

A Comparative Analysis of Transformer and LSTM Models for Detecting Suicidal Ideation on Reddit

BIBLIOGRAPHY AND CASE STUDY, FINAL REPORT, APRIL 19, 2025

Academic Supervisor: Dr. HO Tuong Vinh

1st LUTALA LUSHULI David
option: Intelligent Systems & Multimedia
Institut Francophone International
Hanoi, Vietnam
Report Author

2nd Khalid Hasan
Department of Computer Science
Missouri State University
Springfield, Missouri, USA
Article Author

3rd Jamil Saquer
Department of Computer Science
Missouri State University
Springfield, Missouri, USA
Article Author

Abstract—This report is part of the Bibliography and Case Studies module, which aims to develop students' ability to carry out an in-depth critical analysis of a scientific article. The work presented concerns the automated detection of suicidal thoughts expressed on social networks, a major public health issue in the digital age.

Faced with the exponential growth of online publications and the impossibility of exhaustive human monitoring, the integration of artificial intelligence (AI) technologies, and more specifically automatic natural language processing (NLP), offers promising prospects for suicide prevention.

The main article analyzed rigorously compares two major families of text classification models - Transformers architectures (BERT, RoBERTa, etc.) and LSTM models - applied to messages extracted from Reddit. The experimental results show a clear superiority of the Transformers, particularly RoBERTa, which achieves an F1-score of 93.14%, confirming their effectiveness for this sensitive task.

The study also emphasizes several critical challenges, such as data bias, the lack of clinical validation, and ethical concerns related to privacy and algorithmic transparency. To expand the analysis, three related research works were also reviewed. The first introduces a semi-supervised approach using Su-RoBERTa with data augmentation via GPT-2 to address the scarcity of labeled data. The second proposes a Big Data architecture capable of real-time processing of social media streams. The third focuses on model interpretability using MentalRoBERTa, making predictions more understandable in clinical settings.

This work shows that the different approaches studied are complementary: some are lightweight and computationally inexpensive, others are capable of handling very large volumes of data, and some offer a better understanding of the results. This shows that no single method is sufficient. It is therefore becoming essential to adopt a mixed approach, combining good technical performance, reliability of results, transparency in the decisions made by AI, and respect for ethical principles.

This report therefore concludes that the detection of suicidal ideation through social networks must be based on collaboration between advanced techniques and human expertise, in order to meet the requirements of the different contexts of use.

Index Terms—Artificial intelligence, natural language processing, Transformer models, LSTM (Long Short-Term Memory), suicide prevention, social media, suicidal ideation, RoBERTa, SuRoBERTa, automatic detection, mental health, explainable NLP, Big Data.

I. INTRODUCTION

A. Context

In the digital age, social networks have become privileged spaces of expression, where many people freely share their thoughts, emotions, and sometimes, their deepest distress. In this constant flow of information, warning signs of psychological malaise, or even suicidal ideation, can emerge. Early detection of these signs has become a major public health issue, with the potential to save lives.

However, given the sheer volume and speed of content distribution, manual monitoring is not only ineffective, it's impractical. This realization has prompted the use of advanced technologies such as NLP (Natural Language Processing) and artificial intelligence, capable of automatically analyzing messages posted on social platforms to identify distress signals.

Nevertheless, the implementation of automated detection systems raises significant challenges, both technical and ethical, which need to be carefully addressed to ensure reliable, responsible and privacy-friendly analysis.

B. Problematic

The complexity of detecting suicidal ideation on social networks lies in three main categories of challenges:

1) Challenges related to data extraction:

- Limited data access (platform API restrictions)
- Varied and heterogeneous data formats (abbreviations, emojis, etc.).
- Constraints related to the lack of reliably annotated data.

2) Challenges related to data volume:

- Processing millions of publications a day data storage management
- Need for intelligent filtering mechanisms to avoid false positives

3) Challenges associated with real-time processing:

- Need for instant detection of distress signals
- Minimization of data processing latency
- Continuous adaptation to evolving language and online trends

Faced with these challenges, it is imperative to develop high-performance NLP models that meet these requirements, for rapid and effective management of high-risk situations.

C. Objectives

To meet these challenges, the goal is to design NLP-based solutions capable of :

- Accurately detect suicidal ideas in social network posts, taking into account language specificities (abbreviations, emojis, irony, etc.).
- Guarantee robust classification, minimizing errors to avoid unjustified reports or critical omissions.
- Optimize intervention, by reducing the time between the detection of an alert signal and the activation of assistance in order to effectively prevent further action.

II. PRESENTATION OF THE MAIN ARTICLE

A. Context, issues and specific objectives

The article entitled “*A Comparative Analysis of Transformer and LSTM Models for Detecting Suicidal Ideation on Reddit*” is set in the context of suicide prevention through the analysis of publications on social networks. Reddit, in particular the r/SuicideWatch community, is used as a valuable source of data to identify signals of psychological distress.

Given the complexity of the language and the massive volume of online data, the study aims to compare two

major families of classification models: Transformers (BERT, RoBERTa, etc.) and LSTMs, with various embedding techniques, to automatically detect suicidal publications.

The authors address two key questions:

- 1) Are Transformer models more effective at detecting suicidal ideation?
- 2) How do embedding techniques affect the performance of LSTM models?

B. Methodology and Techniques Used

The article adopts a rigorous multi-step methodology, from dataset construction to comparative evaluation of different classification models.

a) Constitution and annotation of the corpus

The data were collected via the Pushshift API on Reddit between October and December 2022. A total of 37,821 publications were extracted from several subforums, including r/SuicideWatch, r/socialanxiety, r/TrueOffMyChest, r/bipolar, r/confidence and r/geopolitics. Posts from r/SuicideWatch, a forum explicitly dedicated to suicidal thoughts, were automatically labeled as “suicidal”, the others as “non-suicidal”.

To validate this automatic annotation, the authors used two techniques:

- *Thematic Modeling (LDA)*: to check the nature of content on r/SuicideWatch
- *Human annotation* : on a sample of 756 posts, with a Cohen’s Kappa coefficient greater than 0.85, indicating almost perfect reliability.

b) Text Preprocessing

Texts were cleaned and standardized (removal of special characters, emojis, URLs, etc.). Next, lemmatization was applied to simplify linguistic analysis. Several versions of the corpus were used, depending on the task (POS-tagging, topic modeling, extraction of suicidal phrases via TextRank, etc.).

c) Models Evaluated

Two main families of models are compared:

- *Transformers*: BERT, RoBERTa, DistilBERT, ALBERT, ELECTRA
- *LSTM*: With or without attention mechanisms, unidirectional or bidirectional, using Word2Vec, GloVe, or BERT embeddings

All models were trained using 5-fold stratified cross-validation. Hyperparameter tuning was conducted using Ray-Tune.

C. Experimental Results and Performance

The results show that all Transformers models achieve excellent performance, with accuracy and F1-score above 91

- The RoBERTa model is the top performer, with an accuracy of 93.22% and an F1-score of 93.14%
- The ELECTRA, BERT and ALBERT models follow close behind, within a margin of around 1%.
- DistilBERT, although lighter, achieves an F1-score of 91.87% and considerably reduces training time.

As for the LSTM models:

- Those using BERT embeddings achieve scores comparable to Transformers (up to 92.69% F1-score).
- Models LSTM using GloVe or Word2Vec drop considerably in performance (often below 75% F1-score).

Experimental results show that Transformers models outperform LSTMs in terms of accuracy and ability to detect suicidal signals, thanks in particular to their ability to analyze the overall context of a message. However, the study highlights a number of challenges, including sensitivity to training data and the risk of false positives.

D. Critical Analysis and Limitations

Despite the solid results, several limitations should be highlighted:

- *Platform bias*: All the data comes from Reddit and American subreddits, which may pose problems of generalization to other social networks (Twitter, TikTok, etc.) or other cultural contexts.
- *Simplified annotation method*: The choice to automatically label all r/SuicideWatch messages as suicidal, while pragmatic, could include some false positives (e.g., supportive messages may be classified as suicidal).
- *Lack of clinical validation*: Mental health professionals were not involved in evaluating model outputs.
- *Limited discussion of ethical issues*: The question of confidentiality, consent, and the actual use of these systems is only superficially addressed.

E. Study contributions

Apart from its limitations, the study represents an important advance in the field of automatic detection of suicidal ideation, offering important contributions such as :

- It provides an annotated Reddit dataset useful for the research community
- It delivers a robust comparative evaluation of modern NLP models
- It confirms the suitability of Transformer models for this critical task

- It shows that LSTM models can still perform well when enhanced with contextual embeddings like BERT

These contributions offer concrete prospects for the development of automated preventive intelligence systems in the field of mental health.

III. RELATED WORK

Following on from the main article, several recent works explore complementary approaches to the detection of suicidal ideation on social networks. These studies differ as much in their methodologies as in the types of models mobilized, the technical architecture deployed, or the way in which they address issues of explicability.

Three major contributions are presented here:

- A semi-supervised approach leveraging Su-RoBERTa and GPT-2 for low-resource settings.
- A Big Data-oriented architecture for real-time and high-volume stream processing.
- An interpretable NLP pipeline designed for clinical validation and decision support.

A. Article 1: “SU-RoBERTa: A Semi-supervised Approach to Predicting Suicide Risk through Social Media using Base Language Models”

This paper addresses one of the most persistent limitations in mental health detection: the scarcity of high-quality annotated data. To overcome this, the authors propose Su-RoBERTa, a streamlined model based on RoBERTa, trained using a semi-supervised learning framework enhanced with synthetic text generation.

The training begins with 500 manually labeled Reddit posts, which are expanded with 1,500 unlabeled samples using progressive pseudo-labeling. To mitigate class imbalance, synthetic posts reflecting various levels of suicide risk (from mild ideation to explicit attempts) are generated using GPT-2. The approach provides a pragmatic compromise between performance and resource efficiency.

Methodology:

- Implementation of a compact 355M-parameter model (Su-RoBERTa).
- Semi-supervised learning pipeline with pseudo-labeling for data expansion.
- Controlled text augmentation using GPT-2 to address imbalance and enrich diversity.

Results:

Su-RoBERTa achieves a weighted F1-score of 69.84%. While not outperforming larger models, it demonstrates competitive accuracy with significantly lower computational cost.

Critical analysis:

This approach highlights the feasibility of deploying suicide risk detection in environments with limited data and infrastructure. The combined use of weak supervision and synthetic augmentation aligns well with real-world constraints such as privacy-preserving data access or institutional resource limitations. However, its moderate performance indicates the need for fine-tuning or expert-validated pseudo-labeling.

B. Article 2: “Big Data Analytics for Suicide Prevention: Analyzing Online Discussions and Social Media Data”

This study shifts the focus from data scarcity to scalability. By embracing a Big Data paradigm, it builds an end-to-end real-time system capable of processing massive volumes of social media messages to flag suicidal content with minimal latency.

The proposed architecture couples Apache Spark (for distributed parallel processing) with Apache Kafka (for real-time data streaming). It processes messages from Reddit and Twitter, enabling near-instantaneous risk assessments.

Methodology:

- Real-time stream processing via Kafka and distributed learning via Spark.
- Evaluation of multiple traditional classifiers (Naïve Bayes, Logistic Regression, MLP, SVM).
- Feature extraction and sentiment analysis to strengthen input signal quality.

Results:

The multilayer perceptron (MLP) achieves the best accuracy—93.47%—and consistently provides predictions in under two seconds, supporting large-scale monitoring scenarios.

Critical analysis:

This architecture demonstrates the operational potential of real-time suicide prevention systems, especially for use by governments, crisis centers, or public health platforms. However, it remains predominantly reactive and relies on conventional ML models, lacking the contextual nuance and semantic depth of more modern Transformers. The model’s effectiveness in detecting subtle, metaphorical, or culturally specific expressions of distress remains uncertain.

C. Article 3: “Conceptualizing Suicidal Behavior: Utilizing Explanations of Predicted Outcomes to Analyze Longitudinal Social Media Data”

This paper tackles a different but equally critical concern: explainability. In sensitive fields such as mental health, transparency in AI decision-making is not optional—it is essential. This study introduces an explainable NLP framework using MentalRoBERTa, a domain-specific variant of RoBERTa trained on mental health-related corpora.

The model is augmented with explainability tools—LIME, SHAP, and Layer Integrated Gradients (LIG)—which highlight the most influential tokens in each prediction. These tools aim to make the model’s output more interpretable for clinicians, enabling responsible integration into psychological or psychiatric workflows.

Methodology Highlights:

- Use of MentalRoBERTa fine-tuned on mental health forums.
- Deployment of explainability tools (LIME, SHAP, LIG) to visualize inference logic.
- Integration of TF-IDF for signal validation and redundancy checks.

Results:

The model yields a precision of 63.05%, lower than state-of-the-art results, but provides rich interpretability features that are critical in clinical contexts.

Critical analysis:

This study represents a vital step toward ethical AI, where trust, auditability, and collaboration with human experts are paramount. However, the performance gap and computational overhead of interpretability layers pose real challenges for real-time or large-scale use. More importantly, the approach calls for stronger user validation—ideally through pilot testing with healthcare professionals to assess usability and trust.

Synthesis and Comparative Insights

These three articles exemplify the diversity of strategic responses to suicidal ideation detection:

- The first emphasizes feasibility under limited resources through smart learning techniques.
- The second showcases industrial-scale potential using high-throughput systems.
- The third prioritizes transparency and clinical relevance for decision support.

Together, they reinforce the notion that suicide detection is a multifaceted challenge requiring a fusion of technological robustness, ethical sensitivity, and human oversight. No single method suffices. Instead, future systems must integrate real-time processing, interpretability, data efficiency, and psychological insight—an interdisciplinary convergence that is both

a challenge and a necessity in building safe, effective, and socially responsible AI.

IV. COMPARATIVE STUDY

Criteria	Main Article (Transformers vs LSTM)	Related Work 1: SU_RoBERTa
Approach	Supervised	Semi-supervised
Data Source	Reddit (38k posts)	Reddit (2k posts)
Main Model	RoBERTa, LSTM, BERT	Su-RoBERTa + GPT-2
F1 / Accuracy	93.14% (RoBERTa)	69.84%
Infrastructure	Standard (GPU)	Lightweight
Strengths	Strong performance, solid evaluation	Resource-friendly, easy to train
Limitations	Reddit-focused, lacks ethical depth	Small dataset, reliant on synthetic data
Best Use Case	Academic benchmarking	Low-resource scenarios

Table 1 : Comparison of the main article and related studies

V. PERSPECTIVES AND FUTURE IMPROVEMENTS

The comparative analysis of the different approaches to detecting suicidal ideation on social media highlights several promising avenues for improvement, both at the model level and in implementation practices. These perspectives involve enhancing technical performance, integrating human-centered considerations, and aligning with ethical and societal constraints.

Technological Enhancements

First, model performance could be improved by integrating multimodal approaches. So far, most studies have focused solely on textual data. However, valuable signals may also be extracted from images, videos, or metadata associated with social media posts. Combining text and image analysis, especially on platforms like Instagram or TikTok, could increase detection sensitivity without compromising precision.

Additionally, Transformer-based models are continuously evolving. Using newer or more specialized versions—such as DeBERTa, BioBERT (for medical contexts), or LLaMA—may enhance contextual understanding, especially in underrepresented languages or culturally specific contexts.

Data Enrichment and Robustness

Another key area lies in improving the quality and diversity of training data. It would be beneficial to expand beyond Reddit to include other platforms (e.g., Twitter, YouTube comments). Incorporating data from multiple countries, languages, and cultures would lead to more inclusive and representative models.

Controlled data augmentation techniques, combined with expert-guided pseudo-labeling, can also improve robustness while reducing the annotation burden.

VI. GENERAL CONCLUSION

Detecting suicidal ideation on social media is a critical challenge at a time when these platforms have become a primary outlet for individuals in distress. Given the growing volume and diversity of online content, automating this task is not just necessary—it is also a complex technological, methodological, and ethical undertaking.

Through analysis of the main article, we observed that Transformer-based models, especially RoBERTa, outperform traditional LSTM approaches, offering high accuracy in binary classification on Reddit posts. The rigorous methodology, validated annotations, and comparative evaluation make a strong contribution to the state of the art.

The additional related studies offer a broader perspective. The semi-supervised Su-RoBERTa approach shows that effective solutions can be built even with limited labeled data. The Big Data architecture demonstrates the feasibility of real-time processing at industrial scale. Lastly, the explainable EXPLICA framework emphasizes the need for interpretability, particularly when models are deployed in clinical settings.

This study highlights that no single approach can address all needs. The future of suicidal ideation detection on social media likely lies in intelligently combining multiple dimensions: performance, explainability, scalability, and ethical responsibility. Ongoing research in this area must continue to evolve in an interdisciplinary spirit, drawing on the complementary strengths of technical, medical, and human-centered approaches.

REFERENCES

- [1] Khalid Hasan, and Jamil Saquer, “A Comparative Analysis of Transformer and LSTM Models for Detecting Suicidal Ideation on Reddit” IEEE ICMLA, vol. 1, Page 1-7, November 2024.
- [2] Chayan Tank, Shaina Mehta, Sarthak Pol, Vinayak Katoch, Avinash Anand, Raj Jaiswal and Rajiv Ratn Shah, “Su-RoBERTa: A Semi-supervised Approach to Predicting Suicide Risk through Social Media using Base Language Models”, IEEE Big Data, vol. 2 Page 1-8, December 2024.
- [3] Mohamed A. Allayla and Serkan Ayvaz, “A Big Data Analytics System for Predicting Suicidal Ideation in Real-Time Based on Social Media Streaming Data” IEEE Computation and Language, vol. 3 Page 1-23, March 2024.
- [4] Van Minh Nguyen, Nasheen Nur, William Stern, Thomas Mercer, Chiradeep Sen, Siddhartha Bhattacharyya, Victor Tumbiolo and Seng Jhing Goh, “Conceptualizing Suicidal Behavior : Utilizing Explanations of Predicted Outcomes to Analyze Longitudinal Social Media Data”, IEEE ICMLA, vol. 2 Page 1-8, December 2023.

Toward More Explainable Models

In sensitive domains like mental health, integrating explainability mechanisms such as LIME, SHAP, or Integrated Gradients should become standard practice. These tools promote algorithmic transparency and allow for closer collaboration between AI systems and mental health professionals.

Further effort is needed to make these explanations both accessible and usable by non-experts. Intuitive visualizations or automatically generated summaries could accompany each prediction, increasing trust and interpretability.

Ethical and Regulatory Challenges

Beyond performance, future directions must consider deep reflection on ethical and legal issues. Ensuring privacy, managing consent, and preventing misuse of detection technologies are all critical concerns.

Closer collaboration among AI researchers, legal experts, clinicians, and psychologists is essential to develop responsible deployment frameworks that comply with regulations like the GDPR.

Toward Hybrid Systems

In the long run, designing hybrid systems that combine high-performance models with explainability, flexible architecture for batch or real-time processing, and human supervision represents a promising direction. These systems could serve as the backbone for future mental health assistance platforms.