Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

Uncovering Depression on Social Media using BERT Model

Siti Nurulain Mohd Rum[1,*], Nur Fatin Aqilah Saharudin[1], Nor Azura Husin[1], Ahmad Akbar[2]

[1] Faculty of Computer Science & Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2] Faculty Science and Technology, University of Pembangunan Panca Budi, Medan, Sumatera Utara 20122, Indonesia

ARTICLE INFO

Article history:
Received 3 October 2024
Received in revised form 11 March 2025
Accepted 4 April 2025
Available online 30 April 2025

Keywords:
Depression detection; sentiment analysis; mental health early intervention; deep learning; BERT model

ABSTRACT

The lack of early detection and effective treatment programs for depression has left millions struggling with mental illnesses, including anxiety, sleep disorders and, in severe cases, self-harm or suicide. Adolescents and young adults, who are among the most active users of social media platforms like Twitter, represent a particularly vulnerable demographic. With Twitter often serving as a digital diary where individuals share thoughts and emotions, its data offers a unique opportunity for mental health research. This study explores the potential of leveraging Twitter data to detect depression and identify at-risk individuals. Using advanced machine learning techniques, including Natural Language Processing (NLP) with the BERT (Bidirectional Encoder Representations from Transformers) model, this research builds a predictive system that analyses linguistic cues, sentiment and user interactions. The methodology involves integrating and preprocessing data from multiple sources, ensuring a comprehensive and reliable approach to model development. The findings demonstrate the BERT model's superior accuracy, outperforming traditional machine learning methods like Logistic Regression and Support Vector Machines. This study underscores the effectiveness of social media as a tool for early detection and intervention, offering an innovative approach to mitigating the global burden of mental health disorders.

1. Introduction

In today's digital age, social media has become an integral part of daily life, providing a platform for self-expression, connection and information sharing. Beyond its social and informational roles, social media holds immense potential for mental health research and intervention. This potential is particularly crucial in light of the alarming rise in mental health issues, such as depression, that have increasingly affected individuals worldwide in recent years. The World Health Organization (WHO) defines mental health as a state of emotional and psychological well-being that enables individuals to cope with life's challenges, realise their potential, acquire knowledge, achieve personal growth and contribute meaningfully to society. Despite the importance, mental health disorders, especially depression, remain widespread, affecting a significant portion of the global population. This growing prevalence of mental health challenges underscores the urgency of leveraging innovative tools and

* *Corresponding author*
E-mail address: snurulain@upm.edu.my

https://doi.org/10.37934/ard.129.1.4659

approaches to address them. Social media, with its vast reach and real-time nature, offers a unique opportunity to enhance mental health research, facilitate early detection and provide timely interventions to improve emotional well-being and quality of life for individuals around the world [1]. The WHO estimates that 280 million individuals worldwide or 5% of adults globally and 5.7% of persons over 60 struggles with depression [2]. Depression is a common mental disorder described by persistent melancholy, a loss of interest or enjoyment in activities, disrupted sleep, irregular eating and feelings of fatigue and poor concentration. It can have a significant negative impact on someone's everyday functioning and general well-being, with long-term or recurring effects that make it difficult for people to live happy lives [3]. Tragically, depression, if left untreated, can escalate the risk of suicide among individuals. Suicide remains a prevalent global concern, with over 700,000 people losing their lives to suicide each year, accounting for approximately one death every 40 seconds [4].

Surprisingly, suicide stands as the second leading cause of death among individuals between the ages of 10 and 34 [4]. The findings shed light on the effects of suicide, highlighting the need for comprehensive support systems and interventions to address the broader impact on individuals and communities affected by suicide. As we combat the severe repercussions of untreated mental diseases, it becomes increasingly important to investigate precise techniques for predicting, identifying and aiding those in distress in a timely manner. Self-reported surveys, professional interviews and observations have traditionally been the main methods used in mental health assessment [5], but these methods are frequently biassed and have a small sample size.

The increasing influence of Generative AI on digital communication has reshaped interactions on social media, introducing both opportunities and challenges in detecting mental health indicators. Recent studies have highlighted how AI-generated content affects sentiment, engagement and the authenticity of online discourse [6]. This shift underscores the need for advanced Natural Language Processing (NLP) models, such as Bidirectional Encoder Representations from Transformers (BERT), to accurately differentiate between genuine distress signals and AI-influenced content in mental health analysis. Leveraging social media data has become essential for research and intervention in the field of mental health, particularly on Twitter. Due to its accessible application programming interface (API) and user-friendly design, which provide researchers with easy access to data for their studies, Twitter has become a popular research platform [7]. As a large percentage of young adults and people in their middle years use social media as a digital diary, they are open about sharing their ideas, feelings and challenges. This abundance of data provides a valuable chance to identify warning signs early, prevent suicide and lessen the consequences of depression, ultimately leading to better mental health outcomes. The purpose of this study is to investigate how Twitter, a popular social media site, may be used to diagnose depression and identify those who are at risk of developing depression. The study’s goal is to create a robust predictive model that increases the accuracy, reliability and scalability of mental health prediction by examining patterns, linguistic signals, sentiment and user interactions in Twitter data. The findings of this study will eventually lead to the development of instant treatments and support networks, achieving the growing demand for mental health care, including depression and suicide prevention. Given the severe implications of untreated mental health concerns, early diagnosis, assessment and follow-up of depressed individuals are critical components of comprehensive suicide prevention initiatives. Using social media data and machine learning techniques, significant progress can be made in identifying at-risk individuals and providing the aid and attention they require. This study intends to encourage mental health intervention and research by harnessing the potential of social media data, with the goal of improving wellbeing and saving lives affected by depression.

2. Related Work

Depression is a pervasive mental health condition characterized by persistent sadness, loss of interest in activities and significant impairment in daily functioning. Over the years, researchers have explored the potential of social media data, particularly from platforms like Twitter, to detect and understand mental health disorders. This section critically reviews key studies, highlights their contributions and limitations and situates the current research within the broader body of literature. Artificial intelligence (AI) and machine learning (ML) have gained significant attention in recent years as transformative tools for enhancing mental health services. For example, Jain *et al.,* [8] explored the application of AI and ML in identifying and diagnosing mental health disorders, developing AI-powered therapeutic interventions and improving access to mental health care services. These technologies hold immense promise in bridging gaps in traditional healthcare delivery systems. The study by Pachouly *et al.,* [9], proposes a depression analysis and suicidal ideation detection system using machine learning techniques to predict the likelihood of depression based on a Twitter user's tweets. Various machine learning classification algorithms, such as Support Vector Machine (SVM) and Naïve Bayes, are utilized to differentiate whether a user is depressed or not, based on features extracted from their activities within the tweets. Expressions of pessimism or negative remarks have been consistently associated with depressive symptoms [10]. Several studies have delved into the relationship between depressive symptoms and the use of negative language, providing substantial evidence to support this link [11]. For instance, a study by Chancellor *et al.,* [12] examined social media data and identified a strong correlation between the vocabulary used in tweets and the prevalence of depressive symptoms. The findings revealed that individuals experiencing severe depression were more likely to share tweets with negative or unpleasant content. This highlights the potential of analysing linguistic patterns in social media posts as an effective approach for detecting early signs of depression and guiding timely interventions.

NLP and ML techniques have been widely adopted in the detection of depression on social media. For instance, Kim *et al.,* [13] developed a deep learning model using social media posts and demonstrated the efficacy of NLP methods in identifying potential mental health issues. While their research provided valuable insights into linguistic patterns associated with depression, the model's reliance on small, homogeneous datasets limited its generalizability across demographic and linguistic diversity. Ansari *et al.,* [14] employed ensemble hybrid models, integrating deep learning approaches with lexicon-based methods for depression detection. Their study achieved improved classification performance by combining multiple feature sets. However, the use of public datasets raised concerns about data sparsity and the lack of contextual understanding, particularly for tweets that include slang, emojis or code-mixed languages. Similarly, Pachouly *et al.,* [15] applied traditional ML algorithms, such as Support Vector Machines (SVM) and Naïve Bayes, to classify Twitter users based on their likelihood of being depressed. While their approach was computationally efficient, it struggled to capture nuanced language features, such as sarcasm or implicit expressions of distress, which are common on social media.

A recent study by Ansari *et al.,* [14] explored the use of attention mechanisms within Long Short-Term Memory (LSTM) networks to detect mental health markers in tweets. The attention mechanism significantly improved the model’s ability to identify key linguistic cues. However, the lack of integration with user-specific metadata, such as interaction patterns or posting frequency, limited the scope of their analysis. The study by Uban *et al.,* [16] aims to understand whether monitoring language in social media could help with early detection of mental disorders, using deep learning models to analyse linguistic markers of disorders at different levels of language. The research

investigates the manifestation of mental disorders in language, focusing on three disorders: depression, anorexia and self-harm tendencies. The study analyses the importance of modelling mental disorders with computational methods, considering the knowledge provided by psychology in this area. The importance of including contextual data from social media platforms has also been highlighted by other studies [17-19]. There are many studies that have investigated the correlation between depression detection and social context [20-22]. Many achieved this by examining tweet content and the social network linkages among users. The advantages of taking the social context into account while analysing social media indicators of mental health were brought to light by their research. Collectively, these research works show the promise of machine learning methods for identifying depression in user tweets.

A common limitation in previous studies is the lack of diverse and large-scale datasets representative of real-world populations. Many studies rely on publicly available datasets, which may not capture the diversity of language use across different demographics. Furthermore, the overrepresentation of English-language data limits the applicability of these models to non-English-speaking populations. Another methodological challenge is the absence of contextual analysis. While some studies incorporate linguistic features, few consider user-specific metadata, such as temporal patterns, interaction behaviours or the frequency of depressive posts. Additionally, many models fail to account for the high variability in tweet lengths and the prevalence of sarcasm, slang and abbreviations, which can impact classification accuracy.

This research aims to address these limitations by integrating a larger and more diverse dataset, combining Twitter and Reddit data to enhance representativeness. The use of the BERT model, specifically its distilled version (DistilBERT), enables the study to capture bidirectional context while maintaining computational efficiency. Advanced preprocessing techniques, such as normalization, slang expansion and emoji handling, are employed to manage the noisy nature of social media data. Additionally, this research incorporates user-level metadata, such as posting frequency and interaction patterns, to provide a more holistic understanding of depressive behaviours on Twitter. By addressing these gaps, this study contributes to the growing body of literature on depression detection and offers an innovative, scalable approach for leveraging social media data in mental health diagnostics.

3. Methodology

This section outlines the systematic approach undertaken to achieve the objectives of this research. Figure 1 illustrates the step-by-step processes involved in the study, encompassing data collection, preprocessing, feature engineering, model development and evaluation. Each step is designed to ensure the reliability and validity of the results while addressing the complexities associated with analysing social media data for mental health applications. The following subsections provide a detailed explanation of each stage, highlighting the techniques, tools and strategies employed to develop an effective prediction model. The methodology begins with data integration, where datasets from multiple sources are combined to create a robust and diverse corpus of text data. This is followed by data preprocessing, which involves cleaning, normalizing and transforming the raw text data to remove noise and enhance quality for further analysis. Key preprocessing steps include data cleaning, transformation and normalization, such as handling emojis, URLs and non-alphanumeric characters, as well as tokenization and stop word removal. After preprocessing, the focus shifts to model development, where advanced machine learning and natural language processing techniques are applied. Various traditional models such as Naïve Bayes, Decision Trees and SVM are evaluated alongside state-of-the-art deep learning models, particularly BERT. Feature

extraction and vectorization methods, like TF-IDF, are employed to convert text data into numerical representations suitable for computational processing. Finally, the models are rigorously evaluated using standard performance metrics, including accuracy, precision, recall and F1-score. The results provide insights into the effectiveness of the models and guide the selection of the best-performing approach for detecting depression in social media data. The step-by-step methodology ensures a comprehensive and structured analysis, paving the way for developing innovative tools for mental health research and intervention.

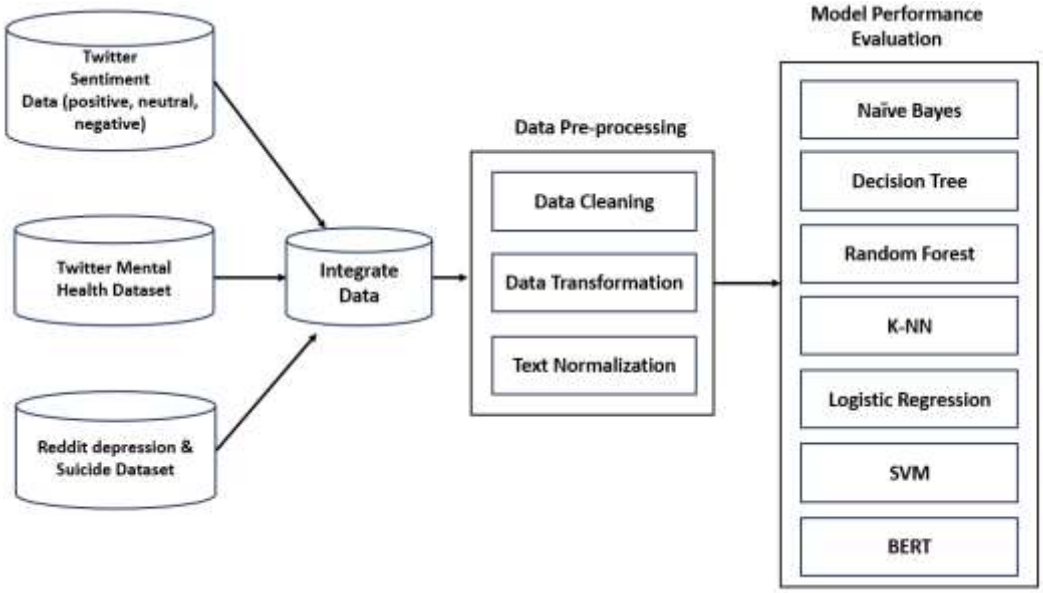

Fig. 1. Methodology of research work

3.1 Integrated Data

The integrated dataset used in this study draws from multiple distinct sources, each providing unique insights into the discourse surrounding mental health on social media platforms. The integration of these datasets ensures a robust and comprehensive foundation for analysing and modelling mental health indicators.

i. Twitter Sentiment Dataset (Sentiment 140): This dataset consists of a vast collection of tweets categorized into negative, neutral and positive sentiments. For this study, only tweets labelled as '0'—representing normal or negative sentiments—were utilized. These tweets serve as critical training data, enabling the model to differentiate between normal/negative sentiment texts and those potentially indicative of depression.
ii. Twitter Mental Health Dataset (Kaggle): This dataset categorizes Twitter users into "depressed" and "non-depressed" groups, offering valuable linguistic and behavioural insights. It sheds light on how individuals experiencing depression express themselves on social media, providing the groundwork for building predictive models.
iii. Reddit Depression and SuicideWatch Dataset: This dataset includes posts from Reddit forums dedicated to discussing mental health topics, specifically depression and suicide. Text data from these forums are categorized into two labels: "depression" and "SuicideWatch," offering an in-depth understanding of the language and emotional expressions of individuals engaging in such discussions.

iv. Reddit Depression Datasets (2019 and 2018-2019): These datasets focus on posts from Reddit's depression forum, capturing discussions about depression over time. By analysing posts from multiple years, this dataset provides a longitudinal perspective on the evolution of language and themes in mental health discourse.
v. Reddit SuicideWatch Dataset: This dataset comprises posts from Reddit's SuicideWatch forum, classified as either "suicidal" or "non-suicidal." It provides critical insights into individuals expressing suicidal ideation and seeking support, making it a valuable resource for understanding the nuances of suicidal behaviour online.

To ensure a unified and consistent structure, these datasets were merged into a single integrated dataset. Each instance in the dataset was labelled according to the nature of its content: '0' for normal/non-depressed/negative sentiment, '1' for depressed and '2' for suicidal. This comprehensive dataset allows for detailed analysis and modelling of mental health discourse across social media platforms, forming the foundation for the development of effective tools and interventions to support individuals in distress. By incorporating diverse data sources, this study captures a broad spectrum of mental health-related content, enabling the creation of a more reliable and generalizable prediction model. The unified dataset also facilitates the exploration of patterns, trends and correlations across platforms, ultimately contributing to advancements in mental health research and intervention.

3.2 Data Pre-Processing

During the data preprocessing phase, various techniques were employed to clean and transform the Twitter and Reddit data for analysis.

3.2.1 Data cleaning

During data cleaning, several key transformations were applied to enhance the quality and consistency of the text data. Firstly, URLs were replaced with a generic placeholder ("<url>") to eliminate any web addresses present in the tweets. This ensures that the analysis focuses solely on the textual content without being influenced by external links. Additionally, emojis and emoticons were replaced with corresponding text placeholders (e.g., "<smile>" and "<sadface>") to standardize emotional expressions within the text. Furthermore, usernames or mentions were anonymized by replacing them with a common placeholder ("<user>"). Punctuation marks were removed to streamline the text and simplify subsequent analysis. Regular expression patterns were compiled to efficiently identify and replace URLs, usernames and emojis, contributing to the overall cleaning process. Lastly, non-alphanumeric characters were removed to further refine the text and prepare it for subsequent processing steps.

3.2.2 Data transformation

Following data cleaning, the text underwent various transformations to ensure uniformity and consistency. Firstly, all text was converted to lowercase to prevent discrepancies caused by case sensitivity during analysis. This lowercase conversion standardizes the text and reduces the likelihood of duplicate tokens due to differing capitalizations. Additionally, contractions (e.g., "don't" for "do not") were expanded using a predefined dictionary to ensure clarity and uniformity in the text. Spaces were added around slashes ("/") to separate words joined by slashes, aiding in accurate tokenization.

These transformations collectively contribute to the normalization and standardization of the text data, facilitating further analysis and interpretation.

3.2.3 Text normalization

In the final stage of preprocessing, the text underwent normalization processes aimed at refining its structure and enhancing its suitability for analysis. Stopword removal was employed to eliminate common, non-informative words (e.g., "and", "the", "is"), which could potentially skew analysis results. This step helps to focus on meaningful content and improve the accuracy of subsequent natural language processing tasks. Tokenization was then applied to segment the text into individual words or tokens, enabling granular analysis and feature extraction. By breaking down the text into smaller units, tokenization facilitates the identification of patterns and sentiments within the data. These normalization techniques ensure that the text data is appropriately prepared for advanced analysis, such as sentiment analysis and machine learning modelling.

4. Building Prediction Model

The process of building prediction models for sentiment analysis involves several key steps.

4.1 Conventional Machine Learning Models

After the data has been processed, it will be divided into training and testing sets using techniques like train-test split, allowing the model to learn patterns from the training data and evaluate its performance on unseen data. Once the data is prepared, feature engineering techniques are applied to convert the text data into numerical feature vectors. In this case, TF-IDF vectorization is utilized, which transforms the text into numerical values based on term frequency-inverse document frequency, capturing the importance of words in the documents relative to the entire corpus. After feature engineering, various machine learning algorithms are selected and defined for building the prediction models. These algorithms include Naive Bayes, Decision Trees, Random Forests, K-Nearest Neighbours (K-NN), Logistic regression and SVM. Once the models are defined, they are trained using the training data and evaluated using evaluation metrics such as accuracy, confusion matrix and classification report. This evaluation process helps assess the performance of each model and identify any areas for improvement. Additionally, techniques like class weighting are applied for Logistic Regression and SVM model to address class imbalance issues. This involves computing class weights, which are used to adjust the importance of each class during model training, ensuring that the models can effectively capture patterns from all classes in the dataset.

4.2 BERT Model

The proposed model for this research work is the BERT model. BERT is a machine learning model for Natural Language Processing (NLP). The main technological advancement of BERT is the application of Transformer, a well-known attention model, to language modelling through bidirectional training, which reads text all at once rather than sequentially, making it adept at understanding context from all around a word [23]. It was trained on tasks like predicting hidden words (Masked Language Modelling) and determining if sentences follow one another (Next Sentence Prediction), making it excellent at comprehending language [24,25]. In this proposed work, DistilBERT, a distilled version of BERT, is utilized as the primary model. DistilBERT is characterized by

its smaller, faster, cheaper and lighter architecture, making it specifically tailored for practical applications with limited computational resources. Despite its streamlined design, DistilBERT retains approximately 95% of BERT's performance, making it an efficient and effective choice of this research. This contrasts with earlier research that examined a text sequence either from the left to the right or by combining training from the left to the right. The study's findings demonstrate that bidirectionally trained language models can capture a richer sense of language context and flow than single-direction models.

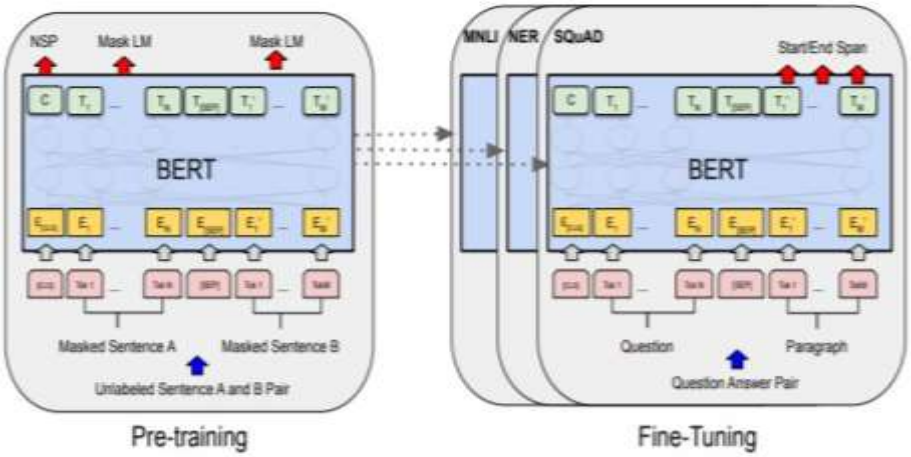

Fig. 2. BERT model

5. Results and Discussion

The evaluation of various machine learning models for depression detection in Twitter posts reveals diverse performances across different algorithms. Table 2 presents the comparison prediction model achievement based on accuracy, F1-Scores, precision and recall for all different models of machine learning, including BERT model.

Table 2
Comparison of prediction model

	Naïve Bayes	Decision Tree	Random Forest	Logistic Regression	Support Vector Machines	BERT
Accuracy	76.7%	68.23%	77.63%	79.97%	80.33%	84.8%
F1-Scores	76.7	67.7	76.7	79.3	79.7	84.7
Precision	77.7	67.7	76.7	79.3	79.7	86.3
Recall	77.7	67.7	77.3	79.7	80.0	85.6

The following sections describe the result achieved for each model.

5.1 Naïve Bayes

The Naïve Bayes model achieved an accuracy of 76.70%, indicating that it correctly classified approximately three-quarters of the tweets. Specifically, it demonstrated high precision for class 0 (non-depressed tweets), correctly identifying 95% of them, while maintaining moderate precision for classes 1 (depressed tweets) and 2 (suicidal tweets) at 70 and 68%, respectively. Moreover, it exhibited respectable recall scores across all classes, with values of 0.84, 0.69 and 0.77 for classes 0, 1 and 2, respectively.

5.2 Decision Tree

The decision tree model achieved an accuracy of 68.23%, indicating its ability to correctly classify around two-thirds of the tweets. While it demonstrated high precision for class 0 (non-depressed tweets) at 88%, it showed lower precision for classes 1 (depressed tweets) and 2 (suicidal tweets) at 56 and 59%, respectively. Additionally, it exhibited relatively balanced recall scores across all classes, with values of 0.89, 0.56 and 0.58 for classes 0, 1 and 2, respectively.

5.3 Random Forest

The random forest model achieved an accuracy of 77.63%, indicating its ability to correctly classify nearly four-fifths of the tweets. It demonstrated high precision for class 0 (nondepressed tweets) at 87%, while maintaining moderate precision for classes 1 (depressed tweets) and 2 (suicidal tweets) at 72 and 71%, respectively. Moreover, it exhibited high recall for class 0 (non-depressed tweets) at 98%, indicating its capability to accurately identify the majority of non-depressed tweets, although it showed slightly lower recall for classes 1 and 2 at 64 and 70%, respectively.

5.4 K-Nearest Neighbour(K-NN)

The K-Nearest Neighbour (K-NN) model achieved an accuracy of 34.83%, indicating its limited ability to correctly classify tweets. It exhibited low precision for classes 1 (depressed tweets) and 2 (suicidal tweets) at 31 and 80%, respectively, while demonstrating high precision for class 0 (non-depressed tweets) at 35%. However, its recall scores were notably low across all classes, particularly for classes 1 and 2, with values of 0.01 each, indicating its difficulty in accurately identifying depressed and suicidal tweets.

5.5 Logistic Regression

The Logistic Regression model achieved an accuracy of 79.97%, indicating its capability to correctly classify approximately four-fifths of the tweets. It demonstrated high precision for class 0 (non-depressed tweets) at 90%, while maintaining moderate precision for classes 1 (depressed tweets) and 2 (suicidal tweets) at 73 and 75%, respectively. Moreover, it exhibited high recall scores for class 0 (nondepressed tweets) at 97%, indicating its ability to accurately identify the majority of non-depressed tweets, although it showed slightly lower recall for classes 1 and 2 at 69 and 73%, respectively.

5.6 Support Vector Machine (SVM)

The SVM model achieved an accuracy of 80.33%, indicating its capability to correctly classify approximately four-fifths of the tweets. It demonstrated high precision for class 0 (non-depressed tweets) at 90%, while maintaining moderate precision for classes 1 (depressed tweets) and 2 (suicidal tweets) at 74 and 75%, respectively. Moreover, it exhibited high recall for class 0 (non-depressed tweets) at 97%, indicating its ability to accurately identify the majority of non-depressed tweets, although it showed slightly lower recall for classes 1 and 2 at 69 and 74%, respectively.

5.7 BERT (distil-BERT)

Table 3 presents the results of 4 epochs for distil-BERT model. The evaluation of the BERT model, conducted over four epochs, yielded promising results, showcasing its efficacy in classifying Twitter posts based on sentiment. With an average accuracy of 84.8%, the model demonstrated a commendable ability to correctly identify the sentiment of tweets. Moreover, the average F1-score of 84.7% reflects a balanced performance in terms of precision and recall, indicating the model's capability to accurately classify both positive and negative sentiment tweets. Precision measures the proportion of correctly predicted positive cases out of all predicted positive cases. With an average precision of 86.3%, the BERT model has a high percentage of correct predictions for all sentiment categories. Recall, also known as sensitivity, measures the proportion of correctly predicted positive cases out of all actual positive cases. With an average recall of 85.6%, the BERT model effectively captures the most positive cases for all sentiment categories. Overall, these results underscore the BERT model's robustness and superior performance compared to traditional machine learning approaches, positioning it as the optimal choice for detecting depression in Twitter posts.

Table 3
Results of 4-epoch for distil-BERT model

Epoch	Training Loss	Accuracy	F1
1	0.415	0.843	0.842
2	0.277	0.855	0.856
3	0.205	0.841	0.849
4	0.137	0.854	0.854

5.8 Best Model

After evaluating the performance of various machine learning models for classifying tweets based on sentiment, it is evident that each model exhibits strengths and limitations in accurately predicting depression-related tweets. However, considering the achieved accuracy, precision and recall scores, the BERT model stands out as the most promising candidate for this task. With an average accuracy of 84.8%, coupled with balanced precision and recall scores across all classes, BERT demonstrates superior performance in identifying depressive sentiments in tweets compared to other models evaluated in this study. Figure 3 is the side navigation bar for web applications.

5.9 Dashboard Development

The dashboard was developed using Streamlit, a popular Python framework for creating web applications with minimal effort. The development process involved coding in Python within the Visual Studio Code (VSCode) environment. The dashboard comprises a side navigation bar with three pages. Firstly, ‘Data Used’ page provides users with information about the datasets used in the prediction system. The ‘Help and Insights’ page is page where users can access resources and information about seeking help for depression and related mental health issues. Lastly the ‘Twitter Depression Prediction’ page (Figure 4) is the main function of the dashboard, where users can input a Twitter username and specify the number of pages of tweets to fetch for analysis.

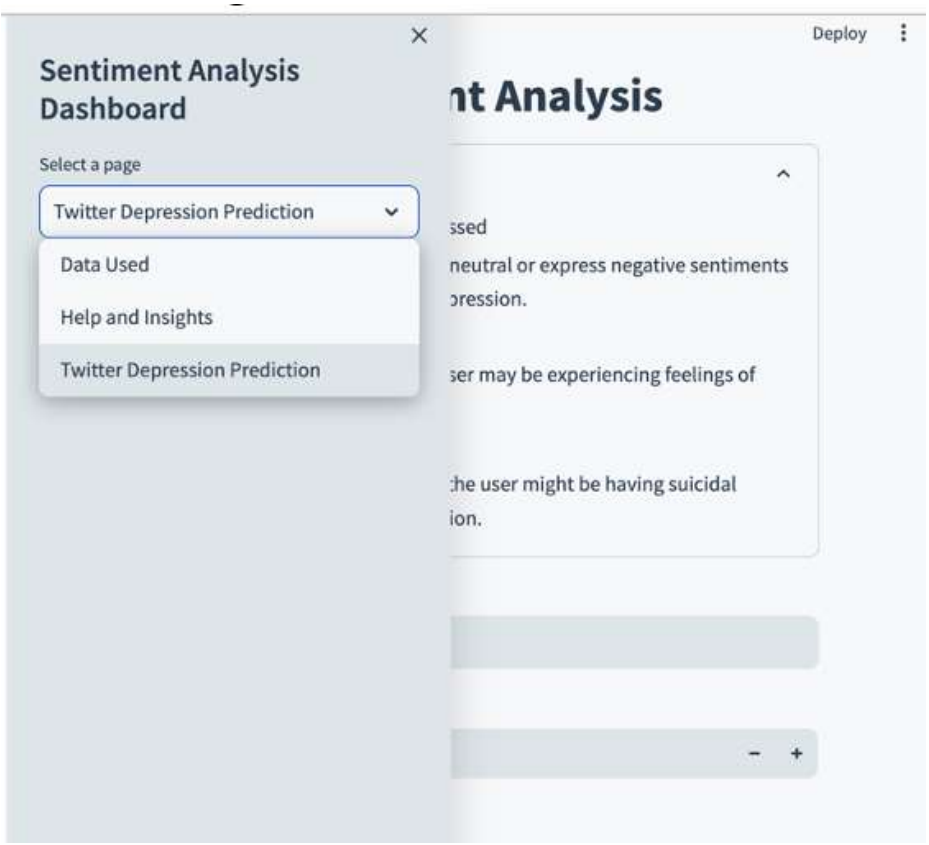

Fig. 3. Side bar navigation

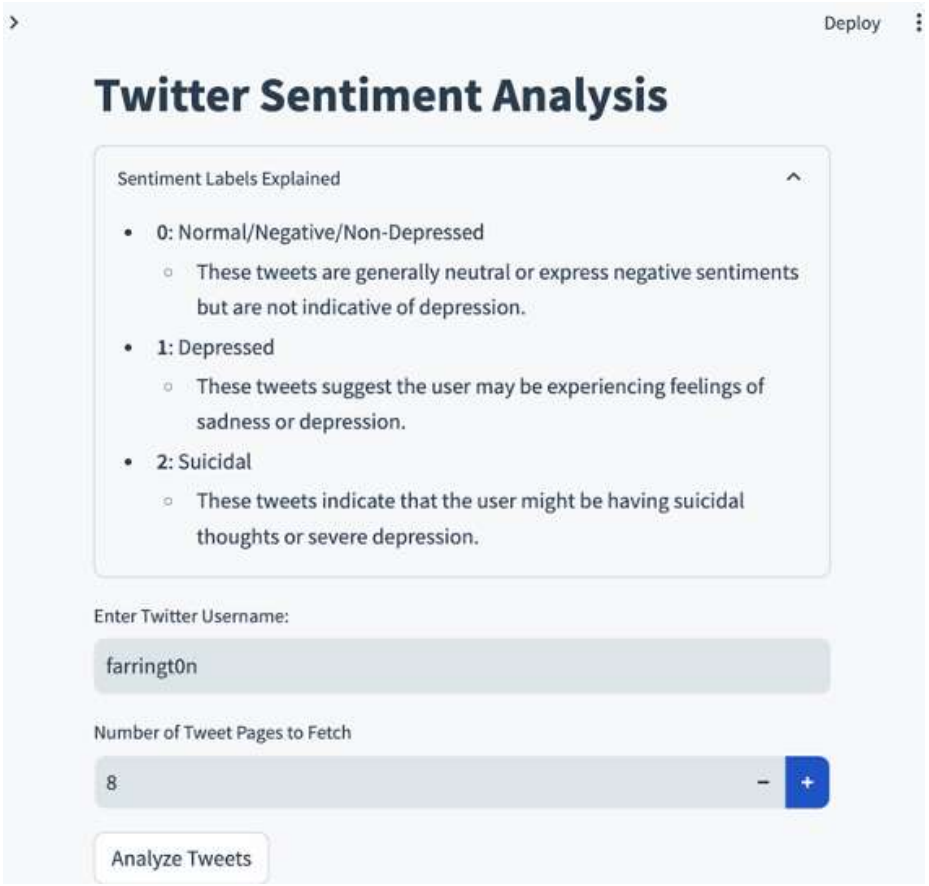

Fig. 4. Twitter depression prediction page

In the ‘Twitter Depression Prediction’ section, users will find an explanation of the sentiment labels used in the prediction system, which are labels 0 (normal, negative and nondepressed), 1 (depressed) and 2 (suicidal), enabling them to comprehend the sentiment analysis results effectively. Firstly, users need to input a Twitter username into the designated field, which will enable users to specify the Twitter account for which they want to analyse the tweets. Users have the option to specify the number of pages of tweets they want to fetch for analysis. Each page typically consists of 20 tweets. This flexibility allows users to control the volume of data retrieved for analysis, depending on their preferences and requirements. Once users have input the desired Twitter username and specified the number of pages to fetch, they can click the "Analyse Tweets" button to initiate the sentiment analysis process. This action triggers the system to retrieve the specified tweets, analyse their sentiment using the predefined models and present the analysis results to the user. Once the data is fetched, the BERT model will process the tweets and classify them. A pie chart illustrated in Figure 5 presents the distribution of sentiments across the analysed tweets. This visualisation provides users with a quick overview of the proportion of tweets classified into different sentiment categories (0, 1 and 2). Boxplot visualisations summarise the distribution of tweet lengths for each sentiment category. This graphical representation helps users understand how tweet length varies across different sentiment labels, offering insights into the characteristics of tweets associated with varying sentiment levels.

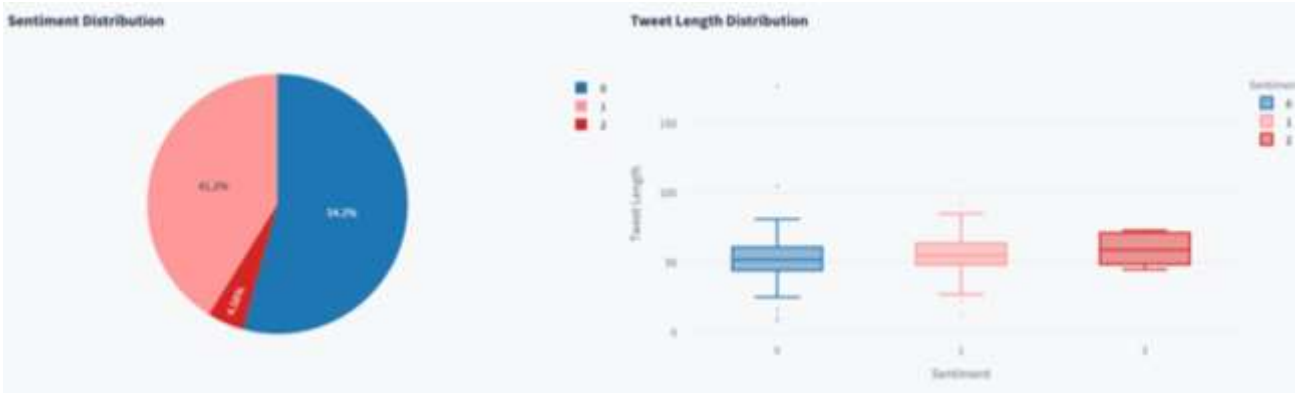

Fig. 5. Visualization of sentiment distribution and Tweet length distribution

For mental health professionals, Figure 6 is an example of a visualisation that is able to provide an insightful backdrop for initial consultations. It allows them to quickly catch emotional fluctuations and potential patterns of concern without relying solely on a patient’s recollection. From the patient’s perspective, this chart serves as a reflective tool, enabling them to visualise and articulate their experiences more accurately during sessions. It can jog memories about their emotional state at specific times, which is often challenging to recall in detail. Moreover, these visual cues can prompt further exploration during consultations. The health professionals are able to look at ‘depressed’ or 'suicidal'’ sentiments and this can help them make a diagnosis of the patient about what was happening during that time, which provides a starting point for conversations.

Fig. 6. Visualization of sentiment label over time

6. Conclusions

This study demonstrates the potential of leveraging Twitter data for early depression detection through advanced machine learning techniques. Among the models tested, the BERT model achieved the highest accuracy at 84.8%, significantly outperforming traditional machine learning approaches. This result highlights the effectiveness of deep learning in capturing linguistic cues and sentiment patterns associated with depression, offering a valuable tool for mental health monitoring. Beyond its technical success, this research carries meaningful real-world implications. By enabling early detection of depressive tendencies, this system could support timely intervention strategies, helping mental health professionals identify individuals who may need assistance. Integrating such models into digital health platforms or mental health support services could provide real-time insights, ultimately improving the well-being of those at risk. However, this study has some limitations. The analysis is based solely on textual data, overlooking other modalities such as images, videos or voice notes, which could offer a more comprehensive understanding of emotional distress. Additionally, while the dataset is diverse, it may not fully capture linguistic and cultural variations in how depression is expressed across different demographics. Future research should focus on expanding dataset diversity and incorporating multi-modal analysis to enhance accuracy and generalizability. Furthermore, ethical considerations such as data privacy, consent and the potential risks of misdiagnosis must be carefully addressed before deploying this technology in real-world applications. Ultimately, this study highlights the transformative role of AI and social media analytics in mental health research. With further refinement and responsible implementation, such tools could help bridge the gap in mental health care, ensuring more individuals receive the support they need, when they need it most.

References

[1] World Health Organization. *World mental health report: Transforming mental health for all*. World Health Organization, 2022.
[2] National Institute of Mental Health. "Depression," (2019). https://www.nimh.nih.gov/health/topics/depression
[3] World Health Organization. "Depressive Disorder (Depression) World Health Organization." *Geneva, Switzerland* (2023).
[4] Shoib, Sheikh. *Handbook Of Suicide Prevention*. Lulu Publication, 2020.
[5] Driscoll, Dana Lynn. "Introduction to primary research: Observations, surveys and interviews." *Writing spaces: Readings on writing* 2, no. 2011 (2011): 153-174.
[6] Ghani, Miharaini Md, Wan Azani Wan Mustafa, Mohd Ekram Alhafis Hashim, Hafizul Fahri Hanafi and Durratul Laquesha Shaiful Bakhtiar. "Impact of generative AI on communication patterns in social media." *Journal of Advanced Research in Computing and Applications* 26, no. 1 (2022): 22-34.
[7] Sohail, Shahab Saquib, Mohammad Muzammil Khan, Mohd Arsalan, Aslam Khan, Jamshed Siddiqui, Syed Hamid Hasan and M. Afshar Alam. "Crawling Twitter data through API: A technical/legal perspective." *arXiv preprint arXiv:2105.10724* (2021).
[8] Jain, Shikha, Kavita Pandey, Princi Jain and Kah Phooi Seng, eds. *Artificial intelligence, machine learning and mental health in pandemics: a computational approach*. Academic Press, 2022.
[9] Pachouly, S. J., Gargee Raut, Kshama Bute, Rushikesh Tambe and Shruti Bhavsar. "Depression detection on social media network (Twitter) using sentiment analysis." *Int. Res. J. Eng. Technol* 8, no. 1 (2021): 1834-1839.
[10] Garg, Muskan. "Mental health analysis in social media posts: a survey." *Archives of Computational Methods in Engineering* 30, no. 3 (2023): 1819-1842. https://doi.org/10.1007/s11831-022-09863-z
[11] Zhang, Tianlin, Kailai Yang, Shaoxiong Ji and Sophia Ananiadou. "Emotion fusion for mental illness detection from social media: A survey." *Information Fusion* 92 (2023): 231-246. https://doi.org/10.1016/j.inffus.2022.11.031
[12] Chancellor, Stevie and Munmun De Choudhury. "Methods in predictive techniques for mental health status on social media: a critical review." *NPJ digital medicine* 3, no. 1 (2020): 43. https://doi.org/10.1038/s41746-020-0233-7
[13] Kim, Jina, Jieon Lee, Eunil Park and Jinyoung Han. "A deep learning model for detecting mental illness from user content on social media." *Scientific reports* 10, no. 1 (2020): 11846. https://doi.org/10.1038/s41598-020-68764-y

[14] Ansari, Luna, Shaoxiong Ji, Qian Chen and Erik Cambria. "Ensemble hybrid learning methods for automated depression detection." *IEEE transactions on computational social systems* 10, no. 1 (2022): 211-219. https://doi.org/10.1109/TCSS.2022.3154442

[15] Pachouly, S. J., Gargee Raut, Kshama Bute, Rushikesh Tambe and Shruti Bhavsar. "Depression detection on social media network (Twitter) using sentiment analysis." *Int. Res. J. Eng. Technol* 8, no. 1 (2021): 1834-1839.

[16] Uban, Ana-Sabina, Berta Chulvi and Paolo Rosso. "An emotion and cognitive based analysis of mental health disorders from social media data." *Future Generation Computer Systems* 124 (2021): 480-494. https://doi.org/10.1016/j.future.2021.05.032

[17] Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee and Huan Liu. "Fakenewsnet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media." *Big data* 8, no. 3 (2020): 171-188. https://doi.org/10.1089/big.2020.0062

[18] Zafar, Abaid Ullah, Jiangnan Qiu, Ying Li, Jingguo Wang and Mohsin Shahzad. "The impact of social media celebrities' posts and contextual interactions on impulse buying in social commerce." *Computers in human behavior* 115 (2021): 106178. https://doi.org/10.1016/j.chb.2019.106178

[19] Madhu, Hiren, Shrey Satapara, Sandip Modha, Thomas Mandl and Prasenjit Majumder. "Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments." *Expert Systems with Applications* 215 (2023): 119342. https://doi.org/10.1016/j.eswa.2022.119342

[20] Lin, Chenhao, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou and Henry Leung. "Sensemood: depression detection on social media." In *Proceedings of the 2020 international conference on multimedia retrieval*, pp. 407-411. 2020. https://doi.org/10.1145/3372278.3391932

[21] McCarthy, Peter A. and Nexhmedin Morina. "Exploring the association of social comparison with depression and anxiety: A systematic review and meta-analysis." *Clinical psychology & psychotherapy* 27, no. 5 (2020): 640-671. https://doi.org/10.1002/cpp.2452

[22] Vidal, Carol, Tenzin Lhaksampa, Leslie Miller and Rheanna Platt. "Social media use and depression in adolescents: a scoping review." *International Review of Psychiatry* 32, no. 3 (2020): 235-253. https://doi.org/10.1080/09540261.2020.1720623

[23] Alaparthi, Shivaji and Manit Mishra. "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey." *arXiv preprint arXiv:2007.01127* (2020).

[24] Chakkarwar, Vrishali, Sharvari Tamane and Ankita Thombre. "A review on BERT and its implementation in various NLP tasks." In *International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*, pp. 112-121. Atlantis Press, 2023. https://doi.org/10.2991/978-94-6463-136-4_12

[25] Zhang, Zhuosheng, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou and Xiang Zhou. "Semantics-aware BERT for language understanding." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, pp. 9628-9635. 2020. https://doi.org/10.1609/aaai.v34i05.6510