

Detecting depression using natural language processing (NLP) techniques

Problem Statement:

Depression is described as one of the biggest burdens affecting a human life, with approximately 332 million people affected worldwide. Depression can affect the quality of everyday life, relationships with family and friends, and even work and school life (World Health Organization, 2021). With social media platforms, individuals can express, seek advice, or form communities. As a result, these online platforms have become a major source of data, providing valuable insights for investigation. For this amount of data, it will require having an advanced collecting tool for extracting, such as natural language process (NLP), data mining, and machine learning (ML) (Alghamdi et al., 2020).

Detecting depression using NLP has been studied in numerous research. Sourcing data from social media platforms such as Reddit (Tadesse et al., 2019) and from medical records (Vaci et al., 2020). Nonetheless, there is still a critical gap between the high accuracy achieved in controlled research settings and the adaptability required for real-world clinical deployment. According to Veci et al. (2020), even with data from 526 annotated papers, the model did not perform as expected with complicated medical entities. Similarly, detecting depression via Reddit (Tadesse et al., 2019) was taken back by the limited data size, which led to overfitting. This proves the high need for a larger dataset to improve testing and overall model performance.

This research aims to conduct a systematic literature review to analyze the gap between controlled accuracy and real-world generalizability in depression detection using NLP, with a focus on identifying the key factors that contribute to performance degradation and proposing directions for future research.

Research objectives:

- **Address the gap between social media platform and real-world detecting technique (Controlled accuracy vs. ecological validity)**
- 1. Acknowledge the difference between the real-life diagnosing technique and how can it affect detecting depression.

These two differ in more ways in which:

- Clinical diagnosis relies on Psychiatric evaluation, physical exams, lab tests, and DSM-5 (Mayo Clinic).
- NLP-based detection from social media typically involves textual data, which lacks the clinical context.

This may affect the **accuracy of detection** as in models trained on social media data can learn correlation pattern (Alghamdi et al., 2020), but not clinical indicators, this may lead to: Fals Positives (e.g., someone expressing sadness after a loss), False negatives (e.g., a depressed person post positively) and poor generalizability when models trained on social media applied to clinical populations.

- 2. Propose and discuss methods to improve the accuracy of detecting techniques.

To bridge this gap, future research should adopt multiple testing strategies such as **cross-domain validation, multi-modal integration, longitudinal modeling, clinician-in-the-loop evaluation, robustness testing** by implementing more approaches NLP systems can improve their accuracy.

- **Establish a comprehensive evaluation framework that prioritizes clinical relevance.**

- Explain how NLP can assist in the medical field.

NLP offers several practical ways to help in a clinical setting, longitudinal monitoring of symptom progression, identification of subtle linguistic markers of depression, and reduction of clinician bias through objective data-driven insights. However, NLP is a tool to assist the clinical field, not a replacement.

- Discuss the need for a clinically approved model.

The current NLP models for depression detection are not clinically approved, as they lack explainable reasoning, regulatory approval, fail to generalize across diverse population, and lack of standardized evaluation metrics. For a system that meets clinical approval must require interpretable output, external validation, fairness and bias testing, and compliance with healthcare regulations to avoid medical errors.

- **Systematically identify and select relevant studies on NLP-based depression detection.**

- Define the search strategy.

Relevant studies will be identified by searching academic databases including PubMed, IEEE Xplore, Scopus, and the ACM Digital Library using search terms such as "depression detection" AND "natural language processing" OR "NLP" OR "computational linguistics," with publication dates limited to the last ten years to capture recent advances.

- **Critically analyze the methodologies employed in existing studies.**

- Overreliance on single platforms.

Most of the study relies on a single dataset source, such as reddit or twitter, this simplifies the data collections, and it severely limits generalizability. A model trained on social media platform will not perform well on a clinical source.

- Lack of external validation.

This matter weakens the models as in there is absence on the clinical approval, which overestimates real-world performance. As demonstrated by Vaci et al. (2020), even well-trained models degrade substantially when applied to complex, real-world medical entities.

- Choice of NLP features.

Methodologies vary widely, from simple bag-of-words to deep learning embeddings (BERT, RoBERTa) to linguistic feature sets (LIWC, part-of-speech tagging) (Tadesse et al., 2019). Deep learning gives a high performance but operates as a black box, limiting clinical trust. Linguistic features are interpretable but may miss complex semantic patterns. A balanced approach combines interpretable features with explainability tools such as SHAP or LIME.

- **Evaluate the datasets used in existing research.**

- Labeling methodology and its limitations.

Most study use a labeling method where they classify data as “depressed” for individual are diagnosed or upload frequent in depression-focused posts. This is useful for a large-scale data collection, but it is not the same as its not clinically verified.

- Dataset size and overfitting concerns.

Datasets often include only a few hundred to thousands of labeled textual data. These data are limited in size, as noted by Tadesse et al. (2019), a small-sized dataset leads to overfitting, which will disturb the quality performance on real-world data. Larger datasets (e.g., tens of thousands of examples) are urgently needed.

- **Identify and synthesize the key limitations affecting real-world applicability.**

While NLP and machine learning have demonstrated considerable promise in enhancing the detection technique, there are still multiple barriers that restrict the real-world results. These limitations persist of:

- Relying on a small or a small demographic group (social media platforms) dataset that fails to capture the real-world complexity (Vaci et al., 2020).
- Poor quality of testing and training, which may lead to overfitting or overgeneralization.
- Developing and maintaining such models may lead to high computational cost.
- The time constraints of data collection, annotation, model development, and clinical validation may slow the deployment to the real world.

Reference

Depressive disorder (depression). World Health Organization. Mar 31, 2023.
URL: <https://www.who.int/news-room/fact-sheets/detail/depression> [accessed 2023-07-21]

N. S. Alghamdi, H. A. Hosni Mahmoud, A. Abraham, S. A. Alanazi and L. García-Hernández, "Predicting Depression Symptoms in an Arabic Psychological Forum," in *IEEE Access*, vol. 8, pp. 57317-57334, 2020, doi: 10.1109/ACCESS.2020.2981834. keywords: {Depression;Feature extraction;Natural language processing;Social networking (online);Mental health;Mental disorders;Supervised learning;semi-supervised learning;machine learning;predictive models;depression;lexicon;text analysis}
<https://ieeexplore.ieee.org/document/9040556/authors#authors>

M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," in *IEEE Access*, vol. 7, pp. 44883-44893, 2019, doi: 10.1109/ACCESS.2019.2909180. keywords: {Feature extraction;Task analysis;Linguistics;Twitter;Natural language processing;Support vector machines;Natural language processing;machine learning;Reddit;social networks;depression},
<https://ieeexplore.ieee.org/document/8681445/authors#authors>

Vaci N, Liu Q, Kormilitzin A, De Crescenzo F, Kurtulmus A, Harvey J, et al. Natural language processing for structuring clinical text data on depression using UK-CRIS. *Evidence Based Mental Health*. 2020;23:. <https://doi.org/10.1136/ebmental-2019-300134>

Mayo Clinic, "Depression (major depressive disorder)," Mar. 14, 2026. [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/depression/diagnosis-treatment/drc-20356013>. [Accessed: Apr. 2, 2026].