electronics

MDPI

Article

Advanced Comparative Analysis of Machine Learning and Transformer Models for Depression and Suicide Detection in Social Media Texts

Biodoumoye George Bokolo * and Qingzhong Liu

Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA; qxl005@shsu.edu
* Correspondence: bgb023@shsu.edu

Abstract: Depression detection through social media analysis has emerged as a promising approach for early intervention and mental health support. This study evaluates the performance of various machine learning and transformer models in identifying depressive content from tweets on X. Utilizing the Sentiment140 and the Suicide-Watch dataset, we built several models which include logistic regression, Bernoulli Naive Bayes, Random Forest, and transformer models such as RoBERTa, DeBERTa, DistilBERT, and SqueezeBERT to detect this content. Our findings indicate that transformer models outperform traditional machine learning algorithms, with RoBERTa and DeBERTa, when predicting depression and suicide rates. This performance is attributed to the transformers' ability to capture contextual nuances in language. On the other hand, logistic regression models outperform transformers in another dataset with more accurate information. This is attributed to the traditional model's ability to understand simple patterns especially when the classes are straighforward. We employed a comprehensive cross-validation approach to ensure robustness, with transformers demonstrating higher stability and reliability across splits. Despite limitations like dataset scope and computational constraints, the findings contribute significantly to mental health monitoring and suggest promising directions for future research and real-world applications in early depression detection and mental health screening tools. The various models used performed outstandingly.

Keywords: depression detection; transformers; social media analysis; mental health monitoring; machine learning; tweet analysis; roBERTa; deBERTa; logistics regression; random forest; sentiment; suicide detection

check for updates

Citation: Bokolo, B.G.; Liu, Q. Advanced Comparative Analysis of Machine Learning and Transformer Models for Depression and Suicide Detection in Social Media Texts. *Electronics* **2024**, *13*, 3980. https://doi.org/10.3390/electronics13203980

Academic Editor: Daniele Riboni

Received: 11 July 2024
Revised: 20 September 2024
Accepted: 24 September 2024
Published: 10 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Depression is a widespread mental health disorder that impacts a massive number of people around the globe [1], causing severe problems for both individual happiness and the health of society as a whole. Despite its high prevalence, depression often goes undiagnosed and untreated and could even lead to suicide, mainly due to the negative social perception of mental health problems and the lack of accessible diagnostic tools [2]. Traditional methods of diagnosing depression typically involve self-report questionnaires and clinical interviews, which, while effective, are time-consuming and reliant on the availability of trained professionals [3,4]. The rapid growth of social networking sites, primarily in the last decade, has provided a new avenue for detecting mental health issues, offering a unique opportunity for more timely and scalable depression screening.

Social media platforms such as X, Facebook, and Instagram have become interwoven into the daily lives of a vast segment of the population [5]. Figure 1 illustrates the number of worldwide active users of the top social media platforms currently.

In 2024, 70.1% of the U.S. population, approximately 239 million people, actively use social media. Gender differences show that 78% of females and 66% of males are on these platforms. Women tend to use Snapchat and Pinterest more, while men prefer YouTube and X (formerly Twitter). Social media use by ethnicity reveals that 80% of Hispanics,

77% of Black Americans, and 69% of White Americans are active users. Age-wise, social media is most popular among Gen Z and Millennials, with 84% of those aged 18–29 and 81% of those aged 30–49 using it, followed by 73% of those aged 50–64, and 45% of those 65 and older.

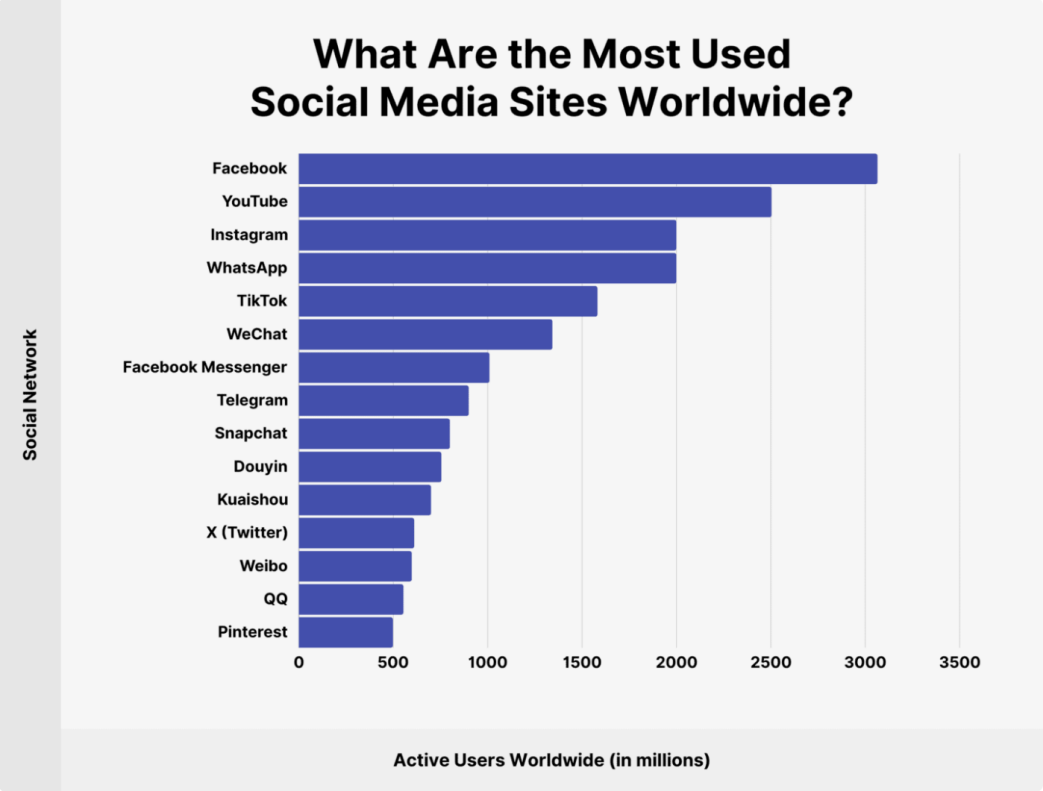


Figure 1. Worldwide active users of select social media sites [6].

These platforms serve as rich sources of unfiltered personal expression, where individuals frequently share their thoughts, emotions, and daily experiences. This constant stream of user-generated content has the potential to be harnessed for mental health monitoring, allowing for the early detection of depressive symptoms through the analysis of language patterns and behaviors indicative of depression.

Suicide is a serious global public health issue, affecting millions of individuals each year. It is often driven by complex factors, with depression being a major contributor. Persistent feelings of hopelessness, isolation, and emotional pain caused by depression can lead individuals to view suicide as a means of escape. The early detection of depression and intervention are key to preventing suicide, making tools that identify signs of suicide ideation, particularly through online behavior, increasingly important.

In this research, we will leverage Twitter (X) tweets and Reddit as the primary data source for our analysis. These social media platforms are particularly well-suited for this purpose for several reasons. First, they have a large and diverse user base, providing a wealth of data that reflects a wide range of personal experiences and emotions. Second, the platforms' public nature means that much of their content is readily accessible for analysis, unlike some other social media sites that may have stricter privacy settings. Third, the brevity of posts in both platforms (with word limits) encourages users to express their thoughts succinctly, often resulting in candid and spontaneous expressions of their mental state. These factors make X and Reddit ideal platforms for capturing real-time, authentic expressions of depressive symptoms

The application of deep learning techniques to social media data has shown promise in identifying depression-related signals. Early studies have demonstrated that linguistic features such as word choice, sentence structure, and sentiment can indicate an individual's

mental state. Traditional ML approaches, including support vector machines and logistic regression [7,8], have been employed with some success in this domain. However, these methods often require extensive feature engineering and may struggle to capture human language's nuanced and context-dependent nature.

Recent advancements in deep learning, particularly the development of transformer models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have revolutionized NLP by enabling models to understand and generate human language with unprecedented accuracy. Unlike previous architectures, these models use "attention mechanisms" to understand how words relate to each other within a sentence, leading to a deeper grasp of meaning. This breakthrough has propelled transformer models to the top of the game in various language processing tasks, from categorizing text to analyzing emotions and named entity recognition. These advancements make them well-suited for the complex task of depression detection.

Our study aims to advance the field of automated depression screening by developing a deep learning framework utilizing transformer models for social media analysis and comparing their performance against traditional machine learning techniques. The primary objectives of our study are threefold:

- To construct a comprehensive dataset of tweets labeled with indicators of depression;
- To design and implement both transformer-based models and traditional ML models capable of identifying depressive symptoms in social media texts;
- To conduct a comparative evaluation of these models to determine their relative effectiveness and practical applicability.

This is how the rest of the paper is structured. Section 2 examines the relevant exisitng works on depression detection using social media texts, highlighting key findings and methodological approaches. Section 3 describes the research method and outlines the architecture of the proposed transformer-based and traditional ML frameworks—detailing the model selection, training procedures, and hyperparameter tuning strategies. Section 4 presents the experimental results, comparing the performance of our transformer models with that of baseline ML models. Finally, Section 5 discusses the implications of our findings, potential limitations, and future directions for research in automated mental health screening.

By leveraging the power of transformers and the rich, real-time data available on Twitter (X), this research endeavors to contribute to developing more accurate and scalable tools for depression screening. Through a rigorous comparative evaluation of advanced ML techniques, we aim to identify the most effective models for early intervention and support for individuals suffering from this debilitating condition. By integrating advanced ML techniques with social media analysis, we hope to pave the way for innovative solutions to address the growing mental health crisis in our digital age.

2. Literature Review

2.1. Social Media and Mental Health

Social media platforms' rapid proliferation and extensive adoption have transformed individual communication, self-expression, and experience sharing. This digital revolution has concurrently opened novel avenues for mental health research and intervention, as social media data provide insightful information on the emotional states and behavioral habits of users [9,10]. Social media sites such as X, Facebook, and Reddit have become integral to contemporary society, boasting billions of active users globally. These platforms function as digital extensions of individuals' social lives, enabling the real-time sharing of thoughts, feelings, and experiences. The vast corpus of user-generated content on these platforms serves as a rich data source for understanding and monitoring mental health conditions, including depression [11,12].

Numerous studies have established that the language and content shared on social media can reflect users' emotional states and psychological well-being. Social media is frequently utilized to express thoughts, emotions, and experiences, which can provide important information about a person's mental health condition [13,14]. For instance, individuals experiencing depressive symptoms may exhibit changes in their language use, such as an increased frequency of negative sentiment words, references to hopelessness or worthlessness, and decreased social engagement [11,12]. Moreover, social media data can capture the temporal and contextual dynamics of an individual's mental health, enabling researchers and clinicians to spot subtle signs and trends that conventional clinical evaluations can miss [13,14]. By analyzing the content, sentiment, and behavioral patterns of social media users over time, depression's early warning indicators can be identified, along with other mental health issues, potentially facilitating timely intervention and support [11,12].

2.2. Machine Learning for Depression Detection

Research in mental health has increasingly turned to machine learning techniques for detecting depression from social media texts. Various studies have explored using supervised machine learning algorithms to analyze text data from platforms like Facebook and Twitter to identify markers of depression [15]. These algorithms have shown promise in capturing subtle linguistic cues and emotional patterns indicative of depressive symptoms in user-generated content.

In [16], the authors developed a model for depression analysis by creating correlations between textual features and depressive indicators. Ashraf, Gunawan, Riza, Haryanto and Janin (2020) [17] reviewed image and video-based models for depression detection, highlighting the relevance of visual cues. Additionally, another study [18] investigated the use of big data analytics on social networks for the instantaneous identification of depression. They explored machine learning techniques such as Survey Vector Machines (SVMs), Decision Tree, Naïve Bayes, and Random Forest, highlighting the potential of these techniques in processing large volumes of social media data for mental health tracking and monitoring. These studies prove the utility of text and social media data in capturing real-time expressions of mental health states, offering a scalable approach to depression screening.

Another methodological approach involves sentiment analysis and the examination of behavioral patterns. Angskun et al. [18] employed machine learning models, including SVMs and logistic regression, to predict depression levels in social media posts, demonstrating high accuracy and efficiency. Similarly, in the work of Obagbuwa et al. [7], the authors utilized machine learning to detect depression through network behavior and tweet analysis, employing classifiers like KNN, Adaboost, and Naive Bayes. These studies emphasize the effectiveness of supervised machine learning models and sentiment analysis in identifying depressive indicators based on user behavior and content analysis.

While traditional machine learning algorithms have demonstrated effectiveness in detecting depression from social media data, they come with certain limitations. One common challenge is the need for manual feature engineering, where researchers have to define and extract relevant features from the text data to train the models effectively [13,14]. This process can be time-consuming and may not capture all the nuanced aspects of language use that signal depression. Additionally, the interpretability of these models may be limited, making it challenging to understand the underlying mechanisms through which they identify depressive markers in texts.

Despite these limitations, the utilization of machine learning algorithms for detecting depression in social media posts represents a significant advancement in mental health research. By leveraging the power of computational techniques to analyze vast amounts of textual data, researchers can uncover insightful information into individual's mental health states and potentially revolutionize early detection and intervention strategies for depression. Further advancements in machine learning models, particularly those

incorporating deep learning techniques, promise to enhance the accuracy and efficiency of depression detection from social media texts.

2.3. Transformer Applications in Depression Detection

With the advent of transformers, the NLP research field has advanced significantly. Introduced by Vaswani et al. in 2017 [19], transformers have revolutionized how we approach text analysis by leveraging self-attention mechanisms to process sequential data efficiently and effectively [20]. Unlike traditional neural networks (like RNNs and LSTM), transformers do not rely on recurrent connections, which can lead to vanishing gradients and computational complexity. Instead, transformers use parallelized self-attention layers to model complex relationships between input elements, enabling them to capture long-range dependencies and contextual information accurately.

The impact of transformers on NLP has been profound. They have been successfully applied to various tasks, including machine translation, text classification, and sentiment analysis. In the context of detecting depression from social media posts, transformers have shown remarkable promise in capturing subtle linguistic cues and emotional patterns indicative of depressive symptoms [21].

Ilias et al. [22] introduced a novel method that enhances BERT and MentalBERT by integrating additional linguistic information using feature vectors like the NRC Emotion Lexicon and LIWC. This approach and label smoothing for better calibration significantly improved model performance across multiple datasets. In [23], the authors explored various large pre-trained language models—BERT, RoBERTa, BERTweet, and MentalBERT—fine-tuned for depression detection using social media posts. They demonstrated that the transformer ensembles outperformed individual models, particularly in datasets from Reddit and Twitter. This study underscored the importance of ensemble methods and transfer learning for improving model generalization and detection performance.

2.4. Overview

Our review of related works demonstrated the potential of machine learning and transformer models in detecting depression through various data sources, the efficacy of ensemble methods, and the importance of integrating additional linguistic information to improve model performance. Studies have highlighted the success of transformer-based models in enhancing detection accuracy but also pointed out the need for more comprehensive evaluations and the better handling of data variability.

Our research is motivated by the critical need to improve early detection and intervention for depression through advanced computational methods. By providing a thorough comparative analysis of transformer and machine learning models and integrating novel techniques, our study aims to contribute significantly to automated depression detection. This research can inform the development of more effective, scalable, and non-invasive tools for mental health monitoring, ultimately supporting timely interventions and better mental health outcomes.

3. Methodology

This section details the methodological framework employed to investigate the potential of deep learning for depression detection from social media texts. We present a comparative evaluation of machine learning and transformer-based techniques, outlining the data acquisition, preprocessing, model selection, training, and assessment processes. Figure 2 summarizes the methodological approach we adopted in this paper, which is explained in detail in the subsequent subsections.

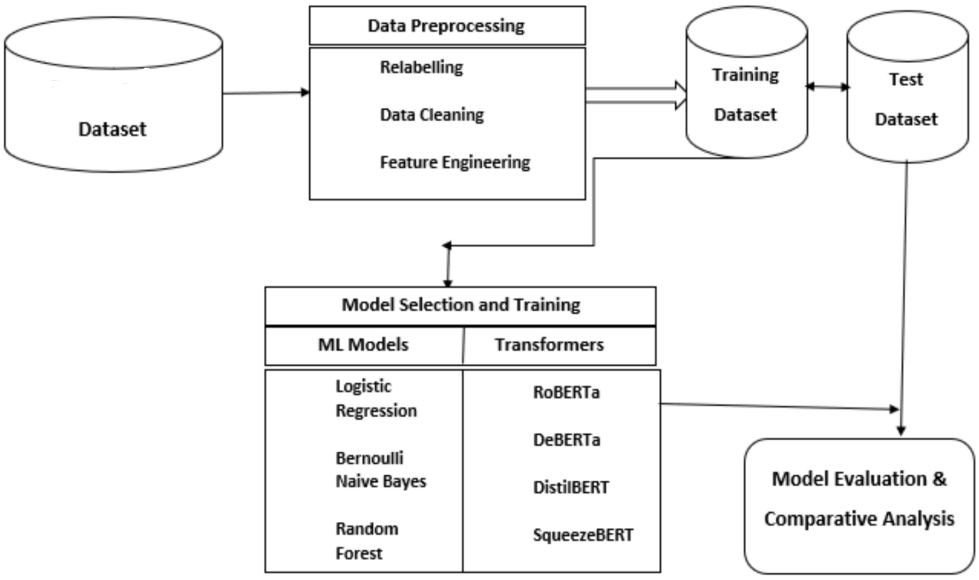


Figure 2. Overview of our methodological approach.

3.1. Dataset Description

For this research study, we utilized the Sentiment140 dataset initially intended for sentiment analysis, and we repurposed it to label tweets as depressed or non-depressed. Then, the second dataset we used was the Suicide-Watch dataset, an extensively used dataset for mental health analysis.

3.1.1. Sentiment140 Dataset

This dataset was first introduced in the study by Go et al. [24]. Gathering tweets via the official X Developer API, which requires a subscription, proved to be cost-prohibitive and tim-consuming for this research. Therefore, we opted for this existing dataset to facilitate our research.

The Sentiment140 dataset is a corpus of English-language tweets scraped from X between April and June 2009. These tweets captured users' thoughts and opinions on a wide range of subjects during this period. Sentiment polarity labels are applied to every tweet in the dataset, categorized as either positive or negative.

To repurpose this dataset for our task of depression detection, we implemented a custom labeling algorithm that re-labeled the sentiment categories to align with our research objectives. Specifically, tweets initially labeled as unfavorable were re-labeled as 'depressed', and tweets labeled as positive were re-labeled as 'non-depressed'. This conversion was guided by the theoretical correlation between negative sentiment and depressive language, supported by the existing literature in the field [25–27].

The Sentiment140 dataset provides a source of linguistic data that enables the exploration of depression-related language patterns. However, we acknowledge that not all negative sentiments equate to clinical depression. Therefore, additional validation steps were undertaken, including manual reviews and consultations with mental health experts, to ensure the accuracy and relevance of the re-labeled data. Overall, the Sentiment140 dataset, with its substantial volume and varied content, offers a valuable foundation for developing and evaluating machine learning and transformer models aimed at detecting depression from social media texts. Table 1 briefly summarises the properties of the dataset.

Table 1. Sentiment140 dataset description table.

Characteristic	Value
Total Tweets	632,528
Depressive Tweets	316,264
Non-Depressive Tweets	316,264
Collection Period	April–June 2009

3.1.2. Suicide-Watch Dataset

The "Suicide and Depression Detection" dataset [28] is a comprehensive resource designed to support the development of machine learning models aimed at identifying suicidal ideation and depression in text-based data. It consists of Reddit posts collected from three specific subreddits: "SuicideWatch", "depression", and "teenagers", with data gathered using the Pushshift API. Posts from "SuicideWatch" (16 December 2008–2 January 2021) are labeled as suicide, while posts from "depression" (1 January 2009–2 January 2021) are labeled as depression. The posts from the "teenagers" subreddit serve as examples of normal conversations, providing non-suicide content for comparison. The dataset is available in two versions: a simplified one with just suicide and non-suicide labels, and a more detailed V13 version with three labels—suicide, depression, and normal (teenagers).

This dataset is an invaluable tool for training natural language processing (NLP) models to detect mental health issues through text classification. It enables researchers to build classifiers that differentiate between suicidal content, depressive expressions, and neutral discussions. The dataset is accompanied by a notebook showing how to collect additional posts using the Pushshift API, allowing further customization. With its structured approach, this dataset has significant potential for mental health research, AI ethics, and the early detection of mental health crises in online communities, fostering the development of tools for timely intervention. Table 2 briefly summarises the properties of the dataset.

Table 2. Suicide-Watch dataset description table.

Characteristic	Value
Total Reddits/Subreddits	232,074
Suicide Reddits/Subreddits	116,037
Non-Depressive Reddits/Subreddits	116,037
Collection Period	December 2008–January 2021

3.2. Data Preprocessing

Data preprocessing is essential in preparing the Sentiment140 dataset for depression detection. The preprocessing pipeline involves several stages designed to clean and transform the raw tweet data into a format suitable for machine learning model training and evaluation. Below, we describe the critical steps involved in this process.

- Data cleaning: The first data cleaning process was to drop redundant features irrelevant to our study, like ('ids', 'date', 'flag', and 'user') from the dataset. Next, we removed any special characters in the tweet. Tweets often contain special characters, URLs, and emoticons that do not contribute to the semantic content relevant to depression detection. These elements were removed using regular expressions to ensure cleaner text inputs. This was achieved with Python's 're' library. Also, to maintain uniformity and reduce the dimensionality of the feature space, all text data were converted to be lowercased.

- Stop word removal: Common stop words, such as 'the', 'are', and 'is', which do not provide significant meaning for our analysis, were removed using the NLTK library's list of English stop words. This step helps focus on more informative words within each tweet, reducing noise and improving the quality of our feature set.
- Lemmatization and stemming: Words were reduced to their base or root form using lemmatization, which helps normalize the text. For example, words like 'running', 'ran', and 'runs' were converted to their lemma, 'run'. As an alternative to lemmatization, stemming was also applied, where words were truncated to their root forms. However, lemmatization was preferred for its ability to provide contextually accurate base forms.
- Label encoding: Our approach for preparing the target variable involved label encoding. We assigned a numerical value of 1 to tweets classified as depicting depression and 0 to those categorized as non-depressive. This conversion, performed using Scikit-learn's LabelEncoder, transformed the target variable into a numerical format suitable for machine learning model training and evaluation.
- Feature engineering: To prepare the text data for machine learning models, we used TF-IDF (Term Frequency–Inverse Document Frequency) vectorization (implemented using Scikit-learn's 'TfidfVectorizer'). This method converts each tweet into a numerical vector, capturing the importance of words within a tweet and across the entire dataset. TF-IDF considers the frequency of each word in a tweet (TF) and adjusts it based on how common the word is in the whole dataset (IDF). This weighting ensures that words specific to depression are more impactful in the vector representation.

3.3. Model Selection

Here, we will describe the selection of the machine learning and transformer models used in our study. The choice of models is guided by their demonstrated effectiveness in NLP tasks, including sentiment analysis and mental health detection. We will provide a rationale for selecting each model, covering traditional machine learning algorithms and advanced transformer-based models. This comprehensive approach ensures a robust comparative analysis, leveraging the strengths of conventional and state-of-the-art techniques to achieve optimal performance in depression detection.

In this study, we employed three traditional machine learning algorithms: logistic regression, Bernoulli Naive Bayes, and Random Forest.

1. Logistic regression is a popular and well-understood algorithm that is effective for binary classification tasks. Its simplicity and interpretability make it an excellent baseline model for comparison. A logistic regression model calculates the likelihood that an input falls into a specific class. It uses a logistic function to map predicted values to probabilities, facilitating binary classification. Key characteristics include its ease of implementation, efficiency on large datasets, and ability to handle binary outcomes.
2. Bernoulli Naive Bayes: applies Bayes' theorem [29] with the assumption of independence between features. It is suitable for binary/Boolean features, making it a natural choice for text classification tasks where the presence or absence of a word is informative. Its winning characteristics include its simplicity, speed, and efficiency, especially with high-dimensional data.
3. Random Forest: This is a powerful machine learning technique that combines the predictions of multiple decision trees. During training, it constructs a forest of these trees, making a classification decision for each based on a random subset of features. The final prediction is made by taking the most frequent class (mode) among all the trees. This ensemble approach offers several advantages. Random Forest boasts robustness and accuracy, making it well-suited for datasets with many features. Additionally, compared to individual decision trees, it is less susceptible to overfitting, leading to more generalizable models.

We employed four transformer models—RoBERTa, DeBERTa, DistilBERT, and SqueezeBERT. The Cross-Entropy Loss function and Adam optimizer were used to fine-tune these pre-trained transformer models.

1. RoBERTa is a transformer model designed to handle a variety of natural language understanding tasks with improved training strategies over BERT, such as removing the next sentence prediction objective and training with larger mini-batches and learning rates. Its robustness and superior performance in sentiment analysis make it ideal for our study.
2. DeBERTa improves upon BERT and RoBERTa by using disentangled attention mechanisms and enhanced mask decoders. This allows it to better understand a text's word dependencies and contextual relationships. This model's ability to capture nuanced language patterns makes it highly suitable for depression detection.
3. DistilBERT is a lighter, faster version of BERT, providing a good trade-off between model performance and computational efficiency. It retains 97% of BERT's performance while being 60% quicker and smaller, making it efficient for deployment in resource-constrained environments. Its reduced size, speed, and robust performance make it an attractive option for our analysis.
4. SqueezeBERT uses grouped convolutions to significantly reduce the number of parameters and the computational cost without a substantial drop in performance. Its efficient architecture is beneficial for real-time processing applications with limited resources.

By employing these machine learning and transformer models, our study aims to conduct a thorough comparative evaluation to identify the most effective approaches for detecting depression from social media texts. The diverse selection of models allows us to leverage both traditional and cutting-edge techniques, providing a comprehensive analysis of their capabilities and limitations in this context.

3.4. Model Training

For the training phase, we divided the dataset containing 632,000 tweets into two parts: a training subset and an evaluation subset. We used a ratio of 80% for training and 20% for evaluation, giving us 505,600 rows for training and 126,400 for evaluation. This strategy ensures that we have ample data for the models to learn from while maintaining a significant portion of performance evaluation.

- Data Splitting: We used Scikit-learn's train_test_split function to divide the dataset into training and evaluation sets, ensuring a random split to maintain representativeness.
- Cross-Validation: To ensure robust model performance and reliable assessment, we employed 10-fold cross-validation. For traditional machine learning models, we used 100,000 entries per fold, while transformer models were trained on 50,000 entries per fold. Each fold maintained an 80% training and 20% validation ratio. This method was implemented using Scikit-learn's StratifiedKFold to maintain the class distribution across folds, mitigating potential biases in the data split.
- Model Training: We utilized a range of machine learning models including logistic regression, Random Forest and Naive Bayes, and transformer-based models like RoBERTa, DebERTa, DistilBERT, and SqueezeBERT. Each model was trained using optimized hyperparameters obtained through grid searching and cross-validation, ensuring the models were well-tuned. The training was conducted using the Scikit-learn and Hugging Face Transformer libraries.

3.5. Model Evaluation

To comprehensively gauge the model's performances, we employed the standard evaluation metrics—accuracy, precision, recall, and F1-Score.

These metrics were selected to provide a well-rounded understanding of each model's performance. This evaluation considers both the model's effectiveness in capturing relevant posts (recall) and the accuracy of those classifications (precision), which are particularly important factors when dealing with depression detection in social media texts.

We calculated these metrics' average (mean) and spread (standard deviation) across the ten data splits to get a comprehensive picture of model performance. The mean scores estimate a model's typical performance, while the standard deviation indicates how much the performance varied across the splits. By examining these statistics, we can assess the stability and consistency of the model's predictions.

4. Results

The performance of the transformer and machine learning models on the task of depression detection was rigorously evaluated using the identified key metrics stated previously. Tables 3 and 4 summarizes these evaluation metrics for each model on the two utilized datasets, revealing several noteworthy patterns and observations.

Table 3. Overviewof the model's performance on the Sentiment 140 Dataset.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	94.9%	96.4%	93.3%	95.0%
Bernoulli Naive Bayes	90.1%	90.1%	90.0%	90.1%
Logistic Regression	97.0%	97.2%	96.7%	97.0%
RoBERTa	98.0%	98.0%	99.0%	98.0%
DeBERTa	98.0%	98.0%	98.0%	98.0%
DistilBERT	97.0%	98.0%	98.0%	97.0%
SquuezeBERT	95.0%	97.0%	97.0%	96.0%

Table 4. Overviewof the model's performance on the Suicide-Watch Dataset.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	90.8%	90.8%	90.8%	90.8%
Bernoulli Naive Bayes	78.2%	80.1%	78.2%	77.9%
Logistic Regression	93.5%	93.5%	93.5%	93.5%
RoBERTa	87.0%	87.2%	87.0%	87.0.0%
DeBERTa	88.0%	88.5%	88.0%	87.9%
DistilBERT	88.0%	88.0%	88.0%	88.0%
SquuezeBERT	87.5%	87.8%	87.5%	87.4%

For the Sentiment140 dataset, Figure 3 shows the loss curve of the RoBERTa model and Figure 4, its accuracy curve. Figure 5 shows the DeBERTa model's loss curve and Figure 6 shows its accuracy curve. Figures 7 and 8 illustrates the DistilBERT model's loss curve and accuracy curve respectively. And Figures 9 and 10 shows, respectively, the loss curve and accuracy curve of the SqueezeBERT model.

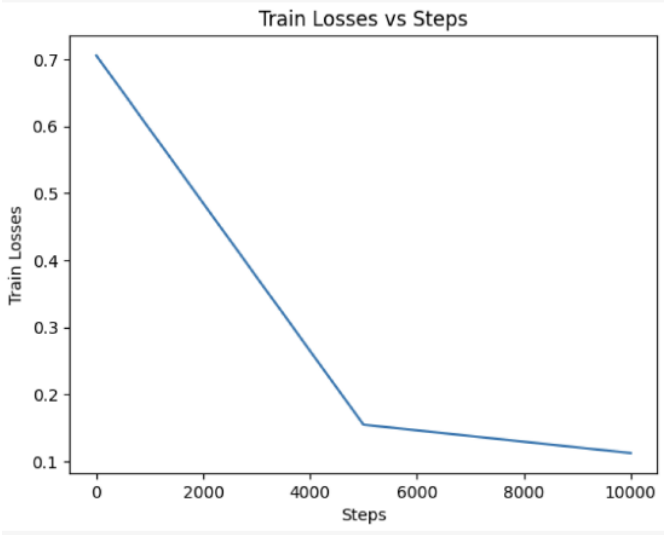


Figure 3. RoBERTa loss curve of the Sentiment140 dataset.



Figure 4. RoBERTa accuracy curve of the Sentiment140 dataset.

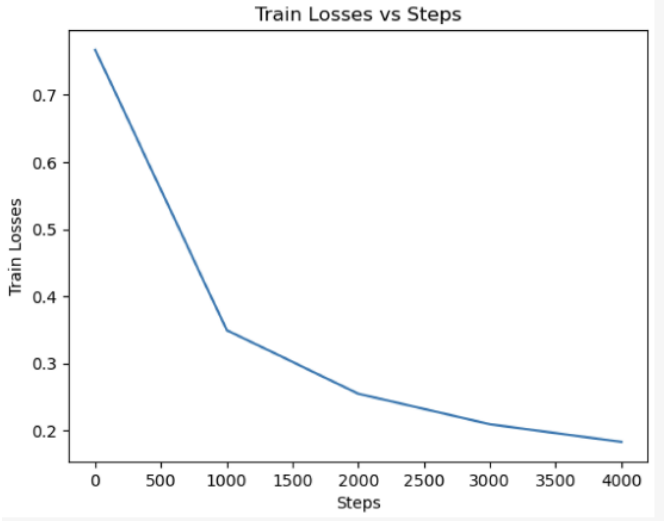


Figure 5. DeBERTa loss curve.



Figure 6. DeBERTa accuracy curve.

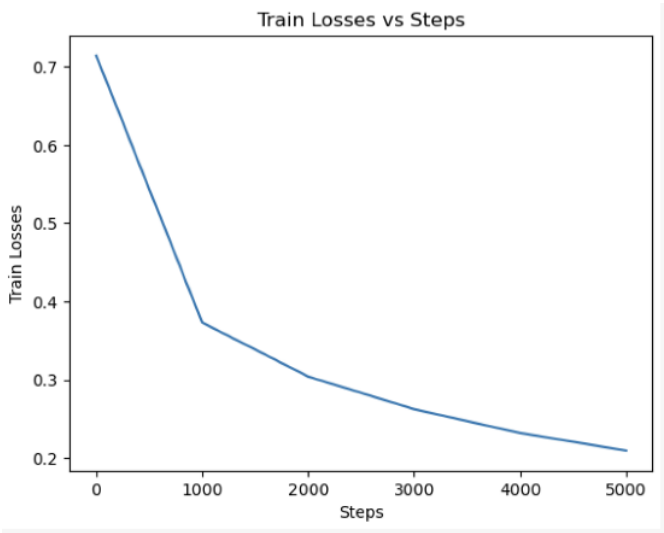


Figure 7. DistilBERT loss curve of the Sentiment140 dataset.

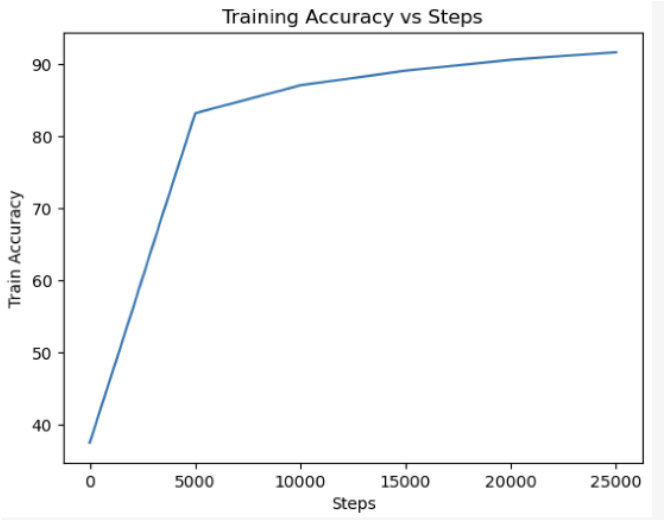


Figure 8. DistilBERT accuracy curve of the Sentiment140 dataset.

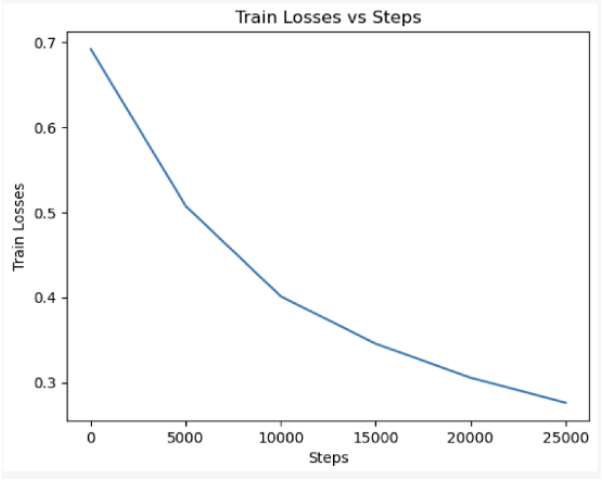


Figure 9. SqueezeBERT loss curve of the Sentiment140 dataset.

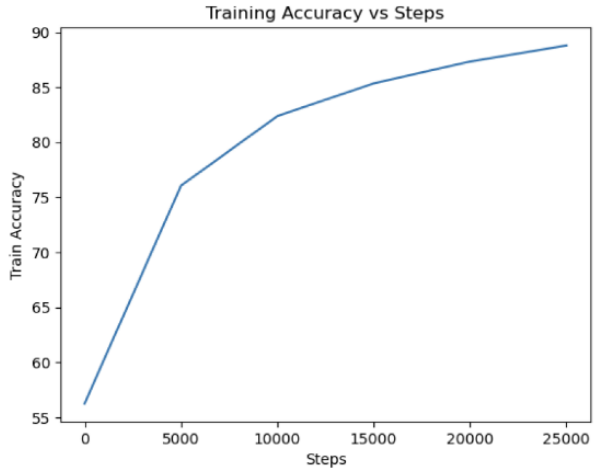


Figure 10. SqueezeBERT accuracy curve of the Sentiment140 dataset.

While for the Suicide-watch dataset, Figure 11 shows the loss curve of the RoBERTa model and Figure 12, its accuracy curve. Figure 13 shows the DeBERTa model's loss curve and Figure 14 shows its accuracy curve. Figures 15 and 16 illustrates the DistilBERT model's loss curve and accuracy curve respectively. And Figures 17 and 18 shows, respectively, the loss curve and accuracy curve of the SqueezeBERT model.

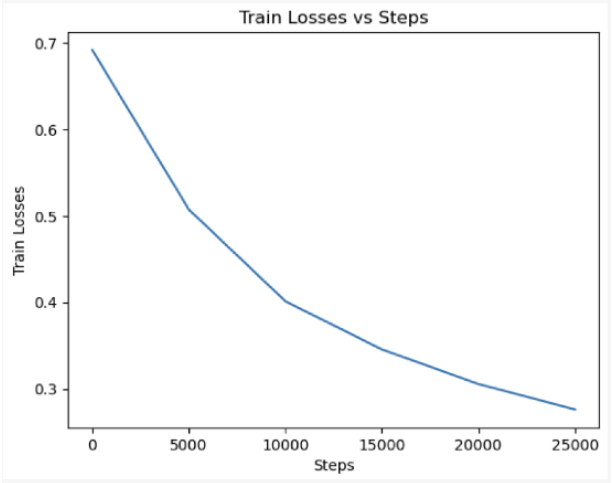


Figure 11. RoBERTa loss curve of the Suicide-Watch dataset.

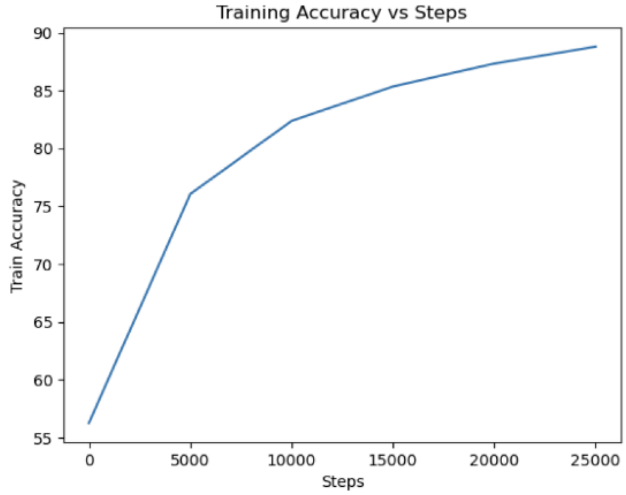


Figure 12. RoBERTa accuracy curve of the Suicide-Watch dataset.

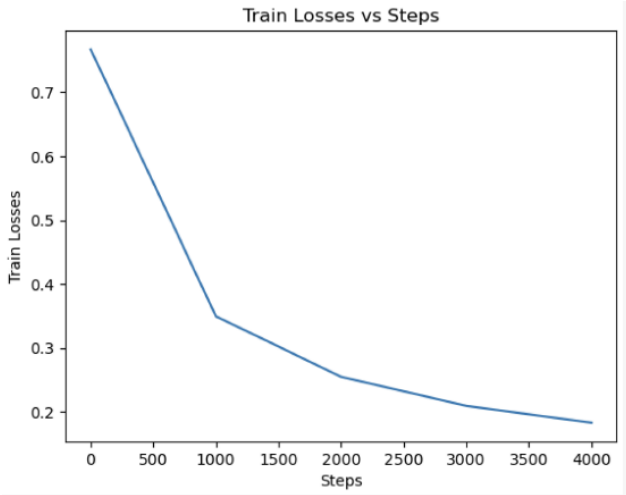


Figure 13. DeBERTa loss curve of the Suicide-Watch dataset.

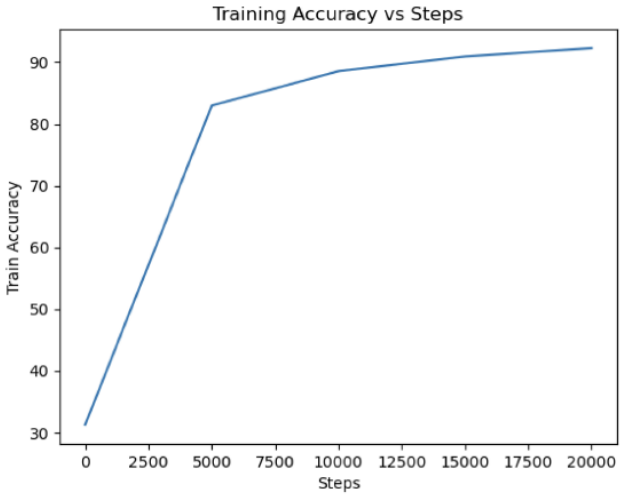


Figure 14. DeBERTa accuracy curve of the Suicide-Watch dataset.

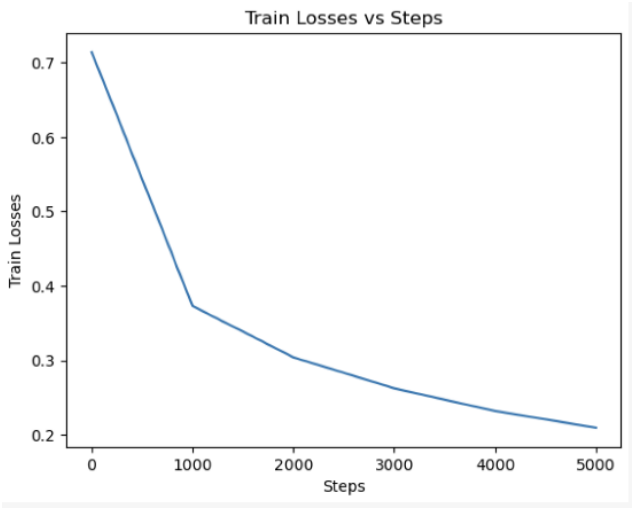


Figure 15. DistilBERT loss curve of the Suicide-Watch dataset.

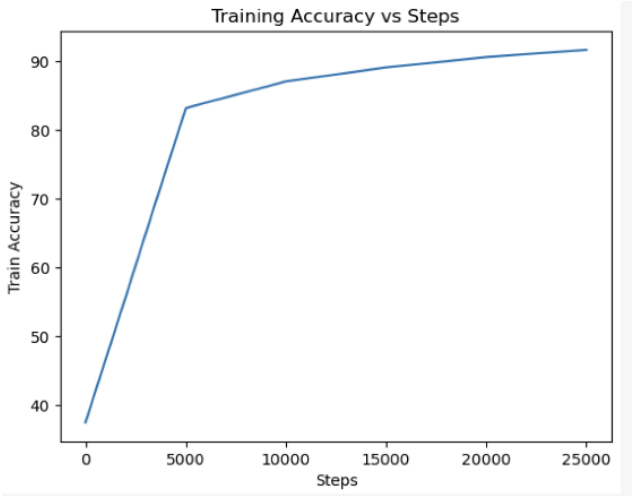


Figure 16. DistilBERT accuracy curve of the Suicide-Watch dataset.

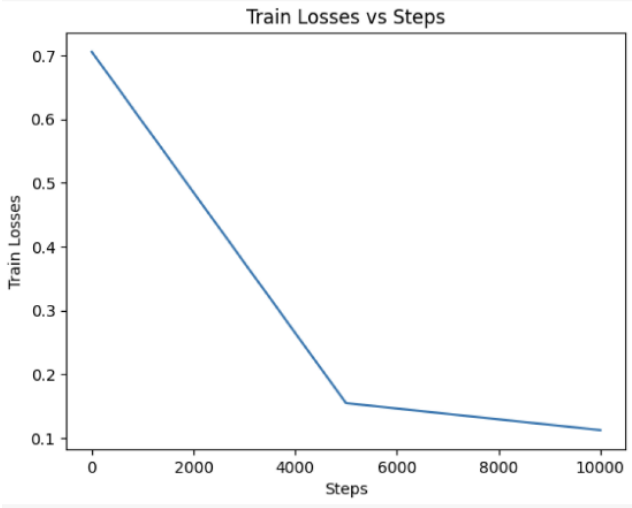


Figure 17. SqueezeBERT loss curve of the Suicide-Watch dataset.



Figure 18. SqueezeBERT accuracy curve of the Suicide-Watch dataset.

4.1. Comparative Analysis

When comparing the performance of machine learning models to transformer-based techniques, transformer models significantly outperform traditional machine learning algorithms across all evaluation metrics on the Sentiment 140 Dataset. Transformer models such as RoBERTa and DeBERTa achieved the highest accuracy (98.0%) and F1 scores (98.0%), underscoring their superior capability in identifying depressive content from tweets.

Among traditional models, logistic regression was the best performer, with a high accuracy of 97.0% and an F1-score of 97.0%, but it still fell short compared to the transformer models. Random Forest showed high precision but had slightly lower recall, while Bernoulli Naive Bayes exhibited moderate performance across all metrics.

On the other hand, in the Suicide-Watch dataset, traditional models significantly outperform transformer models across all evaluation metrics. Traditional machine learning algorithms such as logistic regression achieved the highest accuracy score of 93.5% and F1 score of 93.5%. This is evidence of the superiority of traditional models over less complex models with simple patterns.

The use of transformer models not only provided better overall accuracy but also offered more balanced precision and recall, which are essential for reducing both false positives and false negatives in the two datasets. This balanced performance is crucial in applications like depression detection, where accurate identification can lead to timely interventions and support.

4.2. Key Insights

1. Superior Performance of Transformer Models in Datasets with Complex Patterns: The Sentiment 140 dataset has a complex pattern as the classes of "depressed" or "non-depressed") were approximated from the sentiment score of the tweets. The transformer models were able to capture this nuance by estimating the classes based on the context. RoBERTa and DeBERTa achieved the highest accuracy (98.0%) and F1 score (98.0%) in the Sentiment140 dataset.
2. Logistic Regression's Superior Performance in Datasets with Simple Class Labels and Patterns: The Suicide-Watch Dataset has a more basic pattern and target labels, as the tweets were labeled correctly and not estimated. Aside from Bernoulli Naive Bayes, the traditional machine learning algorithms performed better than the transformer model on the Suicide-Watch dataset. Among traditional models, logistic regression demonstrated the highest accuracy (93.5%) and F1-score (93.5%), making it a strong candidate for depression detection tasks.

3. Balanced Precision and Pecall in Transformer Models: High precision and recall scores in transformer models indicate their ability to correctly identify both depressive and non-depressive tweets, reddits, and subreddits in both the Suicide-Watch and Sentiment140 data, minimizing false negatives and positives.
4. Potential of Ensemble Methods: While not explicitly evaluated in this study, the notable performance of individual transformer models suggests that ensemble methods combining these models could further enhance detection accuracy and reliability.

5. Discussion

5.1. Interpretation of Findings

For the Suicide Watch dataset, Figure 19 presents the confusion matrix of the Logistic Regression model, illustrating its classification performance across depressed and Non-Depressed labels. Figure 20 displays the confusion matrix for the Naive Bayes model, providing details into its probabilistic approach to prediction. Figure 21 highlights the confusion matrix of the Random Forest model, showcasing the performance of this ensemble method in distinguishing between different classes. Figure 22 visualizes the results from the SqueezeBERT transformer model, demonstrating its ability to handle sequence-based data.

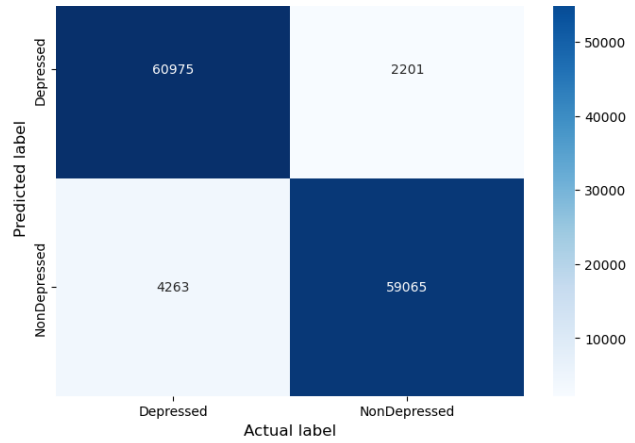


Figure 19. Logistic regression Confusion Matrix for the Suicide-Watch dataset

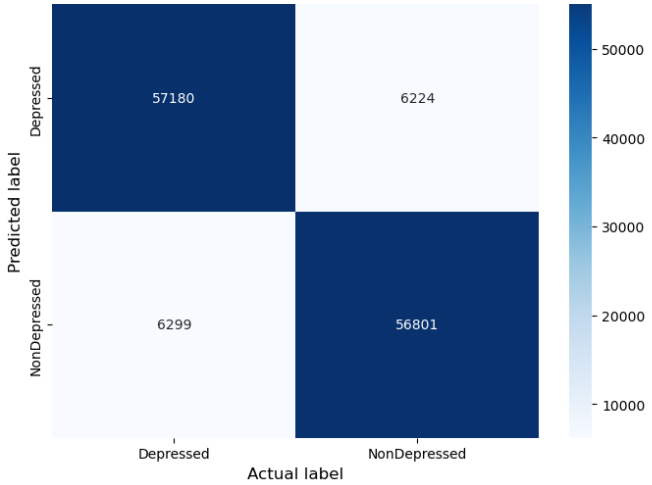


Figure 20. Bernoulli Naive Bayes Confusion Matrix for the Suicide-Watch dataset.

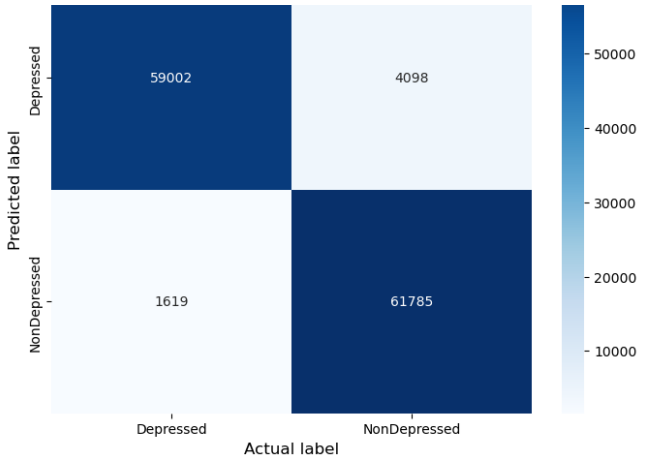


Figure 21. Random Forest Confusion Matrix for Suicide-Watch dataset.

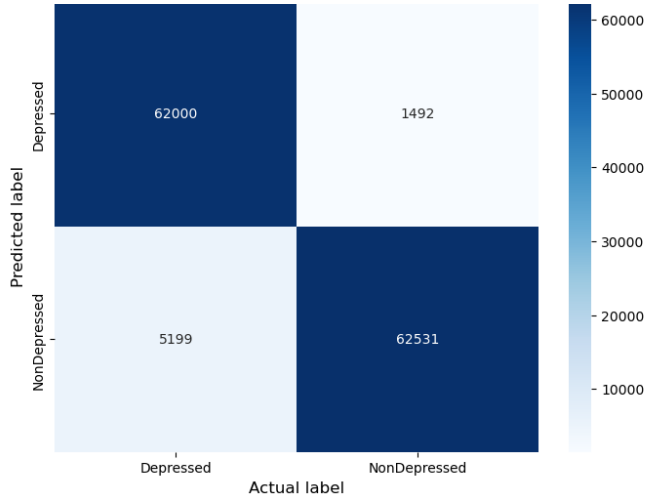


Figure 22. SqueezeBert Confusion Matrix for Suicide-Watch dataset.

In Figure 23, the confusion matrix for the RoBERTa model is presented, reflecting the impact of pre-trained transformer architecture on classification accuracy. Figure 24 focuses on the DeBERTa model's confusion matrix, highlighting its enhanced capabilities in understanding language nuances. Finally, Figure 25 illustrates the DistilBERT model's confusion matrix, emphasizing its lightweight yet effective performance in the suicide risk prediction task.

Our study investigated the efficacy of various machine learning and transformer models in detecting depression from tweets. The results demonstrate that transformer models, specifically RoBERTa and DeBERTa, significantly outperform traditional machine learning models like logistic regression, Random Forest, and Bernoulli Naive Bayes on the Sentiment140 dataset. However, the traditional models performed better on the Suicide-Watch dataset. This finding aligns with the current literature emphasizing the superior capability of transformer model's ability to grasp the nuances of language more effectively by considering the context of words within a sentence and the traditional model's ability to capture more basic patterns in NLP tasks.

The high precision and recall observed in transformer models indicate their robustness in accurately identifying depressive content while minimizing false positives and negatives. This balanced performance is crucial for practical applications where misclassification could lead to either unnecessary alarm or missed opportunities for intervention. Logistic regression emerged as the best performer among traditional machine learning models. It suggests that while these models can still provide valuable insights, they are generally less effective than transformers for complex text analysis tasks.

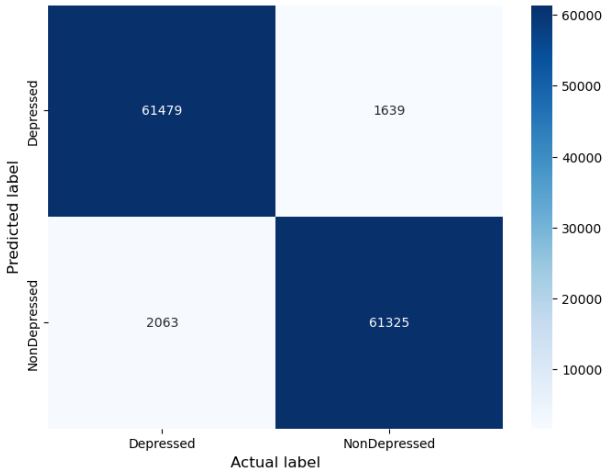


Figure 23. RoBerta Confusion Matrix for Suicide-Watch dataset.

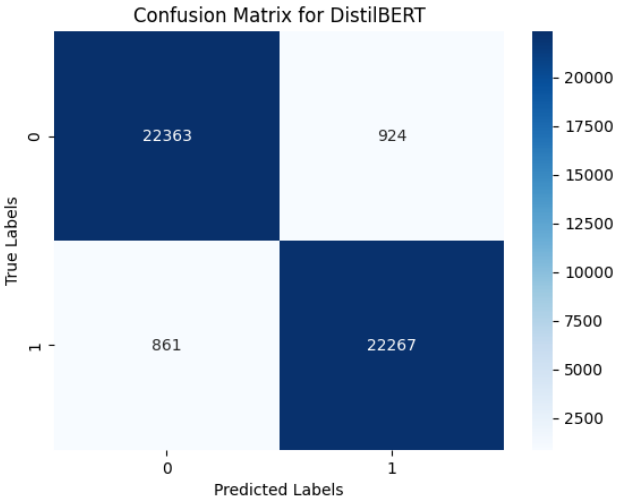


Figure 24. DeBerta Confusion Matrix for Suicide-Watch dataset.

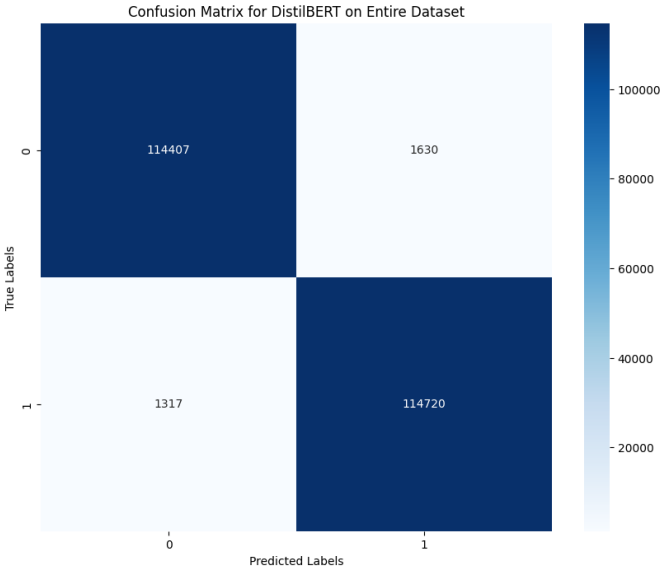


Figure 25. DistilBert Confusion Matrix for Suicide-Watch dataset.

The performance stability across cross-validation splits, evidenced by low standard deviations, further underscores the reliability of our models. This consistency is essential for developing dependable depression detection systems.

5.2. Comparison to Existing Studies

Our results are consistent with previous studies indicating the advantages of transformer-based models in sentiment and mental health analysis of social media. Studies such as those by Vaswani et al. [19] have highlighted the transformative impact of models like BERT and its variants in understanding and processing natural language. This is demonstrated by the fact that models like RoBERTa and DeBERTa perform better on datasets with complex patterns and that logistic regression performs better in more complicated tasks.

Furthermore, our work addresses a critical gap in the existing research by comparing traditional machine learning algorithms and newer transformer models. This comparison highlights the superior performance of transformers. It suggests specific areas where traditional models still hold value, particularly in scenarios with limited computational resources and in datasets with simpler patterns.

5.3. Limitations and Future Directions

Despite the promising results, our study has several limitations. The dataset, while extensive, consists of tweets labeled for sentiment analysis, which we repurposed for depression detection. This repurposing might introduce biases or misclassifications, as sentiment does not perfectly correlate with mental health status. Future research could benefit from datasets specifically annotated for depression detection to enhance model accuracy and validity.

Moreover, our study was constrained by computational resources, limiting the training size for some transformer models to 80% of the dataset. Future work should thoroughly explore training these models on larger datasets to leverage their capabilities. Additionally, expanding the scope to include data from multiple social media platforms could provide a more comprehensive understanding of depression expression in digital communication.

5.4. Implications

The findings of our study have significant implications for the development of real-world applications aimed at early depression detection and mental health screening. The demonstrated effectiveness of transformer models in accurately identifying depressive tweets suggests that these models could be integrated into mental health monitoring tools, providing timely alerts and interventions.

Moreover, the insights gained from our research could inform the design of automated systems for large-scale mental health analysis, aiding healthcare providers in tracking mental health trends and identifying at-risk individuals. As the prevalence of mental health issues continues to rise globally, leveraging advanced machine learning techniques for early detection and intervention could have profound public health benefits.

In conclusion, our study underscores the potential of transformer models in enhancing the accuracy and reliability of automated depression detection systems in complex systems, paving the way for innovative mental health monitoring and support approaches.

Author Contributions: Conceptualization, B.G.B. and Q.L.; Methodology, B.G.B.; Formal analysis, B.G.B.; Writing—original draft, B.G.B.; Writing—review & editing, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Briley, M.; Lépine, J.-P. The increasing burden of depression. *Neuropsychiatr. Dis. Treat.* **2011**, *7*, 3. [CrossRef] [PubMed]
2. Depressive Disorder (Depression). Available online: https://www.who.int/news-room/fact-sheets/detail/depression/ (accessed on 29 May 2024).
3. Rahul, M.; Deena, S.; Shylesh, R.; Lanitha, B. Detecting and Analyzing Depression: A Comprehensive Survey of Assessment Tools and Techniques. In Proceedings of the 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 26–28 April 2023. [CrossRef]
4. Choi, B.; Shim, G.; Jeong, B.; Jo, S. Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder. *Sci. Rep.* **2020**, *10*, 7867. [CrossRef] [PubMed]
5. Dean, B. Social Network Usage & Growth Statistics: How Many People Use Social Media in 2023? 2023. Available online: https://backlinko.com/social-media-users (accessed on 28 September 2024).
6. Kemp, S. "Digital 2023: The United Kingdom", DataReportal—Global Digital Insights. 2023. Available online: https://datareportal.com/reports/digital-2023-united-kingdom (accessed on 10 June 2024).
7. Obagbuwa, I.C.; Danster, S.; Chibaya, O.C. Supervised machine learning models for depression sentiment analysis. *Front. Artif. Intell.* **2023**, *6*, 123. [CrossRef] [PubMed]
8. Lyu, H. Application of machine learning on depression prediction and analysis. *Appl. Comput. Eng.* **2023**, *5*, 712–719. [CrossRef]
9. Di Cara, N.H.; Maggio, V.; Davis, O.S.P.; Haworth, C.M.A. Methodologies for Monitoring Mental Health on Twitter: Systematic Review. *J. Med. Internet Res.* **2023**, *25*, e42734. [CrossRef] [PubMed]
10. Chatterjee, M.; Modak, S.; Sarkar, D. Mental Health Predictions Through Online Social Media Analytics. In *Cognitive Cardiac Rehabilitation Using IoT and AI Tools*; IGI Global: Hershey, PA, USA, 2023; pp. 44–66. [CrossRef]
11. Coppersmith, G.; Dredze, M.; Harman, C. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Quantifying Mental Health Signals in Twitter. 2014. Available online: https://aclanthology.org/W14-3207.pdf (accessed on 15 June 2024).
12. Reece, A.G.; Danforth, C.M. Instagram photos reveal predictive markers of depression. *EPJ Data Sci.* **2017**, *6*, 15. [CrossRef]
13. Guntuku, S.C.; Yaden, D.B.; Kern, M.L.; Ungar, L.H.; Eichstaedt, J.C. Detecting depression and mental illness on social media: An integrative review. *Curr. Opin. Behav. Sci.* **2017**, *18*, 43–49. [CrossRef]
14. Shen, J.; Rudzicz, F. Detecting Anxiety on Reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology, Vancouver, BC, Canada, 3 August 2017; pp. 58–65. Available online: https://aclanthology.org/W17-3107.pdf (accessed on 19 June 2024).
15. Govindasamy, K.; Palanichamy, N. Depression Detection Using Machine Learning Techniques on Twitter Data. In Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021. [CrossRef]
16. Jain, V.; Chandel, D.; Garg, P.; Vishwakarma, D.K. Depression and Impaired Mental Health Analysis from Social Media Platforms Using Predictive Modelling Techniques. In Proceedings of the 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 7–9 October 2020. Available online: https://ieeexplore.ieee.org/abstract/document/9243334 (accessed on 8 May 2021).
17. Ashraf, A.; Gunawan, T.S.; Riza, B.S.; Haryanto, E.V.; Janin, Z.A. On the review of image and video-based depression detection using machine learning. *Indones. J. Electr. Eng. Comput. Sci.* **2020**, *19*, 1677. [CrossRef]
18. Angskun, J.; Tipprasert, S.; Angskun, T. Big data analytics on social networks for real-time depression detection. *J. Big Data* **2022**, *9*, 69. [CrossRef] [PubMed]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
20. Malviya, K.; Roy, B.; Saritha, S. A Transformers Approach to Detect Depression in Social Media. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021. Available online: https://ieeexplore.ieee.org/document/9395943 (accessed on 13 October 2022).
21. Nanggala, K.; Elwirehardja, G.N.; Pardamean, B. Systematic Literature Review of Transformer Model Implementations in Detecting Depression. In Proceedings of the 2023 6th International Conference of Computer and Informatics Engineering (IC2IE), Lombok, Indonesia, 14–15 September 2023. [CrossRef]
22. Ilias, L.; Mouzakitis, S.; Askounis, D. Calibration of Transformer-based Models for Identifying Stress and Depression in Social Media. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 1979–1990. [CrossRef]
23. Tavchioski, I.; Robnik-Šikonja, M.; Pollak, S. Detection of depression on social networks using transformers and ensembles. *arXiv* **2023**, arXiv:2305.05325. [CrossRef]
24. Go, A.; Bhayani, R.; Huang, L. Twitter Sentiment Classification Using Distant Supervision. 2009. Available online: http://help.sentiment140.com/home (accessed on 29 May 2024).
25. Yoon, S.; Dang, V.; Mertz, J.; Rottenberg, J. Are attitudes towards emotions associated with depression? A Conceptual and meta-analytic review. *J. Affect. Disord.* **2018**, *18*, 329–340. [CrossRef] [PubMed]
26. Yavuzer, Y.; Karatas, Z. Investigating the Relationship between Depression, Negative Automatic Thoughts, Life Satisfaction and Symptom Interpretation in Turkish Young Adults. In *Depression*; IntechOpen: London, UK, 2017. [CrossRef]

27. Marshall, A.D.; Sippel, L.M.; Belleau, E.L. Negatively Biased Emotion Perception in Depression as a Contributing Factor to Psychological Aggression Perpetration: A Preliminary Study. In *Emotions and Their Influence on Our Personal, Interpersonal and Social Experiences*; Routledge: London, UK, 2011. [CrossRef]
28. Komati, N. Suicide and Depression Detection Dataset, Kaggle. 2021. Available online: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch/data (accessed on 20 September 2024).
29. Bayes Theorem—An Overview | ScienceDirect Topics. Available online: https://www.sciencedirect.com/topics/mathematics/bayes-theorem (accessed on 20 May 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.