BERT vs. Logistic Regression: Classifying Mental Health-Related Text using Machine Learning and Natural Language Processing

Amila Hrnjić
Department of Information Technologies Faculty of Engineering, Natural and Medical Sciences International Burch University Sarajevo, Bosnia and Herzegovina

Zerina Altoka
Department of Information Technologies Faculty of Engineering, Natural and Medical Sciences International Burch University Sarajevo, Bosnia and Herzegovina

ABSTRACT
With the rise of mental health discussions in online spaces, the ability to automatically detect emotionally sensitive content has become increasingly important. This study compares traditional machine learning (ML) methods with deep learning models to evaluate their effectiveness in classifying mental health-related texts. Two models were tested: Logistic regression (LR) with TF-IDF features and the BERT transformer model. A balanced dataset containing labeled text samples was used, with standard natural language processing (NLP) preprocessing applied. Model performance was evaluated using precision, recall, F1-score, and AUC. Results show that BERT outperforms logistic regression across all metrics, achieving an F1-score of 0.95 and an AUC of 0.99. Confusion matrices and ROC curves confirmed BERT's superior accuracy and its ability to reduce false classifications. These findings highlight the strength of deep learning models in understanding nuanced language, which is crucial in the mental health domain. Overall, the study confirms that transformer-based models like BERT offer a more reliable approach to classifying emotionally sensitive content, with promising applications in early detection tools and mental health support systems.

General Terms
Machine Learning, Natural Language Processing

Keywords
Mental health, Text classification, BERT, Machine Learning, NLP.

1. INTRODUCTION
In recent years, the analysis of text data related to mental health has become increasingly important due to the growing prevalence of mental health issues and the rise of online platforms where users frequently express emotional and psychological states. Automatic classification of such texts, particularly those containing emotionally sensitive content, plays a crucial role in early detection, intervention, and support. Traditional text classification methods, such as logistic regression combined with TF-IDF vectorization, have been widely used due to their simplicity and interpretability. However, these models often fall short when it comes to understanding the deeper semantic and contextual meanings of language, which are especially important in domains like mental health. With the advancement of deep learning and NLP, pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated remarkable performance improvements across various NLP tasks. These models are capable of capturing complex linguistic patterns and context, which traditional models often miss.

This research focuses on evaluating whether deep learning models like BERT outperform traditional ML approaches in classifying sensitive mental health-related texts. Specifically, this study investigates whether the application of deep learning techniques can improve classification accuracy compared to traditional TF-IDF and logistic regression methods, hypothesizing that BERT's contextual understanding will yield significantly higher performance metrics due to its ability to capture semantic relationships in mental health discourse.

The study addresses a critical gap in understanding how different computational approaches handle the nuanced language patterns characteristic of mental health discourse, where context, implicit meaning, and emotional subtleties play crucial roles in accurate classification. Given the complexity of emotional expression in text, particularly regarding mental health content, the study hypothesizes that BERT will show superior performance in detecting emotionally sensitive content, especially in cases involving implicit emotional expressions and contextual sentiment that traditional bag-of-words approaches typically fail to capture. By systematically comparing the performance of both approaches, this study aims to determine which method is more effective for the reliable detection of emotionally charged content in text, providing evidence-based recommendations for mental health text classification applications.

This paper is structured as follows: section 2 presents a literature review, summarizing recent studies that have applied NLP and ML techniques for detecting mental health-related content. Section 3 provides a detailed description of the dataset used in this study, including its structure and class distribution. Section 4 outlines the methodology, covering text preprocessing, cleaning, TF-IDF vectorization, and model training procedures. Section 5 presents the experimental results, comparing the performance of the Logistic Regression and BERT models through various evaluation metrics. Finally, Section 6 concludes the paper by discussing key findings, addressing limitations, and proposing directions for future research.

2. RELATED WORKS
Automatically detecting emotionally sensitive content in mental health-related texts is challenging but crucial for early intervention and support, given the increasing prevalence of mental health issues online.

Given the challenges in accurately detecting emotionally sensitive content, it is important to examine previous research that has applied ML and NLP techniques in the mental health domain. Understanding the methods, datasets, and findings of

these studies provides a foundation for evaluating and improving current classification approaches.

Le Glaz et al. conducted a systematic review to investigate the application of ML and NLP in mental health research. From an initial set of 327 articles, they selected 58 studies that focused on tasks such as symptom detection, illness classification, and risk factor identification. The majority of studies used clinical records or social media data, employing techniques like n-gram modeling and classifiers such as SVMs and random forests. Their review highlighted the potential of these methods for mental health analysis, while also noting challenges related to bias, ethical considerations, and the interpretation of social media data [1].

Bajaj et al. investigated the application of BERT for suicide risk prediction, addressing a critical global public health issue. Their study focused on classifying suicide-related textual data by extracting relevant textual biomarkers. The proposed BERT-based model achieved state-of-the-art accuracy, surpassing 97%, highlighting its effectiveness in handling complex text classification tasks. Additionally, the study discussed challenges specific to AI applications in mental health, including the absence of established biological markers for suicide risk, reliance on subjective data, potential biases in training datasets, and ethical considerations regarding data privacy. Overall, Bajaj et al. (2023) demonstrated the significant potential of BERT and similar NLP techniques for improving early detection and intervention strategies in mental healthcare, providing a relevant foundation for further research in text-based mental health analysis [2].

Yeskuatov et al. investigated the early detection of suicidal ideations using Reddit posts through various approaches, including traditional ML, deep learning, and fine-tuned transformer models. The study utilized a dataset of 187,943 Reddit posts labeled as suicidal or non-suicidal based on community affiliation. For ML, algorithms such as logistic regression, support vector machines, naïve Bayes, and XGBoost were trained on psycholinguistic features extracted with LIWC and NRC lexicons, with XGBoost achieving the highest accuracy of 97.05% and an F1-score of 97.16%. In the deep learning experiments, CNN, LSTM, and a combined LSTM+CNN architecture were trained on GloVe embeddings. CNN slightly outperformed LSTM in accuracy and F1-score, while LSTM achieved marginally better recall. The combined LSTM+CNN model had the highest recall of 98.17%. Finally, three variations of pre-trained BERT models—base BERT, small BERT, and ALBERT—were fine-tuned for suicidal ideation detection.

The base BERT model outperformed all other methods, reaching 98.97% accuracy and a 99.01% F1-score, while ALBERT achieved a slightly higher recall of 99.19%. The study demonstrated that fine-tuned BERT models surpass both traditional ML and deep learning approaches for detecting suicidal ideations on Reddit, without the need for task-specific feature engineering [3].

Jain et al. explored the use of ML and NLP to analyze depression and suicidal ideation in posts from Reddit's r/depression and r/SuicideWatch subreddits. Using a dataset of over 60,000 posts, they applied Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and Random Forest to classify posts as related to depression or suicide. The study found that Logistic Regression and Random Forest had the highest accuracy (77.29%) and f1-scores (0.77), while Naïve Bayes performed the least effectively with 75.87% accuracy. The results suggest that ML models can effectively identify mental health risks in online discussions, providing valuable insights for mental health professionals [4].

Cook et al. investigated the potential of NLP and ML to predict suicidal ideation and psychiatric symptoms among adults recently discharged from psychiatric inpatient or emergency care settings in Madrid, Spain. The study utilized both structured survey data—such as questions related to sleep and general well-being—and unstructured text responses to a single open-ended question, "How do you feel today?" The researchers compared the performance of NLP-based models using these free-text responses with traditional logistic regression models based on structured data. Results showed that models using structured data achieved higher predictive accuracy, with a sensitivity of 0.76 compared to 0.56 for the NLP-based models. Similarly, the positive predictive value for the NLP model predicting suicidal ideation reached 0.61, while the structured data-based model achieved 0.73. These findings indicate that although traditional logistic regression models demonstrated superior performance, NLP-based models were still capable of generating relatively strong predictive values using minimal textual input. The study highlights the potential of language-based models as a low-cost and scalable tool for identifying individuals at risk of suicide or psychological distress, particularly in situations where extensive structured surveys are not feasible [5].

Bokolo, B. G., & Liu, Q. investigated the application of various ML and transformer models for detecting depressive content on social media. Using the Sentiment140 and Suicide-Watch datasets, they compared the performance of Logistic Regression with transformer architectures such as RoBERTa and DeBERTa. On the Sentiment140 dataset, RoBERTa and DeBERTa achieved the highest performance, with 98% accuracy and an F1-score of 98%, outperforming traditional models.

Conversely, on the Suicide-Watch dataset, Logistic Regression performed better (93.5% accuracy, 93.5% F1-score) compared to transformer models, which was attributed to simpler patterns in the data. These results demonstrate that transformer models effectively capture contextual nuances in complex text, while classical models like Logistic Regression can be more efficient for straightforward classification tasks. The study highlights the potential of combining classical and advanced approaches for early detection of depression and suicidal ideation in online content [6].

Viani et al. explored the use of NLP to identify the onset of psychosis from mental health electronic health records (EHRs). They manually annotated a corpus of EHRs and applied a paragraph classification approach to detect disease onset mentions. Temporal expressions were extracted using a rule-based system (SUTimeMentalHealth), and a supervised ML model was developed for classification. When tested on 31 patients, the model correctly identified the onset date among the top three predictions for 71% of cases, demonstrating high reliability in temporal extraction. Applied at scale, the approach estimated onset dates for 2,483 patients, revealing that patients seen by early intervention teams experienced shorter delays in treatment. This study highlights the potential of NLP to enhance the identification of psychosis onset, enabling more timely interventions. In the current work, the focus has been on text processing using traditional NLP techniques, where text cleaning and vectorization were performed using the TF-IDF approach [7].

Spiliotis conducted a comprehensive comparative study on predicting depression and suicidal ideation from social media

posts using ML and transformer-based NLP models. The research analyzed Twitter and Reddit datasets containing depressive and suicidal content and applied extensive preprocessing techniques, including tokenization, lemmatization, TF-IDF, and word embeddings. Several ML models were evaluated, including Logistic Regression, Random Forest, BERT, and DistilBERT. The results demonstrated that transformer-based models substantially outperformed traditional ML methods. In particular, DistilBERT achieved an F1-score of 0.99 on the Depression Twitter dataset, indicating extremely strong capability in identifying depressive language. Although Random Forest also performed well, it was slightly inferior to transformer models, confirming the advantage of context-aware architectures like BERT in mental health prediction tasks. The study further emphasized the need for ethical considerations, multimodal approaches, and continuous model adaptation to evolving linguistic patterns.logies for better mental health care [8].

Overall, the reviewed studies demonstrate the significant potential of ML and NLP techniques in detecting mental health-related content and predicting mental health outcomes. Transformer-based models, such as BERT and RoBERTa, generally outperform traditional ML approaches due to their ability to capture contextual and linguistic nuances in text. However, challenges remain, including limited dataset sizes, annotation bias, ethical concerns, and the need for scalable solutions that can operate effectively across diverse data sources. The reviewed research highlights the importance of combining advanced NLP architectures with practical considerations, such as model interpretability and computational efficiency.

Building on these findings, the present work investigates the classification of emotionally sensitive content in mental health-related texts by comparing traditional ML methods, specifically Logistic Regression with TF-IDF features, with transformer-based models like BERT. This comparison aims to evaluate the effectiveness of both approaches, highlighting the advantages of context-aware deep learning models while considering the accessibility and efficiency of traditional NLP techniques for early detection and intervention.

3. METHODOLOGY

This study follows a structured pipeline for sentiment analysis, including data collection, preprocessing, text cleaning, and splitting, followed by model training and evaluation using Logistic Regression and BERT. The goal is to compare the performance of traditional ML models and deep learning techniques. In addition, the preprocessing stage ensures that the textual data is transformed into a standardized format suitable for both traditional machine learning algorithms and transformer-based models. Particular attention is paid to tokenization and input representation, as these steps differ significantly between LR and BERT. The methodology steps are summarized in Fig. 1.

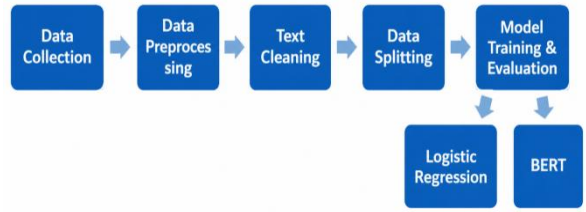

Fig 1: Overview of the proposed methodology for mental health text classification

3.1 Dataset

The dataset used in this study is the Mental Health Classifier - NLP dataset, obtained from Kaggle [9]. The dataset consists of 27,972 unique text samples, each labeled for the presence (1) or absence (0) of mental health-related content. The data is primarily used to train and evaluate models for binary text classification in the context of mental health detection. The dataset was selected due to its practical relevance for NLP applications in mental health screening.

Fig.2 shows the class distribution in the dataset. From the chart, it can be observe that there are:

- 14,139 samples with label 0 (neutral),
- 13,838 samples with label 1 (mental health concern).

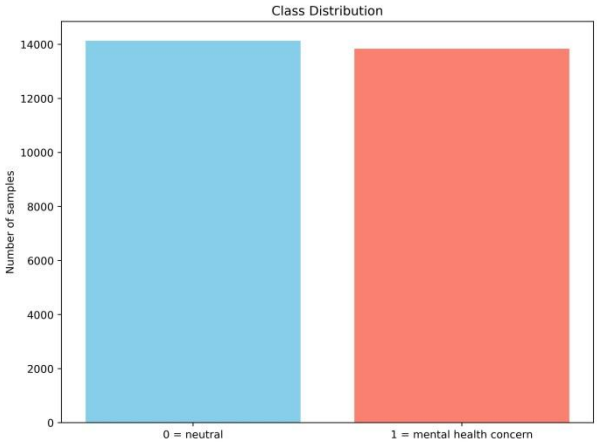

Fig 2: Distribution of class labels (0 – non-depression, 1 – depression) in the dataset

This balanced distribution is beneficial because it reduces the risk of model bias toward the majority class and supports more reliable evaluation.

Table 1 presents the structure of the Student Mental Health dataset, including the column names and their corresponding descriptions.

Table 1. Analysis of the Mental Health Dataset

Column Name	Description
text	A user-generated text sample representing a thought or message
label	1 = mental health concern present; 0 = absent

3.2 Data Preprocessing

Prior to model training, the dataset underwent several preprocessing steps to ensure that it was suitable for text classification.

This phase aimed to prepare the raw data for further analysis by addressing basic formatting and structural issues. The preprocessing included:

- Exploratory Analysis: The dataset was inspected using functions such as data.head() and data.info() to understand its structure and check for any irregularities, such as missing or null values.
- Label Verification: The target variable (label) was

examined to confirm that it consisted of binary values (0 and 1), indicating whether a text contained mental health-related content.
- Dataset Consistency: It was verified that all rows contained text data and corresponding labels. No severe anomalies or inconsistencies were detected

To further improve the quality of the textual data, a text cleaning procedure was applied. Raw text data often contains noise such as URLs, user tags, punctuation, and numbers, which can negatively impact model performance.

3.3 Text Cleaning

To prepare the text data for further processing and analysis, a thorough text cleaning procedure was applied. Raw text data often contains noise such as URLs, user tags, punctuation, and numbers, which can negatively impact model performance. Figure 3 illustrates the sequence of preprocessing steps applied to the raw text, including lowercasing, removing URLs, eliminating user tags, stripping punctuation, and reducing extra spaces.

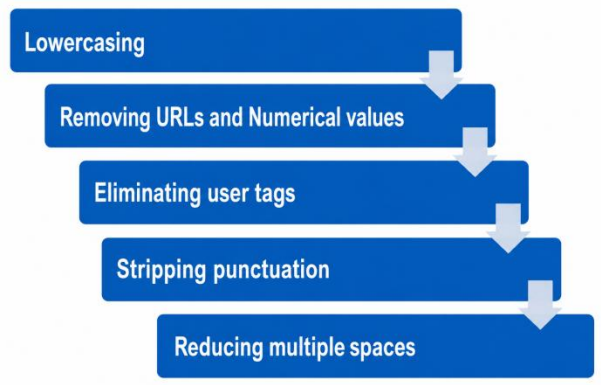

Fig 3: Text preprocessing pipeline applied to the dataset

The following cleaning steps were performed:

- Lowercasing all text to eliminate case sensitivity (e.g., "Happy" and "happy" become the same).
- Removing URLs and Numerical Values. This step eliminates URLs and any numerical values from the text. URLs are typically irrelevant in the context of text classification, while numbers are usually not informative for sentiment or mental health classification tasks.
- Eliminating user tags (e.g., @username) and hashtags (e.g., #topic) as they do not contribute meaningfully to text classification.
- Stripping punctuation to reduce noise and standardize the text.
- Reducing multiple spaces to a single.

These transformations were implemented using a custom clean_text() function, which was then applied to all text data in the text column. The result was stored in a new column called clean_text, which contains the cleaned version of the original text and is used for vectorization and model training.

3.4 Data Splitting

To properly evaluate the performance of the machine learning models, the dataset was divided into two parts: a training set and a test set. The training set is used to train the model, while the test set is used to assess the model's generalization ability on unseen data.

In this study, the dataset was split with 70% of the data used for training and 30% used for testing.

This split ensures that the model has a sufficient amount of data to learn from, while also providing a fair evaluation using a separate, unseen test set. After the split, the sizes of the sets were as follows:

- Training set: 19,583 samples (70% of the dataset)
- Test set: 8,394 samples (30% of the dataset)

This ensures a balanced and robust evaluation framework for the subsequent training and testing of the models.

3.5 Model Training and Evaluation

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7").

3.5.1 Logistic Regression

Before training the Logistic Regression model, it is necessary to convert the text data into numerical form, since machine learning models cannot work directly with raw text. One common method for this is TF-IDF (Term Frequency - Inverse Document Frequency).

TF-IDF is a statistical measure used to evaluate how important a word is to a document in a collection or corpus [10]. It combines two components:

1. Term Frequency (TF): Measures how frequently a term appears in a document, normalized by the total number of words in that document [10].

$$TF=\frac{Totalapperianceofawordindocument}{Totalwordsinadocument} \quad (1)$$

2. Inverse Document Frequency (IDF): Measures how important a term is across all documents. It reduces the weight of common words that appear in many documents and increases the weight of rare ones [10].

$$IDF = log\frac{AllDocumentNumber}{DocumentFrequency} \quad (2)$$

The TF-IDF score for each term is computed as the product of Term Frequency and Inverse Document Frequency [10].

$$TF - IDF = TF \cdot IDF \quad (3)$$

This technique helps reduce the influence of commonly used words (like "the", "and", "is") and highlights more meaningful terms, which can improve model performance [10].

In this project, TF-IDF is applied to the training text data to transform it into numerical feature vectors. These TF-IDF vectors are then used as input for training the Logistic Regression model.

After transforming the text into TF-IDF vectors, the next step is to train the Logistic Regression model. Logistic regression learns a function [11]:

$$h: \mathbb{R}^d \rightarrow [0,1] \quad (4)$$

which represents the probability that an input example belongs to class 1 [11]. The hypothesis is constructed by applying a sigmoid activation function on top of a linear function. The sigmoid (logistic) function is defined as [11]:

$$\sigma(z) = \frac{1}{1+\exp(-z)} \quad (5)$$

Here,

$$z = w \cdot x \quad (6)$$

is the dot product between the weight vector and the input features. is large, approaches 1; when it is a large negative value, it approaches 0. When w·x≈0, the model outputs a probability close to 0.5, indicating uncertainty.

To train the model, logistic regression uses the logistic loss (also called log loss or binary cross-entropy). For a training example with label y∈{1,−1} ,the loss is [11]:

$$l(h_w(x), y) = \log(1 + \exp(-y(w \cdot x))) \quad (7)$$

This loss increases when the model assigns a low probability to the true class and decreases when the predicted probability aligns with the correct label. The output of the sigmoid function can be interpreted as the likelihood that the input belongs to class 1. When the predicted probability is close to 1, the model is confident the instance is positive; when it is close to 0, the model is confident the instance is negative. Predictions near 0.5 indicate uncertainty

The logistic loss is a convex function, meaning it can be optimized efficiently using gradient descent or related optimization algorithms. Minimizing this loss is also equivalent to performing Maximum Likelihood Estimation (MLE) under a Bernoulli distribution assumption. In summary, logistic regression provides a probabilistic and interpretable approach to binary classification by learning parameters that minimize a convex loss function [11].

3.5.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a transformer-based language representation model that has substantially advanced the field of NLP [12]. Unlike traditional embedding methods that generate static word representations, BERT produces contextual embeddings, allowing the meaning of each word to be interpreted based on its surrounding context. The model's bidirectional training enables it to capture relationships from both left and right contexts, resulting in improved understanding of semantic and syntactic structures in text [13].

The architecture of BERT is built on multiple transformer encoder layers that rely on self-attention mechanisms to process textual input efficiently. During pre-training, BERT is trained using masked language modeling and next sentence prediction, which allows the model to learn deep contextual dependencies and linguistic patterns. Once pre-trained, BERT can be fine-tuned on specific tasks with minimal adjustments, achieving strong performance across diverse NLP applications such as sentiment analysis, document classification, question answering, and domain-specific text processing [13].

Despite its strong performance, BERT's large computational requirements present practical challenges, motivating the development of optimized variants and more efficient transformer-based architectures. Nevertheless, BERT remains one of the foundational models in modern NLP, influencing subsequent research and becoming a standard approach for many language understanding tasks [13].

For this study, the BERT model was fine-tuned using the bert-base-uncased pretrained architecture. The model was adapted for binary text classification by adding a classification layer with two output classes corresponding to the labels in the dataset. Before feeding the text into the model, each sentence was tokenized using the BertTokenizer, which converts text into subword tokens and generates the following input features:

- input_ids – token indices corresponding to the BERT vocabulary
- attention_mask – binary mask indicating which tokens should be attended to
- token_type_ids – disabled in this case (set to default), as the task uses single-sentence inputs

A maximum sequence length of 128 tokens was used, and all sequences were padded or truncated to this fixed size to ensure uniformity across batches.

During fine-tuning, the following hyperparameters were applied:

- Learning rate: 2e-5
- Batch size: 8
- Optimizer: AdamW, which is recommended for transformer models
- Number of epochs: 1 (sufficient due to the size of the dataset and computational constraints)
- Loss function: Cross-entropy loss, computed automatically through BertForSequenceClassification

The model was fine-tuned on the cleaned training dataset using backpropagation, where both the transformer layers and the classification head were updated. After training, the model was evaluated on the test set by generating logits and selecting the predicted class using the argmax function. The resulting predictions were compared with the true labels to calculate accuracy and classification metrics.

This configuration enabled BERT to learn task-specific patterns from the mental health dataset, achieving high classification performance while preserving the model's contextual understanding of language.

4. RESULTS AND DISCUSSION

4.1 Evaluation on Original Dataset

The following part describes and interprets the results obtained from the experiments, showing how well the proposed approach performs in text classification. For the purpose of a detailed evaluation of text classification performance, an experimental comparison was conducted between a traditional machine learning model — Logistic Regression — and a modern deep learning model — BERT. This evaluation aimed to quantitatively assess how much the deep learning model (BERT) improves performance compared to the traditional method, using standard metrics such as accuracy, precision, recall, F1-score, and AUC (Area Under the Curve). Below are the evaluation results and a discussion of the performance achieved by both models.

To evaluate the ability of the models to correctly classify textual data, the class-wise F1-score results for the LR and BERT models were compared.

Table 2 presents the comparative performance results of both models across four standard evaluation metrics.

Table 2. Comparison of classification performance metrics

Model	Precision	Recall	F1-score	AUC
Logistic Regression	0.92	0.92	0.92	0.97
BERT	0.95	0.95	0.95	0.99

As presented in Table 2, the BERT model demonstrates superior performance compared to the traditional Logistic Regression model across all three core metrics: precision, recall, and F1-score.

In addition to the reported metrics, both models achieved high accuracy on the test dataset. Logistic Regression achieved an accuracy of 0.92, while BERT achieved a higher accuracy of 0.95, further confirming the superiority of the transformer-based approach.

BERT achieved a precision and recall of 0.95, indicating that it not only correctly identifies relevant instances but also minimizes false positives and false negatives more effectively than Logistic Regression. The F1-score of 0.95 further confirms the strong balance between precision and recall in BERT's predictions. On the other hand, Logistic Regression, while still performing well with values around 0.91–0.92, shows slightly reduced capability in capturing the full complexity of the data, likely due to its simpler linear nature. These results suggest that transformer-based models like BERT are better suited for tasks involving nuanced text classification, offering improved generalization and deeper contextual understanding. To further assess the models' classification capabilities, confusion matrices were generated for both the Logistic Regression and BERT models. These matrices provide a visual representation of the number of correct and incorrect predictions for each class, allowing for a more detailed understanding of the strengths and weaknesses of each approach in classifying textual data.

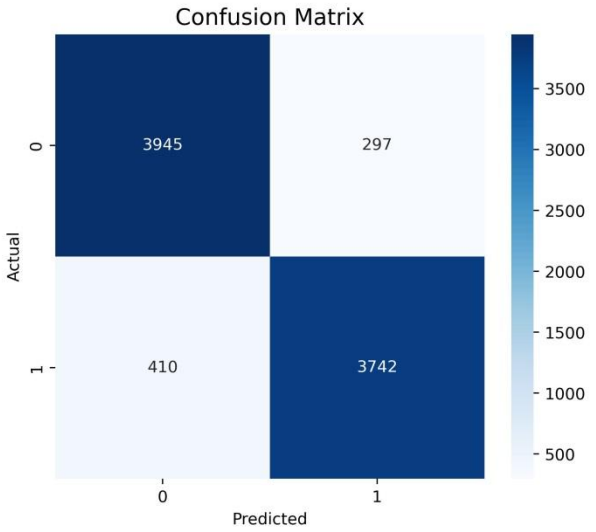

Fig 4: Confusion matrix of the Logistic Regression model on the test dataset

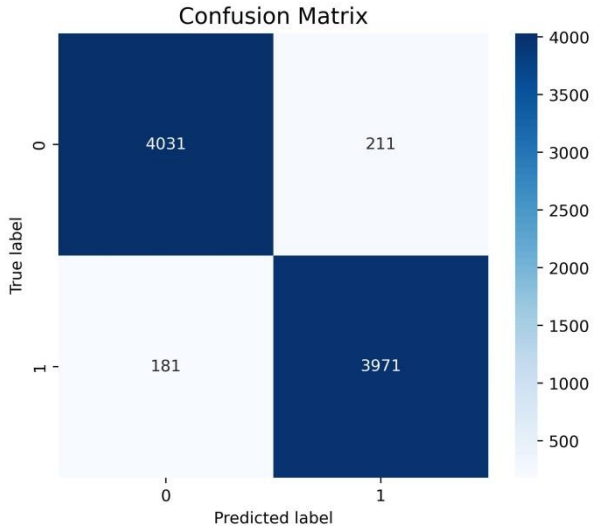

Fig 5: Confusion matrix of the BERT model on the test dataset

Figure 4 shows the confusion matrix for the Logistic Regression model. As illustrated, the model correctly classified 3,945 negative samples and 3,742 positive samples. However, it misclassified 297 instances as false positives and 410 as false negatives. These errors suggest that while the model performs well overall, there are still significant misclassifications in both classes.

Figure 5 displays the confusion matrix for the BERT model. The results show a notable improvement, with 4,031 correctly classified negative samples and 3,971 correctly classified positive samples. The number of false positives is reduced to 211, and false negatives to 181. This significant reduction in misclassifications confirms that BERT achieves higher classification accuracy and overall reliability compared to the traditional logistic regression model. These visual insights reinforce the conclusion that BERT is more effective in accurately classifying both classes in the dataset, especially in scenarios where minimizing false predictions is critical.

Figure 6 shows the ROC AUC curve comparing Logistic Regression and BERT. Both models performed well, with Logistic Regression achieving an AUC of 0.97 and BERT slightly better with 0.99

The curves are close to the top-left corner, indicating high accuracy and low false positive rates. Overall, BERT outperformed Logistic Regression and proved to be more effective for this classification task. The confusion matrices and ROC AUC results further confirm the stronger classification capability of BERT, particularly in reducing false predictions and improving overall reliability in mental health text classification tasks.

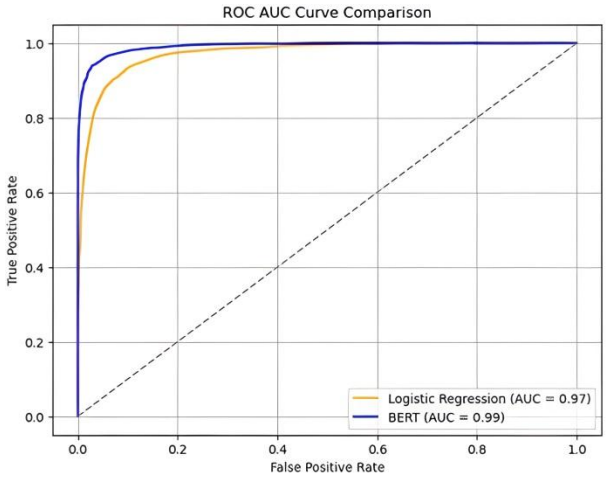

Fig 6: ROC AUC curve comparison between Logistic Regression and BERT

4.2 Evaluation on Unseen Dataset

To further assess the generalization capability of the proposed models, an additional evaluation was conducted on an unseen sentiment analysis dataset [14]. The dataset consists of labelled textual data representing positive and negative sentiment and was not used during the training phase of either model. The same preprocessing pipeline applied in the original experiment was used, including text cleaning and normalization. In order to address class imbalance, oversampling was performed to ensure equal representation of both classes.

Both the Logistic Regression and BERT models, previously trained on the original mental health dataset, were directly evaluated on this unseen dataset without any retraining or fine-tuning. The results show that Logistic Regression achieved an accuracy of 0.723, while BERT achieved an accuracy of 0.748. These findings indicate that both models retain a certain level of performance on unseen data; however, the observed decrease in accuracy compared to the original dataset confirms the impact of domain shift between different text sources. Despite this, BERT consistently shows slightly better robustness and generalization ability, reinforcing its effectiveness in handling diverse and previously unseen textual inputs.

The main goal of this study was to examine whether advanced deep learning models, particularly BERT, could outperform traditional machine learning methods such as Logistic Regression in text classification tasks. The findings clearly show that this objective has been met. BERT consistently achieved the highest scores across all key evaluation metrics — precision, recall, F1-score, and AUC — illustrating its stronger ability to understand context and meaning in text compared to traditional models.

In comparison to TF-IDF combined with Logistic Regression, BERT demonstrated a clear improvement in classification accuracy and overall reliability. Its ability to consider word meaning based on surrounding context allowed it to capture subtle differences in language that traditional methods often miss. This resulted in fewer misclassifications and a better overall understanding of emotionally sensitive content, which is particularly important in mental health–related texts.
Although the main experiments were based on a single dataset, additional evaluation on an unseen dataset was conducted to assess the generalization capability of the models. However, the evaluation was limited to a single additional dataset, and further testing on multiple datasets would be beneficial for a more comprehensive assessment of generalization performance.
Finally, while the model achieved high accuracy, it still makes some misclassifications, particularly in cases where the text expresses emotions subtly or unclearly — an area that could be improved with larger and more diverse datasets.

5. CONCLUSION

Understanding and correctly identifying emotionally sensitive content related to mental health has become increasingly vital in today's digital environment, where users frequently express psychological states through text. With the rise of online communication, the potential to automatically classify such content opens new avenues for timely support, intervention, and prevention.

In this study, a comparison was made between two fundamentally different approaches to text classification: a traditional logistic regression model using TF-IDF features and a deep learning approach using the BERT transformer model. The evaluation, based on a balanced dataset of mental health-related and neutral texts, revealed that BERT consistently achieved higher scores across all evaluation metrics, including precision, recall, F1-score, and AUC. Specifically, BERT reached an F1-score of 0.95 and AUC of 0.99, outperforming the logistic regression model, which achieved an F1-score of 0.92 and AUC of 0.97. Confusion matrices further illustrated BERT's advantage by showing fewer misclassifications, particularly in minimizing false negatives — a crucial factor in the context of mental health.

These findings underscore the strength of transformer-based models in handling subtle and context-dependent language.

Unlike traditional models that rely on word frequency and linear decision boundaries, BERT leverages contextual embeddings, allowing it to better understand the semantics and emotional weight behind words. As a result, it is particularly well-suited for complex classification tasks involving psychological and emotional expression.

Beyond model performance, this work highlights the practical potential of applying deep learning techniques to real-world problems. Reliable classification of sensitive content could support the development of digital tools aimed at mental health monitoring and intervention. However, several ethical and technical considerations remain, such as ensuring data privacy, minimizing bias, and avoiding overreliance on automated systems for decision-making in sensitive contexts.

Although the results are encouraging, the study has limitations. The analysis was primarily conducted on a single dataset with binary classification, with an additional evaluation performed on an unseen dataset to assess generalization performance. Future research should explore model adaptability across various domains, including social media platforms, multilingual data, or even specific mental health conditions. Investigating multiclass classification, fine-tuning domain-specific BERT variants, and integrating explainability mechanisms are all promising directions for expanding this work.

The application of BERT in classifying mental health-related text demonstrates a notable improvement over traditional approaches, offering deeper contextual understanding and more reliable predictions. As natural language processing continues to evolve, such models will play an increasingly important role in developing intelligent, ethical, and effective digital mental health tools.

6. REFERENCES

[1] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, and C. Lemey. Machine learning and natural language processing in mental health: Systematic review. Journal of Medical Internet Research, 23(5):e15708, 2021.

[2] K. Bajaj, M. Kumar, S. Jain, V. Bhardwaj, and S. Walia. Enhancing Suicide Risk Prediction through BERT: Leveraging Textual Biomarkers for Early Detection. International Journal of Intelligent Systems and Applications, 17(2):101–111, 2025.

[3] E. Yeskuatov, S.-L. Chua, and L. K. Foo. Detecting suicidal ideations on Reddit with transformer models. Artificial Intelligence and Human-Computer Interaction, IOS Press, 2025.

[4] P. Jain, K. R. Srinivas, and A. Vichare. Depression and Suicide Analysis Using Machine Learning and NLP. Journal of Physics: Conference Series, 2161(1):012034, 2022.

[5] B. L. Cook, A. M. Progovac, P. Chen, B. Mullin, S. Hou, and E. Baca-Garcia. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Computational and Mathematical Methods in Medicine, 2016:8708434, 2016.

[6] B. G. Bokolo and Q. Liu. Advanced Comparative Analysis of Machine Learning and Transformer Models for Depression and Suicide Detection in Social Media Texts. Electronics, 13(20):3980, 2024.

[7] N. Viani, R. Botelle, J. Kerwin, L. Yin, R. Patel, R. Stewart, and S. Velupillai. A natural language processing approach for identifying temporal disease onset information from mental healthcare text. Scientific Reports, 11(1), 2021.

[8] T. A. Spiliotis. Comparative analysis for mental health prediction tasks based on social media posts. Postgraduate diploma thesis, National Technical University of Athens, 2024.

[9] T. Sasaki. Mental Health Classifier – NLP. Kaggle dataset, 2022.

[10] B. Hans Christian, M. P. Agus, and D. Suhartono. Single Document Automatic Text Summarization using Term Frequency–Inverse Document Frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications, 7(4):285–294, 2016.

[11] S. Shalev-Shwartz and S. Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2019.

[13] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri. BERT applications in natural language processing: a review. Artificial Intelligence Review, 58:166, 2025.

[14] Shinigami. Sentimental Analysis. Kaggle dataset, 2021