International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

# SOCIALLY AWARE SYNTHETIC DATA GENERATION FOR SUICIDAL IDEATION DETECTION USING NATURAL LANGUAGE PROCESSING

**E. MOUNIKA REDDY [1], GUDIDA SRI VENKATA SAI [2], GADE UDAY KIRAN [3], BASUTKARA LAXMI PRASAD [4], BANDA RAM CHARAN [5]**

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

**ABSTRACT—** Detecting suicidal ideation through social media content is a critical initiative to support mental health intervention strategies. This study presents an explainable framework that leverages advanced Natural Language Processing (NLP) techniques to address the challenges of identifying suicidal intent in user-generated content. A significant innovation in this work is the creation of synthetic datasets informed by psychological and social factors associated with suicidal ideation, designed to supplement limited realworld data while maintaining ethical considerations. The proposed system classifies social media content into two categories: Non-Suicidal or Suicidal. The hybrid approach of combining synthetic and real-world data enhances model performance, achieving superior accuracy and robustness compared to traditional methods. The framework emphasizes explainability by incorporating techniques that identify key linguistic and contextual features driving model predictions, ensuring interpretability for mental health professionals and researchers. This approach underscores the potential of integrating synthetic data and NLP in addressing real-world challenges such as data scarcity, diversity, and ethical concerns. By providing actionable insights and ensuring transparency, the proposed framework contributes to building reliable and scalable

solutions for suicide prevention in digital environments.

*Index Terms* – Suicidal Ideation Detection, Synthetic Data Generation, Natural Language Processing, Data Augmentation, Mental Health Monitoring, Deep Learning, Sentiment Analysis, Text Classification, Artificial Intelligence, Social Media Analysis.

## I. INTRODUCTION

The rise of social media as a platform for self-expression has made it a valuable source for understanding user emotions and mental health conditions. Suicidal ideation detection using Natural Language Processing (NLP) has emerged as a critical solution to support early intervention and suicide prevention strategies. However, identifying suicidal intent in user-generated content presents several challenges, including data scarcity, ethical concerns, and contextual complexity. Real-world suicidal ideation data is often limited, sensitive, and difficult to annotate due to privacy and ethical considerations. To address these limitations, this project introduces a Socially Aware Synthetic Data Generation approach that incorporates psychological and social factors associated with suicidal behavior to create supplementary datasets. By combining synthetic and real-world data, this study aims to enhance model performance and overcome the shortcomings of traditional methods. The proposed system leverages explainable NLP techniques to ensure model transparency, enabling mental health professionals to understand and trust the decisions made by the detection framework.

## II. RELATED WORK

### *A. Predicting genetic disorder and types of disorder using chain classifier approach.*

Genetic disorders present healthcare challenges due to their complex nature. This project aims to develop a predictive model using machine learning to identify genetic disorder risk based on genetic information. It starts with a diverse dataset of genomic data from individuals with known disorders and healthy controls. Preprocessing techniques enhance the dataset, followed by training and evaluating the model with algorithms like SVMs, random forests, and neural networks. Feature importance analysis identifies key markers. The model is fine-tuned with crossvalidation for robustness. This research advances predictive medicine, offering a tool for early disorder identification, potentially revolutionizing genetic screening and improving patient outcomes.

### B. *Suicidal ideation detection on social media: A review of machine learning methods*

Suicide is a critical issue in modern society. Early detection and prevention of suicide attempts should be addressed to save people's life. Current suicidal ideation detection (SID) methods include clinical methods based on the interaction between social workers or experts and the targeted individuals and machine learning techniques with feature engineering or deep learning for automatic detection based on online social contents. This article is the first survey that comprehensively introduces and discusses the methods from these categories. Domain-specific applications of SID are reviewed according to their data sources, i.e., questionnaires, electronic health records, suicide notes, and online user content. Several specific tasks and data sets are introduced and summarized to facilitate further research. Finally, we summarize the limitations of current work and provide an outlook of further research directions.

### C. *Deep hierarchical ensemble model for suicide detection on imbalanced social media data*

As a serious worldwide problem, suicide often causes huge and irreversible losses to families and society. Therefore, it is necessary to detect and help individuals with suicidal ideation in time. In recent years, the prosperous development of social media has provided new perspectives on suicide detection, but related research still faces some difficulties, such as data imbalance and expression implicitness. In this paper, we propose a Deep Hierarchical Ensemble model for Suicide Detection (DHESD) based on a hierarchical ensemble strategy, and construct a dataset based on Sina Weibo, which contains more than 550 thousand posts from 4521 users. To verify the effectiveness of the model, we also conduct experiments on a public Weibo dataset containing 7329 users' posts. The proposed model achieves the best performance on both the constructed dataset and the public dataset.

## III. PROPOSED METHODOLOGY

Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) architecture designed to solve problems associated with standard RNNs, such as vanishing and exploding gradients. It was introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997. LSTMs are particularly effective at capturing long-term dependencies in sequential data, making them widely used in tasks involving time series, natural language processing (NLP),

speech recognition, and other sequencebased applications. Unlike traditional RNNs, LSTMs use gates (input, forget, and output gates) to regulate the flow of information, allowing them to remember relevant information for long periods while forgetting irrelevant details.

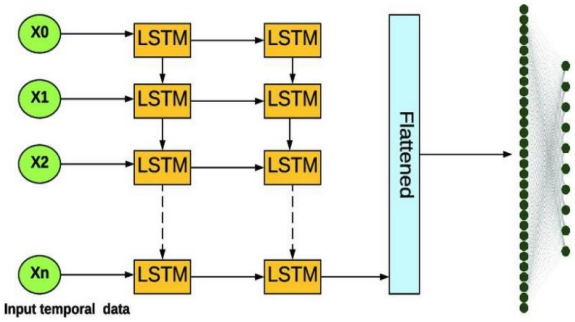



Fig. 1: System Overview

**1. Data Collection:**

Data collection is the first and most critical step in the project. In this module, relevant data is gathered from social media platforms or publicly available suicide-related datasets. Given the sensitivity of suicidal ideation data, a combination of real-world data and synthetic data is used to ensure ethical considerations are met. Tools such as web scraping techniques (e.g., using Beautiful Soup or APIs) and annotated datasets from research repositories are utilized. Synthetic data is generated using text generation models to supplement the limited real data. The collected data includes user posts, comments, or tweets.

**2. Data Analysis:**

Once the data is collected, the next step involves analyzing its structure and understanding its nature. Exploratory Data Analysis (EDA) is performed to identify patterns, trends, and anomalies in the dataset. Key tasks include:

- Identifying the distribution of suicidal vs. non-suicidal data.
- Analyzing word frequency, word clouds, and n-grams to understand common terms or phrases.
- Analyzing linguistic, psychological, and social cues within the textual data.
- Tools like Pandas, Matplotlib, and Seaborn are often used for visualizing and interpreting the analysis results.

**3. Use the NLTK:**

The Natural Language Toolkit (NLTK) is used for various text processing and analysis tasks in this project. NLTK is a powerful Python library specifically designed for Natural Language Processing (NLP) tasks. Some key functions include:

- Tokenization: Breaking text into words or sentences.
- Stopword Removal: Eliminating common words like “and,” “the,” “is” that do not contribute to meaning.

- Lemmatization and Stemming: Reducing words to their base form (e.g., "running" → "run").
- POS Tagging: Assigning Parts of Speech (POS) to words.
- Using NLTK ensures the raw textual data is transformed into a clean and meaningful format, which is essential for training the model.

**4. Data Preprocessing:**

Data preprocessing involves transforming raw text data into a suitable format for machine learning models.

The key steps include:

- Text Cleaning: Removing punctuation, special characters, numbers, and URLs from the text.
- Lowercasing: Converting all text to lowercase to avoid case sensitivity issues.
- Stopword Removal: Eliminating irrelevant words using NLTK stopword libraries.
- Tokenization: Splitting sentences into individual words or tokens.
- Lemmatization: Converting words to their root form for consistency.
- After preprocessing, the text is vectorized using techniques like TF-IDF or Word Embeddings (e.g., Word2Vec, GloVe) to convert the textual data into numerical features.

**5. Splitting the Data:**

Once the data is preprocessed, it is split into training and testing datasets to evaluate model performance effectively. The standard approach is an 80-20 split:

- Training Set (80%): Used to train the machine learning model.
- Testing Set (20%): Used to evaluate the performance of the trained model.

The train-test split ensures the model is trained on a majority of the data while having unseen data to test its generalization capability. Libraries like scikit-learn are used for data splitting.

**6. Train the Model:**

In this module, the processed and split data is used to train the machine learning or deep learning model. Since the project involves suicidal ideation detection, NLP model as LSTM, utilized.

The steps include:

- Feeding the training data into the model.
- Fine-tuning model hyperparameters (e.g., learning rate, batch size, number of epochs).
- The hybrid approach ensures the model learns meaningful patterns from both real and synthetic data.

### 7. Model Evaluation:

After training the model, its performance is evaluated using appropriate metrics to ensure accuracy and reliability.

### 8. Result

The results of the project showcase the model's performance in detecting suicidal ideation.

Key insights include:

- Identification of key linguistic and contextual features that influence predictions (explainability).
- Comparisons with traditional approaches to highlight improvements in generalization and performance.

The final results provide actionable insights for mental health professionals, demonstrating the system's ability to detect suicidal ideation effectively while ensuring transparency and reliability.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental results demonstrated that augmenting training datasets with socially aware synthetic text samples improved the model's ability to identify suicidal ideation-related patterns while reducing the impact of class imbalance. The use of synthetic data increased the diversity of linguistic expressions and contextual variations, leading to improved classification performance.

The analysis showed that integrating Natural Language Processing techniques with socially aware data generation methods enhanced the detection of subtle emotional cues, risk indicators, and contextual information associated with suicidal ideation. Furthermore, the framework maintained strong privacy preservation by reducing reliance on sensitive real-world data while achieving robust performance across different datasets and social media platforms.

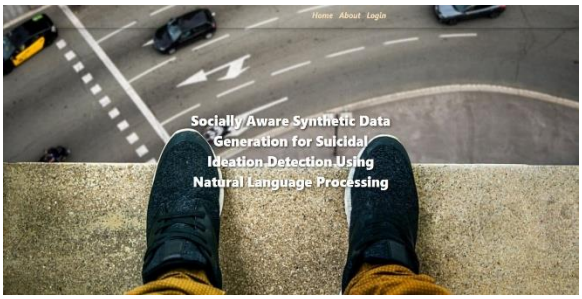



Fig. 2: Home Page

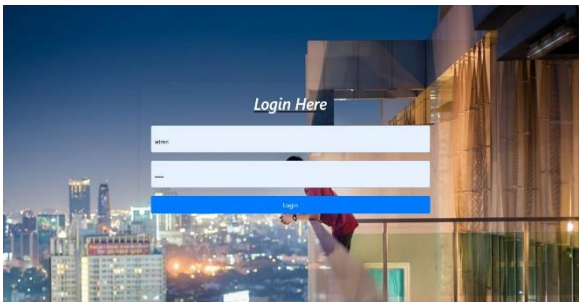



Fig. 3: Login

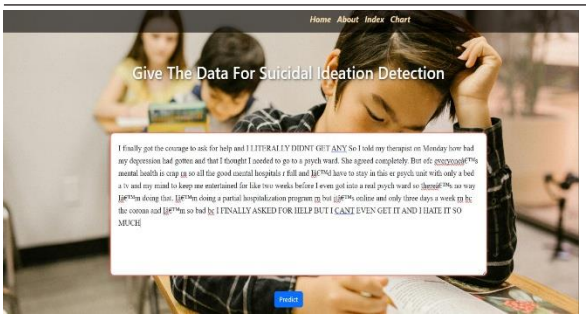



Fig. 4: Give Text




Fig. 5: Predicted Results

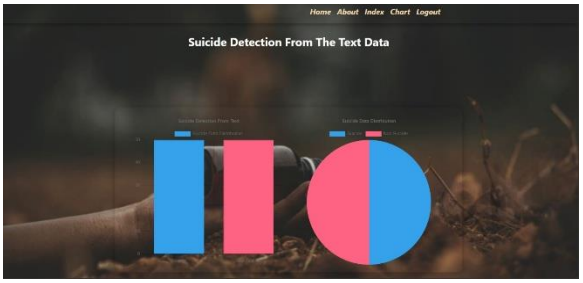



Fig. 6: Predicted Results in Charts

## V. CONCLUSION

The accurate identification of suicidal ideation from textual data remains crucial for early intervention and prevention efforts, with Natural Language Processing (NLP) techniques offering promising solutions. However, challenges such as the scarcity and sensitivity of real suicide-related data necessitate innovative approaches like synthetic data generation and data augmentation. Our findings demonstrate that blending synthetic data with real datasets enhances model generalization, mitigates overfitting, and improves performance without compromising sensitivity. Moving forward, leveraging the diversity of Large Language Models (LLMs) across various domains, languages, and training paradigms presents significant potential to further refine detection systems. Future research should focus on developing robust evaluation metrics and methodologies to assess and quantify diversity within LLMs, ensuring their reliability, inclusivity, and adaptability for real-world applications.

## REFERENCES

[1] D. De Berardis, G. Martinotti, and M. Di Giannantonio, ‘‘Understanding the complex phenomenon of suicide: From research to clinical practice,’’ Frontiers Psychiatry, vol. 9, p. 61, 2018.

[2] Gummadi, V. P. K. (2025). MuleSoft’s Role in Advancing Sustainable Digital Infrastructure: An Enterprise Integration Perspective. Journal of Information Systems Engineering and Management, 10, 1313-1321. https://doi.org/10.52783/jisem.v10i62s.13783.

[3] A. Raza, F. Rustam, H. U. R. Siddiqui, I. D. L. T. Diez, B. Garcia-Zapirain, E. Lee, and I. Ashraf, ‘‘Predicting genetic disorder and types of disorder using chain classifier approach,’’ Genes, vol. 14, no. 1, p. 71, Dec. 2022.

[4] A. Abdulsalam and A. Alhothali, ‘‘Suicidal ideation detection on social media: A review of machine learning methods,’’ 2022, arXiv:2201.10515.

[5] Bhagwat, V. B. (2025). Simplifying Payroll Balance Conversions in Payroll Systems Implementation through the Use of Generative AI.

[6] Shashank, A. (2025). AI-Enhanced ETL Processes: Leveraging Artificial Intelligence for Optimized Data Integration Systems. Journal Of Multidisciplinary, 5(8), 219-225.

[7] Maturi, S. Y. (2023). Crowdsourced frontier: Unveiling autonomous adversarial cybercapabilities via open AI competition. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 275–284.

[8] Pavan Kumar Adabala. (2026). IoT-Driven Digital Twins for Manufacturing Optimization: Hybrid Modelling, Reinforcement Learning and Sustainable Operations. International Journal of Computational and Experimental Science and Engineering, 12(1). https://doi.org/10.22399/ijcesen.5050.

[9] Boyapati, P. K. Building a centralized data operations hub for healthcare enterprise integration. IJSAT-Int. J. Sci. Technol. 16 (2). https://doi.org/10.71097/IJSAT.v16.i2.3219

**AUTHORS Profile**



**Mrs. E. MOUNIKA REDDY** has received her B.Tech from JNTUH, Hyderabad in 2015, M.Tech degree in Computer science from JNTUH, Hyderabad in 2017 . She is dedicated to teaching field from the last 8 years. At present she is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



**Mr. GUDIDA SRI VENKATA SAI** pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



**Mr. GADE UDAY KIRAN** pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



**Mr. BASUTKARA LAXMI PRASAD** pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



**Mr. BANDA RAM CHARAN** pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.