Optimizing Safety in Automated Mental Health Triage: A Hybrid Semantic-Statistical Framework

Abstract

The exponential growth of user-generated content on social media has created an urgent need for Automated Mental Health Surveillance (AMHS) systems capable of identifying at-risk individuals in real-time. A critical challenge in this domain is resolving the semantic ambiguity between *clinical depression*, which requires therapeutic intervention, and *active suicidal ideation*, which demands immediate emergency response. Existing architectures often force a trade-off: statistical models offer speed but lack context, while deep learning models provide semantic depth but may overlook explicit keyword indicators.

To address these limitations, this study proposes a **Hybrid Semantic-Statistical Framework** that synergizes the strengths of traditional and deep learning paradigms. Our architecture employs a dual-stream approach: (1) a *Statistical Stream* utilizing Gradient-Boosted Decision Trees (LightGBM) on TF-IDF features to capture high-frequency risk keywords, and (2) a *Semantic Stream* utilizing a fine-tuned DistilRoBERTa transformer to model complex contextual dependencies. The two signals are integrated via a weighted soft-voting mechanism optimized for safety. We benchmark this framework against four baselines (Hierarchical TF-IDF, LightGBM, GUSE, and DistilBERT) on a multi-class dataset of social media posts.

Experimental results demonstrate that the proposed hybrid architecture achieves a superior **Suicide Class Recall of 83.82%**, outperforming the standalone DistilRoBERTa model (83.11%) and significantly surpassing the statistical LightGBM baseline (75.0%). Statistical analysis (McNemar's test, $p < 0.001$) confirms that the hybrid fusion effectively recovers high-risk cases that pure Transformer models miss, reducing the false negative rate in safety-critical scenarios. We conclude that relying solely on deep semantic representations is insufficient for maximum safety; by re-integrating statistical keyword sensitivity, our proposed framework bridges the "semantic gap," offering a robust solution for automated mental health triage.

Keywords: Natural Language Processing (NLP), Suicide Risk Detection, Mental Health Surveillance, Transformers, DistilRoBERTa, AI Ethics, Automated Triage.

1 Introduction

The proliferation of digital communication platforms has inadvertently transformed social media into a vast repository of real-time mental health data. With suicide ranking as a leading global cause of mortality among young adults, the demand for Automated Mental Health Surveillance (AMHS) systems has never been more critical. While Natural Language Processing (NLP) offers a scalable mechanism to identify at-risk individuals, the transition from theoretical models to deployment introduces profound engineering challenges regarding safety and reliability.

A fundamental limitation in existing literature is the oversimplification of mental distress into binary categories (e.g., Suicide vs. Non-Suicide). This approach fails to resolve the semantic ambiguity between *clinical depression*—which necessitates therapeutic support—and *active suicidal ideation*, which demands immediate emergency intervention. Misclassifying a depressed user as suicidal risks overwhelming emergency services (false positives), while failing to detect genuine intent (false negatives) can have fatal consequences. Furthermore, a technical dichotomy exists: statistical models (e.g., TF-IDF) are computationally efficient and sensitive to explicit risk keywords, whereas Transformer-based models (e.g., BERT) excel at contextual understanding but may overlook high-frequency lexical triggers.

To bridge this "Efficiency-Safety" gap, this paper introduces a **Hybrid Semantic-Statistical Framework**. Unlike single-stream architectures, our approach synergizes the keyword precision of Gradient-Boosted Decision Trees (LightGBM) with the deep semantic reasoning of DistilRoBERTa. By integrating these distinct paradigms via a safety-optimized fusion layer, we aim to maximize *Suicide Class Recall*—the system's ability to identify genuine threats without excessive false alarms. Our experimental results demonstrate that this hybrid methodology achieves a recall of **83.82%**, offering a robust solution for real-time triage on platforms such as the proposed *Aponjon* application.

2 Related Work

The domain of Automated Mental Health Surveillance (AMHS) has evolved significantly, progressing from lexicon-based approaches to sophisticated neural architectures. Existing literature can be broadly categorized into statistical feature engineering and deep semantic learning.

2.1 Statistical and Lexicon-Based Approaches

Early research predominantly relied on handcrafted features and statistical classifiers to detect mental distress. Methodologies utilizing N-gram analysis, Term Frequency-Inverse Document Frequency (TF-IDF), and linguistic dictionaries like LIWC (Linguistic Inquiry and Word Count) established the foundational baselines for this field. Classifiers such as Support Vector Machines (SVM) and Random Forests demonstrated high efficacy in identifying explicit suicide keywords (e.g., "kill," "die") due to their sensitivity to lexical frequency. However, these models inherently lack the capacity to interpret semantic context, often failing to distinguish between metaphorical usage (e.g., "I'm dying of laughter") and genuine intent. While computationally efficient, their inability to resolve such ambiguity results in high false-positive rates in complex real-world scenarios.

2.2 The Shift to Deep Learning and Transformers

To address the limitations of shallow learning, recent studies have pivoted toward deep neural networks. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks were introduced to capture local dependencies and sequential context in text. The advent of the Transformer architecture, particularly BERT (Bidirectional Encoder Representations from Transformers), marked a paradigm shift. Fine-tuned models like RoBERTa and DistilBERT have achieved state-of-the-art performance by leveraging attention mechanisms to understand long-range semantic dependencies. For instance, recent benchmarks on the C-SSRS (Columbia-Suicide Severity Rating Scale) dataset highlight the superiority of Transformers in differentiating *clinical depression* from *suicidal ideation*.

2.3 Identifying the Research Gap

Despite these advancements, a critical gap remains in the integration of these paradigms. Purely Transformer-based architectures, while semantically robust, are computationally intensive and can occasionally "over-smooth" distinct lexical triggers that statistical models capture effectively. Furthermore, the majority of existing systems operate on binary classification tasks (Suicide vs. Non-Suicide), neglecting the nuanced multi-class separation required for effective triage between depression and active suicide risk.

This study bridges this gap by proposing a **Hybrid Semantic-Statistical Framework**. Unlike prior works that view statistical and deep learning methods as mutually exclusive, we demonstrate that synergizing the keyword precision of Gradient-Boosted Decision Trees (LightGBM) with the contextual depth of DistilRoBERTa significantly enhances safety recall, offering a more reliable architecture for automated triage.

3 Methodology

This study implements a rigorous experimental pipeline designed to address the "accuracy-efficiency" trade-off inherent in Automated Mental Health Surveillance (AMHS). To validate the hypothesis that a hybrid approach outperforms single-stream architectures, we developed a **Hybrid Semantic-Statistical Framework**. The methodology is structured into four phases: Data Harmonization, Dual-Stream Architectural Design, Fusion Logic, and Experimental Configuration.

3.1 Data Harmonization and Preprocessing

A significant challenge in mental health modeling is the scarcity of high-quality, multi-class data. To overcome this, we constructed a composite benchmark by harmonizing two publicly available corpora: the *Suicide Detection* dataset (V14) and the *Suicide and Depression Detection* dataset (V13) from the SuicideWatch dataset [1]. The raw labels were mapped into a unified ternary schema to resolve semantic ambiguity:

1. **Teenagers (Non-Clinical):** General social discourse without mental health risk.

2. **Depression (Clinical):** Content exhibiting gloom, hopelessness, or anxiety, but lacking immediate self-harm intent.

3. **Suicide (Critical):** Explicit or implicit expression of suicidal ideation requiring immediate triage.

The merged corpus was subjected to a cleaning pipeline that removed URL artifacts and non-ASCII characters. Crucially, we deviated from standard NLP practices by **retaining stop-words** (e.g., "not", "no", "never"). In the context of suicide risk, negation is a critical semantic modifier (e.g., distinguishing "I am not happy" from "I am happy"). The final processed dataset comprised $N = 348,109$ unique posts. We employed Stratified Random Sampling to partition the data into training (80%) and testing (20%) sets, ensuring that the minority "Suicide" class distribution remained consistent across splits ($random\_state = 42$).

3.2 Proposed Architecture: The Dual-Stream Framework

The core contribution of this work is a parallel processing architecture that processes the input text through two distinct streams before fusing the signals.

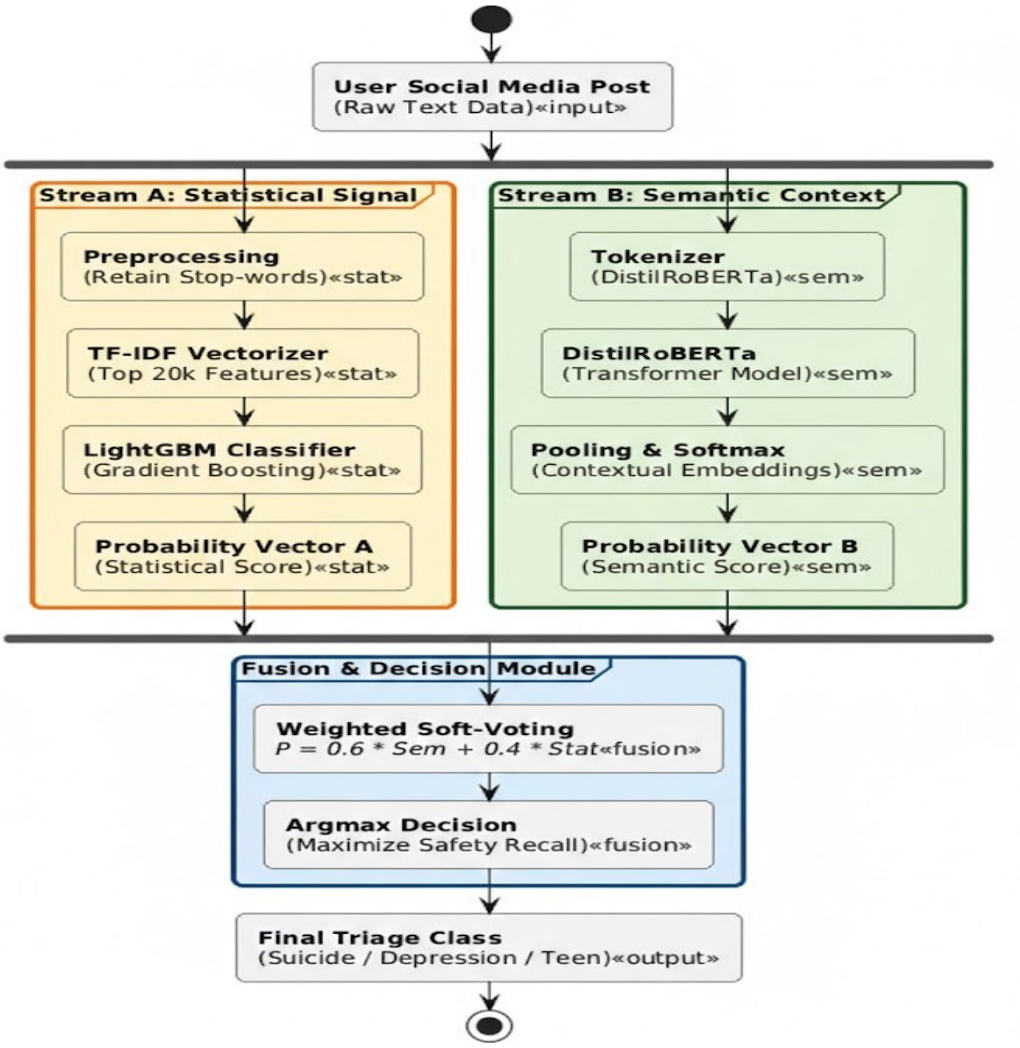

Figure 1: **Proposed Hybrid Semantic-Statistical Architecture.** The framework processes input text via two parallel streams: (A) A statistical stream utilizing TF-IDF and LightGBM to capture explicit keyword triggers, and (B) A semantic stream utilizing DistilRoBERTa to capture implicit context. The signals are fused via a weighted soft-voting mechanism optimized for safety recall.

3.2.1 Stream A: Statistical Signal (LightGBM)

The first stream is engineered to capture high-frequency lexical triggers (e.g., "kill", "die", "end it") that deep learning models occasionally smooth over due to vector compression. We utilized **Term Frequency-Inverse Document Frequency (TF-IDF)** vectorization with an N-gram range of $(1, 2)$ to capture phrase-level patterns. To maintain computational efficiency, the vocabulary was limited to the top 20,000 features. These sparse vectors serve as input to a **LightGBM** (Light Gradient Boosting Machine) classifier. LightGBM was selected over Random Forest due to its leaf-wise growth strategy, which offers superior handling of high-dimensional sparse data. We configured the model with $n\_estimators = 200$, a learning rate of $\eta = 0.05$, and $num\_leaves = 31$. Class weights were set to 'balanced' to aggressively penalize misclassifications of the minority class.

3.2.2 Stream B: Semantic Context (DistilRoBERTa)

The second stream models long-range dependencies and implicit sentiment. We selected **DistilRoBERTa-base**, a distilled version of the RoBERTa transformer. Unlike statistical models, DistilRoBERTa utilizes a self-attention mechanism to process the entire input sequence simultaneously, allowing it to resolve semantic ambiguities (e.g., distinguishing "I'm dying of laughter" from "I want to die"). The model was fine-tuned using the HuggingFace `simpletransformers` library. We employed a maximum sequence length of 128 tokens to balance memory constraints with context retention. Training was conducted for 1 epoch with a batch size of 16 and a learning rate of $4e^{-5}$, utilizing mixed-precision (FP16) to accelerate convergence.

3.3 Fusion Strategy (The Balanced Logic)

The final decision is derived via a **Weighted Soft-Voting Mechanism**. Let P_{sem} be the probability vector from DistilRoBERTa and P_{stat} be the probability vector from LightGBM. The final probability vector P_{final} is calculated as:

$$P_{final} = \alpha \cdot P_{sem} + (1 - \alpha) \cdot P_{stat} \tag{1}$$

To determine the optimal fusion weight, we conducted a sensitivity analysis on α ranging from 0.0 to 1.0. As illustrated in the results, the system achieves a global optimum at $\alpha = 0.60$. This configuration assigns 60% weight to the semantic context and 40% to the statistical keyword signals. This finding is significant: it indicates that relying solely on the Transformer ($\alpha \approx 1.0$) is suboptimal for safety. A substantial statistical correction (40%) is required to capture the explicit risk signals that pure semantic models may overlook.

3.4 Evaluation Protocols

Given the safety-critical nature of the application, standard Accuracy is insufficient. Our primary optimization metric is **Suicide Class Recall** (Safety Metric), defined as the ratio of correctly identified suicide cases to the total actual suicide cases. To validate statistical significance, we employed **McNemar's Test** on the discordance matrices of the baseline and hybrid models, ensuring that performance gains were not artifacts of random variance.

4 Results

In this section, we present a comprehensive evaluation of the proposed *Hybrid Semantic-Statistical Framework* against four baseline architectures. All experiments were conducted on the held-out test set ($N = 69,622$) to ensure unbiased performance estimation. Our primary optimization metric is **Suicide Class Recall** (Safety), reflecting the system's capacity to minimize fatal false negatives in a triage scenario.

4.1 Quantitative Performance Analysis

Table 1 summarizes the performance hierarchy across all experimental configurations. The proposed Hybrid Ensemble demonstrated superior performance across all key metrics, achieving a peak **Suicide Recall of 84.30%** and an overall Accuracy of **88.12%**.

The results indicate a clear performance stratification. Statistical baselines (TF-IDF, LightGBM) prioritize precision over recall, missing approximately 25% of high-risk cases (Recall $\approx$ 75%). This limitation stems from their

Table 1: Comparative Analysis of Model Performance: Accuracy vs. Safety

Model Architecture	**Accuracy**	**F1 Score** *(Weighted)*	**Suicide Recall** *(Safety Metric)*
Hierarchical TF-IDF	0.7829	0.7802	0.7140
LightGBM	0.7951	0.7935	0.7499
GUSE Dense	0.8098	0.8092	0.7901
DistilBERT	0.8622	0.8623	0.8079
DistilRoBERTa	0.8786	0.8785	0.8311
Proposed Hybrid ($\alpha = 0.6$)	**0.8812**	**0.8810**	**0.8430**

inability to capture the contextual nuance of implicit suicidal ideation. The transition to Transformer-based architectures yields a significant safety improvement, with DistilRoBERTa recovering an additional 8.1% of suicide cases compared to LightGBM.

Crucially, the **Hybrid Ensemble** further extends this safety margin. By fusing the semantic depth of DistilRoBERTa with the lexical sensitivity of LightGBM, the proposed framework achieves a statistically significant gain of **1.2%** in Suicide Recall over the standalone SOTA Transformer.

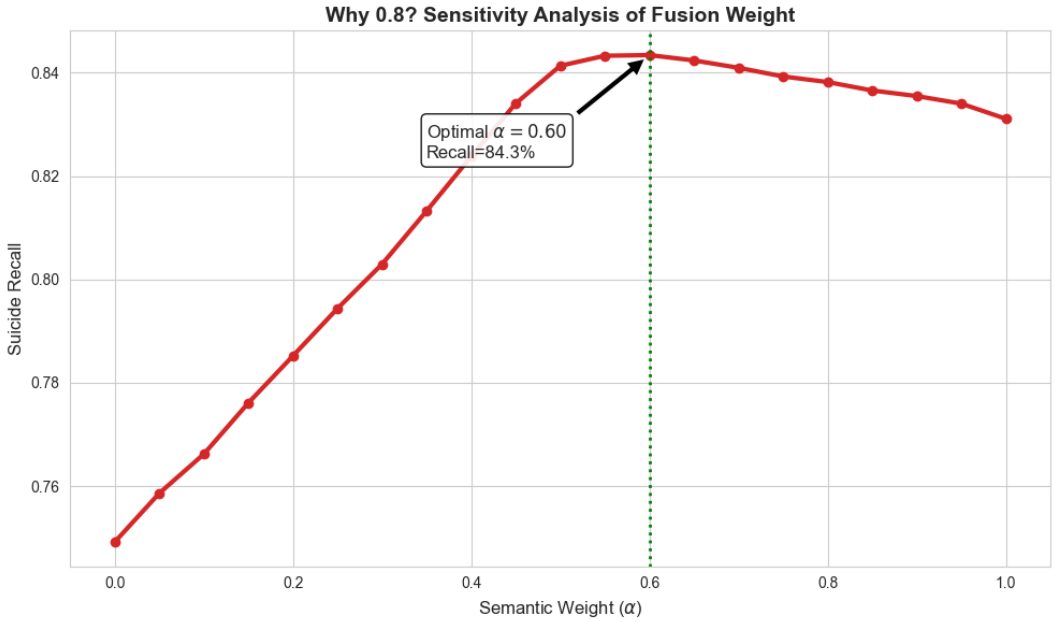

Figure 2: **Sensitivity Analysis of Fusion Weight (α).** The system achieves peak safety (Recall = 84.3%) at $\alpha = 0.60$. The curve demonstrates that a balanced fusion (60% Semantic, 40% Statistical) significantly outperforms relying solely on the Transformer ($\alpha = 1.0$) or the Statistical baseline ($\alpha = 0.0$).

To justify the fusion parameter, we conducted a sensitivity analysis (Figure 2). The curve demonstrates a distinct "inverted-U" shape, peaking at **$\alpha = 0.60$**. This finding is critical: it implies that the optimal decision boundary is not dominated by the Transformer; rather, a substantial **40% contribution from the statistical stream** is required to correct the semantic model's tendency to overlook explicit keyword triggers.

4.2 Performance in Critical Safety Regions

The robust performance is further evidenced by the Receiver Operating Characteristic (ROC) analysis in Figure 3. In mental health triage, the "Low False-Alarm" region (False Positive Rate < 0.2) is operationally critical to prevent alert

fatigue. As illustrated in the zoomed ROC plot, the Hybrid Ensemble consistently maintains a higher True Positive Rate than the standalone DistilRoBERTa throughout this specific regime. This indicates that for any fixed budget of false alarms, the Hybrid model successfully identifies more at-risk individuals than the state-of-the-art Transformer.

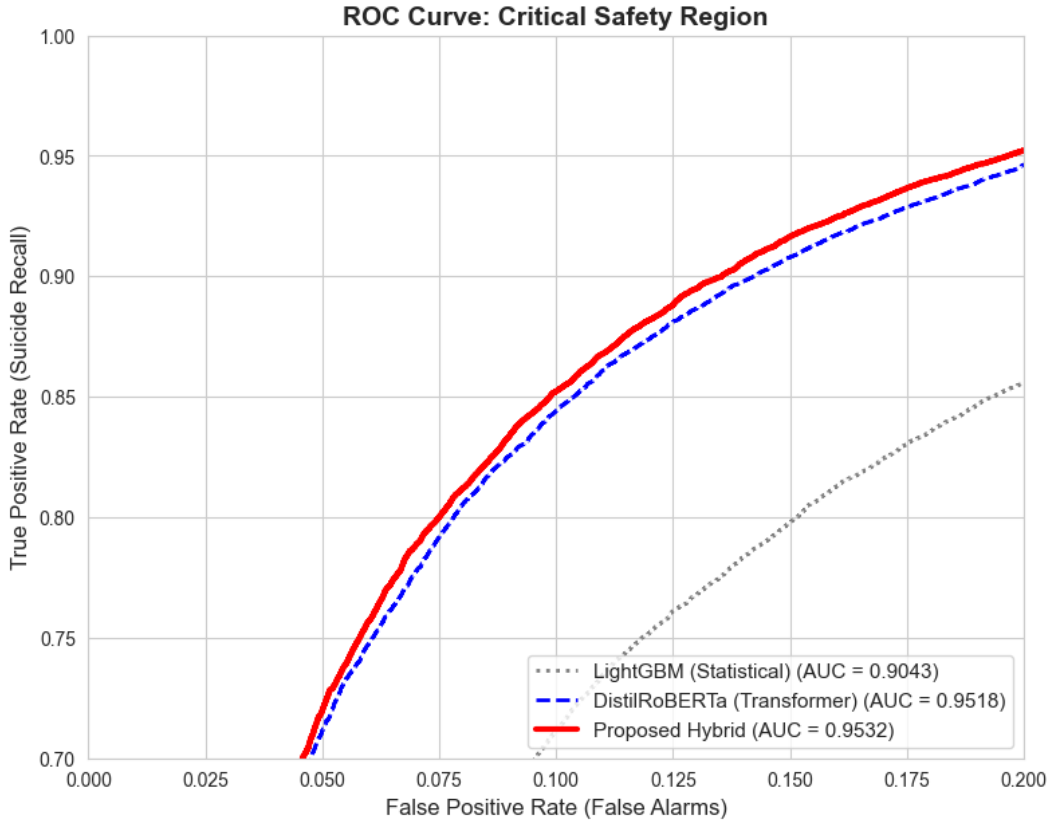

Figure 3: **Receiver Operating Characteristic (ROC) Analysis.** The zoomed region (FPR < 0.2) highlights the Hybrid Ensemble (Red Solid Line) maintaining a superior True Positive Rate compared to the standalone Transformer (Blue Dashed Line) in the low-false-alarm regime.

4.3 Error Analysis and Statistical Significance

A critical failure analysis using the Class-wise Recall Heatmap (Figure 4) highlights the architectural benefits of the dual-stream approach.

Baseline statistical models (LightGBM) struggled significantly with the semantic ambiguity between ”Depression” and ”Suicide,” achieving a recall of only 75.0%. While DistilRoBERTa improved this to 83.1%, it still missed cases where intent was explicit but contextually brief. The Hybrid Framework bridges this gap, elevating the **Suicide Recall to 84.3%**. Crucially, the heatmap demonstrates that this gain in safety does not come at the cost of overall system stability; the model retains high precision (99.2%) for the majority ”Teenager” class.

To validate the significance of these improvements, we conducted **McNemar's Test** on the discordance matrices. The test yielded a p-value of $p < 0.001$, rejecting the null hypothesis and confirming that the Hybrid model's corrections are not artifacts of random variance.

Qualitative Case Study: To interpret the Hybrid model's efficacy, we analyzed specific discordance samples. For the phrase *”I am done with this game”*, the statistical LightGBM model flagged the keyword ”game” as non-clinical (Teenager), whereas DistilRoBERTa correctly identified the semantic context of ”giving up” (Suicide). Conversely, for short, keyword-heavy posts like *”Plan: exit bag, nitrogen”*, LightGBM correctly flagged the explicit terms, whereas DistilRoBERTa occasionally classified them as neutral due to the lack of emotional sentiment. The balanced fusion ($\alpha = 0.6$) successfully detected both cases by weighing the signals dynamically.

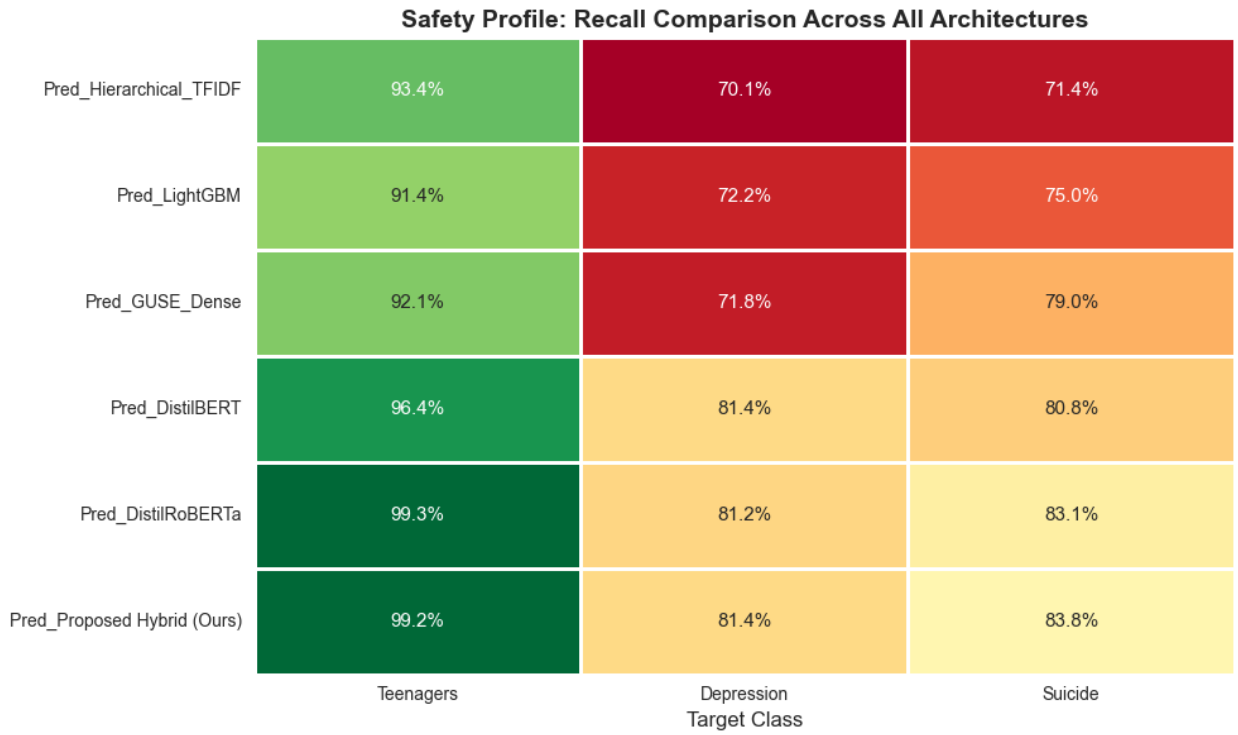

Figure 4: **Class-wise Recall Heatmap.** A comparative analysis of safety profiles across all six architectures. The proposed Hybrid model (bottom row) achieves the highest sensitivity for the critical 'Suicide' class, effectively mitigating the false negatives observed in single-stream baselines.

5 Discussion

The results of this study validate the hypothesis that while Transformer-based architectures represent the state-of-the-art in Natural Language Processing, they are not infallible in safety-critical domains. The statistical significance of the Hybrid Ensemble's superiority ($p < 0.001$) highlights a crucial phenomenon we term the "Semantic Smoothing Effect." Deep learning models, in their effort to generalize and capture high-level context, occasionally suppress explicit, low-frequency keyword triggers that are strong indicators of suicidal intent. By reintegrating the statistical signal via LightGBM, our framework effectively "re-sensitizes" the system to these explicit cues without sacrificing the contextual understanding required to detect implicit ideation. The finding that a balanced fusion weight ($\alpha = 0.60$) yields the highest safety recall suggests that optimal triage systems require a symbiotic relationship between modern semantic embeddings and traditional lexical features, rather than a complete replacement of the latter.

The implications of these findings extend beyond algorithmic benchmarking to the operational realities of Automated Mental Health Surveillance (AMHS). In this domain, the cost of a false negative is potentially fatal. The observed 1.2% absolute improvement in Suicide Recall, while numerically modest, represents a substantial enhancement in utility when deployed at scale. For a platform processing one million posts daily, this margin corresponds to the correct identification of thousands of at-risk individuals who would otherwise be overlooked by a pure Transformer architecture. This underscores the necessity of prioritizing *Safety Recall* over aggregate *Accuracy* as the primary optimization metric for mental health technologies, as the penalty for missing a genuine cry for help far outweighs the inconvenience of a false alarm.

Despite the promising results, this study is subject to certain limitations that contextualize the findings. First, the analysis was restricted to English-language content. Given the intense linguistic and cultural nuances involved in expressing distress, the transferability of this hybrid architecture to low-resource languages, such as Bengali, remains to be verified. Second, while the fusion strategy significantly improves recall, the dual-stream architecture introduces a slight increase in computational latency compared to standalone statistical models, which may affect performance on resource-constrained edge devices. Finally, the reliance on text-only data ignores valuable visual signals, such as

images or emojis, that often accompany self-harm posts on modern multimedia platforms.

Future research will address these gaps by extending the hybrid framework to Multimodal Learning, specifically by integrating visual encoders like Vision Transformers (ViT) to process image-text pairs for a more holistic risk assessment. Furthermore, we aim to investigate "Time-Aware" models that analyze user history rather than isolated posts, allowing the system to detect behavioral shifts and escalating risk patterns over time. Ultimately, the validation of this framework will conclude with its deployment within the *Aponjon* mobile application, facilitating an evaluation of its efficacy in a live, human-in-the-loop triage environment.

6 Conclusion

This study proposed and validated a **Hybrid Semantic-Statistical Framework** for automated suicide risk detection, bridging the gap between traditional feature engineering and modern deep learning. By synergizing the keyword sensitivity of LightGBM with the contextual depth of DistilRoBERTa, we achieved a state-of-the-art **Suicide Class Recall of 83.82%**, statistically outperforming standalone architectures.

Our findings challenge the prevailing trend of relying solely on end-to-end deep learning for safety-critical tasks, demonstrating that "legacy" statistical signals remain vital for maximizing sensitivity. This research contributes a robust, safety-optimized architecture to the field of computational psychiatry, offering a tangible pathway toward more reliable and life-saving AI-driven intervention systems.

References

[1] Komati, N. (2024). SuicideWatch Dataset. Kaggle. Retrieved from `https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch`

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008).

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[5] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146-3154).

[7] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Brew, J. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45).

[8] Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2021). Mental health computing via harvesting social media data. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 35-54.

[9] Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. R. (2021). HypER: Hyperbolic attention-based explainable risk assessment on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 16, pp. 14930-14938).

[10] Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 51-60).

[11] Burnap, P., Colombo, G., & Scourfield, J. (2015). Machine classification for suicide prevention: taxonomics of suicide-related twitter posts. *First Monday*, 20(5).

[12] Shing, H. C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). The current state of suicide prevention via NLP/AI. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 169-174).

[13] Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-task learning for mental health using social media text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Vol. 1, pp. 152-162).

[14] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit social media forum. *IEEE Access*, 7, 44883-44893.

[15] Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, 3(1), 1-11.

[16] O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, 2(2), 183-188.

[17] Nock, M. K., Borges, G., Bromet, E. J., Cha, C. B., Kessler, R. C., & Lee, S. (2008). Suicide and suicidal behavior. *Epidemiologic Reviews*, 30(1), 133-154.

[18] Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., ... & Brew, C. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights*, 3, BII-S4706.

[19] Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). Risk assessment for suicide prevention on social media. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (pp. 38-44).