High-Performance Suicide Ideation Detection Using a Transformer-Based Architecture

FRIKI IMED SALAH EDDINE[1], NAWAF R. ALHARBE[2], and SALEM MOHAMMED[1]

[1]LISYS Laboratory, University of Mascara, Mascara, Algeria
friki.imed@univ-mascara.dz, salem@univ-mascara.dz
[2]Department of Computer Scienceand Engineering, Madinah, Saudi Arabia
nrharbe@taibahu.edu.sa

Abstract

Suicide represents a major global public health problem, with hundreds of thousands of deaths each year. Early identification of at-risk individuals is a fundamental prevention strategy, but it is complicated by the significant gap between suicidal ideation, which is relatively common, and suicide attempts, which are rarer. Social media platforms, having become outlets for expressing psychological distress, provide an unprecedented source of data for early detection.This paper demonstrates the superiority of a fine-tuned Transformer architecture (BERT) for high-performance suicidal ideation detection. By comparing BERT against a range of classical machine learning and deep learning baselines, we show that our model achieves state-of-the-art results on a large, balanced dataset from Reddit. The performance of our approach is rigorously evaluated using a suite of standard metrics, with our best model achieving an **F1-Score of 0.977** and a **ROC AUC of 0.997**, significantly surpassing other deep learning and classical machine learning baselines. Beyond quantitative performance, we explore the model's interpretability through an analysis of self-attention mechanisms. Finally, we conduct an in-depth discussion on the limitations of our approach and the fundamental ethical considerations that the deployment of such technologies entails. We conclude that while advanced language models offer considerable potential for high-accuracy suicide risk classification, their integration into prevention strategies must occur within a responsible, human-centered, and clinically validated framework.

Keywords: Suicidal Ideation Detection, Natural Language Processing(NLP), Deep Learning, BERT, Machine Learning, AI in Mental Health, Information Technology Applications.

1 Introduction

Despite decades of prevention efforts, suicide remains one of the leading causes of death worldwide—and one of the most complex to predict. Suicide is both a profound human tragedy and a major global public health challenge. According to the World Health Organization (WHO), approximately 727,000 people died by suicide in 2021, an increase from 703,000 in 2019 [1]. Other estimates place the annual toll even higher, at nearly 740,000 deaths—equivalent to one death every 43 seconds [3]. Although the global age-standardized suicide rate declined slightly from 9.0 per 100,000 in 2019 to 8.9 in 2021, this trend obscures regional and demographic disparities and masks a worrying increase in the absolute number of deaths.

Certain populations are disproportionately affected. Suicide is the third leading cause of death among individuals aged 15–29 worldwide [4]. Moreover, over 73% of suicides occur in low- and middle-income countries, where mental health care is often severely limited or absent. However, this crisis is not limited to resource-constrained regions. High-income countries also face significant challenges. For example, France reports a suicide rate of 13.3 per 100,000 inhabitants—well above the European Union

average of 10.2 [5,6]. This illustrates the limitations of current prevention strategies and the urgent need for complementary technological solutions.

A major obstacle to effective prevention is the clinical phenomenon known as the *suicidal transition*. This refers to the significant gap between the relatively high prevalence of suicidal ideation and the much lower prevalence of suicide attempts or deaths. Epidemiological data illustrate this clearly: in France, the 12-month prevalence of suicidal ideation among adults was 4.2% in 2021, while the prevalence of suicide attempts was just 0.5% [5]. Despite being a statistically significant risk factor, suicidal ideation alone is a poor predictor of action. A meta-analysis of 71 studies found positive predictive values between only 0.3% and 3.9% [9]. In practice, identifying ideation is not sufficient—what matters is detecting the subtle cues that suggest a transition to behavior. This makes the task a highly *imbalanced classification problem*, akin to finding a "needle in a haystack."

Meanwhile, the rise of digital technologies has profoundly transformed how psychological distress is communicated. Social media platforms such as Reddit, Twitter, and Facebook have become major venues where individuals—especially young people—openly share their emotions and experiences, including signs of mental suffering. While this phenomenon presents risks such as contagion and exposure to harmful content [39], it also provides a unique, real-time window into the psychological state of large populations. These large volumes of unstructured text offer unprecedented opportunities for early detection.

Natural Language Processing (NLP)—a key domain within artificial intelligence—provides the tools to analyze this data and detect emotional tone, linguistic patterns, and contextual signals associated with suicidal ideation [10, 11]. The use of NLP and Machine Learning (ML) in analyzing social media communication aligns closely with the WHO's "LIVE LIFE" strategy, which emphasizes early identification, assessment, and follow-up of at-risk individuals [1]. Thus, this work constitutes not merely a technological application of AI, but a direct response to a well-defined public health imperative.

In this study, we develop a high-performance deep learning model based on the Transformer architecture (BERT), specifically fine-tuned for suicidal ideation detection. We evaluate its effectiveness using a rigorous experimental protocol and diverse performance metrics. Finally, we explore the ethical implications of deploying such technologies in real-world clinical settings and propose a human-centered framework for responsible implementation.

The remainder of this paper is structured as follows: Section 2 reviews existing work on suicide risk detection. Section 3 details our methodology, focusing on the fine-tuning of the BERT model. Section 4 presents the experimental results, benchmarking BERT against other models. Section 5 discusses the results, along with the study's limitations and ethical considerations.

2 Related Work

The first attempts to automate the detection of suicide risk relied on traditional machine learning (ML) models. Algorithms such as Support Vector Machines (SVM), Random Forest, and Logistic Regression were applied to structured data, such as clinical and demographic characteristics from electronic health records (EHRs). [17] Systematic reviews of the literature have shown that these approaches can achieve "good" prediction performance, with Area Under the Receiver Operating Characteristic Curve (AUC-ROC) values frequently between 0.80 and 0.89. [17, 37]

However, these studies also reveal considerable heterogeneity in performance. The effectiveness of a model strongly depends on the specific outcome it seeks to predict (e.g., ideation, attempt, or death by suicide) and the algorithm used. [17] For example, one systematic review showed that boosting algorithms achieved good predictions for suicidal thoughts and deaths, while neural networks were more effective for predicting suicide attempts. [17] Similarly, SVMs were effective for predicting suicidal thoughts but less so for attempts. [17] Despite promising results, with accuracies often exceeding 70%, these traditional approaches struggle to capture the complex relationships and semantic nuances present in natural language, which is the primary vehicle for expressing suicidal ideation online. [17]

The introduction of Transformer architectures, and particularly the BERT (Bidirectional Encoder

Representations from Transformers) model by Devlin et al. (2019), marked a major turning point in the field of natural language processing and, by extension, in its application to mental health. [19, 20] The fundamental power of these models lies in their ability to pre-train deep linguistic representations that are *bidirectional*. [20] Unlike previous sequential models (like RNNs or LSTMs) that read text from left to right or right to left, Transformers analyze the entire sentence simultaneously through a self-attention mechanism. [19] This allows them to understand the context of a word by considering both the words that precede and follow it, thereby capturing complex and long-range semantic relationships. [19, 21]

Recent systematic reviews on the use of large language models (LLMs), including BERT and its derivatives, for suicide detection have confirmed their remarkable effectiveness. [32] These studies report that LLMs are highly performant, often outperforming human experts in early detection and prediction capabilities. [32] However, this power comes with new limitations. The "black box" nature of these complex models makes their decisions difficult to interpret, a major obstacle for clinical adoption. [32] Furthermore, studies have shown inconsistent performance compared to clinicians in certain evaluation scenarios and have raised concerns about potential biases encoded in these models during their pre-training on vast internet text corpora. [32]

The vast majority of research in this area relies on text data collected from social media platforms, with Reddit being the most predominant source. [11, 14, 22] Specific communities (subreddits) like `r/SuicideWatch` and `r/Depression` have become prime grounds for building large-scale corpora, as they bring together users openly discussing their mental health. [22, 23, 40] Several benchmark datasets have thus been created, such as the *UMD Reddit Suicidality Dataset* and the *Reddit SuicideWatch and Mental Health Collection (SWMH)*. [22, 23]

The major challenge lies not only in data collection but especially in its annotation. The quality of a supervised learning model's predictions fundamentally depends on the quality of the labels it was trained on. A common but limited practice is to use the source subreddit as a proxy label (e.g., a post from `r/SuicideWatch` is labeled "suicidal," a post from another subreddit is labeled "non-suicidal"). [22] This method is fast but imprecise. To address this problem, efforts have been made to create expert-annotated datasets. One of the most accomplished examples is the *Reddit C-SSRS Suicide Dataset*. [22] This dataset is distinguished by its use of an annotation scheme based on a validated clinical rating scale, the *Columbia-Suicide Severity Rating Scale* (C-SSRS). This allows for much finer and clinically relevant labels, such as "supportive indicator," "suicidal ideation," "suicidal behavior," and "suicide attempt". [22]

The analysis of the state of the art reveals a concerning trend that could lead to a crisis of reproducibility and generalization. While the most powerful models, like LLMs, show increasingly impressive performance [32], they are trained and evaluated on an increasingly homogeneous set of data, primarily English-language texts from Reddit. [22] A major systematic review on NLP for mental health interventions explicitly identified this lack of linguistic diversity (87.3% of the analyzed studies were exclusively in English) and low reproducibility (due to a lack of code and data sharing) as critical limitations of the field. [10, 24] This convergence towards a narrow benchmark risks producing a generation of "state-of-the-art" models whose performance is actually fragile and not generalizable to other linguistic, cultural, or demographic contexts. This systemic critique of the current research trajectory fully justifies the emphasis in our own work on increased methodological rigor and a thorough discussion of limitations and ethical deployment.

3 Methodology

3.1 Corpus Constitution and Preprocessing

For our experiments, we used the public dataset "Suicide and Depression Detection." This corpus, collected from the Reddit platform, is specifically designed for this task and contains a total of 232,074 posts. [25] It is structured in a perfectly balanced manner to avoid biases related to class size during training:

- **Positive Class (`suicide`):** This class consists of 116,037 posts extracted from the `r/SuicideWatch` subreddit, a community where users explicitly discuss their suicidal thoughts and experiences. [25]
- **Negative Class (`non-suicide`):** To create a representative control set of everyday language, this class is composed of 116,037 posts from subreddits not related to mental health, notably `r/teenagers`. [25]

The dataset is provided as a CSV file containing two columns: `text` for the post content and `class` for the corresponding label. This deliberate balancing, while not reflective of real-world prevalence, was chosen to establish a robust performance baseline. This approach allows for an evaluation of each model's core semantic differentiation capabilities without the confounding factor of natural class imbalance, which could otherwise obscure the model's ability to learn from the positive class.

Before being used for training, the text corpus underwent a standard preprocessing pipeline. [26] This pipeline includes the following steps:

1. **Lowercasing:** Converting all text to lowercase to normalize the vocabulary.
2. **Cleaning:** Removing non-textual elements that could introduce noise, such as URLs, usernames, special characters, and excessive punctuation.
3. **Stopword Removal:** Removing very frequent but semantically poor words (e.g., "the," "a," "is").
4. **Tokenization:** The final step involves splitting the cleaned text into individual tokens.

3.2 Exploratory Data Analysis (EDA)

An exploratory analysis phase was conducted to better understand the characteristics of the dataset. This step included visualizing the class distribution to confirm the perfect balance of the corpus. Analyses of text length and word count per post were performed for each class to identify any potential structural differences. Additionally, techniques like sentiment analysis were applied to assess the general emotional polarity of the texts in both categories. Finally, visualizations using Word Clouds and N-gram analysis (bigrams, trigrams) helped to highlight the most frequent and discriminating terms and phrases for the `suicide` and `non-suicide` classes, thus providing qualitative insights into potential linguistic markers before modeling. [26]

3.3 Transformer Model Training

Transformer-Based Model: A pre-trained **BERT (Bidirectional Encoder Representations from Transformers)**. For our best-performing model, we adopted a fine-tuning approach based on the Transformer architecture. We used the pre-trained bert-base-uncased model via the BertForSequenceClassification class of the Hugging Face Transformers library. This architecture consists of the base BERT model, whose set of weights is updated during training, and an added classification head, consisting of a single linear layer that maps the token representation [CLS] to our two target classes (suicide, non-suicide).

Data preparation followed BERT's standard protocol. Texts were tokenized using the BertTokenizer (WordPiece) corresponding to the model, then each sequence was truncated or padded to reach a fixed length of 128 tokens. The special tokens [CLS] and [SEP] were added at the beginning and end of each sequence respectively.

Fine-tuning was performed over 3 epochs with a batch size of 16. We used the AdamW optimizer with a weight decay of 0.01 and an initial learning rate of 2e-5. A linear learning rate planner with 500 warmup steps was used to stabilize model convergence.to minimize the Cross-Entropy Loss function. [19, 20]

To validate the effectiveness of this fine-tuned model, its performance was benchmarked against several other architectures, which are detailed in the following section

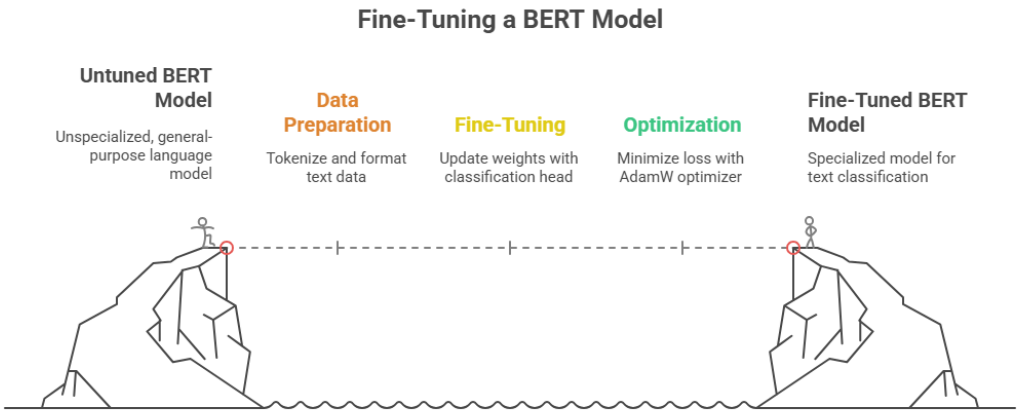

Figure 1: A fine-tuning approach based on the Transformer architecture.

3.4 Evaluation Metrics

The choice of evaluation metrics is of paramount importance for a rigorous assessment. While our dataset is balanced, the real-world application of such a model would face a severe class imbalance, making a nuanced evaluation critical. For this reason, we used a suite of metrics to provide a comprehensive view of model performance:

- **Accuracy:** The proportion of total correct predictions. While simple, it can be misleading in imbalanced scenarios but serves as a useful baseline for a balanced dataset.
- **Precision:** The proportion of true positive predictions among all positive predictions ($TP/(TP+FP)$). High precision indicates a low false positive rate, which is important for avoiding unnecessary interventions.
- **Recall (Sensitivity):** The proportion of actual positives that were correctly identified ($TP/(TP+FN)$). In suicide risk detection, recall for the 'suicidal' class is arguably the most critical metric, as failing to identify a person at risk (a false negative) has the most severe consequences.
- **F1-Score:** The harmonic mean of Precision and Recall ($2 \times (Precision \times Recall)/(Precision + Recall)$). It provides a single score that balances the concerns of both precision and recall, making it an excellent metric for evaluating overall model effectiveness, especially when there is an uneven cost associated with false positives and false negatives. [1, 9]
- **ROC AUC (Area Under the Receiver Operating Characteristic Curve):** This metric evaluates a model's ability to distinguish between the positive and negative classes across all possible classification thresholds. An AUC of 1.0 represents a perfect classifier, while 0.5 represents a model with no discriminative ability. [9]

By analyzing these metrics together, we can gain a more complete and reliable understanding of each model's strengths and weaknesses in the context of this critical task.

4 Experimental Results

4.1 Parameters settings

Table 1: Hyperparameters for BERT Model Training

Hyperparameter	Value
Base Model	`bert-base-uncased`
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Number of Epochs	3
Learning Rate Scheduler	Linear scheduler with 500 warmup steps

The corpus was divided into three distinct sets: 80% for training, 10% for validation (used to tune hyperparameters), and 10% for testing (used for the final evaluation and not seen by the model during training). The models were trained using the AdamW optimizer with a low learning rate, a common practice for fine-tuning Transformer models. [21] The specific hyperparameters for the BERT model, chosen after an experimentation phase on the validation set, are detailed in Table 1.

4.2 Comparative Model Performance

To rigorously evaluate the effectiveness of our fine-tuned BERT model, we benchmarked its performance against a comprehensive suite of models. These baselines included traditional machine learning approaches (Logistic Regression, SVM, Random Forest, KNN) as well as several deep learning architectures such as Simple RNN, LSTM, GRU, and hybrid CNN-recurrent models. This comparative framework allows us to contextualize BERT's results and empirically validate its superiority for this task. The performance of all evaluated models on the test set is summarized in**Table 2**.

Table 2: Comparative Performance of Models on the Test Set

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
BERT	0.977	0.979	0.975	0.977	0.997
GRU	0.962	0.961	0.964	0.963	0.994
CNN-LSTM	0.962	0.957	0.967	0.962	0.994
CNN-GRU	0.961	0.954	0.968	0.961	0.994
LSTM	0.960	0.954	0.965	0.960	0.993
Logistic Regression	0.937	0.944	0.929	0.936	0.982
Random Forest	0.856	0.894	0.808	0.849	0.932
Simple RNN	0.586	0.904	0.193	0.318	0.584

The graph in Figure 2 illustrates the superiority of the BERT model, which achieves the highest score, followed by recurrent neural architectures (GRU, LSTM).

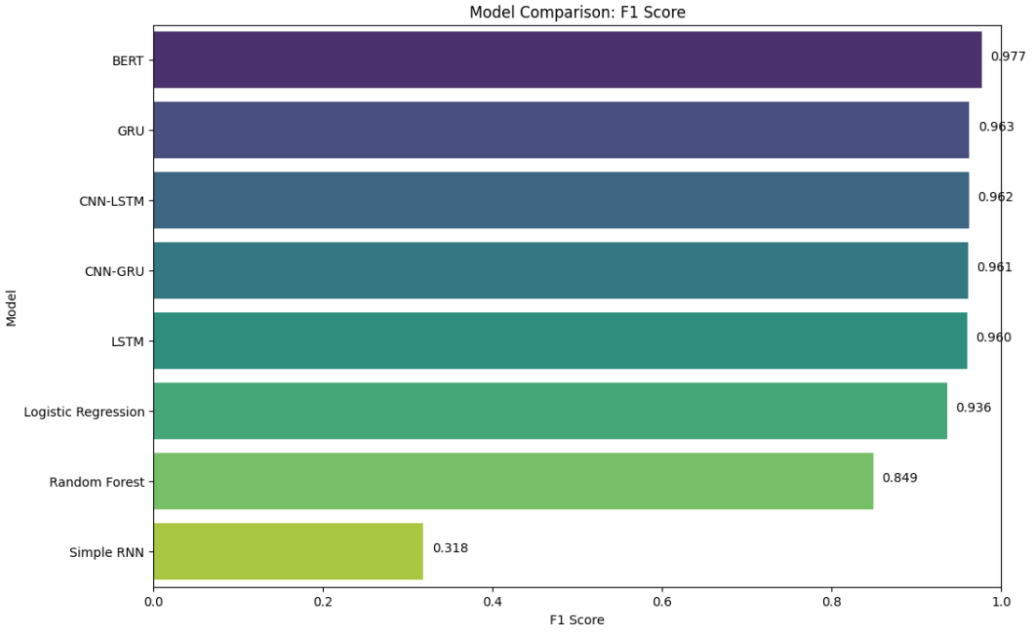

Figure 2: Comparison of Models by F1-Score Metric.

The visualization in Figure 3 confirms the BERT model's first-rate performance. All advanced deep learning models show an excellent ability to distinguish classes, while the simple RNN model is only slightly better than a random classifier.

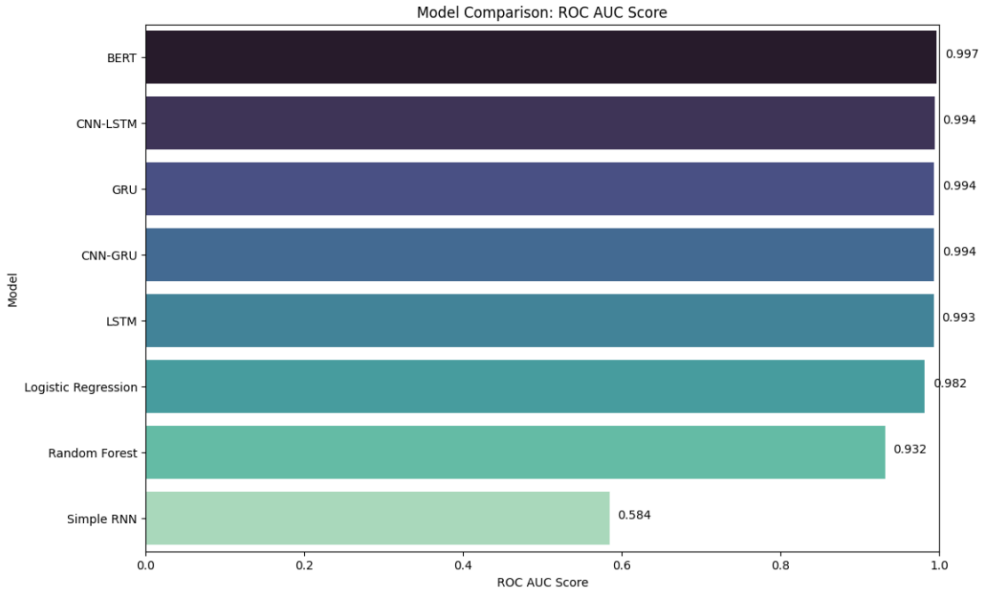

Figure 3: Comparison of Models by ROC AUC Metric.

4.3 Validation of the BERT Model's Training

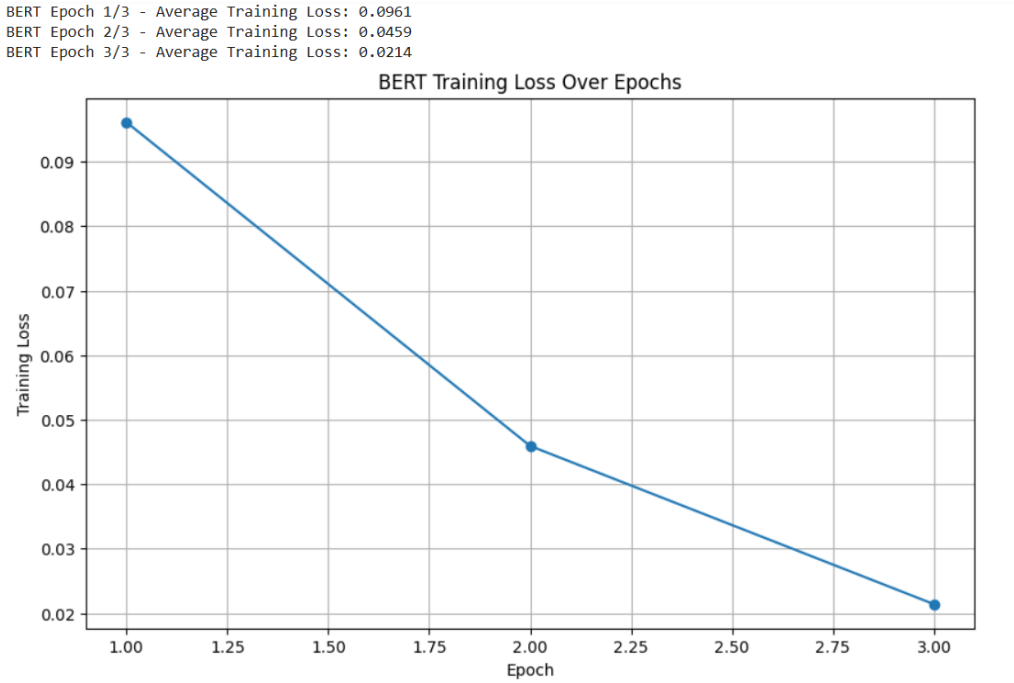

Figure 4: Evolution of the BERT Model's Training Loss Over 3 Epochs.

Analysis: Before evaluating the final performance of our models, it is essential to validate the training phase of our best-performing architecture, BERT. Figure 4 illustrates the evolution of the average loss function on the training dataset over the three epochs. We observe a clear and steady decrease in the loss, dropping from an average value of 0.0961 in the first epoch, to 0.0459 in the second, and reaching a low value of 0.0214 at the end of training. This monotonic downward curve indicates that the model successfully converged, effectively learning the distinctive linguistic patterns of the 'suicidal' and 'non-suicidal' classes in the data. The absence of fluctuations or divergence confirms the stability of the learning process and justifies confidence in the performance metrics obtained on the test set.

4.4 Performance Analysis

The analysis of the results, summarized in **Table 2**, reveals a clear and instructive performance hierarchy among the different model families.

Superiority of Advanced Neural Architectures: Deep learning models, with the notable exception of the simple RNN, significantly outperform classical machine learning approaches. The Logistic Regression model achieves respectable results (F1-Score of 0.936), even outperforming the Random Forest (F1-Score of 0.849). However, models based on complex recurrent architectures like GRU (F1-Score of 0.963), CNN-LSTM (F1-Score of 0.962), and LSTM (F1-Score of 0.960) reach a higher performance tier, demonstrating their ability to better model the sequential dependencies of language.

The Dominance of the Transformer Architecture: The BERT model clearly stands out, achieving an F1-Score of 0.977 and a ROC AUC of 0.997. This state-of-the-art performance confirms the superiority of the Transformer architecture and its bidirectional self-attention mechanism for nuanced language understanding tasks. BERT's ability to analyze the global context of a sentence gives it a decisive advantage over sequential models (LSTM, GRU) which, although performant, process information directionally.

Analysis of Less-Performing Models: The case of the simple RNN is particularly illuminating regarding evaluation challenges. Its ROC AUC score (0.584) is barely better than a random classifier. In more detail, we observe high precision (0.904) a critically low recall (0.193). This indicates that the model, while often correct when it predicts the "suicidal" class, misses the vast majority (over 80%) of actual

cases. The F1-Score (0.318) captures this imbalanced and clinically unacceptable performance much better, highlighting the limitations of an evaluation based on a single metric and the crucial importance of recall in this context.

4.5 Error Analysis and Model Interpretability

Beyond quantitative metrics, a qualitative analysis of the model's errors is essential to understand its strengths and weaknesses.

- **False Positives:** These are texts that the model incorrectly classified as suicidal. Manual analysis of these errors reveals recurring patterns. The model tends to be triggered by texts using metaphorical language related to death or finality ("I just want it all to stop," "I feel dead inside"), by expressions of dark humor, or by philosophical discussions about mortality. Although these texts share a lexical field with suicidal ideation, they do not express a direct intent to commit suicide.
- **False Negatives:** These are the most critical errors, where the model failed to detect a text expressing suicidal ideation. As illustrated by the performance of the simple RNN model, low recall results in a large number of false negatives, which is clinically unacceptable. These errors often occur when the language is subtle, coded, or allusive. Users may express their distress indirectly ("I won't be a burden to anyone for much longer," "I've found a way to find permanent peace") or use euphemisms. Even the best model, BERT, can fail with new or highly metaphorical formulations it has not encountered during training.

To counter the "black box" criticism often leveled at deep learning models [32], we leveraged the very architecture of our model to improve its interpretability. Specifically, we used the self-attention mechanism to visualize the model's decisions. [19] By extracting the attention weights from the last layer of the Transformer for a given prediction, we can highlight the words or tokens that contributed most to the final decision. For example, in a sentence correctly classified as suicidal such as "No one will care if I disappear tomorrow, I've already planned everything," the analysis of attention weights reveals that the model focuses heavily on terms like "disappear," "no one will care," and especially "planned." This ability to identify the semantic "keywords" that drive the classification provides a valuable bridge between computational output and clinical understanding, showing that the model learns concepts aligned with known risk factors.

5 Discussion

The experimental results unequivocally demonstrate the exceptional performance of the fine-tuned BERT model, which significantly surpasses all other tested architectures. Achieving an F1-Score of 0.977 and a ROC AUC of 0.997 places this work among the highest-performing to date, confirming the immense potential of Transformer models and significantly advancing beyond the typical performance of classical ML models reported in systematic reviews [37, 38]. This high performance is not merely a statistical artifact; interpretability analysis reveals that the model learns to identify clinically relevant linguistic markers of suicide risk. By assigning high attention weights to phrases expressing hopelessness, burdensomeness, social isolation, and planning or intentionality [34, 40, 45], the model autonomously captures key constructs from established clinical theories of suicidality, suggesting a capture of deep semantic meaning rather than superficial pattern matching.

However, a critical assessment of this study's limitations is essential to contextualize these findings. A primary limitation is the binary classification scope ('suicidal' vs. 'non-suicidal'), which, while effective for high-level screening, lacks the granular analysis needed for nuanced clinical utility, a task better served by clinically validated frameworks like the Columbia-Suicide Severity Rating Scale (C-SSRS) [33]. Furthermore, the use of a perfectly balanced dataset, though methodologically sound for

comparing model capabilities, likely inflates performance metrics and does not reflect the 'needle in a haystack' nature of real-world data. Generalizability is also constrained, as the model was trained exclusively on English-language texts from the Reddit platform, a significant limitation shared by much of the research in this field [45]. This data bias makes its performance on other platforms or in diverse cultural contexts unknown and potentially diminished, a known risk for models trained on unrepresentative data [32, 39]. Finally, it is crucial to recognize the model's functional constraints: it detects the textual expression of suicidal *ideation*, not the imminent risk of *action*, failing to bridge the critical 'suicidal transition' gap [36, 39]. Moreover, its static analysis of individual posts does not account for the dynamic, temporal evolution of a user's mental state, a critical element for a more comprehensive risk model [45].

6 Ethical Considerations and Implications for Prevention

The development of AI tools for suicide detection, while driven by a public health imperative, raises ethical questions of paramount importance, demanding that the technology's power be balanced by robust safeguards to prevent harm [41, 42]. Navigating this landscape requires addressing three fundamental challenges. First, the issue of **Privacy and Consent** is critical, as using data from social media, even if public, operates in an ethical gray area where users do not provide explicit, informed consent for their highly sensitive personal writings to be analyzed [41, 43]. Second, **Bias and Equity** pose a significant threat; as AI models reflect the data they are trained on, an algorithm trained primarily on a specific demographic is likely to underperform for underrepresented groups, potentially creating or exacerbating health inequalities [43, 44]. Third, the question of **Accountability and Harm** is complex, as the lines of responsibility for model errors—whether a false negative with tragic consequences or a false positive triggering unnecessary and traumatic interventions—are blurred between developers, platforms, and clinicians [42, 43].

Consequently, responsible clinical deployment is not merely an algorithmic challenge but a sociotechnical one, forcing a confrontation between the utilitarianism of public health and the deontological duty to "do no harm" to an individual patient [41]. A framework for responsible implementation must therefore be established on core principles. The foremost is the **Human-in-the-Loop** model, where these tools serve as clinical decision support systems that augment, not replace, the judgment of a qualified professional [42, 44]. This is intrinsically linked to **Transparency and Explainability**, as a clinician must be able to understand, at some level, why a model generated an alert to trust it and use it critically [42]. Finally, any intervention must follow the **Principle of the "Least Restrictive Alternative,"** ensuring that the response to an AI-generated alert is graduated and minimally intrusive, respecting the individual's autonomy by starting with options like providing resources or suggesting a consultation rather than immediately escalating to severe measures [42].

7 Conclusion

In this paper, we addressed the critical problem of detecting suicidal ideation from social media, proposing a fine-tuned Transformer (BERT) model that demonstrated state-of-the-art performance through a rigorous evaluation. Our main contribution lies not only in achieving high scores but also in the comprehensive methodology, the in-depth discussion of the inherent limitations of our approach, and the complex ethical landscape. The fundamental conclusion of our work is twofold: on one hand, artificial intelligence holds immense potential to positively transform suicide prevention; on the other, this technological power must be wielded with a keen sense of responsibility. The path to beneficial application is not through fully autonomous systems, but through the creation of carefully integrated, transparent, and human-supervised decision support tools that empower clinicians and respect patient dignity. Looking ahead, realizing this potential requires directly addressing the challenges stemming from this study's limitations. It is imperative that future research focuses on improving **linguistic and cultural diver-**

sity to create equitable and generalizable tools; moving beyond static analysis to embrace **longitudinal modeling** to capture suicide risk as a dynamic process; and continuing to advance **model interpretability** to strengthen clinician trust and bridge the gap between computational science and mental health practice [45]. Ultimately, the goal is not to predict suicide with absolute certainty, but to build intelligent and ethical early warning systems that can direct limited human resources where they are most needed, thereby offering a chance for intervention and hope.

References

[1] World Health Organization. (2021). *Suicide worldwide in 2019: Global Health Estimates*. Geneva: World Health Organization.

[2] Elyoseph, Z., Levkovich, I., & Kalish, Y. (2024). Evaluating of BERT-based and Large Language Mod for Suicide Detection, Prevention, and Risk Assessment: A Systematic Review. *Journal of Medical Systems, 48*(1).

[3] Weaver, N., et al. (2025). Global, regional, and national burden of suicide, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*.

[4] International Association for Suicide Prevention. (2025). New WHO Suicide Data Reaffirms Urgent Need for Global Prevention.

[5] Observatoire National du Suicide. (2025). National Suicide Observatory Report.

[6] Macrotrends. (2021). France Suicide Rate 2000-2024.

[7] TradingEconomics. (2021). France - Suicide Mortality Rate (per 100,000 Population).

[8] Mizan News Agency. (2024). Suicide rate in French prisons hitting a record.

[9] Nock, M. K., et al. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans, and attempts. *The British Journal of Psychiatry, 192*(2), 98-105.

[10] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: systematic review and research framework. *Translational Psychiatry, 13*(1), 309.

[11] Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering, 23*(5), 649-685.

[12] Sedgwick, M., et al. (2019). The influence of social media on suicidal ideation: a systematic literature review. *Journal of Affective Disorders, 259*, 14-29.

[13] Zomick, J., et al. (2022). Social Media Use and Suicidal Thoughts and Behaviors in Adolescents in Intensive Outpatient Programming. *Journal of the American Academy of Child & Adolescent Psychiatry, 61*(1), 84-95.

[14] Low, D. M., et al. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19. *Journal of medical Internet research, 22*(10), e22635.

[15] Li, T. M. H., et al. (2023). Detection of Suicidal Ideation in Clinical Interviews for Depression Using Natural Language Processing and Machine Learning: Cross-Sectional Study. *JMIR Medical Informatics*.

[16] Alowais, S. A., et al. (2023). Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice. *Journal of Multidisciplinary Healthcare*.

[17] Pigoni, A., et al. (2024). Machine learning and the prediction of suicide in psychiatric populations: a systematic review. *Molecular Psychiatry*.

[18] O'Dea, B., et al. (2022). The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review. *Journal of psychiatric research, 155*, 579-588.

[19] Vaswani, A., et al. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

[20] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).

[21] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[22] Gaur, M., Al-Hasan, A., Al-Hasan, H., & Al-Tawfiq, J. A. (2019). A Systematic Review of the Use of Machine Learning for Suicide Ideation Detection on Reddit. *Frontiers in Psychiatry*.

[23] Ji, S., Li, X., Huang, Z., & Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*.

[24] Coppersmith, G., et al. (2018). Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*.

[25] Nikhileswar, P., et al. (2020). A balanced dataset of suicidal and non-suicidal posts from Reddit. *Data in Brief, 31*, 105853.

[26] Tadesse, M. M., et al. (2019). Detection of depression-related posts in Reddit social media forum. *IEEE Access, 7*, 44883-44893.

[27] Joiner, T. E. (2005). *Why people die by suicide*. Harvard University Press.

[28] D'Hotman, D., et al. (2021). Ethical considerations for AI-driven suicide prediction. *BMJ Leader, 5*(2), 102–107.

[29] Benton, A., et al. (2017). Ethical considerations for using AI to predict suicide risk. *The Hastings Center Report*.

[30] Thakkar, S., et al. (2024). Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being. *Social Sciences, 13*(7), 381.

[31] Simbo. (2024). Ethical Implications of Using AI in Mental Health. *Simbo.ai Blog*.

[32] Elyoseph, Z., Levkovich, I., & Kalish, Y. (2024). Evaluating of BERT-based and Large Language Mod for Suicide Detection, Prevention, and Risk Assessment: A Systematic Review. *Journal of Medical Systems*, 48(1).

[33] Gaur, M., Al-Hasan, A., Al-Hasan, H., & Al-Tawfiq, J. A. (2019). A Systematic Review of the Use of Machine Learning for Suicide Ideation Detection on Reddit. *Frontiers in Psychiatry*.

[34] Joiner, T. E. (2005). *Why people die by suicide*. Harvard University Press.

[35] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: systematic review and research framework. *Translational Psychiatry*, 13(1), 309.

[36] Nock, M. K., et al. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans, and attempts. *The British Journal of Psychiatry*, 192(2), 98-105.

[37] O'Dea, B., et al. (2022). The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review. *Journal of psychiatric research*, 155, 579-588.

[38] Pigoni, A., et al. (2024). Machine learning and the prediction of suicide in psychiatric populations: a systematic review. *Molecular Psychiatry*.

[39] Sedgwick, M., et al. (2019). The influence of social media on suicidal ideation: a systematic literature review. *Journal of Affective Disorders*, 259, 14-29.

[40] Zomick, J., et al. (2022). Social Media Use and Suicidal Thoughts and Behaviors in Adolescents in Intensive Outpatient Programming. *Journal of the American Academy of Child & Adolescent Psychiatry*, 61(1), 84-95.

[41] Benton, A., et al. (2017). Ethical considerations for using AI to predict suicide risk. *The Hastings Center Report*.

[42] D'Hotman, D., et al. (2021). Ethical considerations for AI-driven suicide prediction. *BMJ Leader*, 5(2), 102-107.

[43] Simbo. (2024). Ethical Implications of Using AI in Mental Health. *Simbo.ai Blog*.

[44] Thakkar, S., et al. (2024). Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being. *Social Sciences*, 13(7), 381.

[45] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: systematic review and research framework. *Translational Psychiatry*, 13(1), 309.