Depression Detection from Social Media with Sarcasm Awareness

Anuj Prakash
Indiana University Bloomington
asprakas@iu.edu

December 19, 2025

Abstract

Social media platforms contain rich linguistic signals that can reveal early signs of mental health distress. Recent advances in Natural Language Processing (NLP) have enabled automated depression detection from user-generated text; however, sarcastic and ironic language remains a major source of error in such systems. In this work, I present a sarcasm-aware depression detection pipeline based on fine-tuned DistilBERT models. Our approach combines a depression classifier with an independent sarcasm detection model and an interpretation layer that flags potentially unreliable predictions. Experiments on Reddit datasets demonstrate that while a DistilBERT-based depression classifier achieves high accuracy, many remaining errors involve sarcastic language. Incorporating sarcasm detection improves interpretability and reliability, highlighting the importance of multi-model NLP pipelines for sensitive real-world applications.

Introduction

Mental health disorders such as depression affect millions of individuals worldwide, yet early detection remains challenging. Social media platforms like Reddit provide a space where users openly discuss emotions, struggles, and personal experiences, often revealing early linguistic indicators of mental distress.

Recent NLP research has explored automated depression detection using social media text. While these systems can achieve high predictive performance, they often struggle with figurative language, particularly sarcasm and irony. Sarcastic expressions can invert sentiment or mask true emotional intent, leading to confident but incorrect predictions.

In this project, I address this limitation by proposing a sarcasm-aware depression detection framework. Instead of relying on a single classifier, I combine two fine-tuned transformer models—one for depression detection and one for sarcasm detection—along with an interpretation layer that contextualizes predictions. The goal is not to provide clinical diagnosis, but to improve model reliability and transparency in research-oriented mental health analysis.

Related Work

Depression Detection from Social Media

Early work in depression detection relied on lexical features and traditional machine learning models. De Choudhury et al. Choudhury et al. (2013) demonstrated that linguistic patterns on social media could predict depression risk. Subsequent studies expanded this line of research using neural models and deep learning architectures.

With the advent of transformers, models such as BERT have significantly improved performance on mental health classification tasks. Studies have shown that fine-tuned transformer models outperform traditional feature-based approaches on depression detection benchmarks Ji et al. (2022).

Sarcasm Detection

Sarcasm detection is a well-known challenge in NLP due to its reliance on context, pragmatics, and world knowledge. Early approaches used rule-based and feature-based methods Riloff et al. (2013), while more recent work leverages deep neural networks and transformers Ghosh and Veale (2016); Potamias et al. (2020).

Reddit has become a popular source for sarcasm datasets due to explicit user annotations, enabling large-scale supervised learning Khodak et al. (2018).

Sarcasm in Mental Health NLP

Despite advances in both depression detection and sarcasm detection, relatively little work has explicitly combined the two. Most mental health NLP systems implicitly assume literal language, which can lead to misinterpretation when sarcasm is present. Our work contributes to this gap by integrating sarcasm detection as a reliability signal rather than attempting to override depression predictions.

Data

Depression Dataset

I use a cleaned Reddit-based depression dataset consisting of posts labeled as depressed or non-depressed. The dataset includes user-generated text from mental health-related subreddits and control groups. Labels are normalized to a binary format, and empty or malformed entries are removed.

Sarcasm Dataset

For sarcasm detection, I use a large-scale Reddit sarcasm corpus containing explicitly labeled sarcastic and non-sarcastic comments. The dataset is balanced and provides sufficient data for fine-tuning transformer-based models.

Methodology

Our approach consists of three main components: a depression classifier, a sarcasm classifier, and an interpretation layer that combines the outputs of both models to improve reliability.

Baseline Model

As an initial baseline, I implemented a TF-IDF representation combined with Logistic Regression for depression classification. While this approach achieved reasonable performance, it was limited in its ability to capture contextual and semantic nuances, particularly in figurative language.

Transformer-Based Depression Detection

I fine-tune DistilBERT for binary depression classification. DistilBERT is a compressed version of BERT that retains much of its representational power while significantly reducing computational cost. This makes it suitable for training on moderately sized social media datasets.

Text is tokenized using the DistilBERT tokenizer with a maximum sequence length of 128 tokens. The model is fine-tuned using cross-entropy loss, with macro-averaged F1 score used for model selection.

Sarcasm Detection Model

Sarcasm detection is treated as a separate binary classification task. I fine-tune an independent DistilBERT model on a large-scale Reddit sarcasm dataset. Training the sarcasm model separately allows it to specialize in linguistic cues such as irony, exaggeration, and contextual incongruity.

Interpretation Layer

Rather than directly modifying depression predictions, I introduce an interpretation layer that combines outputs from both classifiers. When the depression model predicts high likelihood of depression while the sarcasm model also predicts high sarcasm confidence, the system flags the prediction as potentially unreliable. This design prioritizes transparency over forced decision-making.

Experiments and Results

Model	Accuracy	Precision	Macro F1
TF-IDF + Logistic Regression	0.90	0.90	0.90
DistilBERT (Depression)	0.98	0.98	0.98
DistilBERT (Sarcasm)	0.78	0.78	0.78

Table 1: Performance comparison of models on test datasets.

Depression Detection Performance

The DistilBERT depression classifier achieves approximately 98% accuracy and a macro F1-score of 0.98 on the test set. Error analysis reveals only 29 misclassified samples, indicating strong overall performance. As shown in the below Figure 1, the depression classifier achieves strong performance across both classes, with very few false positives and false negatives.

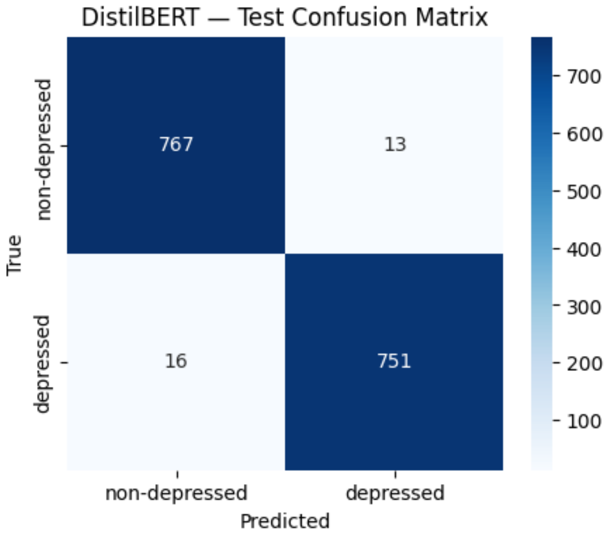


Figure 1: Confusion matrix for the DistilBERT depression classifier on the test set.

Sarcasm Detection Performance

The sarcasm classifier achieves approximately 77.7% accuracy with a macro F1-score of 0.78. While lower than the depression model, this performance is consistent with prior work on sarcasm detection and sufficient for flagging potentially unreliable cases. Figure 2 illustrates that sarcasm detection remains a challenging task, motivating its use as a contextual reliability signal rather than a strict decision rule.

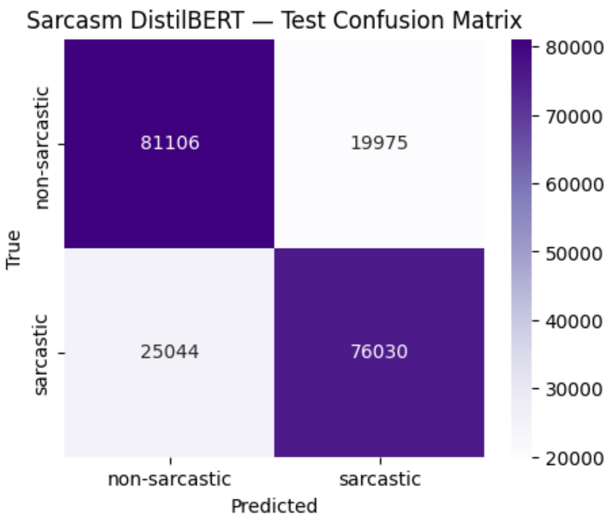


Figure 2: Confusion matrix for the DistilBERT sarcasm classifier on the Reddit sarcasm dataset.

Error Analysis

To better understand model behavior beyond aggregate metrics, I conducted a detailed error analysis of the depression classifier. I examined false positives, false negatives, and high-confidence misclassifications. Figure 3 highlights that while overall misclassifications are rare, a subset of errors occur with high model confidence, posing a risk for misinterpretation in sensitive applications.

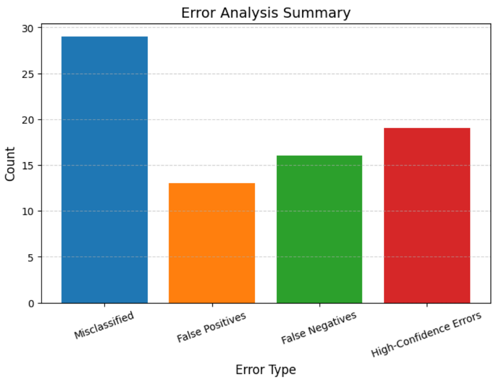


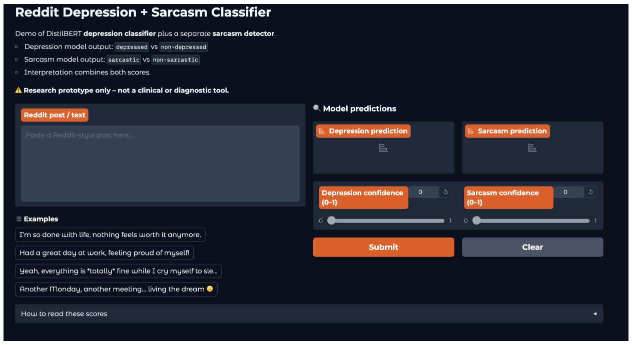
Figure 3: Distribution of classification outcomes for the depression model, including false positives, false negatives, and high-confidence misclassifications.

Despite strong overall performance, the model produced 29 misclassified samples in the test set. Manual inspection revealed that a substantial portion of these errors involved sarcastic or ironic language, where surface sentiment conflicted with the underlying emotional intent.

I further analyzed predictions with confidence scores greater than 0.9 that were nevertheless incorrect. These high-confidence errors are particularly concerning in mental health applications, as they may convey unjustified certainty. Sarcasm was a recurring factor in such cases, motivating the integration of an explicit sarcasm detection module.

Demo System

I implement an interactive Gradio-based web demo that allows users to input Reddit-style text and receive depression and sarcasm predictions along with confidence scores. The system provides an explanation that contextualizes predictions and explicitly warns when sarcasm may reduce reliability. The demo is designed for research and educational purposes only. Figure 4 shows the user-facing interpretation layer, which presents both depression and sarcasm predictions alongside confidence scores to encourage cautious interpretation.



Reddit Depression + Sarcasm Classifier

Demo of DistilBERT **depression classifier** plus a separate **sarcasm detector**.

- Depression model output: `depressed` vs `non-depressed`
- Sarcasm model output: `sarcastic` vs `non-sarcastic`
- Interpretation combines both scores.

⚠ **Research prototype only – not a clinical or diagnostic tool.**

Paste a Reddit-style post here...

Examples

- I'm so done with life, nothing feels worth it anymore.
- Had a great day at work, feeling proud of myself!
- Yeah, everything is *totally* fine while I cry myself to sle...
- Another Monday, another meeting... living the dream 🙃

How to read these scores

Figure 4: Screenshot of the Gradio-based interpretation layer combining depression and sarcasm predictions with confidence scores.

Discussion

Our findings demonstrate that high classification accuracy does not necessarily imply reliable interpretation in sensitive domains. Sarcasm presents a fundamental linguistic challenge that cannot be resolved through sentiment polarity alone. By explicitly modeling sarcasm, our system acknowledges uncertainty rather than obscuring it.

This work emphasizes the importance of error-aware NLP design, especially in mental health research, where misinterpretation may carry ethical risks. Rather than replacing human judgment, such systems should serve as decision-support tools that highlight uncertainty and ambiguity.

Future Work

Future extensions of this work include multi-task learning for joint depression and sarcasm detection, incorporating emotion and stress-level classification, and exploring explainability techniques such as attention visualization or SHAP values. Additionally, adapting the system to other platforms and languages would further improve generalizability.

Conclusion

I present a sarcasm-aware depression detection framework that combines multiple transformer-based models and error analysis to improve reliability and transparency. Our work demonstrates the importance of contextual understanding in NLP systems for mental health research and provides a foundation for future multi-model approaches.

References

Choudhury, M. D., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Ghosh, A. and Veale, T. (2016). Fracking Sarcasm using Neural Network. In Balahur, A., van der Goot, E., Vossen, P., and Montoyo, A., editors, *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Khodak, M., Saunshi, N., and Vodrahalli, K. (2018). A Large Self-Annotated Corpus for Sarcasm. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Potamias, R. A., Siolas, G., and Stafylopatis, A.-G. (2020). A Transformer-based approach to Irony and Sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320. arXiv:1911.10401 [cs].

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.