December 2018 to December 2019 and the post-COVID period covering January 2020 to April 2020.

To create our training data, we randomly select 5,000 positive posts from r/suicidewatch and 5,000 negative posts from the other subreddits, all during the pre-COVID period, for a total of 10,000 posts. To create our test data, we randomly select 2,500 positive posts from r/suicidewatch and 2,500 negative posts from the other subreddits, all during the post-COVID period, for a total of 5,000 posts. This creates a balanced dataset for which we can establish 0.5 accuracy as the bare minimum baseline.

We use the BERT, MentalBERT, and DisorBERT transformers available on Huggingface to perform the NLP classification. For each transformer, we train a classifier with all the transformer layers frozen and a classifier with all the transformer layers tunable, for a total of six classification models. We tokenize the input text in accordance with the respective tokenizer for each transformer model, padding to a maximum sequence length of 256 tokens.

To build the full classification model, we feed the CLS token from the transformer output into a hidden layer of size 100, a dropout layer of 0.3, and a binary classification layer with a sigmoid activation function. The final classification model is compiled

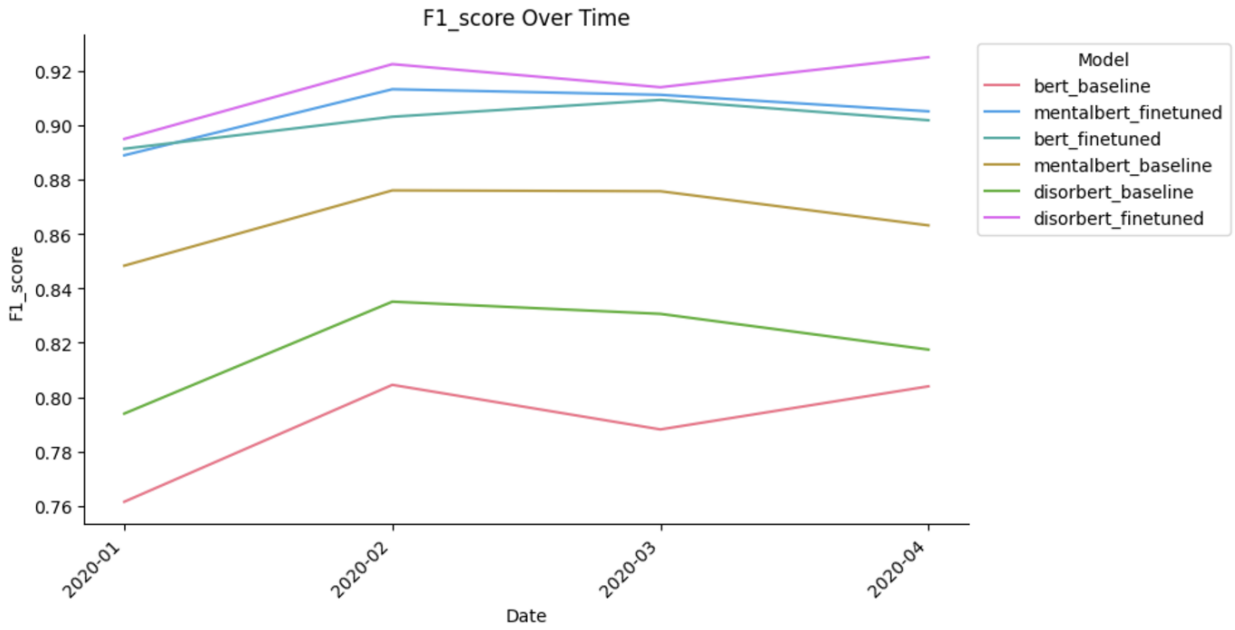

Figure 2: F1 scores of baseline and fine-tuned models (BERT, MentalBERT, DisorBERT) across months in early 2020, showing consistent performance despite the onset of the COVID-19 pandemic.

with the TensorFlow Adam optimizer with the binary cross entropy loss function and a learning rate of 0.00005. With higher learning rates (e.g. 0.0001), we observe exploding gradients in the form of increasing loss and decreasing accuracy after each epoch.

Each model is trained with a validation split of 0.2 and a batch size of 16, resulting in 500 calculations per epoch. Reducing the batch size to eight does not improve the results enough to justify doubling the number of calculations. We train the model for only two epochs because we do not observe enough improvement in performance after the second epoch to justify the additional computational load.

We evaluate the fitted models on the test dataset based on accuracy, precision, recall, and F1 score. We consider two criteria to measure success:

1. Each transformer must have an accuracy that is significantly higher than 0.5.

2. Each transformer must perform better with its layers trainable than with its layers frozen.

In this study, we evaluate model performance using accuracy, precision, recall, and F1 score. These metrics comprehensively capture the nuances of the task. Accuracy provides a general overview of the model's ability to classify posts correctly. However, given the critical nature of suicide risk detection, where both false positives and false negatives have significant implications, precision and recall are particularly important. Precision measures the proportion of correctly identified suicide-risk posts among all posts classified, which in turn helps minimize false positives that could misclassify general mental health issues as suicide risk. Recall measures the proportion of actual suicide-risk posts correctly identified, prioritizing the reduction of false negatives, as missing such cases could have severe consequences. To balance the trade-off between precision and recall, we use the F1 score to provide a single, balanced evaluation. This combination ensures a robust assessment of model performance, addressing both the need for accurate detection and the prevention of misclassifications which is critical for real-world applications in mental health.

In addition to evaluating the performance of each model as a whole, we also perform sub-group evaluations and experiments:

1. To evaluate model accuracy over time, we group the test set by month (i.e. January 2020, February 2020, March 2020, April 2020) and assess whether performance degrades over time as the

COVID-19 pandemic spreads globally.

2. We group the negative posts by subreddit to identify which subreddits are the most likely to have false positives, and whether they vary by model.

3. We identify the negative posts written by users who have posted in r/suicidewatch during the same timeframe and compare the false positive rates between posts written by suicidal authors and those written by non-suicidal authors.

5 Results and Discussion

We consider the overall experiment a success, as each transformer model produces an accuracy that is significantly higher than 0.5. Without fine-tuning, the MentalBERT baseline model performs the best, with an F1 score eight points higher than the BERT baseline model. The DisorBERT model performs three points better than BERT, but not as well as MentalBERT (Table 1).

After fine-tuning, all three transformer models have similar performance results, with all accuracy and F1 scores within one point of each other. Each fine-tuned model outperforms its baseline counterpart by at least three points.

None of our six models exhibit performance degradation over time. The F1 score in a given month of our test period remains relatively constant compared to its overall F1 score. Therefore, we can conclude that the onset of the COVID-19 pandemic did not have a short-term effect on how suicidal thoughts are expressed in Reddit posts.

The three subreddits r/depression, r/addiction, and r/lonely are the most likely to generate false positives (Table 2). Within these subreddits, interestingly, the BERT model performs better than the MentalBERT and the DisorBERT models in both the baseline and the fine-tuned cases. This phenomenon can be explained because the MentalBERT and DisorBERT models during their respective pretraining and domain adaptation processes do not distinguish between general mental health issues and warning signs specific to suicide risk. This also explains why the BERT fine-tuned model have better precision than the MentalBERT and DisorBERT fine-tuned models even after segmenting for suicidal vs non-suicidal authors.

The use of additional fine-tuning, employing a balanced dataset of r/suicidewatch and r/depression with two extra training epochs, significantly improved model performance, particularly in reducing false positives. MentalBERT Fine-Tuned showed the most notable gains, with false positives for suicidal authors decreasing by 16.66 percentage points and for non-suicidal authors by 10.58 points. This highlights MentalBERT's enhanced ability to differentiate between suicide risk and general mental health issues. BERT Fine-Tuned also saw a substantial 10.25 percentage point reduction in false positives for suicidal authors, though false positives for non-suicidal authors slightly increased by 0.66 points, reflecting a minor trade-off.

DisorBERT Fine-Tuned reduced false positives for suicidal authors by 11.17 percentage points, but saw a 1.29 percentage point increase for non-suicidal authors. Among the baseline models, DisorBERT Baseline achieved the most significant reduction in false positives for suicidal authors, dropping from 54.95% to 14.10%, demonstrating the impact of the balanced dataset without fine-tuning (Table 3).

Overall, the fine-tuning improved precision across all models, with MentalBERT Fine-Tuned showing the largest improvement. BERT Fine-Tuned maintained consistent high performance despite minor trade-offs, while DisorBERT Fine-Tuned showed mixed results, excelling at identifying suicidal authors but facing challenges with non-suicidal classifications. These findings underscore the importance of targeted training datasets for refining model precision and distinguishing nuanced mental health issues from suicide risk indicators.

Overall Performance	**Accuracy**	**Precision**	**Recall**	**F1 Score**
BERT Baseline	0.7836	0.7733	0.8024	0.7876
MentalBERT Baseline	0.8618	0.8420	0.8908	0.8657
DisorBERT Baseline	0.8144	0.7993	0.8396	0.8190
BERT Fine-Tuned	0.9044	0.8960	0.9060	0.9010
MentalBERT Fine-Tuned	0.8972	0.8470	0.9696	0.9041
DisorBERT Fine-Tuned	0.9084	0.8734	0.9552	0.9143

Table 1: Overall performance metrics (Accuracy, Precision, Recall, and F1 Score) for baseline and fine-tuned versions of BERT, MentalBERT, and DisorBERT, highlighting improvements with fine-tuning.

False Positive Rate	r/depression	r/addiction	r/lonely	Suicidal author	Non-suicidal author
BERT Baseline	0.67	0.71	0.63	0.5385	0.2237
MentalBERT Baseline	0.70	0.62	0.57	0.5165	0.1540
DisorBERT Baseline	0.70	0.86	0.63	0.5495	0.1980
BERT Fine-Tuned	0.51	0.34	0.29	0.3846	0.0946
MentalBERT Fine-Tuned	0.70	0.50	0.46	0.4615	0.1644
DisorBERT Fine-Tuned	0.62	0.36	0.40	0.4066	0.1283

Table 2: Comparison of false positive rates across subreddits for baseline and fine-tuned versions of BERT, MentalBERT, and DisorBERT models. Subreddits like r/depression, r/addiction, and r/lonely exhibit higher false positive rates, with fine-tuned models demonstrating improved performance by reducing these rates.

False Positive Rate	Suicidal author	Non-suicidal author
BERT Baseline	0.71	0.64
MentalBERT Baseline	0.49	0.50
DisorBERT Baseline	0.14	0.14
BERT Fine-Tuned	0.28	0.10
MentalBERT Fine-Tuned	0.29	0.05
DisorBERT Fine-Tuned	0.29	0.14

Table 3: Comparison of false positive rates for baseline and fine-tuned versions of BERT, MentalBERT, and DisorBERT models after additional fine-tuning. Fine-tuned models demonstrated improved performance by reducing these rates when trained solely on r/depression and r/suicidewatch data with a 50/50 split.

6 Conclusion

This study demonstrates the effectiveness of transformer-based models—BERT, MentalBERT, and DisorBERT—in identifying suicide risk in social media posts. Among the baseline models, MentalBERT showed the highest performance with an F1 score of 0.8657, outperforming BERT and DisorBERT by 8 and 4.66 percentage points, respectively. After fine-tuning, all three models achieved accuracy scores exceeding 90 percent, with DisorBERT Fine-Tuned emerging as the top performer (accuracy: 0.9084, F1 score: 0.9143). These results underscore the potential of fine-tuning to significantly enhance model effectiveness, with all fine-tuned models improving their baseline performance by at least three percentage points across key metrics.

Fine-tuning using a balanced dataset from r/suicidewatch and r/depression further reduced false positive rates, particularly for MentalBERT Fine-Tuned, which achieved a notable reduction of 16.66 percentage points for suicidal authors and 10.58 points for non-suicidal authors. BERT Fine-Tuned demonstrated a 10.25 percentage point improvement for suicidal authors, though its false positive rate for non-suicidal authors increased slightly by 0.66 points. DisorBERT Fine-Tuned achieved an 11.17 percentage point reduction for suicidal authors but encountered a minor increase of 1.29 points for non-suicidal classifications. These mixed results highlight the importance of refining models to balance precision and recall.

Interestingly, the three subreddits most prone to generating false positives—r/depression, r/addiction, and r/lonely—revealed a unique performance trend. In these contexts, the baseline and fine-tuned BERT models outperformed MentalBERT and DisorBERT, likely due to the latter models' pre-training emphasis on broad mental health issues rather than distinguishing suicide risk. This phenomenon also explains why BERT Fine-Tuned achieved higher precision when segmenting suicidal and non-suicidal authors.

Temporal analysis showed no significant performance degradation over time, suggesting that the onset of the COVID-19 pandemic did not affect the expression of suicidal thoughts in Reddit posts. While these findings are encouraging, the results also expose areas for improvement. False positive rates remain higher than desirable for specific subreddits, underscoring the need for targeted fine-tuning strategies. Future work should explore more granular domain adaptation using subreddit-specific posts to refine the ability to differentiate between general mental health issues and suicide risk indicators. By addressing these challenges, transformer-based models can be further optimized to provide robust, accurate tools for identifying suicide risk in social media contexts.

7 References

Coppersmith G, Leary R, Crutchley P, Fine A (2018). Natural Language Processing of Social Media as Screening for Suicide Risk. Biomedical Informatics Insights. 2018;10. DOI: 10.1177/1178222618792860

Low D, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh S (2020). Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. J Med Internet Res 2020; 22(10): e22635. DOI: 10.2196/22635

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7184–7190, Marseille, France. European Language Resources Association.

Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15305–15318, Toronto, Canada. Association for Computational Linguistics.

World Health Organization (2024). Suicide: Key Facts. Retrieved from https://www.who.int/news-room/fact-sheets/detail/suicide on 29 August 2024.

8 Appendix

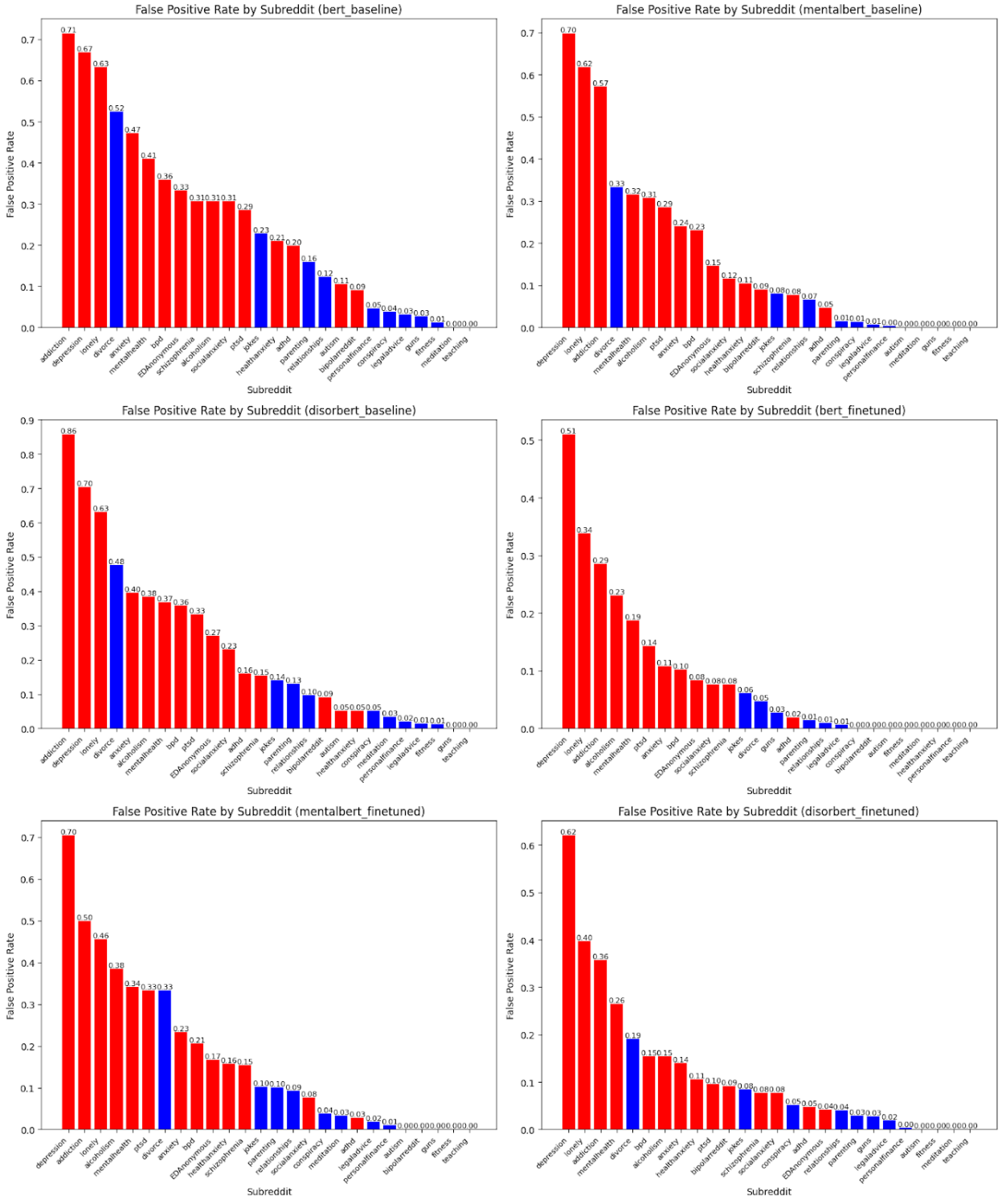

Figure 3: Bar charts comparing the false positive rates of BERT, MentalBERT, and DisorBERT models (baseline and fine-tuned) across various subreddits, with higher rates observed in r/depression, r/addiction, and r/lonely.

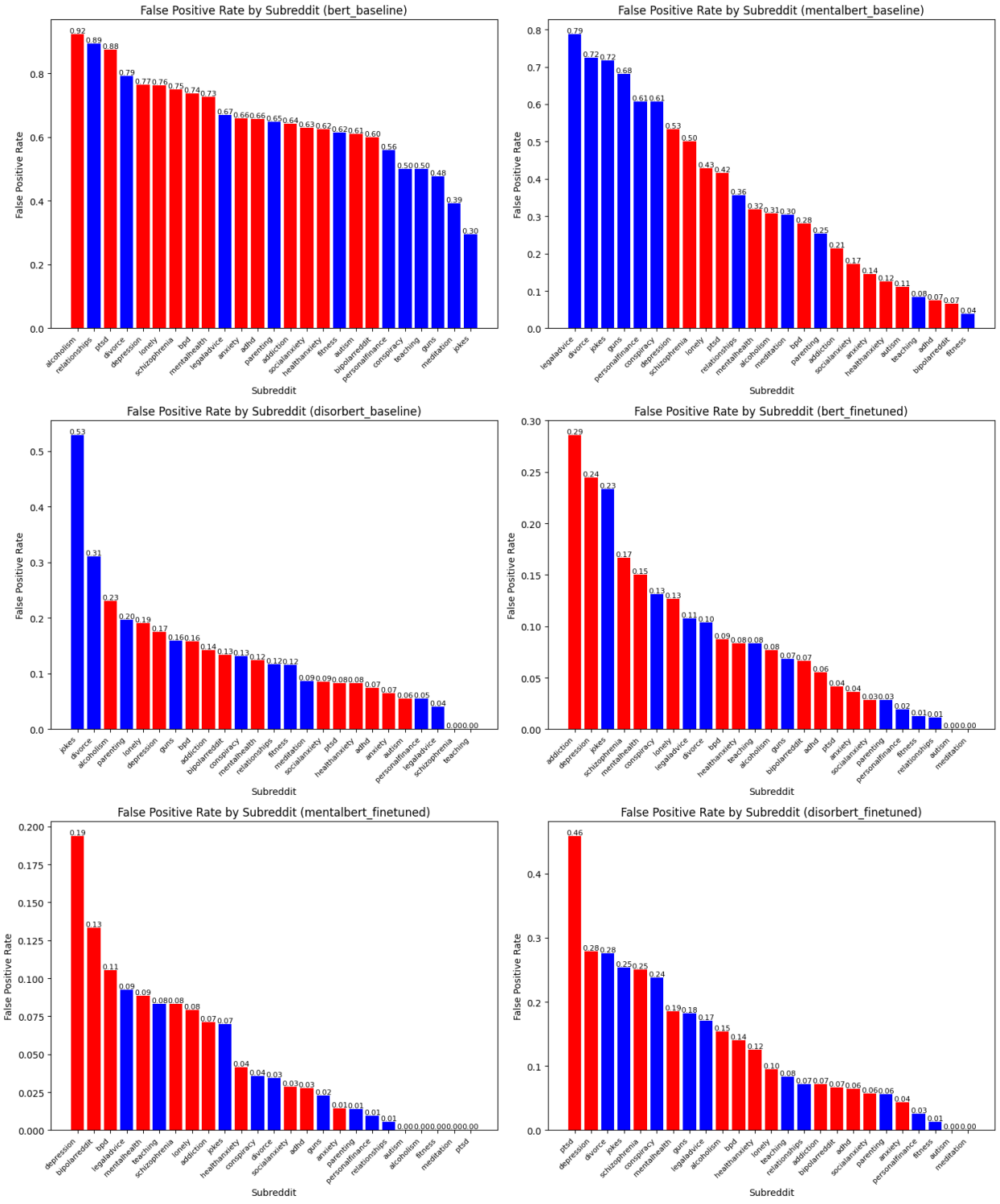

Figure 4: Bar charts compare the false positive rates of BERT, MentalBERT, and DisorBERT models (baseline and fine-tuned) across various subreddits. The charts show a greater distribution of higher false positive rates across subreddits for all models when using the additional fine-tuning training set.