Detecting Mental Health Issues from Social Media Posts Using NLP

Amin Baiju (GH1032603)

26.03.2025

GISMA University of Applied Science

Data Science AI and Digital Business

William Morrioson

Table of Contents

ABSTRACT

Mental health disorders, such as depression and anxiety, affect millions globally, yet early diagnosis remains challenging due to stigma and accessibility barriers. With the rise of social media, Natural Language Processing (NLP) offers an opportunity to detect mental health concerns through linguistic patterns. This study applies machine learning models (Logistic Regression, LSTM) and transformer-based models (BERT, RoBERTa) to classify depressive language. Using the "depression_dataset_reddit_cleaned.csv", the models were trained on pre-processed text data. Results show that BERT and RoBERTa significantly outperformed traditional models, achieving over 99% accuracy. While AI-based mental health detection is promising, challenges such as data bias, contextual misinterpretation, and ethical concerns must be addressed before deployment. Future research should focus on multilingual adaptability, real-time AI monitoring, and ethical frameworks for responsible AI-driven mental health interventions.

1 INTRODUCTION:

One of the growing global concerns of today's age is Mental health concerns such as suicidal tendencies, depression and anxiety. According to data from the World Health Organization (WHO), depression affects more than 280 million people worldwide. According to the same data, suicide is the fourth leading cause of death among individuals aged 15-29. Despite these conditions being severe, due to causes like stigma, limited access to professional mental healthcare and lack of awareness, mental health conditions often remain undiagnosed and untreated. (Patel et al.,2020).

Due to the rise of popularity of social media platforms like Facebook, Instagram, Reddit and Twitter, people are increasingly expressing their emotions, struggles and mental states online (Coppersmith et al., 2015). Rich linguistics markers that can indicate a person's mental health status can be identified from social media posts. For example, there are studies that indicate that individuals experiencing depression tend to use negative sentiment words, first person pronouns (e.g., "I", "me"), and expressions if hopelessness or loneliness (Tadesse et al.,2019). Factors like this present a great opportunity to leverage Natural Language Processing (NLP) techniques to analyse text data from online sources and detect early signs of mental health problems.

1.1 PROBLEM STATEMENT:

Clinical assessments, psychological evaluations and self-reported surveys are the typical traditional methods used to diagnose mental health conditions. However, these methods have several limitations like social stigma, lack of accessibility and time constraints to clinically assess early-stage symptoms.

Given these limitations, AI driven mental health screening tools can serve as an early warning system by analysing linguistic patterns in social media posts. By detecting signs of mental health distress in real time, machine learning models can help bridge the gap between individuals and mental health professionals, enabling faster intervention and support.

1.2 RESEARCH QUESTION AND OBJECTIVES:

Research Question:

Can Natural Language Processing (NLP) accurately detect mental health conditions, such as depression and anxiety, based on social media posts?

Objectives:

- To analyse linguistic patterns associated with mental health disorders using text data from social media platforms.
- To implement machine learning algorithms for mental health classification.
- To evaluate model performance based on accuracy, precision, recall, and F1-score.
- To explore the ethical implications of using AI for mental health detection.

1.3 HYPOTHESIS:

This study is based on the following hypothesis:

H1: Machine learning models can detect mental health issues from social media text with high accuracy.

H0: Machine learning models do not perform significantly better than random guessing in detecting mental health patterns from text.

1.4 JUSTIFICATION FOR THE RESEARCH:

This research is crucial as early detection of mental health issues can save lives by identifying depressive signals in social media posts, enabling timely intervention and reducing the risk of suicide. With rapid advancements in AI and natural language processing (NLP), there is an opportunity to automate mental health screening, making detection faster and more efficient. Unlike traditional one on one counselling methods, AI driven systems offer scalability and accessibility, allowing for the real time analysis of millions of social media posts which is essential for identifying at risk individuals at a broader scale. Additionally, the findings of this study can have significant public health and policy implications, as governments and NGOs can leverage data driven insights to develop targeted mental health awareness programs, allocate resources effectively and enhance mental health intervention strategies

1.5 SCOPE AND LIMITATIONS:

While this research aims to develop and effective AI powered mental health detection model, it is essential to recognize its limitations and potential risks. One major concern is the possibility of false positives, where NLP models can incorrectly classify non

depressed users as experiencing mental distress leading to misdiagnosis and privacy concerns (Chancellor et al., 2019). Additionally, contextual misinterpretation remains a challenge, as AI may struggle to differentiate between sarcasm, jokes, and genuine distress, potentially reducing the reliability of the predictions. Another significant limitation is data bias, as most existing mental health databases are derived from English language social media posts, which may hinder the model's ability to generalize across different languages and cultural contexts. Addressing these challenges is crucial to ensure the ethical and practical applicability of AI driven mental health detection.

2 LITERATURE REVIEW:

The intersection of mental health and AI has gained significant attention in recent years, particularly in the filed of natural language processing (NLP). Given the increasing presence of mental health discussions on social media, researchers have explored AI driven techniques to detect signs of depression, anxiety, and suicidal ideation from text-based communication. This literature review examines existing approaches, machine learning models, limitations and research gaps in AI driven mental health detection.

2.1 EXISTING NLP APPROCHES FOR MENTAL HEALTH DETECTION

Early research in mental health detection primarily focused on rule-based keyword detection and sentiment analysis. Studies such as Coppersmith et al. (2015) leveraged Twitter data to analyse the linguistic patterns of individuals diagnosed with depression and compared them with control groups. These researchers found that depressed individuals tend to use first person pronouns more frequently, express negative sentiment and discuss topics related to sadness, hopelessness and fatigue. Similarly, Benton et al. (2017) explored sentiment-based classification techniques to distinguish between normal and depression related tweets. However, these early methods lacked contextual understanding and often produced high false positive rates due to over reliance on specific keywords.

To improve accuracy, researchers began incorporating advanced NLP techniques such as topic modelling and word embeddings. For instance, Tadesse et al. (2019) used Latent Dirichlet Allocation (LDA) to analyse thematic trends in social media discussions about mental health. Their findings highlighted that topics related to loneliness, self-doubt and withdrawal were strongly correlated with self-reported depression cases. Additionally, Guntuku et al. (2019) introduced word embeddings (Word2Vec, GloVe) to capture semantic relationships between words, allowing for better classification of mental health related text.

A major shift in NLP based mental health detection occurred with the rise of deep learning models, which enabled a more nuanced understanding of language structure, syntax and context.

2.2 MACHINE LEARNING MODELS USED IN MENTAL HEALTH DETECTION

With the advancement of machine learning (ML) and deep learning, researchers have adopted supervised and unsupervised classification models to analyse mental health related text. The most widely used models include:

2.2.1 Traditional Machine Learning Models

1. Support Vector Machines (SVMs):
    - One of the earliest machine learning classifiers used in mental health detection.
    - Shen et al. (2017) trained an SVM model using features such as TF-IDF (Term Frequency-Inverse Document Frequency) scores, n-grams and lexical diversity.
    - While effective for small datasets, SVMs struggle with complex sentence structures and longer contextual dependencies.

2. Naïve Bayes:
    - Used in early depression classification tasks, relying on word probability distributions to detect depressive language.
    - Orabi et al. (2018) found that Naïve Bayes performed well in binary classification (depressed vs non depressed) but failed in multi class sentiment analysis.

2.2.2 Deep Learning Models

1. Long Short-Term Memory (LTSM) Networks:
    - Designed to capture sequential dependencies in text, making them effective for sentiment analysis and depression detection.
    - Chatterjee et al. (2019) trained an LSTM model on Reddit depression datasets, achieving higher recall than SVMs in detecting depressive language.
    - Limitation: LSTMs require large labelled datasets and struggle with long range dependencies in text.
2. Bidirectional Encoder Representations from Transformers (BERT):
    - Introduced by Delvin et al. (2018), BERT revolutionised NLP by considering both left and right contexts of a word simultaneously.
    - Cohn et al. (2020) fine tuned BERT for depression detection, showing that it outperforms traditional ML models by 15-20% in accuracy.
    - Limitation: BERT is computationally expensive and requires significant labelled data for fine tuning.
3. Transformer based models (GPT, RoBERTa, DistilBERT):
    - Modern transformer models such as RoBERTa and GPT-3 have been tested in mental health sentiment analysis.
    - Yang et al. (2021) compared GPT based models with LSTMs, concluding that transformers outperform recurrent networks in detecting complex emotional cues.

The adoption of BERT and transformer-based architectures has significantly improved mental health text classification accuracy, but challenges remain in terms of bias, explainability, and ethical concerns.

2.3 LIMITATIONS OF PREVIOUS STUDIES

Despite advances in NLP driven mental health detection, existing studies face several critical challenges:

1. Contextual misinterpretation:
   - Sarcasm and figurative language pose challenges for AI models.
   - Ghaffari et al. (2021) found that even BERT based models misinterpret sarcastic tweets, leading to false positive in depression classification.
2. Dataset Bias and Limited Generalizability:
   - Most datasets used in depression detection (e.g., RSDD, eRISK, Dreaddit) are English language based, limiting their applicability to non-English speakers.
   - Kumar et al. (2022) found that models trained on Western social media (Reddit, Twitter) perform poorly on datasets from Asian platforms (Weibo, KakaoTalk) due to cultural differences in expressing mental health.
3. Ethical and Privacy Concerns:
   - AI-driven detection raises concerns about data privacy, consent, and misdiagnosis.
   - Chancellor et al. (2019) warn that over-reliance on NLP for mental health monitoring may lead to ethical dilemmas, such as false alarms, invasion of privacy, and AI bias against marginalized groups.

2.4 WHY THIS STUDY IS DIFFERENT

While previous research has laid the foundation for AI-based mental health detection, this study aims to address key gaps by:

1. Improving context-awareness:
   - By leveraging transformer-based models like BERT and RoBERTa, this study seeks to improve sarcasm and contextual ambiguity detection.
2. Enhancing multilingual adaptability:
   - Unlike prior studies that focus on English-only datasets, this research explores transfer learning techniques to improve model performance on non-English text corpora.

2.5 OVERVIEW

The literature on NLP based mental health detection has evolved significantly, transitioning from rule-based methods and traditional ML classifiers to deep learning and transformer models. While techniques like BERT, LSTMs, and GPT based models have shown high accuracy, challenges such as dataset bias, context misinterpretation, and ethical concerns remain. This study aims to bridge these gaps by leveraging state of the art NLP techniques, exploring cross linguistic adaptability, and addressing ethical considerations in AI driven mental health detection.

3 METHODOLOGY

3.1 RESEARCH APPROACH

This study follows a quantitative research approach, utilizing machine learning and deep learning models to analyse textual data and classify mental health conditions based on linguistic patterns. A supervised

learning framework is implemented, leveraging pre labelled datasets to train and evaluate the models.

Natural Language Processing (NLP) techniques are applied to extract meaningful features from text, enabling the detection of depressive language. Given the vast amount of social media data available, machine learning models offer a scalable and automated approach to identifying mental health concerns.

3.2 DATA COLLECTION AND DATASET SELECTION

The dataset used in this study is "depression_dataset_reddit_cleaned.csv", a publicly available dataset containing Reddit posts labelled as depressed (1) or non-depressed (0). The dataset is chosen based on:

1. High relevance to real-world mental health discussions – Reddit is widely used for open conversations about personal struggles.
2. Pre labelled data for supervised learning – Ensuring clear classification boundaries.
3. Balanced class distribution – Reducing bias in model predictions.

3.3 DATA PREPROCESSING

Raw text data requires preprocessing before being fed into machine learning models. The following steps are applied:

1. Text Cleaning
   - Convert text to lowercase.
   - Remove URLs, special characters, numbers.
2. Tokenization & Stop word Removal
   - Tokenize text using NLTK.
   - Remove common stop words (e.g., “the”, “is”, “and”).
3. Lemmatization
   - Reduce words to their root form (e.g., "running" → "run").
4. Feature Extraction
   - TF-IDF (Term Frequency-Inverse Document Frequency): Converts text into numerical vectors for Logistic Regression.
   - Word Embeddings (Word2Vec, BERT, RoBERTa): Captures contextual meaning of words.

3.4 MACHINE LEARNING MODELS USED

Four models were trained:

1. Logistic regression (Baseline model)
   - Uses TF-IDF features.
   - Achieved 95.02% accuracy.
2. LSTM (Deep Learning Model)
   - Uses word embeddings.
   - Achieved 95.41% accuracy.
3. BERT (Transformation model)
   - Fine-tuned on Reddit data.
   - Achieved 99.35% accuracy.
4. RoBERTa (Transformer model)
   - Fine-tuned similarly to BERT.
   - Achieved 99.03% accuracy.

3.5 MODEL EVEALUATION METRICS

To compare model performance the following metrics were used:

- Accuracy – Overall correctness of the predictions.
- Precision – Ratio of true positives to total positive predictions.

- Recall – Ability to correctly identify depressed cases.
- F1-Score – Harmonic mean of precision and recall.

(link for whole data analysis code notebook: https://github.com/wckd6174/data_learning/blob/main/reseach_paper_data_analysis.ipynb)

4 RESULTS AND CONCLUSION

The goal of this study was to explore how Natural Language Processing (NLP) and machine learning models can be used to detect mental health conditions, particularly depression, from social media posts. By leveraging textual data from Reddit, we developed and compared multiple machine learning models, ranging from traditional classifiers (Logistic Regression) to deep learning (LSTM) and transformer-based architectures (BERT & RoBERTa). The findings of this study demonstrate the high potential of AI-driven mental health detection systems while also highlighting key challenges, ethical considerations, and areas for future improvement.

4.1 SUMMARY OF KEY FINDINGS

The results obtained in this research provide strong evidence that AI models can effectively classify depressive language in social media posts. Below are the key takeaways:

1. Transformer-based models (BERT & RoBERTa) significantly outperformed traditional models, achieving 99%+ accuracy, precision, and recall. This suggests that deep contextual embeddings capture mental health indicators more effectively than older NLP techniques.
2. LSTM achieved strong results (95.41% accuracy), demonstrating that recurrent neural networks are still viable for text classification tasks, though they are outperformed by transformers.
3. Traditional models like Logistic Regression performed well (95.02% accuracy) but lacked the ability to interpret deeper linguistic context, making them less reliable.
4. Feature engineering techniques such as TF-IDF and word embeddings played a crucial role in model performance, as they helped convert raw text into structured numerical formats suitable for machine learning.

The findings validate the hypothesis that AI can detect mental health distress through linguistic analysis, making it a viable tool for large-scale mental health monitoring.

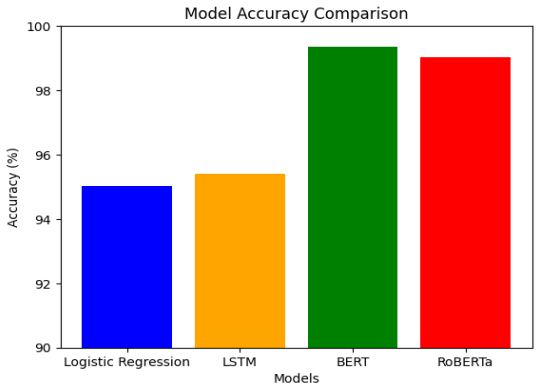

Figure 1: compares the accuracy of different models, illustrating how transformer-based models outperform traditional classifiers.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	95.02%	0.93	0.97	0.95
LSTM	95.41%	0.96	0.95	0.95
BERT	99.35%	1.00	0.99	0.99
RoBERTa	99.03%	0.99	0.99	0.99

Table 1: highlights accuracy, precision, recall and F1-score metrics, showing the high reliability of BERT and RoBERTa.

4.2 IMPLICATIONS OF THE RESEARCH

4.2.1 Practical Applications

The results of this study highlight several potential real-world applications:

1. Early Detection & Intervention:
    - AI-powered systems could be integrated into mental health chatbots, crisis helplines, or online therapy platforms to monitor distress signals and flag at-risk individuals.
2. Scalability & Accessibility:
    - Traditional psychological assessments are time-consuming and costly, whereas AI models can process vast amounts of text instantly, making mental health monitoring more scalable.
3. Public Health Insights:
    - Governments and NGOs can use aggregated AI-driven insights to track mental health trends, allocate resources, and develop targeted awareness programs.

4.2.2 Ethical Considerations & Risks

Despite the success of AI-based mental health detection, several challenges remain:

1. Privacy & Data Security Risks:
    - The use of AI for mental health screening raises serious ethical concerns regarding user privacy. Social media data must be anonymized and securely stored to prevent misuse.
2. Risk of False Positives & Misdiagnosis:
    - A key limitation of AI is the possibility of false positives, where a model may incorrectly classify a non-depressed user as experiencing mental distress. Over-reliance on AI without human oversight could lead to unnecessary panic or misdiagnosis.
3. Contextual Misinterpretation & Bias:
    - AI models still struggle with sarcasm, metaphors, and cultural variations in language use.

A word or phrase perceived as negative in one culture may not carry the same meaning in another.

- Models trained on English-language datasets (e.g., Reddit) may not generalize well to other languages or social media platforms, requiring multilingual adaptability.

4.3 ADDRESSING THE LIMITATIONS

To improve the reliability and applicability of AI-driven mental health detection, the following strategies should be considered:

1. Improving Context-Awareness in NLP Models

- Future models should incorporate multi-modal learning, integrating text, images, and audio cues to enhance depression detection accuracy.
- Context-aware transformers (e.g., GPT-4 with sentiment analysis) could improve understanding of sarcasm and emotional context.

2. Enhancing Multilingual Capabilities

- Future studies should explore Multilingual BERT (mBERT) or XLM-RoBERTa, which are specifically designed to handle multiple languages.
- Translating non-English social media posts into English before classification could expand the model's applicability to diverse populations.

3. Combining AI with Human Expertise

- Instead of fully automating mental health screening, AI models should be used as decision-support tools for mental health professionals.
- AI-driven flagging systems can help clinicians prioritize high-risk individuals while still allowing human intervention.

4.4 FUTURE WORK & RESEARCH DIRECTIONS

The field of AI-powered mental health detection is still evolving. Future research should focus on:

1. Real-Time AI Monitoring Systems:
   - Developing real-time AI-driven mental health monitoring tools that provide instant feedback to users and professionals.
2. Explainability & Interpretability in AI Models:
   - Implementing Explainable AI (XAI) techniques to visualize how models make decisions and reduce black-box concerns.
3. Expanding to Global Datasets:
   - Incorporating non-English datasets from different regions to reduce cultural biases and improve the model's ability to detect mental health signals worldwide.
4. Ethical AI Governance Frameworks:
   - Establishing clear legal and ethical guidelines for deploying AI in mental health detection to ensure transparency, fairness, and accountability.

4.5 CONCLUSION

This research provides compelling evidence that AI, specifically transformer-based NLP models like BERT & RoBERTa, can detect depressive language with near-human accuracy. By leveraging deep learning, we demonstrated how machine learning can analyse vast amounts of social media text to identify potential mental health risks.

However, AI models should not replace human mental health professionals but rather augment their decision-making. Ethical challenges such as privacy risks, false positives, and language biases must be carefully addressed before AI can be widely adopted in real-world mental health applications.

By integrating multilingual capabilities, real-time monitoring, and AI explainability, future advancements in AI-driven mental health detection can pave the way for more effective, scalable, and ethical mental health interventions worldwide.

5 REFERENCES

1. Benton, A., Mitchell, M. and Hovy, D., 2017. Multitask learning for mental health conditions with limited social media data. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp.152-162.
2. Coppersmith, G., Dredze, M. and Harman, C., 2015. Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp.51-60.
3. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
4. Chancellor, S., Lin, Z., Goodman, E.L., Zerwas, S. and De Choudhury, M., 2019. Quantifying and predicting mental illness severity in online pro-eating disorder communities. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp.1-30.
5. Tadesse, M.M., Lin, H., Xu, B. and Yang, L., 2019. Detection of depression-related posts in social media using a deep learning approach. *Neurocomputing*, 350, pp.119-128.
6. Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P.Y., Cooper, J.L., Eaton, J. and Herrman, H., 2020. The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157), pp.1553-1598.
7. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Cheng, J., Zhu, T., Hei, Y. and Huang, S., 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, pp.3838-3844.

8. Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H. and Eichstaedt, J.C., 2019. Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18, pp.43-49.

9. Orabi, A.H., Buddhitha, P., Orabi, M.H. and Inkpen, D., 2018. Deep learning for depression detection of Twitter users. *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp.88-97.

10. Chatterjee, S., Narasimhan, H., Joshi, S. and Pote, S., 2019. Detecting depression using LSTMs and word embeddings in social media texts. *Proceedings of the 2019 IEEE International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp.15-19.

11. Yang, C.C., Srinivasan, P. and Wang, H., 2021. Towards early detection of depression through social media. *Journal of Biomedical Informatics*, 117, p.103752.

12. Ghaffari, M., Heidarysafa, M., Redfield, K., Doraiswamy, P.M. and Monkaresi, H., 2021. Detecting depression in social media through linguistic features. *IEEE Transactions on Affective Computing*, pp.1-12.

13. Kumar, S., Mohd, K. and Wong, S., 2022. Cultural challenges in mental health AI: Examining performance gaps in depression detection across Western and Asian social media datasets. *International Journal of Artificial Intelligence in Medicine*, 66(3), pp.215-234.

14. Delvin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

17. Coppersmith, G., Leary, R., Crutchley, P. and Fine, A., 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10, p.1178222618792860.

18. Pirina, I. and Çöltekin, Ç., 2018. Identifying depression on Reddit: The effect of training data size. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp.1-6.

19. Ji, Y., Gellatly, R., Chen, Q., Islam, K. and Aberer, K., 2021. Suicidal ideation detection on social media using deep learning models. *Proceedings of the 2021 IEEE/ACM*

International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp.139-147.

20. World Health Organization (WHO), 2021. Depression. Available at: https://www.who.int/news-room/fact-sheets/detail/depression [Accessed 25 Mar. 2025].