Journal of Behavioral Data Science, 2025, *5* (1), 67–102.
DOI: https://doi.org/10.35566/jbds/caoyc

Machine Learning Approaches for Mental Illness Detection on Social Media: A Systematic Review of Biases and Methodological Challenges

Yuchen Cao[1,⋆], Jianglai Dai[2,⋆], Zhongyan Wang[3], Yeyubei Zhang[4], Xiaorui Shen[1], Yunchong Liu[4], and Yexin Tian[5,⋆⋆]

[1] Khoury College of Computer Science, Northeastern University, USA
[2] Department of EECS, University of California, Berkeley, USA
[3] Center for Data Science, New York University, USA
[4] School of Engineering and Applied Science, University of Pennsylvania, USA
[5] Georgia Institute of Technology, College of Computing, USA

Abstract. The global increase in mental illness requires innovative detection methods for early intervention. Social media provides a valuable platform to identify mental illness through user-generated content. This systematic review examines machine learning (ML) models for detecting mental illness, with a particular focus on depression, using social media data. It highlights biases and methodological challenges encountered throughout the ML lifecycle. A search of PubMed, IEEE Xplore, and Google Scholar identified 47 relevant studies published after 2010. The Prediction model Risk Of Bias ASsessment Tool (PROBAST) was utilized to assess methodological quality and risk of bias. The review reveals significant biases affecting model reliability and generalizability. A predominant reliance on Twitter (63.8%) and English-language content (over 90%) limits diversity, with most studies focused on users from the United States and Europe. Non-probability sampling methods (approximately 80%) limit representativeness. Only 23% of studies explicitly addressed linguistic nuances like negations, crucial for accurate sentiment analysis. Inconsistent hyperparameter tuning was observed, with only 27.7% properly tuning models. About 17% did not adequately partition data into training, validation, and test sets, risking overfitting. While 74.5% used appropriate evaluation metrics for imbalanced data, others relied on accuracy without addressing class imbalance, potentially skewing results. Reporting transparency varied, often lacking critical methodological details. These findings highlight the need to diversify data sources, standardize preprocessing protocols, ensure consistent model development practices, address class imbalance, and enhance reporting transparency. By overcoming these challenges, future research can develop more robust and generalizable ML models for depression

[⋆] These authors contributed equally to this work.
[⋆⋆] Corresponding author

detection on social media, contributing to improved mental health outcomes globally.

Keywords: Mental illness · Social media · Bias evaluation with PROBAST · Systematic review · Machine learning and deep learning

1 Introduction

Mental health disorders, including depression, represent a critical global health challenge, impacting approximately 1 in 8 people worldwide—approximately 970 million individuals in 2019 (WHO, 2023). Depression, one of the most prevalent mental health conditions, affects over 280 million individuals globally, including around 23 million children and adolescents. The COVID-19 pandemic has further exacerbated mental health issues, with notable increases in depression and anxiety observed during this period (WHO, 2023). The prevalence of mental health conditions, especially depression, highlights an urgent need for innovative detection methods and interventions. Early identification can lead to more effective treatment outcomes, alleviating the burdens placed on individuals, their families, and healthcare systems (Kessler et al., 2017).

In today's digital age, social media platforms such as Twitter, Facebook, and Reddit play a central role in daily life for millions of people. Studies have shown that individuals often openly express their thoughts, emotions, and mental states on Twitter, making it a valuable platform for examining mental health trends and developing tools for detection and intervention (De Choudhury, Counts, & Horvitz, 2013). The extensive user-generated content on these platforms provides a unique opportunity for mental health research, enabling the real-time analysis of linguistic patterns and behavioral trends, and providing insights that may otherwise be inaccessible (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017).

Advancements in machine learning and deep learning have significantly enhanced the ability to process and analyze large-scale datasets. These technologies are particularly suited for handling the complex and nuanced data found on social media, as they identify patterns and make predictions based on textual and behavioral cues. This capability offers practical tools for mental health detection, allowing researchers to develop models that can potentially identify at-risk individuals based on their social media activity (Shatte, Hutchinson, & Teague, 2019). By leveraging algorithms capable of learning from such diverse and rich datasets, researchers are able to develop models that contribute to early intervention efforts in mental health care.

1.1 Overview of Historical Studies on Machine Learning Approaches for Mental Health Detection in Social Media

A growing body of research has explored the application of machine learning techniques to detect depression through social media platforms. Approaches range from traditional machine learning techniques such as logistic regression and support vector machines to advanced deep learning models and ensemble methods—have been employed to classify user posts and predict mental health conditions based on linguistic features, patterns, and metadata (Calvo, Milne, Hussain, & Christensen, 2017; De Choudhury et al., 2013; Yazdavar et al., 2020). Platforms like Twitter, Facebook, and Reddit are frequently utilized due to their large user bases and the accessibility of publicly available text-based data. In contrast, TikTok, with its short-video format, provides a distinct medium that captures audiovisual cues such as tone, facial expressions, and gestures, providing researchers with additional dimensions for understanding mental health dynamics.

One of the most common approaches within this research involves sentiment analysis, which aims to determine the emotional tone of user-generated content. By assessing positive, negative, or neutral sentiment (Kumar, Khan, & Kalra, 2020), researchers attempt to correlate language patterns with indicators of depression. For instance, several studies have examined the use of first-person singular pronouns and negative emotion words as potential depression signals (Rude, Gortner, & Pennebaker, 2004).

Despite promising results, multiple challenges remain. First, many studies suffer from limited generalizability due to small or homogeneous samples that may not represent the broader population. Data bias is a significant concern, stemming from the overrepresentation of certain demographic groups or linguistic communities while underrepresenting others (Olteanu, Castillo, Diaz, & Kiciman, 2019). Moreover, the dispersion of research in advanced machine learning methods for mental health detection across the literature, combined with a lack of robust sampling methods and standardized protocols, impedes the reliability of findings. Additionally, insufficient handling of complex linguistic nuances, such as context-dependent meanings, further limits the effectiveness of these detection efforts (Calvo et al., 2017).

1.2 Research Gaps and Objectives of the Current Study

While individual studies have provided valuable insights into the application of machine learning for mental health detection, significant gaps persist in the literature. These include the broader implications of biases and limitations across studies and the lack of comprehensive reviews consolidating the effectiveness of machine learning models (Calvo et al., 2017). Additionally, existing research does not consistently address methodological challenges across different stages of machine learning applications, such as sampling, preprocessing, model development, and evaluation (Thieme, Belgrave, & Doherty, 2020). Therefore, a systematic review is essential to unify findings and evaluate the pervasiveness and impact of biases across studies.

To address these gaps, this study aims to conduct a systematic review that synthesizes and evaluates existing machine-learning models for detecting depression on social media. The specific objectives are:

1. Examine the effectiveness of machine learning and deep learning models by focusing on bias present in sampling, data preprocessing, model construction,

fine-tuning, evaluation, and comparison, as well as the challenges associated with model generalizability across different social media platforms.

2. Explore methodological challenges, including those unique to mental health detection—such as handling class imbalance where depressive posts are the minority and preprocessing for sentiment analysis involving negations. Additionally, more general machine learning challenges, like improving model evaluation techniques and addressing data biases related to language and platform-specific factors, also persist. It is important to recognize that most of these biases are unintentional, either from practical challenges or from a lack of standardized guidelines for applying machine learning to mental health detection. By addressing these biases, the review aims to provide insights and strategies to mitigate these unintended biases, advancing the development of more reliable and generalizable models.
3. Provide recommendations for future research to enhance the reliability and applicability of machine learning models in mental health detection. These insights aim to inform strategies that improve early intervention efforts and contribute to the development of more robust, generalizable, and ethically sound machine learning applications. In doing so, the review seeks to provide guidance that fills the gap left by current practice, where a lack of formal guidelines has sometimes led to the persistence of unintended biases.

By addressing these objectives, this review seeks to provide a comprehensive understanding of the current practices and limitations within the field. The findings aim to guide future research and development into more robust, generalizable, and ethical applications of machine-learning models for mental health detection using social media data. In the following sections, we will first examine the methodologies and models used across studies, followed by an analysis of common biases and limitations. We will conclude with a discussion on best practices and recommendations for advancing the field.

2 Methodology

2.1 Search Strategy

The search focused on publications on machine learning and deep learning models for detecting depression and other mental health conditions using social media data, primarily from platforms like Twitter, Facebook, and Reddit. To identify relevant studies, a systematic search was conducted across multiple academic databases including PubMed, ACM, and IEEE Xplore, with Google Scholar used for additional sources. The search included combinations of ‘machine learning,’ ‘deep learning,’ ‘artificial intelligence,’ ‘social media,’ ‘Twitter,’ ‘Facebook,’ ‘Reddit,’ ‘depression,’ ‘sentiment analysis,’ and ‘mental health.’ To broaden the scope of the search, additional terms such as ‘anxiety,’ ‘mental disorders,’ ‘neural networks,’ and ‘supervised learning’ were included. The search process was carried out from June to July 2024.

The search strategy was structured around three main categories: social media platforms (e.g., 'social media,' 'Twitter,' 'Facebook,' 'Reddit'), mental health topics (e.g., 'depression,' 'sentiment analysis'), and machine learning and data analysis techniques (e.g., 'machine learning,' 'deep learning,' 'artificial intelligence'). The comprehensive search query[1] formulated for this review is:

```
((social media OR 'Twitter' OR 'Facebook' OR 'Reddit')
   AND ('depression' OR 'sentiment analysis' OR '
   mental health' OR 'anxiety' OR 'mental disorders')
   AND ('machine learning' OR 'deep learning' OR '
   artificial intelligence' OR 'neural networks' OR '
   supervised learning'))
```

2.2 Inclusion and Exclusion Criteria

To be included in this review, studies needed to meet the following criteria:

- **Publication Date:** Studies published after 2010 were included to ensure contemporary research and methods were considered.
- **Language:** Only studies published in English were included.
- **Research Focus:** The study must use machine learning or deep learning models for detecting depression or other mental health conditions, with a particular focus on analyzing data from social media platforms like Twitter, Facebook, or Reddit.
- **Study Type:** The review included primary research articles, specifically those that involved data-driven analyses.

Studies were excluded based on the following criteria:

- **Publication Type:** Review articles, systematic reviews, conference abstracts, editorials, opinion pieces, and non-peer-reviewed literature were excluded.
- **Scope:** Studies not directly focused on mental health detection through social media or not employing machine learning models were excluded.
- **Methodology:** Studies that did not directly employ machine learning or deep learning and applied solely on quantitative analysis were excluded.

[1] The search query used the term 'Twitter' to align with the naming convention at the time of the review, which covered literature up to June/July 2024. Twitter was rebranded as 'X' after this period. The search algorithm was adjusted to include both 'Twitter' and 'X' where applicable to ensure coverage of relevant results under the new name. However, no additional papers published up to June/July 2024 were identified using the term 'X.' Notably, one manuscript, Jamali, Berger, and Spiteri (2023), included both terms.

2.3 Study Selection Process

The selection process was conducted in three stages to ensure a rigorous and unbiased review of relevant studies. The process, which followed the search process that concluded in July 2024, lasted until August 2024.

1. **Initial Identification:** Duplicates were removed, and an initial screening was conducted based on titles and abstracts to filter out irrelevant studies. All authors contributed to this step.
2. **Title and Abstract Screening:** An independent review was conducted by two authors, Y.T. and J.D., to assess the relevance of studies based on their titles and abstracts. Both authors have expertise in machine learning and mental health research, ensuring a thorough evaluation. Any discrepancies in their assessments were discussed and resolved to ensure a consistent screening process.
3. **Full-Text Screening:** A comprehensive review of the full texts of selected studies was conducted. Any disagreements were resolved through discussion to maintain an unbiased selection process. Additionally, relevant studies identified through references in full-text articles were included for consideration. All authors contributed to this step.

2.4 Data Extraction and Analysis

The data extraction process involved using a standardized form to systematically capture detailed information from each selected study. The form included fields to record author names, study titles, publication journals, and publication years. It also documented the study designs, settings, and sample sizes, alongside specific inclusion and exclusion criteria. In addition, the form provided details on the machine learning models employed, the social media platforms analyzed (such as Twitter, Facebook, and Weibo), and the primary and secondary outcomes measured. Additionally, performance metrics, including accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic (AUROC)[2], which were collected when applicable.

Special attention was given to identifying potential sources of bias, study limitations, and funding sources, ensuring a comprehensive overview of each study's context and reliability. Table 1 below outlines the key categories and details included in the data extraction form.

[2] Accuracy measures the proportion of correctly classified instances among all instances. Precision focuses on the correctness of positive predictions, while recall measures the ability to identify actual positive cases. Both F1-score and Area Under the Receiver Operating Characteristic Curve (AUROC) are composite metrics that combine aspects of precision and recall to evaluate the performance of models. A detailed explanation of these metrics is provided in Section 3.7

Table 1: Key Data Extraction Categories for Systematic Review.

Category	**Details**
Study Details	Title, Authors, Year of Publication, Journal or Source, DOI or URL
Research Objectives	Purpose of the Study, Research Questions or Hypotheses
Methodological Aspects	Study Design, Settings, Sample Sizes, Inclusion and Exclusion Criteria, Data Collection Methods, ML/DL Models Employed
Criteria Applied	Data included, e.g., publicly available tweets, specific language posts. Data excluded, e.g., private or insufficiently detailed posts
Performance Metrics	Metrics Used (e.g., Accuracy, Precision, Recall, F1-score, AUROC, etc.)
Bias Evaluation	Data Collection and Preprocessing, Model Development and Tuning, Model Evaluation and Reporting
Additional Information	Confounding Factors, Study Limitations, Ethical Considerations, Funding Sources

2.5 Analytical Methods Used to Synthesize Findings

The extracted data were synthesized using a narrative approach, systematically examining each aspect of the machine learning lifecycle—sampling, data preprocessing, model construction, tuning, evaluation, comparison, and reporting—across the selected studies. This synthesis involved reviewing how studies approached sampling and data preprocessing, examining their approaches to model construction and tuning, and assessing model evaluation and comparison based on quantitative metrics such as accuracy, precision, recall, F1 scores, and AUROCs. For each stage, we summarized the methodologies employed by the studies and identified potential biases with established tools. This comprehensive approach provided insights into the current state of research, highlighting areas for future investigation to enhance the accuracy, generalizability, and applicability of machine learning models in this field.

2.6 Systematic Review Registration

This systematic review has been registered in the International Prospective Register of Systematic Reviews (PROSPERO) database under the title *Systematic Review of Machine Learning and Deep Learning Algorithms for Detecting Depression and Mental Health Conditions on Social Media* (ID: 617763). The registration has been approved.

3 Results

3.1 Study Selection

The search process began by identifying a total of 328 studies from three databases: 192 from Google Scholar, 101 from PubMed, and 35 from IEEE Xplore. After removing 57 duplicate studies, 271 unique titles and abstracts were retained for screening. During the title and abstract review, 174 studies were excluded. These exclusions were due to issues related to methodology (53 studies), scope (77 studies), and publication type (44 studies). This left 97 full-text studies to be reviewed in detail.

Upon reviewing the full texts, another 50 publications were excluded. The reasons for exclusion included being outside the scope or irrelevant (32 studies), methodological concerns (6 studies), publication type (9 studies), and unavailability (3 studies). Ultimately, 47 studies were included in the final narrative synthesis.

Figure 1 outlines how the initial pool of studies was refined down to the most relevant research for inclusion.

3.2 Characteristics of Included Studies

In this systematic review, key details of all 47 included studies, as summarized in Table 1, are provided in an online supplementary document. The majority of studies focused on Twitter (32 studies), Reddit (8 studies), and Facebook (7 studies). Additionally, one study examined Blued, a platform for MSM communities, and another focused on Indian social networking sites (SNS). Notably, 8 studies (17.02%) analyzed data from multiple platforms. Upon further examination, the datasets used in these 47 studies were found to be independent. The most commonly used models included traditional machine learning approaches such as Support Vector Machines (SVM) (19 studies), tree-based models (e.g., Decision Trees in 6 studies, Random Forests in 13 studies, and eXtreme Gradient Boosting (XGBoost) in 3 studies), and Logistic Regression (6 studies). Some studies also utilized deep learning models, including Convolutional Neural Networks (CNNs) (9 studies), Long Short-Term Memory (LSTM) networks (5 studies), and Bidirectional Encoder Representations from Transformers (BERT) (9 studies) for depression detection.

3.3 Methodological Quality and Risk of Bias

The risk of bias in the studies included in this systematic review was assessed using the Prediction model Risk Of Bias Assessment Tool (PROBAST, Wolff et al., 2019). PROBAST is a structured tool designed to assess the risk of bias and applicability of prediction models. It evaluates four key domains: participants, predictors, outcomes, and analysis, ensuring methodological rigor in studies. This tool provides a systematic framework for identifying biases and limitations

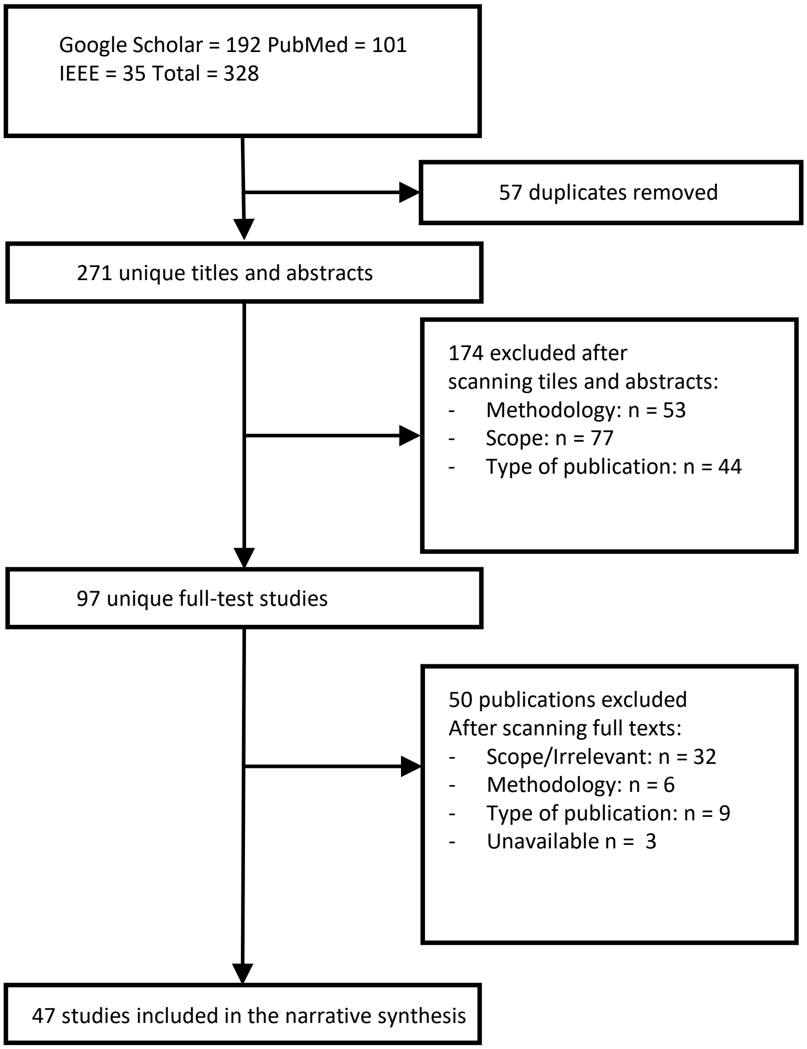
Google Scholar = 192 PubMed = 101
IEEE = 35 Total = 328
57 duplicates removed
271 unique titles and abstracts
174 excluded after
scanning tiles and abstracts:
- Methodology: n = 53
- Scope: n = 77
- Type of publication: n = 44
97 unique full-test studies
50 publications excluded
After scanning full texts:
- Scope/Irrelevant: n = 32
- Methodology: n = 6
- Type of publication: n = 9
- Unavailable n = 3
47 studies included in the narrative synthesis

Figure 1: PRISMA Flow Diagram of Study Selection Process for Systematic Review on Machine Learning Models for Depression Detection Using Social Media Data

in prediction models, offering critical insights into their validity and applicability. PROBAST is particularly relevant in this systematic review as it allows for a comprehensive assessment of potential methodological biases throughout the machine learning lifecycle, including data collection, preprocessing, model development, and evaluation. By identifying biases in these areas, the tool supports a rigorous evaluation of the reliability and generalizability of machine learning models used for mental health detection on social media. In addition to its role in assessing the risk of bias, PROBAST was used to evaluate transparency and completeness in the reporting of study methodologies and findings. The assessment covered 20 structured questions across the four domains, as detailed in Table 2. By incorporating PROBAST, this review identifies methodological weaknesses in the included studies, assesses their implications for the validity of findings, and evaluates the overall applicability of machine learning models used for mental health detection on social media. This ensures a thorough understanding of bias and enhances the reliability of the review's conclusions.

Table 2: Bias Evaluation Questions for Each Domain.

Domain	**Evaluation Questions**
Sample Selection and Representativeness	Q1. What is the sample used in this study, including the platform, sampling criteria, and sampling method? Q2. Does the sample represent the target population of social media users or posts?
Data Preprocessing	Q3. Did the study specify its approach to handling negative words when using traditional or machine learning methods for sentiment analysis?
Model Development	Q4. Did this study report hyperparameters? Q5. If reported, did this study tune (optimize) hyperparameters or use default settings? Q6. If tuned hyperparameters in this study, was this done on all models mentioned in the study?
Model Evaluation	Q7. Did the study divide the dataset into training, validation, and test sets, and were the reported metrics based only on training data? Q8. What evaluation metric was used in this study? Q9. Is the evaluation metric appropriate for this context (i.e., class-imbalanced settings)? Q10. If the study used accuracy as an evaluation metric, did it mention preprocessing steps to address class imbalance?

3.4 Sample Selection and Representativeness (Q1 & Q2)

The reviewed studies employed diverse sampling methods across various social media platforms, primarily focusing on Twitter (63.8%) with additional data from Reddit (23.5%), Facebook (8.5%), and other social media (2.1%). Most studies (around 80%) used non-probability sampling techniques, such as convenience sampling or keyword filtering, often utilizing APIs (e.g., Twitter API, Reddit API) to filter posts by specific mental health-related keywords like 'depression' or '#MentalHealth,' or leveraging pre-existing datasets from repositories like Kaggle.

The diversity in sampling criteria, sample sizes, demographic details, language focus, and geographic regions across the studies introduces potential biases. Sample sizes and levels of representation varied significantly among the studies, from small-scale studies (e.g., Study #46, which analyzed 4,124 Facebook posts from 43 undergraduate students with pre-specified criteria from the U.S.) to large-scale analyses (e.g., Study #5, which analyzed 56,411,200 tweets from 70,000 users across seven major U.S. cities). Many studies lacked detailed demographic information. The majority of studies focused predominantly on English-language posts, which are commonly associated with specific regions such as the U.S., U.K., Japan, Spain, and Portugal (although geographic information was explicitly reported in only about one-third of the studies) limiting the generalizability of the findings. Only a few studies examined posts in other languages, like Study #15, which analyzed Arabic tweets. Even within these regions and language-specific studies, demographic distribution was not always fully balanced. For example, Study #1 reported a mean participant age of 30.5 years (ranging from 18 to 68) and had a slight overrepresentation of female participants at 66.4%.

The non-representative sampling approaches observed across studies suggest limited generalizability to broader social media user populations. The primary biases identified include:

- **Platform Bias:** The predominance of Twitter (63.8%) over other platforms means that findings may not represent behaviors on platforms like Facebook, Instagram, or Reddit. As suggested by Olteanu et al. (2019), utilizing multi-platform data can reduce platform-specific biases and provide a more comprehensive view of user behaviors. However, while multi-platform data broaden the scope and reduce single-platform bias, platform-specific user demographics and engagement patterns may affect generalizability, with some platforms carrying more weight due to larger user bases or data volume.
- **Selection Bias:** Some studies relied on keyword-based sampling, which may overlook users not explicitly mentioning mental health. Study #7, for instance, searched for tweets containing 'I was diagnosed with depression.' As suggested by Morstatter, Pfeffer, Liu, and Carley (2013), combining keyword-based and random sampling can capture a broader range of user behaviors and discussions. Additionally, the limitations of Twitter's API exacerbate platform-specific challenges. As highlighted by Morstatter et al. (2013), Twitter's API does not provide access to all user-generated content, raising concerns about whether sampled data is representative of the platform's overall activity. This issue may lead to incomplete or skewed representations of user behavior, particularly in studies relying solely on API data. Researchers must critically evaluate the validity of conclusions drawn from API-retrieved data and consider combining multiple sampling strategies to mitigate such biases.

- **Language Bias:** The overwhelming focus on English-language content (over 90%) excludes insights from non-English-speaking communities, limiting the generalizability of findings across diverse linguistic groups. For instance, Study #15 was one of the few that analyzed non-English tweets, indicating the rarity of multilingual studies in this field. To address this, Danet and Herring (2007) recommended leveraging multilingual analysis methods, such as machine translation, or employing multilingual research teams to capture a more diverse linguistic landscape.
- **Geographic Bias:** While explicit geographic information was reported in only about one-third of the studies, the predominance of English-language posts suggests an implicit bias toward regions where English is the primary language, such as the U.S., U.K., and other English-speaking countries. Among the studies that reported geographic information, this predominance was evident. For example, Study #5 analyzed tweets from seven major U.S. cities, and Study #19 focused on Twitter users in Spain and Portugal. Hargittai (2015) suggested broadening the geographic scope to better represent global populations and avoid region-specific findings.
- **Self-selection Bias:** Platforms like Mechanical Turk (MTurk) or Clickworker, used in some studies (e.g., Studies #45 and #1, respectively), may attract specific demographic or employment profiles (e.g., higher digital literacy, particular age ranges, or specific socioeconomic statuses), affecting generalizability. While Chandler and Shapiro (2016) assessed the use of MTurk as a crowdsourcing tool, highlighting limitations in participant diversity and representativeness, which may skew results and underscore the need for multiple recruitment sources and stratified sampling for better generalizability.

In summary, no study in the review provided a fully representative sample of all social media users or posts. Key limitations include platform-specific focus (mostly Twitter), heavy reliance on non-probability sampling techniques (e.g., approximately 80% of the studies utilized convenience sampling or keyword filtering), and geographic and linguistic constraints. Notably, over 90% of the studies themselves acknowledged these limitations, recognizing the challenges of achieving representativeness in social media research. These limitations are, to a large extent, unavoidable due to the nature of social media platforms and the constraints of current data collection methodologies. This underscores the need for ongoing efforts to develop more sophisticated sampling techniques and analytical methods to mitigate these biases.

Similarly, some studies explicitly stated that their findings were intended to represent only specific populations. For instance, Study #8 and Study #21 focused on users discussing mental health or particular demographic groups on specific platforms. These limitations significantly impact the generalizability of findings to the broader population of social media users. Future research should strive for more diverse and representative sampling across platforms, languages, and geographic regions to enhance the applicability of results in the field of mental health and social media research.

3.5 Data Preprocessing with Focus on Negative Words Handling (Q3)

Across all studies, several common preprocessing tasks were consistently performed. Tokenization was conducted in all studies to break text into individual words or tokens, and text normalization steps included converting text to lowercase, as well as removing punctuation, URLs, and special characters. Many studies also performed stop-word removal to eliminate common words that are generally not informative for modeling. Additionally, some studies applied stemming and lemmatization to reduce words to their base or root forms, thereby unifying different morphological variants. Feature extraction techniques such as Bag of Words (BoW, Harris, 1954)[3], Term Frequency-Inverse Document Frequency (TF-IDF, Salton & Buckley, 1988)[4], and various word embedding methods were widely used to represent textual data numerically for modeling purposes.

While these standard preprocessing steps were broadly applied, certain aspects of sentiment analysis in mental health detection require additional attention. One such aspect is the effective handling of negative words, which is crucial for accurately interpreting sentiment and emotional tone, especially within this context. Among the 47 reviewed studies, approaches to negative words varied significantly:

First, only a minority of studies (11 out of 47 studies, approximately 23%) explicitly addressed negative words or negations in their preprocessing steps. Methods included standardizing all negative words to a basic form, like 'not," during preprocessing, which simplifies the representation of negations and improves sentiment recognition (e.g., Studies #3 and #34). Some studies quantified negative words as features by calculating metrics such as the user-specific average number of negative words per post. This metric captures the frequency of negative expressions per user and is then used as input for machine learning models to identify depressive emotions (e.g., Study #21). Others (e.g., Study #25) assigned a weight of -1 to negative adverbs to account for their inversion effect on sentence sentiment, ensuring more accurate sentiment quantification. Moreover, several studies employed specific methods for managing negations within their sentiment analysis frameworks. For example, some studies used sentiment analysis tools like TextBlob to determine the polarity of words in context, identifying negative words as indicators of depressive symptoms (e.g., Study #31). Others incorporated linguistic inquiry and word count (LIWC) categories related to

[3] BoW represents text as a vector by creating a vocabulary of all unique words in a corpus and counting the frequency of each word in a document. While simple and effective, BoW disregards word order and context, treating documents as collections of independent words.

[4] TF-IDF evaluates the importance of a word in a document relative to a collection of documents. It combines term frequency (how often a word appears in a document) with inverse document frequency (reducing the weight of common words that appear across many documents). This technique highlights terms that are more informative for classification or clustering tasks.

negations and negative emotions, indirectly addressing negations through predefined lexicon categories (Studies #1, #40, #42, #46, and #47).

The importance of negation handling has also been recognized in studies currently under review. For instance, Study #6 specifically explored the role of negation preprocessing in sentiment analysis for depression detection. By comparing datasets with and without negation handling, the authors demonstrated that addressing negations can significantly improve the accuracy of both sentiment analysis and depression detection, underscoring the need to address them in preprocessing. This study highlights the critical need for comprehensive negation handling in preprocessing to enhance the reliability of machine learning models in mental health contexts.

Second, a subset of studies (9 out of 47 studies, approximately 19%) did not explicitly handle negative words but employed advanced language models capable of inherently managing negations due to their contextual understanding, such as transformer-based models like Bidirectional Encoder Representations from Transformers (BERT, Devlin, Chang, Lee, & Toutanova, 2018) and Mental Health BERT (MentalBERT, Ji et al., 2022) (e.g., Studies #8, #9, #15, #16, and #39). These transformer-based models can capture the context of negations by processing text bi-directionally without explicit preprocessing steps. Other studies used attention mechanisms[5] (Vaswani et al., 2017) with word embeddings, such as attention layers combined with Global Vectors for Word Representation (GloVe) embeddings (Pennington, Socher, & Manning, 2014), allowing models to inherently understand and assign appropriate weights to negations through contextual embeddings (e.g., Studies #7, #10, and #13). Additionally, Embeddings from Language Models (ELMo, Peters et al., 2018), which capture the entire context of a word within a sentence, was also noted as a method that could capture the effect of negative words without explicit handling (Study #45).

However, the majority (27 out of 47 studies, approximately 57%) neither explicitly addressed negative words in their preprocessing nor used models inherently capable of handling negations (i.e., Studies #2, #4, #5, #11, #12, #14, #17, #18, #19, #20, #22, #23, #24, #26, #27, #28, #29, #30, #32, #33, #35, #36, #37, #38, #41, #43, and #44). These studies primarily focused on standard preprocessing tasks (e.g., tokenization, lowercasing, stop-word removal, stemming, and lemmatization), feature extraction methods (e.g., TF-IDF, BoW), and basic word embeddings (e.g., Word to Vector [Word2Vec]), without any special consideration for negations.

The impact on model performance and potential bias varied depending on how negative words were handled. Studies that explicitly addressed negative word handling reported improvements in model accuracy and a more nuanced understanding of sentiment (Helmy, Nassar, & Ramadan, 2024). Proper handling of negations allowed these models to correctly interpret phrases where

[5] Attention mechanisms allow models to focus on specific parts of the input data by assigning different weights to different elements. This enables the model to capture and utilize relevant contextual information more effectively during processing.

negations invert the sentiment (e.g., ‘not happy” versus ‘happy”), leading to more reliable results. In contrast, studies that did not explicitly account for negative words risked misinterpreting negated expressions, introducing bias into their findings. This oversight can cause models to incorrectly assign positive sentiment to negated negative expressions or vice versa, thus skewing the analysis. Such biases can significantly affect the overall performance and generalizability of the models, particularly in sensitive applications like depression detection. While some studies used advanced models capable of inherently handling negations (e.g., Studies #7, #8, #9, #10, #13, #15, #16, #39, and #45), reliance solely on the model's ability without explicit preprocessing might not capture all nuances of negations. Explicitly addressing negations can further enhance model performance, even when using sophisticated language models (Khandelwal & Sawant, 2020). Therefore, integrating both advanced modeling techniques and careful preprocessing of negative words may provide the most effective approach.

In summary, the review highlights a significant gap in the explicit handling of negative words in data preprocessing among studies focused on sentiment analysis and related fields. Proper management of negations is crucial, as it can substantially impact both model accuracy and reliability. Without adequately handling negative words, models may introduce bias and reduce their effectiveness, particularly in applications such as mental analysis and depression detection, where understanding sentiment nuances is critical. Future studies should prioritize the inclusion of explicit negation handling techniques within their preprocessing pipelines to enhance model performance and ensure more accurate interpretations of textual data.

3.6 Model Development

Hyperparameter Tuning (Q3, Q4 & Q5) Hyperparameters are external configurations set before the training process of machine learning models. Unlike model parameters, which are learned from the data during training, hyperparameters govern the learning process itself, such as the learning rate, regularization strength, and the number of hidden layers. Proper hyperparameter tuning ensures optimal model performance by balancing underfitting and overfitting, thus improving the model's ability to generalize to unseen data. Hyperparameter tuning is a critical aspect of optimizing machine learning models, directly impacting their performance and reliability. Our evaluation of the 47 reviewed studies focused on whether the studies reported their hyperparameters, the extent to which these hyperparameters were optimized, and whether tuning was applied consistently across all models within each study.

In particular, 27 studies (approximately 60%) reported using hyperparameters, but not all of them performed proper tuning. Only a limited number of studies ensured consistent tuning across all models, with many opting for default settings or tuning only specific models, leaving significant performance potential unexplored (Yang & Shami, 2020). This practice suggests that while hyperparameters are acknowledged by researchers, there is still a notable gap in their comprehensive and consistent optimization across studies. The breakdown of hyperparameter reporting and tuning practices is presented in Table 3.

Table 3: Hyperparameter Reporting and Tuning Practices in Reviewed Studies.

Hyperparameter Reporting	**Number (%) of Studies**	**Studies #**
Reported & Tuned for All Models	13 (27.7%)	#11, #12, #13, #16, #18, #21, #22, #25, #26, #28, #33, #45, #47
Reported but Partially Tuned	4 (8.5%)	#1, #8, #15, #23
Reported but Not Tuned	11 (23.4%)	#3, #4, #7, #9, #10, #31, #36, #39, #40, #41, #43
Not Reported or Tuned	19 (40.4%)	#2, #5, #6, #14, #17, #19, #20, #24, #27, #29, #30, #32, #34, #35, #37, #38, #42, #44, #46

The absence of consistent hyperparameter tuning can result in suboptimal model performance, reduced generalizability, or biased model comparisons. Key hyperparameters such as learning rate, regularization terms, or the number of hidden layers directly impact a model's training process and final accuracy (Mantovani, Rossi, Vanschoren, Bischl, & de Carvalho, 2015; Probst, Boulesteix, & Bischl, 2019). Without proper tuning, models may overfit, meaning they perform well on training data but poorly on unseen data, or underfit, failing to capture the complexity of the data altogether. For example, Study #2 did not report any tuning, which likely affected its model's ability to generalize to unseen data, leading to reduced model performance.

When only some models are tuned, comparisons across models become biased, as those with optimized hyperparameters gain an undue advantage. In Study #1, for instance, the Elastic Net model had its hyperparameters tuned, while other models, such as random forest, were left with default settings. This discrepancy can misleadingly suggest the superiority of the Elastic Net model due to tuning alone, rather than any inherent advantage in its architecture, leading to biased model comparisons.

A significant proportion of studies did not report hyperparameter tuning (approximately 40%) or failed to consistently tune them across all models (approximately 32%), which compromises the validity of their findings. For example, Studies #2 and #4 used default settings and missed opportunities to enhance performance, while Study #1 tuned hyperparameters for only one model, resulting in biased comparisons. Proper hyperparameter tuning is essential to avoid issues like overfitting or underfitting. Consistent tuning across all models ensures fair comparisons and enhances result validity.

Providing detailed descriptions of hyperparameter settings and optimization processes enhances transparency and reproducibility. Standardized tuning protocols, such as grid search, random search, or Bayesian optimization, should be employed to explore optimal configurations. Clearly documenting tuning strategies and any challenges encountered will provide valuable context for interpreting model performance results and strengthen the credibility of future machine learning studies. Future research should prioritize consistent tuning strategies and detailed reporting to enhance the credibility and reproducibility of their machine learning studies.

Data Partitioning (Q6) Proper data partitioning is fundamental to developing robust machine learning models that generalize well to unseen data. Typically, datasets are divided into three subsets: the training set, used to train the model and learn patterns; the validation set, used to fine-tune hyperparameters and avoid overfitting; and the test set, reserved for evaluating the model's final performance on unseen data. Of the 47 reviewed studies, 32 studies (approximately 68%) adhered to recommended machine learning protocols by appropriately dividing their datasets into training, validation, and test sets or by employing cross-validation techniques. The breakdown of data partitioning practices is summarized in Table 4.

Among the studies that explicitly partitioned their datasets, such as Studies #1, #6, and #7, performance metrics were reported based on the test sets, adhering to the best practices outlined by Goodfellow, Bengio, and Courville (2016). By evaluating their models on unseen data, they ensured that the models' performance accurately reflected their generalizability.

Table 4: Summary of Data Partitioning Practices Across Reviewed Studies.

Data Partitioning Practices	**Number (%) of Studies**	**Studies #**
Training / Validation /Test Split	32 (68.1%)	#1, #6, #7, #8, #10, #11, #13, #15, #16, #17, #18, #19, #21, #22, #23, #25, #26, #28, #29, #30, #32, #33, #34, #35, #36, #40, #41, #42, #43, #45, #46, #47
Cross-validation without Traditional Split	7 (14.9%)	#3, #4, #14, #24, #38, #39, #44
Inadequate or Unreported Partitioning	8 (17.0%)	#2, #5, #9, #12, #20, #27, #31, #37

Seven studies used cross-validation methods instead of a traditional train/-validation/test split. Techniques like k-fold cross-validation provide a robust assessment of a model's ability to generalize by iteratively training and testing on different subsets of the dataset (Hastie, Friedman, & Tibshirani, 2009). For instance, Study #39 utilized 5-fold cross-validation, where the dataset was divided into five subsets, with each subset used as a test set once while the remaining subsets formed the training set. The reported metrics—Positive Predictive Value (PPV), Sensitivity, and F1 Score—were averaged across the five test folds in the cross-validation process, ensuring that evaluation was based on separate test data rather than solely on the training data.

Conversely, as shown in Table 4, approximately 17% of studies (8 out of 47) did not report sufficient details on data partitioning or did not employ partitioning techniques. For example, Study #2 provided limited information about its dataset division and did not elaborate on how model performance was evaluated, while Study #5 applied pre-existing models without conducting new data partitioning or validation within their analysis, thereby limiting the validity of their performance assessments.

Inadequate data partitioning practices introduce significant risks of bias, particularly overfitting. Models that lack proper data division tend to memorize the training data, leading to overly optimistic performance metrics that do not accurately reflect real-world applicability (Bishop, 2006).

According to A. Ng (2018), proper validation and testing sets are crucial for assessing generalization and preventing overfitting. Without these, models may appear overly effective due to inflated performance metrics, misleading when applied beyond the training context. For example, studies that evaluated models solely on training data, such as Studies #2 and #5, likely overestimate their real-world performance.

In summary, while the majority of the reviewed studies adhered to best practices in data partitioning—thereby enhancing the credibility and generalizability of their findings—a significant minority did not. The lack of proper data partitioning in approximately 17% of studies introduces risks of bias, underscoring the need for more rigorous practices. For the development of robust models, future research should consistently apply proper data partitioning and report performance based on validation or test sets to provide accurate, unbiased evaluations. Transparent data partitioning and evaluation reporting, as emphasized by Bishop (2006) and Goodfellow et al. (2016), is fundamental to enhancing reproducibility and reliability in machine learning research. By incorporating these practices, researchers can enhance the reliability of their models, ensure that findings are both valid and applicable in real-world scenarios, and contribute to the advancement of the field.

3.7 Model Evaluation: Evaluation Metrics for Imbalanced Class Scenarios (Q8, Q9 & Q10)

In the domain of depression-related emotion detection, datasets often exhibit significant class imbalance, with non-depressed cases vastly outnumbering depressed ones. This imbalance poses challenges for model evaluation, as traditional metrics like accuracy can be misleading. According to He and Garcia (2009), accuracy may not adequately reflect a model's performance in imbalanced scenarios because a model could achieve high accuracy by simply predicting the majority class. Therefore, metrics such as recall, precision, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUROC or AUC) are preferred, as they provide a more balanced evaluation by accounting for both false positives and false negatives. He and Ma (2013) and Japkowicz and Stephen (2002) further emphasize the necessity of using these metrics, arguing that they are crucial for a comprehensive assessment of model performance in the presence of class imbalance.

In the context of depression detection, recall, measures the proportion of actual positive cases (individuals with depression) that the model correctly identifies (i.e., $\text{Recall} = \frac{\text{True Positive}}{\text{True Positive}+\text{False Negative}}$), is particularly important. A high recall indicates that the model is successfully identifying most individuals who are truly depressed (true positives), although this often comes at the cost of more false positives, where individuals without depression are incorrectly flagged as depressed. Failing to identify someone who is depressed (a false negative) could have serious consequences, as it may result in a missed opportunity to provide help or intervention. Therefore, prioritizing recall ensures that the model captures as many true positive cases as possible, even if it risks increasing false positives. In this context, minimizing false negatives is often a higher priority, given the potential implications for those who might otherwise go undiagnosed and unsupported (Bradford, Meyer, Khan, Giardina, & Singh, 2024).

Precision, on the other hand, measures the proportion of positive predictions that are correct (i.e., $\text{Precision} = \frac{\text{True Positive}}{\text{True Positive}+\text{False Positive}}$), highlighting the model's ability to avoid false positives. In depression detection, a low precision score indicates a high rate of false positives, where individuals who are not depressed are incorrectly labeled as depressed. This could lead to unnecessary concern or even stigmatization for those wrongly flagged. While high precision is desirable to avoid false alarms, an overly strict focus on precision could inadvertently lower recall, leading to more false negatives. Therefore, balancing precision and recall is essential to ensure that the model is not only identifying true cases of depression but also minimizing the number of false alarms. This balance is particularly critical in applications where both false negatives (missing a depressed individual) and false positives (incorrectly flagging someone as depressed) carry significant consequences (Bradford et al., 2024).

The F1 score, representing the harmonic mean of precision and recall, provides a balanced measure of both recall and precision. It is particularly useful in imbalanced datasets, where a balance between recall and precision is essential.

Finally, AUROC measures the model's ability to distinguish between positive and negative classes across different threshold settings, providing a comprehensive view of the model's discriminatory power. A higher AUROC indicates a better capability of distinguishing between depressed and non-depressed individuals, making it a robust metric for evaluating models in this domain. Among the

47 studies reviewed, approximately 35 (Studies #1, #3, #6, #7, #8, #13, #14, #15, #16, #17, #19, #21, #22, #23, #25, #26, #27, #28, #29, #30, #31, #32, #33, #34, #35, #36, #37, #39, #40, #41, #42, #43, #44, #45, #46) utilized these preferred metrics. For example, Study #6, "Depression Detection for Twitter Users Using Sentiment Analysis in English and Arabic Tweets," employed precision, recall, F1 score, and AUC to evaluate their models, acknowledging the importance of these metrics for imbalanced data. Similarly, Study #42, "Classification of Helpful Comments on Online Suicide Watch Forums," emphasized recall as a key metric in evaluating their model's effectiveness in identifying individuals at risk.

Other than the utilization of preferred metrics, an alternative way to address imbalanced data involves implementing data balancing techniques, including resampling and reweighting. For instance, Study #6 employed dynamic sampling methods, such as oversampling the minority class and undersampling the majority class, to balance the dataset. This approach ensured that the model had sufficient exposure to both classes before model construction and evaluation. Similarly, Study #41, "A Deep Learning Model for Detecting Mental Illness from User Content on Social Media," used Synthetic Minority Oversampling Technique (SMOTE) to enhance the representation of the minority class, leading to improved classification performance, particularly for underrepresented classes.

Notably, some studies (Studies #3, #6, #13, #15, #34, #40, #41, #42, #43) applied both data balancing techniques and preferred evaluation metrics together to comprehensively address the class imbalance. For example, 'Explainable Depression Detection with Multi-Aspect Features Using a Hybrid Deep Learning Model on Social Media' (Study #13) first implemented preprocessing steps to balance the dataset, enhancing the model's ability to learn from both classes equally. After addressing the class imbalance, the study then used the F1 score and related metrics to evaluate model performance, ensuring a more accurate and fair assessment. These examples indicate that researchers are increasingly aware of the class imbalance issue and are employing various approaches to address it effectively.

Conversely, some studies primarily relied on accuracy without addressing class imbalance issues. For example, Studies #2, #10, and #24 reported high accuracy but did not mention techniques to mitigate the effects of class imbalance.

In the context of depression detection, addressing class imbalance is essential for achieving reliable model evaluation. When instances of the non-depressed class significantly outnumber those of the depressed class, the resulting imbalance can skew model outcomes if not properly managed. Two primary strategies are commonly employed to mitigate this issue: the use of evaluation metrics that accommodate class imbalance and data preprocessing techniques, such as resampling and reweighting. Japkowicz and Stephen (2002) emphasize that metrics like recall, precision, and F1 score offer a more nuanced evaluation by accounting for both positive and negative classes, thus reducing potential bias. Additionally, data preprocessing methods like reweighting or resampling adjust the dataset to provide a balanced exposure to both classes, enhancing model training on imbalanced data.

While some studies utilized both strategies, demonstrating a thorough approach to handling imbalance, others employed just one—either through preferred evaluation metrics or data balancing. Even when only one strategy is adopted, it can still reduce potential bias to some extent. However, solely relying on accuracy introduces a significant risk of bias, as it often leads the model to favor the majority class, thereby failing to identify depressed individuals accurately. Chawla, Japkowicz, and Kotcz (2004) highlight that this reliance on accuracy alone can lead to misleading conclusions in imbalanced datasets, as it does not accurately reflect the model's ability to detect minority class instances.

Out of the 47 studies analyzed, approximately 35 employed preferred metrics such as F1 score, precision, recall, or AUROC, recognizing their importance in evaluating models on imbalanced datasets. Seven studies explicitly mentioned preprocessing steps like resampling to mitigate class imbalance, even when using accuracy as an evaluation metric. However, several studies relied mainly on accuracy without addressing class imbalance, potentially introducing bias into their evaluations.

In conclusion, while a significant number of studies have adopted appropriate evaluation metrics and techniques to address class imbalance, there remains a need for broader implementation of these practices. Incorporating balanced metrics and addressing class imbalance is essential for reliable and valid model evaluations in depression detection research. As Fernandez et al. (2018) recommended, employing these strategies enhances the robustness of machine learning models in domains characterized by imbalanced datasets.

3.8 Reporting: Transparency and Completeness

Transparency and completeness in reporting are fundamental to the integrity and reproducibility of scientific research. In our examination of the 47 studies, we assessed the extent to which they transparently reported their methodologies, findings, and limitations. Notably, all studies (100%) included a limitation section, indicating an overall acknowledgment of the importance of addressing potential shortcomings. However, the depth and specificity of these disclosures varied significantly across the studies.

While every study mentioned limitations, not all of them fully recognized or disclosed all critical methodological issues that could impact their findings. For instance, as highlighted in our earlier assessments, approximately 23% of the studies (11 out of 47) did not properly partition their data or failed to report their data partitioning methods adequately (Studies #2, #5, #9, #12, #20, #27, #31, and #37). Despite this, only a few of these studies explicitly acknowledged the potential biases introduced by improper data partitioning in their sections of limitations. This suggests that while researchers are generally aware of the necessity to report limitations, there is a gap in fully understanding or disclosing specific methodological shortcomings, such as data partitioning, which is crucial for model generalizability and validity.

Similarly, in the context of hyperparameter tuning, approximately 43% of the studies did not report or properly tune hyperparameters across all models used (e.g., Studies #1, #2, #4, #5, #12, #14, #17, #19, #20, #24, #27, #29, #30, #32, #34, #35, #37, #38, #42, #44, and #46). Only a few acknowledged this limitation in their reports. This lack of comprehensive reporting on hyperparameter tuning can lead to biased model comparisons and affect the reproducibility of the studies.

Incomplete or non-transparent reporting can introduce significant bias and limit the reproducibility and applicability of research findings. When critical methodological details are omitted or underreported, it hinders the ability of other researchers to replicate studies or to understand the context in which the results are valid. For instance, failing to disclose improper data partitioning can lead to overestimation of model performance due to overfitting (Bishop, 2006). Models evaluated on training data or without appropriate validation may appear to perform well, but this performance may not generalize to new, unseen data. This oversight can mislead stakeholders about the efficacy of the models and affect subsequent research or practical applications that build upon these findings.

Similarly, not reporting on hyperparameter tuning practices can result in unfair comparisons between models and misinterpretations of their relative performances (Claesen & Moor, 2015; Zhang et al., 2025). Models with optimized hyperparameters may outperform others not because they are inherently better but because they were given an optimization advantage. Without transparency in reporting these practices, readers cannot assess the fairness of the comparisons or replicate the optimization process.

In conclusion, while all 47 studies recognized the importance of reporting limitations, there remains a notable disparity in the thoroughness and transparency of their reporting. For the field to advance, transparent and comprehensive reporting of methodologies and limitations is essential. Future research should strive for complete disclosure of data collection, preprocessing, model development, hyperparameter tuning, and evaluation metrics. This includes acknowledging specific methodological limitations, such as data partitioning practices and sampling biases, and discussing how these limitations may impact results and generalizability. Such transparency will allow others to interpret findings accurately, replicate studies, and build upon prior work effectively.

3.9 Summary of Findings and Implications for Future Research

This systematic review evaluated biases throughout the entire lifecycle of machine learning and deep learning models for depression detection on social media. In sampling, biases arose from a predominant reliance on Twitter, English-language data, and specific geographic regions, limiting the representativeness of findings. Data preprocessing commonly showed inadequate handling of negations, which can skew sentiment analysis results. Model development was often compromised by inconsistent hyperparameter tuning and improper data partitioning, reducing model reliability and generalizability. Lastly, in model evaluation, an overreliance on accuracy without addressing class imbalance risked favoring majority class predictions, potentially misleading results. These findings highlight the importance of enhancing methodologies to bolster the validity and applicability of future research.

To address these biases, future research should improve practices across all stages of the machine learning lifecycle. Expanding data sources across multiple platforms, languages, and regions will help mitigate platform and language biases and improve representativeness. Standardizing data preprocessing, especially with explicit negation handling, and employing resampling and reweighting techniques will enhance sentiment analysis accuracy and balance datasets. Consistent hyperparameter tuning protocols are essential to ensure fair model comparisons and optimal performance. Lastly, prioritizing evaluation metrics like precision, recall, F1 score, and AUROC in imbalanced datasets, particularly for depression detection, will yield more accurate and insightful assessments. By implementing these improvements, future studies can achieve greater model robustness and generalizability, contributing to more effective mental health detection tools.

4 Discussion

The escalating prevalence of mental health conditions, particularly depression, poses a significant global health challenge. Social media platforms have emerged as rich data sources where individuals express their thoughts and emotions, offering a unique opportunity to detect mental health issues through advanced computational methods. Machine learning and deep learning models hold promise for analyzing this vast, unstructured data to identify patterns indicative of depression. This systematic review aimed to evaluate the effectiveness of these models in detecting depression on social media, focusing on identifying and analyzing biases throughout the ML lifecycle.

4.1 Summary of Key Findings

Our review uncovered several key biases and methodological challenges that impact the reliability and generalizability of machine learning and deep learning models in this domain. Sampling biases emerged due to a predominant reliance on specific social media platforms, particularly Twitter, which was used in 63.8% of the studies. Additionally, most studies focused on English-language content and users from specific geographic regions, primarily the United States and Europe. These biases limit the representativeness of findings, as they do not capture the diversity of global social media users. In data preprocessing, many studies inadequately handled linguistic nuances, such as negations and sarcasm. Only about 23% of the studies explicitly addressed the handling of negative words or negations, which are crucial for accurate sentiment analysis in depression detection.

Model development issues were also prominent. Inconsistent hyperparameter tuning practices were observed, with only 27.7% of the studies properly tuning hyperparameters for all models. Moreover, approximately 17% of the studies did not adequately partition their data into training, validation, and test sets. These practices can lead to overfitting, reducing the models' ability to generalize to new data. Regarding model evaluation, many studies relied heavily on accuracy as the primary evaluation metric without addressing class imbalances inherent in depression detection datasets. While about 74.5% of the studies used metrics suitable for imbalanced data, such as precision, recall, F1 score, and AU-ROC, others did not, potentially skewing the evaluation of model performance. Finally, despite all studies including a limitations section, transparency varied significantly, with critical methodological details like data partitioning methods and hyperparameter settings often underreported. This inconsistency hinders reproducibility and the ability to fully assess the validity of the findings.

4.2 Strengths and Limitations of the Review

This systematic review stands out for its comprehensive scope, examining biases across the entire ML lifecycle, from sampling to reporting, in depression detection on social media. By not limiting the analysis to specific aspects, the review offers a holistic view of how biases can influence model validity. Another strength is the structured methodological approach, adhering to established guidelines with a well-defined search strategy and clear inclusion criteria. Focusing on studies published after 2010, it reflects the latest advancements in ML and DL applications for mental health.

The use of established bias assessment tools, particularly PROBAST, adds rigor by systematically evaluating bias across key methodological domains. Additionally, the review's detailed data extraction process facilitated a structured analysis, allowing for the identification of patterns and providing actionable recommendations, such as diversifying data sources and improving transparency.

However, the review also has limitations. Limited database coverage and the English-only restriction may exclude valuable insights from non-English research, potentially affecting the generalizability of the findings. The focus on recent studies (post-2010) might have overlooked earlier influential works, while heterogeneity in study designs hindered direct comparisons and precluded a quantitative meta-analysis. Moreover, publication bias could skew findings toward positive results, and excluding grey literature means emerging methodologies may not be fully captured. Lastly, while ethical considerations were acknowledged, a deeper exploration of issues like data privacy and informed consent is warranted.

These limitations suggest areas for improvement in future research, such as broadening database and language coverage, including grey literature, and conducting a meta-analysis where feasible. By addressing these areas, future studies can enhance the robustness of ML models for mental health detection and provide a more comprehensive, ethical, and globally relevant understanding of the field.

4.3 Implications for Future Research

To enhance the generalizability and applicability of machine learning and deep learning models in depression detection on social media, addressing identified biases is essential. First, diversifying data sources across multiple social media platforms and incorporating non-English languages and underrepresented regions will improve representativeness and generalizability. Improving sampling methods is crucial. Combining keyword-based sampling with random sampling techniques can help reduce selection bias and capture users who may not explicitly mention depression but exhibit relevant behaviors. In the data preprocessing step, researchers should standardize practices to explicitly handle linguistic nuances like negations and sarcasm, which are vital for accurate sentiment analysis. Additionally, applying resampling or reweighting techniques can help balance datasets, ensuring that both classes—particularly the minority depressive class—are adequately represented. Advanced natural language processing techniques that account for linguistic nuances, such as sarcasm and context-dependent meanings, should be employed.

Consistent and comprehensive hyperparameter tuning across all models is essential to ensure fair comparisons and optimize model performance. Proper data partitioning practices, including the use of validation and test sets, should be implemented to prevent overfitting and assess model generalizability. When evaluating models, researchers should prioritize metrics that account for class imbalance, such as precision, recall, F1 score, and AUROC. These metrics provide a more balanced assessment of model performance and are more informative in the context of detecting depression, where the minority class is of primary interest.

4.4 Concluding Remarks

This systematic review highlights significant methodological limitations in current research on detecting depression through social media analysis using machine learning and deep learning models. Addressing these limitations is critical to developing more accurate, reliable, and generalizable models that can effectively identify individuals at risk of depression. Future research should focus on diversifying data sources, improving sampling methods, enhancing data preprocessing and model development practices, and employing appropriate evaluation metrics to ensure balanced and meaningful assessments.

By advancing these methodological approaches, researchers can contribute to the advancement of mental health detection tools that are ethically sound and effective across diverse populations and platforms. Such advancements hold the potential to facilitate early intervention strategies, ultimately improving mental health outcomes on a global scale.

Acknowledgments

The authors thank Dr. Jin (Veronica) Liu (ORCID: 0000-0001-5922-6643) for her valuable comments and insightful feedback on the development of this manuscript.

Her input has contributed to improving the clarity and overall presentation of the work.

Availability of Data and Materials

The reviewed titles, authors, and publication years of the included studies have been provided in Table A.1. Detailed information on each reviewed paper is hosted on GitHub: `https://github.com/odile1999/Systematic-Review-Machine-Learning-on-Depression`.

Authors' Contributions

Project administration: Y.T. and Y.C.; Conceptualization: Y.T. and Y.C.; Methodology: Y.T., J.D., and Y.C.; Investigation: Y.T., Y.C., J.D., Z.W., Y.Z., X.S., and Y.L.; Formal Analysis: Y.T., Y.C., J.D., Z.W., Y.Z., X.S., and Y.L.; Writing - Original Draft: Y.T., Y.C., and J.D.; Writing - Review and Editing: Y.C., Z.W., Y.Z., X.S., and Y.L.

References

Agarwal, A. K., Mittal, J., Tran, A., Merchant, R., & Guntuku, S. C. (2023). Investigating social media to evaluate emergency medicine physicians' emotional well-being during covid-19. *JAMA Netw Open*, *6*(5), e321708. doi: https://doi.org/10.1001/jamanetworkopen.2023.12708

Angskun, J., Tipprasert, S., & Angskun, T. (2022). Big data analytics on social networks for real-time depression detection. *J Big Data*, *9*, 69. doi: https://doi.org/10.1186/s40537-022-00622-2

Baghdadi, N. A., Malki, A., Magdy Balaha, H., Abdul-Azeem, Y., Badawy, M., & Elhossieni, M. (2022). An optimized deep learning approach for suicide detection through arabic tweets. *PeerJ Comput Sci*, *8*, e1070. doi: https://doi.org/10.7717/peerj-cs.1070

Baird, A., Xia, Y., & Cheng, Y. (2022). Consumer perceptions of telehealth for mental health or substance abuse: A twitter-based topic modeling analysis. *JAMIA Open*, *5*(2), ooac028. doi: https://doi.org/10.1093/jamiaopen/ooac028

Beier, F., Pryss, R., & Aizawa, A. (2023). Sentiments about mental health on twitter—before and during the covid-19 pandemic. *Healthcare (Basel)*, *11*(21), 2893. doi: https://doi.org/10.3390/healthcare11212893

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer. doi: https://doi.org/http://www.loc.gov/catdir/enhancements/fy0818/2006922522-d.html

Borba de Souza, V., Campos Nobre, J., & Becker, K. (2022). Dac stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts. *IEEE J Biomed Health Inform*, *26*(7), 3303-3311. doi: https://doi.org/10.1109/JBHI.2022.3151589

Bradford, A., Meyer, A. N. D., Khan, S., Giardina, T. D., & Singh, H. (2024). Diagnostic error in mental health: a review. *BMJ Quality & Safety*, *33*(10), 663–672. doi: https://doi.org/https://qualitysafety.bmj.com/content/33/10/663

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, *23*(5), 649-685. doi: https://doi.org/10.1017/S1351324916000383

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annu Rev Clin Psychol*, *12*, 53-81. doi: https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chandra, R., & Krishna, A. (2021). Covid-19 sentiment analysis via deep learning during the rise of novel cases. *PLOS ONE*, *16*(8), e0255615. doi: https://doi.org/10.1371/journal.pone.0255615

Chawla, N., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations*, *6*(1), 1-6. doi: https://doi.org/10.1145/1007730.1007733

Chen, L., Gong, T., Kosinski, M., Stillwell, D., & Davidson, R. L. (2017). Building a profile of subjective well-being for social media users. *PLOS ONE*, *12*(11), e0187278. doi: https://doi.org/10.1371/journal.pone.0187278

Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Comput Biol Med*, *135*, 104499. doi: https://doi.org/10.1016/j.compbiomed.2021.104499

Claesen, M., & Moor, B. D. (2015). *Hyperparameter search in machine learning.* Retrieved from `https://arxiv.org/abs/1502.02127`

Danet, B., & Herring, S. C. (2007). *The multilingual internet: Language, culture, and communication online.* Oxford University Press.

Das Swain, V., Ye, J., Ramesh, S. K., Mondal, A., Abowd, G. D., & De Choudhury, M. (2024). Leveraging social media to predict covid-19-induced disruptions to mental well-being among university students: Modeling study. *JMIR Form Res*, *8*, e52316. doi: https://doi.org/10.2196/52316

Davis, B. D., McKnight, D. E., Teodorescu, D., Quan-Haase, A., Chunara, R., Fyshe, A., & Lizotte, D. J. (2020). Quantifying depression-related language on social media during the covid-19 pandemic. *Int J Popul Data Sci*, *5*(4), 1716. doi: https://doi.org/10.23889/ijpds.v5i4.1716

de Anta, L., Alvarez-Mon, M. A., Ortega, M. A., Salazar, C., Donat-Vargas, C., Santoma-Vilaclara, J., ... Alvarez-Mon, M. (2022). Areas of interest and social consideration of antidepressants on english tweets: A natural language processing classification study. *Journal of Personalized Medicine*, *12*(2), 20155. doi: https://doi.org/10.3390/jpm12020155

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement of depression in populations. In *Proceedings of the acm annual web science conference* (p. 47-56). New York, NY, USA. doi: https://doi.org/10.1145/2464464.2464480

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, *abs/1810.04805*. doi: https://doi.org/https://arxiv.org/abs/1810.04805

Doan, S., Ritchart, A., Perry, N., Chaparro, J. D., & Conway, M. (2017). How do you #relax when you're #stressed? a content analysis and infodemiology study of stress-related tweets. *JMIR Public Health Surveill*, *3*(2), e35. doi: https://doi.org/10.2196/publichealth.5939

Fernandez, A., Garcia, S., Galar, M., Prati, R., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets.* Springer. doi: https://doi.org/10.1007/978-3-319-98074-4

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* The MIT Press.

Guntuku, S. C., Schneider, R., Pelullo, A., Young, J., Wong, V., Ungar, L., ... Merchant, R. (2019). Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ Open*, *9*(1), e030355. doi: https://doi.org/10.1136/bmjopen-2019-030355

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49. doi: https://doi.org/10.1016/j.cobeha.2017.07.005

Hargittai, E. (2015). Is bigger always better? potential biases of big data derived from social network sites. *The Annals of the American Academy of Political and Social Science*, *659*, 63-76. doi: https://doi.org/http://www.jstor.org/stable/24541849

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. doi: https://doi.org/10.1080/00437956.1954.11659520

Hastie, T., Friedman, J. H., & Tibshirani, R. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. doi: https://doi.org/10.1109/TKDE.2008.239

He, H., & Ma, Y. (2013). Assessment metrics for imbalanced learning. In *Imbalanced learning: Foundations, algorithms, and applications* (p. Chapter Reference (add specific pages if available)). Wiley-IEEE Press. (Book Chapter) doi: https://doi.org/10.1002/9781118646106.ch8

Helmy, A., Nassar, R., & Ramadan, N. (2024). Depression detection for twitter users using sentiment analysis in english and arabic tweets. *Artificial Intelligence in Medicine*, *147*, 102716. doi: https://doi.org/10.1016/j.artmed.2023.102716

Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhag, A. (2018). Depression detection from social network data using machine learning techniques. *Health Inf Sci Syst*, *6*(1), 8. doi: https://doi.org/10.1007/s13755-018-0046-0

Jamali, A. A., Berger, C., & Spiteri, R. J. (2023). Momentary depressive feeling detection using x (formerly twitter) data: Contextual language approach.

JMIR AI, *2*, e49531. doi: https://doi.org/10.2196/49531

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intell Data Anal*, *6*, 429-449. doi: https://doi.org/10.3233/IDA-2002-6504

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022, June). MentalBERT: Publicly available pretrained language models for mental healthcare. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 7184–7190). Marseille, France: European Language Resources Association. doi: https://doi.org/https://aclanthology.org/2022.lrec-1.778/

Kaur, R., Ahassan, S. U., Alankar, B., & Chang, V. (2021). A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets. *Inf Syst Front*, *23*(6), 1417-1429. doi: https://doi.org/10.1007/s10796-021-10135-7

Kavuluru, R., Williams, A. G., Ramos-Morales, M., Haye, L., Holaday, T., & Cerel, J. (2016). Classification of helpful comments on online suicide watch forums. In *Proceedings of the 7th acm conference on bioinformatics, computational biology, and health informatics (acm-bcb)* (pp. 32–40). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/2975167.2975170

Kelley, S. W., Monaghan, C. N., Burke, L., Whelan, R., & Gillan, C. M. (2022). Machine learning for anxiety language on twitter reveals weak and non-specific predictions. *NPJ Digit Med*, *5*, 35. doi: https://doi.org/10.1038/s41746-022-00576-y

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E. J., Cardoso, G., … Koenen, K. C. (2017). Trauma and ptsd in the who world mental health surveys. *European Journal of Psychotraumatology*, *8*(sup5), 1353383. doi: https://doi.org/10.1080/20008198.2017.1353383

Khandelwal, A., & Sawant, S. (2020, May). NegBERT: A transfer learning approach for negation detection and scope resolution. In N. Calzolari et al. (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 5739–5748). Marseille, France: European Language Resources Association. doi: https://doi.org/https://aclanthology.org/2020.lrec-1.704/

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Sci Rep*, *10*, 18446. doi: https://doi.org/10.1038/s41598-020-68764-y

Kumar, A., Khan, S. U., & Kalra, A. (2020). Covid-19 pandemic: a sentiment analysis. *European Heart Journal*, *41*(39), 3782–3783. doi: https://doi.org/10.1093/eurheartj/ehaa597

Levanti, D., Monastero, R. N., Zamani, M., Eichstaedt, J. C., Giorgi, S., Schwartz, H. A., & Meilxer, J. R. (2023). Depression and anxiety on twitter during the covid-19 stay-at-home period in 7 major u.s. cities. *AJPM Focus*, *2*, 100062. doi: https://doi.org/10.1016/j.focus.2022.100062

Li, Y., Cai, M., Qin, S., & Lu, X. (2020). Depressive emotion detection and behavior analysis of men who have sex with men via social media. *Front Psychiatry*, *11*, 830. doi: https://doi.org/10.3389/fpsyt.2020.00830

Low, D. M., Munoz, F. L., Talkar, T., Torres, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened anxiety on reddit during covid-19. *JMIR Ment Health*, *8*(6), e22635. doi: https://doi.org/10.2196/22635

Mantovani, R. G., Rossi, A. L. D., Vanschoren, J., Bischl, B., & de Carvalho, A. C. P. L. F. (2015). Effectiveness of random search in svm hyper-parameter tuning. In *Proceedings of the 2015 international joint conference on neural networks (ijcnn)* (pp. 1–8). IEEE. doi: https://doi.org/https://ieeexplore.ieee.org/document/7280664

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter firehose. In *Proceedings of the 7th international conference on weblogs and social media, icwsm 2013* (pp. 400–408).

Ng, A. (2018). *Machine learning yearning.* doi: https://doi.org/https://info.deeplearning.ai/machine-learning-yearning-book

Ng, Q. X., Lim, Y. L., Ong, C., New, S., Fam, J., & Liew, T. M. (2024). Hype or hope? ketamine for the treatment of depression: results from the application of deep learning to twitter posts from 2010 to 2023. *Front Psychiatry*, *15*, 1369727. doi: https://doi.org/10.3389/fpsyt.2024.1369727

Obagbuwa, I. C., Danster, S., & Chibaya, O. C. (2023). Supervised machine learning models for depression sentiment analysis. *Front Artif Intell*, *6*, 1230649. doi: https://doi.org/10.3389/frai.2023.1230649

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Front Big Data*, *2*, 13. doi: https://doi.org/10.3389/fdata.2019.00013

Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. *Sci Rep*, *10*(1), 16685. doi: https://doi.org/10.1038/s41598-020-73917-0

Owen, D., Antypas, D., Hassoulas, A., Pardinas, A. F., Espinosa-Anke, L., & De Choudhury, M. (2023). Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation. *JMIR AI*, *2*, e41205. doi: https://doi.org/10.2196/41205

Patel, D., Sumner, S. A., Bowen, D., Zwald, M., Yard, E., Wang, J., … Chen, Y. (2023). Predicting state-level suicide fatalities in the united states with real-time data and machine learning. *NPJ Ment Health Res*, *3*(1), 3. doi: https://doi.org/10.1038/s44184-023-00045-8

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Association for Computational Linguistics. doi: https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 2227–2237). Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N18-1202

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: A good place to detect health conditions. *PLoS One*, *9*(1), e86191. doi: https://doi.org/10.1371/journal.pone.0086191

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(53), 1-32. doi: https://doi.org/http://jmlr.org/papers/v20/18-444.html

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk of suicidal ideation from social media data. *NPJ Digit Med*, *3*, 78. doi: https://doi.org/10.1038/s41746-020-0287-6

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(6), 1121-1133. doi: https://doi.org/10.1080/02699930441000030

Saha, K., Yousuf, A., Boyd, R. L., Pennebaker, J. W., & De Choudhury, M. (2022). Social media discussions predict mental health consultations on college campuses. *Sci Rep*, *12*(1), 123. doi: https://doi.org/10.1038/s41598-021-03423-4

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi: https://doi.org/10.1016/0306-4573(88)90021-0

Shatte, A. B. R., Hutchinson, D.-M., Fuller-Tyszkiewicz, M., & Teague, S. J. (2020). Social media markers to identify fathers at risk of postpartum depression: A machine learning approach. *Cyberpsychol Behav Soc Netw*, *23*(9), 611-618. doi: https://doi.org/10.1089/cyber.2019.0746

Shatte, A. B. R., Hutchinson, D.-M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(8), 1426-1448. doi: https://doi.org/10.1017/S0033291719000151

Singh, A., & Singh, J. (2022). Synthesis of affective expressions and artificial intelligence to discover mental distress in online community. *Int J Ment Health Addict*, 1-26. doi: https://doi.org/10.1007/s11469-022-00966-z

Sun, B., Zhang, Y., He, J., Xiao, Y., & Xiao, R. (2019). An automatic diagnostic network using skew-robust adversarial discriminative domain adaptation to evaluate the severity of depression. *Comput Methods Programs Biomed*, *173*, 185-195. doi: https://doi.org/10.1016/j.cmpb.2019.01.006

Swapnarekha, H., Nayak, J., Behera, H. S., Dash, P. B., & Pelusi, D. (2023). An optimistic firefly algorithm-based deep learning approach for sentiment analysis of covid-19 tweets. *Math Biosci Eng*, *20*(2), 2582-2607. doi:

https://doi.org/10.3934/mbe.2023112

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *27*(5), 1–53. doi: https://doi.org/10.1145/3398069

Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big data approach. *Behav Res Methods*, *51*(4), 1586-1600. doi: https://doi.org/10.3758/s13428-019-01235-z

Trifan, A., Semeraro, D., Drake, J., Bukowski, R., & Oliveira, J. L. (2020). Social media mining for postpartum depression prediction. *Stud Health Technol Inform*, *270*, 1391-1392. doi: https://doi.org/10.3233/SHTI200457

Ueda, M., Watanabe, K., & Sueki, H. (2023). Correction: Emotional distress during covid-19 by mental health conditions and economic vulnerability: Retrospective analysis of survey-linked twitter data with a semi-supervised machine learning algorithm. *J Med Internet Res*, *25*, e759. doi: https://doi.org/10.2196/47549

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems (neurips)* (pp. 5998–6008). Curran Associates Inc. doi: https://doi.org/https://arxiv.org/abs/1706.03762

WHO. (2023). *Mental disorders.* (Retrieved February 2, 2025) doi: https://doi.org/https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., . . . Group, P. (2019). Probast: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. doi: https://doi.org/10.7326/M18-1376

Wongkoblap, A. (2023). Automatic profiles collection from twitter users with depressive symptoms. *Stud Health Technol Inform*, *305*, 419-422. doi: https://doi.org/10.3233/SHTI230520

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2021). Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study. *JMIR Ment Health*, *8*(6), e19624. doi: https://doi.org/10.2196/19824

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316. doi: https://doi.org/10.1016/j.neucom.2020.07.061

Yao, H., Rashidian, S., Dong, X., Duanmu, H., Rosenthal, R. N., & Wang, F. (2020). Detection of suicidality among opioid users on reddit: Machine learning-based approach. *J Med Internet Res*, *22*(21), e15293. doi: https://doi.org/10.2196/15293

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., . . . Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the ieee/acm international conference on advances in social networks analysis and mining* (p. 1191-1198). doi: https://doi.org/10.1145/3110025.3123028

Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., . . . Hitzler, P. (2020). Multimodal mental health analysis in social media. *PLoS One*, *15*(8). doi: https://doi.org/10.1371/journal.pone.0226248

Yin, Z., Fabbri, D., Rosenbloom, S. T., & Malin, B. (2015). A scalable framework to detect personal health mentions on twitter. *J Med Internet Res*, *17*(6), e138. doi: https://doi.org/10.2196/jmir.4305

Zhang, Y., Wang, Z., Ding, Z., Tian, Y., Dai, J., Shen, X., . . . Cao, Y. (2025). *Tutorial on using machine learning and deep learning models for mental illness detection.* Retrieved from https://arxiv.org/abs/2502.04342

Zhou, T. H., Hu, G. L., & Wang, L. (2019). Psychological disorder identifying method based on emotion perception over social networks. *Int J Environ Res Public Health*, *16*(6). doi: https://doi.org/10.3390/ijerph16060953

Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, *25*(1), 281-304. doi: https://doi.org/10.1007/s11280-021-00992-2

Appendix: Reviewed Studies on Machine Learning Models for Depression Detection on Social Media

Table A1: Reviewed Studies on Machine Learning Models for Depression Detection on Social Media

Index	Title of the Paper	Reference
#1	Machine learning of language use on Twitter reveals weak and non-specific predictions	Kelley, Monaghan, Burke, Whelan, and Gillan (2022)
#2	Supervised machine learning models for depression sentiment analysis	Obagbuwa, Danster, and Chibaya (2023)
#3	A textual-based featuring approach for depression detection using machine learning classifiers and social media texts	Chiong, Budhi, Dhakal, and Chiong (2021)
#4	Emotional Distress During COVID-19 by Mental Health Conditions and Economic Vulnerability: Retrospective Analysis of Survey-Linked Twitter Data With a Semisupervised Machine Learning Algorithm	Ueda, Watanabe, and Sueki (2023)

Continued on next page

Index	Title of the Paper	Reference
#5	Depression and Anxiety on Twitter During the COVID-19 Stay-At-Home Period in 7 Major U.S. Cities	Levanti et al. (2023)
#6	Depression detection for Twitter users using sentiment analysis in English and Arabic tweets	Helmy et al. (2024)
#7	Deep Learning With Anaphora Resolution for the Detection of Tweeters With Depression: Algorithm Development and Validation Study	Wongkoblap, Vadillo, and Curcin (2021)
#8	Sentiments about Mental Health on Twitter-Before and during the COVID-19 Pandemic	Beier, Pryss, and Aizawa (2023)
#9	Hype or hope? Ketamine for the treatment of depression: results from the application of deep learning to Twitter posts from 2010 to 2023	Q. X. Ng et al. (2024)
#10	Quantifying depression-related language on social media during the COVID-19 pandemic	Davis et al. (2020)
#11	Predicting state-level suicide fatalities in the United States with realtime data and machine learning	Patel et al. (2023)
#12	Investigating Social Media to Evaluate Emergency Medicine Physicians' Emotional Well-being During COVID-19	Agarwal, Mittal, Tran, Merchant, and Guntuku (2023)
#13	Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media	Zogan, Razzak, Wang, Jameel, and Xu (2022)
#14	Big data analytics on social networks for real-time depression detection	Angskun, Tipprasert, and Angskun (2022)
#15	An optimized deep learning approach for suicide detection through Arabic tweets	Baghdadi et al. (2022)
#16	COVID-19 sentiment analysis via deep learning during the rise of novel cases	Chandra and Krishna (2021)
#17	A Scalable Framework to Detect Personal Health Mentions on Twitter	Yin, Fabbri, Rosenbloom, and Malin (2015)
#18	An automatic diagnostic network using skew-robust adversarial discriminative domain adaptation to evaluate the severity of depression	Sun, Zhang, He, Xiao, and Xiao (2019)
#19	Twitter: a good place to detect health conditions	Prieto, Matos, Alvarez, Cacheda, and Oliveira (2014)
#20	Consumer perceptions of telehealth for mental health or substance abuse: A Twitter-based topic modeling analysis	Baird, Xia, and Cheng (2022)

Continued on next page

Index	Title of the Paper	Reference
#21	Depressive Emotion Detection and Behavior Analysis of Men Who Have Sex With Men via Social Media	Li, Cai, Qin, and Lu (2020)
#22	Areas of Interest and Social Consideration of Antidepressants on English Tweets: A Natural Language Processing Classification Study	de Anta et al. (2022)
#23	An Optimistic Firefly Algorithm-Based Deep Learning Approach for Sentiment Analysis of COVID-19 Tweets	Swapnarekha, Nayak, Behera, Dash, and Pelusi (2023)
#24	How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets	Doan, Ritchart, Perry, Chaparro, and Conway (2017)
#25	Psychological Disorder Identifying Method Based on Emotion Perception over Social Networks	Zhou, Hu, and Wang (2019)
#26	Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach	Jamali et al. (2023)
#27	A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets	Kaur, Ahassan, Alankar, and Chang (2021)
#28	A machine learning approach predicts future risk to suicidal ideation from social media data	Roy et al. (2020)
#29	Studying expressions of loneliness in individuals using Twitter: an observational study	Guntuku et al. (2019)
#30	Automatic Profiles Collection from Twitter Users with Depressive Symptoms	Wongkoblap (2023)
#31	Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media	Yazdavar et al. (2017)
#32	Synthesis of Affective Expressions and Artificial Intelligence to Discover Mental Distress in Online Community	Singh and Singh (2022)
#33	DAC Stacking: A Deep Learning Ensemble to Classify Anxiety, Depression, and Their Comorbidity from Reddit Texts	Borba de Souza, Campos Nobre, and Becker (2022)
#34	A textual-based featuring approach for depression detection using machine learning classifiers and social media texts	Chiong et al. (2021)
#35	Detection of Suicidality Among Opioid Users on Reddit: Machine Learning-Based Approach	Yao et al. (2020)
#36	Social Media Markers to Identify Fathers at Risk of Postpartum Depression: A Machine Learning Approach	Shatte, Hutchinson, Fuller-Tyszkiewicz, and Teague (2020)

Continued on next page

Index	Title of the Paper	Reference
#37	Social Media Mining for Postpartum Depression Prediction	Trifan, Semeraro, Drake, Bukowski, and Oliveira (2020)
#38	Social Media Discussions Predict Mental Health Consultations on College Campuses	Saha, Yousuf, Boyd, Pennebaker, and De Choudhury (2022)
#39	Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation	Owen et al. (2023)
#40	Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study	Low et al. (2020)
#41	A deep learning model for detecting mental illness from user content on social media	Kim, Lee, Park, and Han (2020)
#42	Classification of Helpful Comments on Online Suicide Watch Forums	Kavuluru et al. (2016)
#43	Predicting future mental illness from social media: A big-data approach	Thorstad and Wolff (2019)
#44	Depression detection from social network data using machine learning techniques	Islam et al. (2018)
#45	Deep neural networks detect suicide risk from textual Facebook posts	Ophir, Tikochinski, Asterhan, Sisso, and Reichart (2020)
#46	Leveraging Social Media to Predict COVID-19-Induced Disruptions to Mental Well-Being Among University Students: Modeling Study	Das Swain et al. (2024)
#47	Building a profile of subjective well-being for social media users	Chen, Gong, Kosinski, Stillwell, and Davidson (2017)