ISSN 2007-9737

Emotion-Aware LSTM Networks for Detection of Suicide Notes

María del Carmen García-Galindo, Ángel Hernández-Castañeda,
René Arnulfo García-Hernández, Yulia Ledeneva*

Autonomous University of the State of Mexico,
Mexico

mgarciag832@alumno.uaemex.mx {anhernandezc, reagarciah, ynledeneva}@uaemex.mx

Abstract. Suicide represents a critical global public health issue, causing over 700,000 lives each year, particularly among young people. Emotional distress, dysregulation, hopelessness, and psychache are key risk indicators. In the digital era, individuals increasingly express suicidal thoughts through text on social media, offering valuable insights into mental states. In this study, we propose a computational approach to classify suicide notes based on sentence-level emotional content. To analyze the sequence and evolution of emotions within each note, we employed two LSTM architectures: a standard LSTM and one incorporating an attention mechanism. Our results demonstrate that both architectures achieve F-scores exceeding 80across all classification scenarios. We observed that, although consummated suicide notes often lack explicit indicators of suicidal ideation, both emotional and lexical patterns complement each other, with emotional cues driving classification and lexical features enhancing overall performance. These findings highlight the importance of modeling emotional dynamics for the automated detection of suicide-related texts.

Keywords. Suicide notes, emotion recognition, long short-term memory, attention mechanism.

1 Introduction

Psychologically, suicide reflects intense emotional distress, where negative self-perceptions and distorted cognitions may lead to self-destructive behaviors [15].

Suicide, defined as the intentional act of taking one's own life [10], represents a critical global public health issue [3,7]. According to the World Health Organization [19], more than 700,000 individuals die by suicide each year. Although it affects people of all ages, suicide is particularly prevalent among young individuals [6,17].

The causes of suicide are complex and multifaceted [20], involving biological, psychological, social, cultural, and spiritual factors [1]. Psychologically, suicide reflects intense emotional distress, where negative self-perceptions and distorted cognitions may lead to self-destructive behaviors [15].

Emotional dysregulation and the expression of specific emotions such as sadness, fear, anger, guilt, and hopelessness are strongly associated with elevated suicide risk [21,20]. Hopelessness, in particular, intensifies suicidal ideation and increases the likelihood of attempts.

Extending previous findings, Edwin Shneidman, the father of modern suicidology, argued that all emotional states contribute to unbearable psychological pain (psychache), which he identified as the central mechanism in suicidal behavior [10,20].

Suicidal ideation—recurrent and intentional thoughts of ending one's life—is a key predictor of suicide risk [7,8,20].

Traditionally, suicidal ideation has been assessed through clinical interviews, questionnaires, and psychological scales [3,12,14]. However, these methods often fail to capture the full range of emotions due to biases, fixed response options [3]. Other limitations are that many individuals lack access to these resources or avoid them due to the stigma associated with the topic, consequently, do not receive treatment [3,13]. In addition, some individuals who later died by suicide denied suicidal thoughts during evaluations [12].

Given these limitations, textual expression offers a complementary perspective on mental states [4,17,15]. In particular, suicide notes have

long served as a valuable source for understanding the circumstances and emotions that lead an individual to consider or complete suicide [5,16].

Currently, many individuals—particularly the young—express emotions, thoughts, and suicidal plans on platforms such as Facebook, X, and Reddit, often preferring these to mental health professionals [7, 12, 17]. Reddit hosts active mental health communities, including r/Depression, r/SuicideWatch, and r/Anxiety [3].

Emotions play a critical role in detecting suicidal ideation, as affective fluctuations often precede or accompany suicidal thoughts. Automatic analysis of emotional cues in online posts enables early intervention, potentially saving lives.

Natural language processing (NLP), a subfield of artificial intelligence, facilitates the extraction and analysis of textual and emotional cues from social media. In this field, approaches for detecting suicidal ideation have been developed that often rely on lexical, syntactic, semantic, and affective features, typically analyzing emotions at a global level—either assigning a predominant emotion to the full text or considering long-term historical context. However, they often fail to capture nuanced, dynamic emotional patterns at the sentence level.

To address this, we propose a methodology that models emotional dynamics at the sentence level for suicide risk identification. Our approach uses emotional vectors alone or combined with linguistic features, incorporating attention mechanisms to identify and weigh salient emotional patterns within each sentence. This allows the model to capture the dynamic evolution of emotions throughout a text, providing deeper insights into the emotional signals associated with suicidal ideation, even when suicidal intent is not explicitly expressed.

The remainder of this paper is organized as follows. Section 2 reviews related work on computational suicidal ideation analysis. Section 3 describes datasets and feature extraction techniques. Section 4 details the proposed methodology. Section 5 presents classification experiments, models, and results. Finally, Section 6 summarizes key findings.

2 Related Work

Recent advances in computational methods have enabled the automatic analysis of mental health disorders [4,13] and suicidal behaviors from textual data originating from social media [1,7].

A variety of approaches have been explored, including machine learning and deep learning architectures such as CNNs, LSTMs, BiLSTMs with attention mechanisms, and pre-trained transformers (e.g., BERT, RoBERTa, and AlBERT) [17]. These models leverage multiple feature types, including textual content, linguistic style (e.g., LIWC categories), semantic representations [16], and emotional signals [16] derived from lexicons such as Plutchik's eight basic emotions and the NRC Emotion Lexicon [17]. Temporal modeling and user-level aggregation are frequently employed to capture the evolution of emotional [12] and cognitive states [17,16], which are critical indicators of mental health disorders.

Data sources include social media posts, blogs, and genuine suicide notes, offering complementary perspectives on the expression of psychological distress [16]. These studies demonstrate that integrating content, stylistic, and emotional features [11], along with temporal dynamics, substantially improves the early detection and understanding of suicidal behavior.

Uban et al. [17] focused on the automatic detection of depression, anorexia, and self-harm tendencies using social media data. Their approach combines representations of content, linguistic style (LIWC 2007), and emotions (Plutchik and NRC Emotion Lexicon), considering both static and temporal dynamics. Classification was performed using BILSTM networks with attention at the post and user levels, CNNs, and pre-trained transformers. Results indicate that users with self-harm tendencies display distinctive patterns: increased anger when discussing the causes of their emotions, while fear and sadness are heightened when such discussions are infrequent. Moreover, first-person pronouns and

the word" I" correlate with higher negative emotions, differentiating affected users from healthy controls and supporting early detection efforts.

Teixeira et al. [16] examined the semantic and emotional structure of genuine suicide notes through cognitive network science. The methodology reconstructs knowledge structures and analyzes interconnections between ideas and emotional states via measures such as emotional balance, semantic prominence, and affective profiling. Findings reveal that positive and negative terms interact to create higher emotional balance than randomized models and that suicide notes exhibit affective compartmentalization, with clusters of positive concepts dominating the network. Central nodes, such as "love," integrate self-other relations and multiple emotions, reflecting theoretical perspectives from narrative psychology and meaning-making. These results provide a quantitative framework for understanding the cognitive-emotional states of individuals at risk and inform suicide prevention strategies.

Sawhney et al. [12] proposed PHASE, a framework for detecting suicidal ideation by modeling emotional phase-aware representations from historical social media posts annotated for suicide risk. The approach encodes sequences of textual content alongside inferred emotional states using BILSTM networks with attention, emphasizing posts' relative importance within a user timeline. Both textual embeddings and emotion-aware representations are incorporated. Experiments show that integrating temporal emotional dynamics significantly improves detection performance over baseline models that treat posts independently, highlighting the relevance of longitudinal emotional patterns and user-level context for early identification.

Ren et al. [11] introduced a complex emotion topic (CET) model to examine accumulated emotional traits in suicide blogs. Emotional features are derived from eight basic emotions and five intensity levels, capturing accumulation, covariance, and transition of consecutive emotions. Generalized linear regression was used to assess predictive power for discriminating suicidal from non-suicidal blogs. The study also identified recurring emotional themes, such as hopelessness and isolation. Results demonstrate that emotional transitions are significant predictors of suicide risk and that combining the three cumulative traits enhances discrimination between suicidal and non-suicidal content.

Emotional content analysis in social media has been proposed as a relevant source of information to identify suicidal ideation. However, most existing approaches focus on reporting whether an emotion is present or absent (binary classification) or identifying its polarity (positive or negative emotion), limiting the depth of insight into users' affective states. A deeper analysis of emotional dynamics is necessary to extract more informative features capable of capturing suicidal thinking.

To address this limitation, the present study proposes an approach that analyzes emotional changes at the sentence level within each note. By segmenting documents and extracting temporal emotional patterns, we aim to identify notes indicative of suicidal ideation. The proposed system consists of preprocessing, feature extraction, latent emotion identification, temporal analysis, and classification. This framework allows for a deeper understanding of the interplay between emotional variations and suicidal thoughts.

3 Materials and Methods

3.1 Datasets

In order the validate of the proposed approach, in this study three types of datasets-consummated suicide, suicide ideation, and general domain-were considered. Each dataset is described below:

Consummated suicide notes: this set consists of 309 notes authored by individuals who died by suicide, with each case verified. The notes were collected from diverse sources, including news articles, online publications, and books. Its primary strength lies in its confirmed content, directly associated with documented cases of completed suicide, which ensures high reliability and scientific validity for research purposes.

Suicidal ideation notes: this dataset comprises 309 notes collected from Reddit, specifically from

the subreddits r/Suicide Watch and r/Depression, where users openly share their thoughts and emotions associated with suicide. Although the authenticity of these notes cannot always be verified, they provide unique access to the spontaneous and direct expression of suicidal ideation in digital environments. This corpus facilitates the identification of emotional patterns and early warning signals, which represent a valuable resource for advancing preventive research and intervention strategies.

General domain notes: This collection includes 255 publicly available notes covering a wide range of non-suicidal topics, including culture, economy, politics, music, food, and sports. These texts were also collected from Reddit and serve as a control group, to differentiate between suicide-related and non-suicide-related notes. A description of each dataset is provided below (Table 1).

3.2 Feature Extraction Methods

Feature vectors provide numerical representations of text at different linguistic levels, enabling machine learning algorithms to classify data effectively. In this study, we evaluate the relevance of lexical, semantic, and contextual features using five representation methods: Term Frequency-Inverse Document Frequency (TF-IDF), Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), Doc2Vec (D2V), and Bidirectional Encoder Representations from Transformers (BERT). With the exception of TF-IDF and BERT, all feature generation models were trained on the Wikipedia corpus. A summary of these methods is presented below. TF-IDF: is a statistical measure that evaluates the importance of a term within a document relative to an entire corpus by combining two main components: term frequency (TF) and inverse document frequency (IDF). Mathematically, TF-IDF is defined as:

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D), \quad (1)$$

where t represents the term and d the document.

Term Frequency (TF): Term frequency measures the local relevance of a term within a document, calculated as the ratio of the number of times the term appears in the document to the total number of terms in the same document:

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \quad (2)$$

where $f_{t',d}$ is the number of occurrences of term t in document d, and the denominator is the sum of frequencies of all terms of the document d. Inverse Document Frequency (IDF): IDF measures the relevance of a term across the entire corpus, assigning higher weights to terms that are uniques and lowering the importance of terms that appear frequently across many documents. Following the formula:

$$IDF(t,D) = \log\left(\frac{|D|}{|d \in D : t \in d|}\right), \quad (3)$$

where D is the total number of documents in the corpus, and the denominator is the number of documents containing term t. To represent documents numerically using TF-IDF, a term-document matrix is constructed in which each row corresponds to a document d, each column corresponds to a unique term t in the corpus vocabulary, and each cell contains the TF-IDF weight $TF - IDF(t,d,D)$ indicating the relevance of term t in document d. This representation enables the identification of relevant and distinctive lexical features, making TF-IDF a widely adopted baseline for text mining and classification tasks.

LDA [2]: is a probabilistic topic model that represents documents as mixtures of latent topics, where each topic is defined by a probability distribution over a set of words. This model assumes that documents are generated from these hidden topics, and that each word assigned to a topic according to a certain probability. LDA produces numerical vectors that capture both the probability of word-topic associations and the proportion of topics within each document. This allows the transformation of texts into semantic representations that capture the underlying thematic relationships within the corpus.

In simple terms, the generative process assumed by LDA consists of the following steps: 1. Determine the number N of words in the

Table 1. Description of the corpus used

Class	#Docs	Avg. words	Avg. sent
Consummated suicide	309	50	5
Suicidal ideation	309	143	10
General topics	255	110	6

document according to a Poisson distribution. 2. Choose a mixture of topics for the document from a fixed set of K topics according to a Dirichlet distribution. 3. Generate each word in the document as follows:

(a) Select a topic (b) Select a word within that topic D2V [9]: is a artificial neural network-based algorithm designed to generate fixed-length numeric vectors that capture the semantic content of entire documents, independent of their length. Inspired by word embedding techniques, Doc2Vec aims to predict words based on their contextual surroundings while simultaneously learning a unique vector for each document.

Unlike traditional frequency or count-based techniques, Doc2Vec incorporates contextual information by considering the order and semantic relationships between words within a document. To achieve this, it trains a neural network that predicts words within the document's context, while the document vector is updated to represent the overall semantic content. Formally, given a sequence of training word w_1, w_2, ..., w_n, the model maximizes the average log probability:

$$\sum_{t=-k}^{T-k} \log p(w_t | w_{t-k}, \cdots, w_{t+k}; W), \quad (4)$$

where prediction is carried out by a multiclass classifier (softmax), following the formula: y_i p(WtWt-k…. Wt+k) = is calculated as: egwt (5):

$$y = b + Uh(w_{t-k}, \cdots, w_{t+k}; W), \quad (5)$$

where h is obtained by concatenating vectors from the word matrix W, and U and b are softmax parameters. A hierarchical softmax is often used to speed up training.

The model is trained using stochastic gradient descent with backpropagation. Upon convergence, words with similar meanings are represented by vectors in close proximity, and each document vector encodes both semantic and contextual information, making Doc2Vec a powerful tool for document-level representation in downstream machine learning tasks.

BERT [18] is a language representation model pretrained on unlabeled text using a bidirectional approach. By processing context from both the left and right directions simultaneously, BERT is able to capture not only the semantic relationships between words but also the complete contextual meaning of sentences within a document.

In this study, we employ DistilBERT, a compact and efficient variant of BERT. DistilBERT preserves the fundamental architecture of BERT while reducing the number of layers from 12 to 6, maintaining a hidden size of 768 and 12 attention heads, resulting in approximately 66 million parameters compared to the 110 million in BERT-base. This reduction significantly decreases computational cost and training time without substantially compromising performance on NLP tasks.

Like BERT, DistilBERT processes input sequences that can consist of a single sentence or sentence pairs, starting each sequence with the special [CLS] token. The final hidden state of this token, extracted from the model's last layer, acts as a contextualized feature vector summarizing the information contained in the entire sequence.

4 Proposed Method

This study presents an approach for detecting suicidal ideation by analyzing the emotional changes present in each note, considering them a key factor. Documents are segmented at the sentence level, enabling temporal extraction and analysis of emotional dynamics across the text. Subsequently, these emotional patterns are used to identify notes focused on suicidal ideation, providing insights into the relationship between emotional variations and suicidal thoughts.

In general, our system consists of five main modules: preprocessing, feature generation, latent emotion identification, temporal analysis of

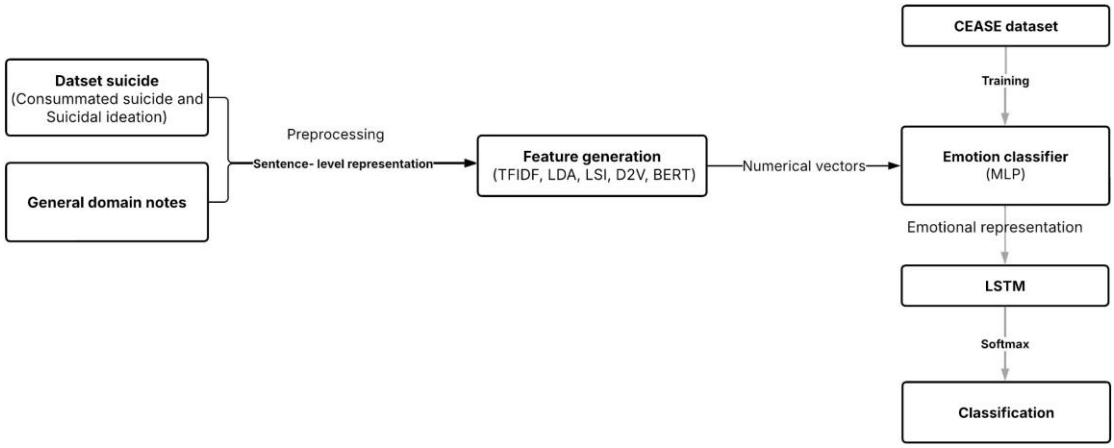

Fig. 1. General scheme of the proposed method

emotional dynamics, and classification. During preprocessing, the dataset is cleaned by removing URLs and user mentions, followed by word- and sentence-level tokenization using the NLTK library. Subsequently, feature extraction is performed using multiple methods such as TF-IDF, LDA, LSI, Doc2Vec, and BERT model. Sentence-level emotional content is quantified through the emotion identification module, revealing latent affective patterns. These emotional representations are fed into an LSTM network that analyzes the relationships within a note's emotional vectors to finally predict its class.

In the tokenization stage, suicide notes are divided into individual sentences, as each sentence is considered a complete unit reflecting an idea or thought of the author. Thus, each document D is composed of a set of sentences$D = \{s_1, s_2, \ldots, s_n\}$. Figure 1 provides a general overview of our proposed method.

4.1 Emotional Representation

The emotional distribution quantifies the degree to which each emotion is expressed in every sentence $s_n \in D$ This distribution is represented as a continuous value between 0 and 1, providing a fine-grained representation of the affective content. Consequently, each sentences s_n is represented by a normalized 15-dimensional probability vector:

$$e_d = (e_1, e_2, \ldots, e_{15}), \quad (6)$$

where $e_1 \in [0, 1]$ denotes the normalized intensity of emotion in sentence d, and $\sum_i e_1 = 1$.

To estimate this distribution, an emotion classifier was trained using the CEASE corpus, a dataset specifically developed for sentiment analysis in the context of suicide. The corpus comprises textual data from multiple sources like websites, newspapers, blogs, and books.

For this study, a subset of 499 sentences from the CEASE dataset was employed, each annotated with one of 15 emotional categories. This subset was subsequently partitioned into training (70%), validation (20%), and test (10%) sets.

Sentences were transformed into numerical vectors using a set of feature extraction methods (Section 3.2) designed to capture both lexical and semantic information relevant to emotion detection.

These vector representations were then processed by a feedforward artificial neural network (ANN), specifically a Multilayer Perceptron (MLP) which consists of a single hidden layer with 50 neurons and an output layer of 15 neurons. The output of the perceptron corresponded to each of the predefined emotional categories. The MLP produced normalized probability distributions using the predict probability method in scikit-learn, ensuring that the emotion scores for each sentence sum to one.

Table 2. Performance of the emotion classifier (per class)

Class	Precision	Recall	F-score
Abuse	0.2100	0.8000	0.3300
Anger	1.0000	0.5000	0.6700
Blame	1.0000	0.3300	0.5000
Fear	1.0000	0.7500	0.8600
Forgiveness	0.6700	0.6700	0.6700
Guilt	0.6000	0.5000	0.5500
Peacefulness	0.3300	0.3300	0.3300
Hopefulness	0.0000	0.0000	0.0000
Hopelessness	0.2000	0.2500	0.2200
Information	0.5000	0.4000	0.4400
Instruction	0.3300	1.0000	0.5000
Love	0.5000	0.3300	0.4000
Pride	0.0000	0.0000	0.0000
Sorrow	1.0000	0.6000	0.7500

The performance of the classifier was evaluated in terms of precision, recall, and F-score.

The results, presented in Table 2, reveal heterogeneous performance across emotion categories. Emotions such as fear (F1 = 0.86) and sorrow (F1 = 0.75) achieved the best results, demonstrating the model's ability to recognize emotions strongly associated with suicidal ideation.

In contrast, emotions such as hopefulness and pride, obtained an $F1 = 0.00$, reflecting the model's inability to detect these categories. These findings suggest that expanding and balancing the training dataset, particularly for less frequent emotions, could substantially improve performance of emotion classification.

After training, the model was applied to infer emotions in sentences from both suicide-related and non-suicide documents. The MLP processes sentence embeddings-obtained using the methods in (Section 3.2) and generates a 15-dimensional vector for each sentence, where each dimension corresponds to a predefined emotional category.

Probabilities for each category are estimated using the softmax function, yielding the emotional state expressed in the text.

4.2 Identification of Suicide Notes

A document D is composed of multiple sentences $\boldsymbol{D} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_n}\}$, where order is essential to preserve meaning. Emotional content may vary across sentences, and such shifts can indicate suicidal ideation. To capture these temporal patterns, a Long Short-Term Memory (LSTM) network is employed, preserving both context and sentence order while processing the sequence of emotional states across sentences.

Each sentence sn is represented as a 15-dimensional vector $\boldsymbol{e}\pm \in \mathbb{R}^{\boldsymbol{15}}$ encoding its emotional distribution. The sequence of emotional states $\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_n}$ defines the temporal steps of the LSTM, with each time step t corresponding to the sentence position within the document.

At each step, the LSTM updates its internal gates by processing the input emotional vector et along with the previous hidden state $\boldsymbol{h_{t-1}}$ and cell state $\boldsymbol{c_{t-1}}$.

The forget gate$\boldsymbol{f_t}$controls which information is retained; the input gate $\boldsymbol{i_t}$regulates incorporation of new information into the memory $\boldsymbol{c_t}$ and the output gate $\boldsymbol{O_1}$determines the portion of memory exposed as the hidden state $\boldsymbol{h_t}$. This architecture allows the LSTM to effectively model temporal dependencies and emotional transitions across sentences, preserving the sequential structure of the author's discourse.

The final hidden state, encoding the document level representation, is passed through a fully connected layer followed by a softmax function to estimate class probabilities $\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{D})$ for binary classification between suicide-related notes and notes from other topics.

5 Experimentation

Figure 2 illustrates the sentence-level emotional distributions for one document from each dataset, allowing a comparison across different text types.

Each emotion is visualized as a set of bars representing the emotional change throughout the sentences. In the graphs, emotions are plotted on the horizontal axis and their corresponding probabilities on the vertical axis.

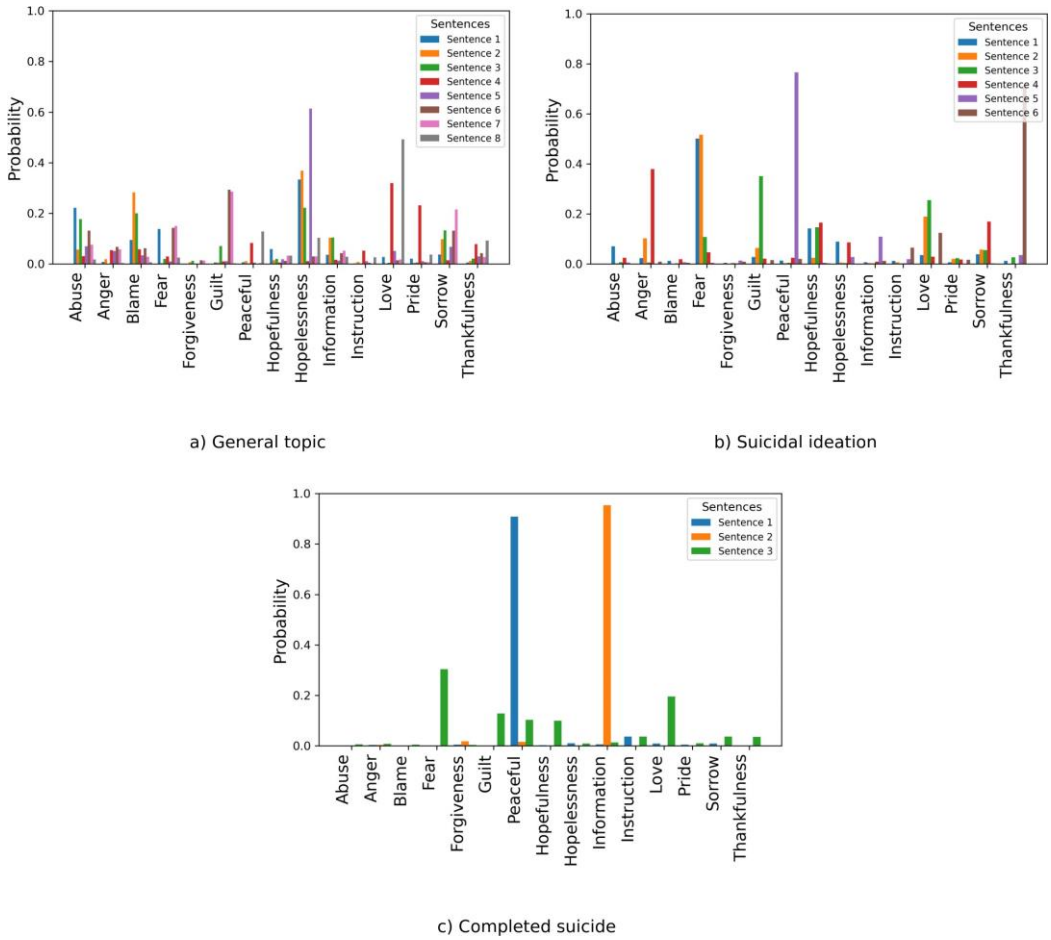

Fig. 2. Probability of emotions in general topics, suicidal intention, and consummated suicide notes

As can be observed, in the case of the general topic (Figure 2a), emotional probabilities are broadly dispersed across sentences, reflecting high diversity and variability among categories. In the suicidal ideation note (Figure 2b), multiple emotions remain present, but distributions exhibit more pronounced peaks in specific categories, indicating reduced dispersion relative to general texts.

Finally, in completed suicide note (Figure 2c), the distribution is highly concentrated: one or two emotions dominate with probabilities exceeding 0.8, while remaining categories are near zero.

Overall, these findings indicate a progressive decrease in emotional variability and an increase in concentration of specific emotions from general texts to completed suicide notes. Such patterns constitute a relevant discriminative feature for the automatic detection of suicidal ideation in texts. To assess the impact of emotional dynamics and the integration of linguistic features on these models are described in the following subsections.

5.1 Model A: Standard LSTM (Emotion-Only)

Detection of suicide ideation notes, two sequential neural network architectures were designed and compared. Both process sequences of sentence-level vectors but differ in input representations and the use of attention mechanisms. The details.

In this model, each document D is represented as a sequence of emotional vectors $S = (e_1, e_2, \ldots, e_n$. where $e_n \ \mathbb{R}^{15}$ represents the

Table 3. Consummated suicide vs General topics

Method	Model A P	Model A R	Model A F	Model B P	Model B R	Model B F
TF-IDF	0.7522 ± 0.060	0.7305 ± 0.0375	0.7244 ± 0.0398	0.8239 ± 0.0155	0.8227 ± 0.0162	0.8223 ± 0.0166
LSI	0.7432 ± 0.0208	0.7305 ± 0.0250	0.7289 ± 0.0300	0.8958 ± 0.02288	0.8883 ± 0.0207	0.8874 ± 0.0185
LDA	0.7481 ± 0.0265	0.7376 ± 0.0217	0.7385 ± 0.0230	0.8089 ± 0.0393	0.8014 ± 0.0472	0.7975 ± 0.0480
Doc2Vec	0.7328 ± 0.0332	0.7233 ± 0.0302	0.7235 ± 0.0300	0.7329 ± 0.0479	0.7252 ± 0.0502	0.7259 ± 0.0474
BERT	0.8695 ± 0.0445	0.8695 ± 0.0445	0.8637 ± 0.0463	0.9248 ± 0.0270	0.9220 ± 0.0445	0.9210 ± 0.0399

Table 4. Top ten most frequent words per dataset

Consummated suicide		Suicidal ideation		General Topics	
Word	**Freq.**	**Word**	**Freq.**	**Word**	**Freq.**
life	86	want	280	people	172
love	59	life	250	like	136
one	52	suicide	222	get	94
like	47	like	208	one	89
would	42	know	197	even	76
time	38	feel	169	would	74
know	38	even	156	really	69
good	38	one	143	time	65
please	37	people	138	good	63
people	37	die	133	know	62

probability distribution over the predefined emotional categories for sentence $s_n \in D$ The sequence emotional is processed by an LSTM, and the final hidden state h_t is regularized using dropout and passed through a softmax layer to perform binary classification.

5.2 Model B: LSTM with Attention (Feature Concatenation)

Model B extends Model A by (i) concatenating emotional vectors with the original document-level linguistic vectors and (ii) incorporating an attention mechanism to dynamically weight the relevance of each sentence. Each sentence sn is encoded as $x_i = [e_t; l_i] \in \mathbb{R}^{15+P}$ combining emotional and linguistic features. The sequence $X = (x_1, x_2, \ldots, x_N)$ is processed by an LSTM with attention, producing a context vector that emphasizes the most informative segments, which is then passed through a sigmoid layer for binary classification.

The main advantage of Model B lies in its hybrid input representation and attention mechanism, which overcome Model A's limitations of ignoring linguistic information and treating all segments equally. By assigning differentiated weights to each segment, highlighting the most informative and task-critical fragments, Model B captures the interactions between emotional dynamics and linguistic features, enhancing classification performance.

5.3 Results

In this section, we evaluate the proposed approach for detecting suicide notes by comparing two architectures: Model A (emotion-only) and Model B (emotion+linguistic with attention). Evaluations were conducted across datasets from diverse contexts (Section 3.1) and multiple feature representations (Section 3.2) to examine the robustness and generalizability of each model under varying linguistic and emotional conditions.

All models were evaluated using five-fold cross- validation to ensure balanced class distributions and robust estimates. Results are reported in terms of precision (P), recall (R), F-measure (F), and standard deviation (SD).

In the first scenario, consummated suicide notes were classified against general-topic documents. Table 3 presents the results. The relevance of this experiment lies in evaluating the model's ability to discriminate between texts produced in extreme psychological contexts preceding a suicide and everyday writings. The challenge is heightened by shared vocabulary across both types of documents (Table 4), which limits purely lexical approaches. Furthermore, the diversity of authors in terms of nationality, gender, socioeconomic background, culture, and personal motivations-adds complexity, reflected in differences in writing style, word choice, and emotional patterns.

Beyond lexical content, this scenario highlights the discriminative power of emotional variability. The results show that Model B consistently outperforms Model A, particularly for TF-IDF (+10 points) and LSI (+16 points), demonstrating that the integration of lexical-emotional features and attention, mitigates the limitations of approaches based solely on lexical or emotional analysis. On

Table 5. Suicidal ideation vs General topics

Method	Model A P	Model A R	Model A F	Model B P	Model B R	Model B F
TF-IDF	0.6865 ± 0.0559	0.6683 ± 0.0592	0.6683 ± 0.0592	0.9279 ± 0.0077	0.9279 ± 0.0137	0.9277 ± 0.0106
LSI	0.5099 ± 0.1271	0.5496 ± 0.348	0.5095 ± 0.0894	0.8479 ± 0.0375	0.8206 ± 0.0604	0.8085 ± 0.0578
LDA	0.6480 ± 0.0637	0.5797 ± 0.0566	0.5647 ± 0.0703	0.8322 ± 0.0334	0.8219 ± 0.0324	0.8122 ± 0.0318
Doc2Vec	0.5023 ± 0.1577	0.5196 ± 0.0629	0.4658 ± 0.1169	0.4530 ± 0.1062	0.6298 ± 0.0354	0.5196 ± 0.0761
BERT	0.8528 ± 0.0442	0.8473 ± 0.0411	0.8475 ± 0.0431	0.9660 ± 0.0141	0.9660 ± 0.0165	0.9659 ± 0.0154

Table 6. Consummated suicide vs Suicidal ideation

Method	Model A P	Model A R	Model A F	Model B P	Model B R	Model B F
TF-IDF	0.7524 ± 0.0511	0.7362 ± 0.0468	0.7341 ± 0.0458	0.8929 ± 0.0089	0.8923 ± 0.0139	0.8921 ± 0.0113
LSI	0.7344 ± 0.0355	0.7249 ± 0.0383	0.7210 ± 0.0381	0.8296 ± 0.0276	0.8149 ± 0.0243	0.8096 ± 0.0298
LDA	0.7714 ± 0.0353	0.7508 ± 0.0430	0.7454 ± 0.0448	0.8048 ± 0.0221	0.7939 ± 0.0378	0.7895 ± 0.0344
Doc2Vec	0.7429 ± 0.0295	0.7298 ± 0.0320	0.7271 ± 0.0369	0.7519 ± 0.0464	0.7427 ± 0.0530	0.7318 ± 0.0563
BERT	0.8316 ± 0.0292	0.8269 ± 0.0255	0.8270 ± 0.0273	0.8957 ± 0.0270	0.8702 ± 0.0445	0.8675 ± 0.0399

the other hand, BERT embeddings achieved highest performance across both architectures. Model B, by integrating BERT, hybrid features, and attention mechanism, achieved the highest overall F-score (0.921 ± 0.040), also exhibiting the lowest variability, reflecting enhanced robustness and stability.

In the second experiment, notes of suicidal ideation were classified against general-topic texts. Table 5 presents the results. BERT markedly outperformed traditional methods, achieving F-score of 0.8475 (Model A) and 0.9659 (Model B), while TF-IDF achieved intermediate results (0.6633 and 0.9277). LSI, LDA, and Doc2Vec performed lower with greater variability.

Lexical analysis revealed that suicidal ideation notes contain explicit, emotionally charged vocabulary, such as ”want,” ”suicide,” ”feel,” and ”die,” whereas general texts employ neutral terms (”people,” ”like,” ”good”). This contrast allows BERT, particularly Model B, to combine emotional and lexical cues via attention, maximizing the relevance of the most informative segments.

Finally, consummated suicide notes were classified against suicidal ideation notes. Although both datasets reflect suicide-related texts, ideation notes often contain explicit expressions of intent, whereas consummated suicide notes may lack such direct indicators. The aim of this experiment was to differentiate texts reflecting actualized suicidal behavior from those denoting suicidal ideation without implying an imminent act.

The results, summarized in Table 6, show that text representation and model architecture significantly affect classification performance. Under Model A with BERT, an F-measure of 0.8270 was achieved, reflecting the model’s capacity to capture nuanced emotional signals. Model B, leveraging emotional vectors enriched with linguistic features and attention, achieved the highest performance (F-measure = 0.8921), surpassing BERT alone (F − measure = 0.8675). TF-IDF contributed by emphasizing key lexical cues indicative of suicidal intent. By combining TF-IDF with emotional vectors in an attention-based LSTM, the model effectively focused on critical patterns, enhancing classification performance.

6 Conclusions

Suicide represents a critical public health concern, strongly influenced by emotions as expressed in text. This study proposes a computational method for classifying suicidal texts based on sentence-level emotional analysis. Two long short-term memory (LSTM) architectures were evaluated: a standard LSTM and an LSTM with an attention mechanism, which combines emotional vectors with document-level features.

Evaluation was conducted on three types of notes: consummated suicide, suicidal ideation, and general-domain texts. Analysis of emotional distributions shows clear differences across categories. General-domain texts exhibit high variability and a wide emotional range. Suicidal ideation notes maintain multiple emotions but with more pronounced peaks in specific categories. Consummated suicide notes display strong emotional concentration, with probabilities focused on one or two dominant emotions. These

results indicate that sentence-level emotional patterns are highly discriminative for identifying suicidal content.

Lexical analysis revealed that suicidal ideation notes often use explicit, emotionally charged language, unlike consummated suicide and general-domain texts, which are more neutral. This highlights both the complexity of the task and the importance of emotional features for detecting subtle risk signals.

Combining emotional vectors with document-level representations in an LSTM with attention improved classification across all note types. The attention mechanism highlights relevant text segments, allowing the model to capture dynamic, fine-grained emotional cues that enhance detection. While integrating multiple vectors reduces interpretability, emotional and lexical patterns complement each other, with emotional signals guiding classification and lexical features boosting overall performance.

References

1. **Abdulsalam, A., Alhothali, A. (2024).** Suicidal Ideation Detection on Social Media: Review of Machine Learning Methods. Social Network Analysis and Mining, Vol. 14, No. 1, pp. 1–16. doi: 10.1007/s13278-024-01348-0.
2. **Blei, D.M., Ng, A.Y., Jordan, M.I. (2003).** Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, No. Jan, pp. 993–1022.
3. **Dan, E., Zhu, J., Jin, R. (2025).** Exploring Suicide Factors in Online Discourse: Sentiment and Thematic Analysis of Reddit. ACM Transactions on the Web, Vol. 19, No. 2. doi: 10.1145/3716546.
4. **D'Alfonso, S. (2020).** AI in Mental Health. Current Opinion in Psychology, Vol. 36, pp. 112–117. doi: 10.1016/j.copsyc.2020.04.005.
5. **Fata, I.A., Yusuf, Y.Q., Kamal, R., Namaziandost, E. (2021).** The Characteristics of Linguistic Features Enfolded in Suicide Notes. Journal of Language and Linguistic Studies, Vol. 17, No. 2, pp. 720–735. doi: 10.52462/jlls.50.
6. **Ghosh, S., Roy, S., Ekbal, A., Bhattacharyya, P. (2022).** CARES: Cause Recognition for Emotion in Suicide Notes. In Advances in Information Retrieval, Springer, pp. 128–136. doi: 10.1007/978-3-030-99739-7_15.
7. **Haque, R., Islam, N., Islam, M., Ahsan, M.M. (2022).** A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies, Vol. 10, No. 3, pp. 57. doi: 10.3390/technologies10030057.
8. **Ji, S. (2022).** Towards Intention Understanding in Suicidal Risk Assessment with Natural Language Processing. Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 4028–4038. doi: 10.18653/v1/2022.findings-emnlp.297.
9. **Le, Q., Mikolov, T. (2014).** Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, PMLR, pp. 1188–1196.
10. **Leenaars, A.A. (2010).** Edwin S. Shneidman on Suicide. Suicidology Online, Vol. 1, No. 1, pp. 5–18.
11. **Ren, F., Kang, X., Quan, C. (2015).** Examining Accumulated Emotional Traits in Suicide Blogs with an Emotion Topic Model. IEEE Journal of Biomedical and Health Informatics, Vol. 20, No. 5, pp. 1384–1396. doi: 10.1109/JBHI.2015.2459683.
12. **Sawhney, R., Joshi, H., Flek, L., Shah, R. (2021).** PHASE: Learning Emotional Phase-Aware Representations for Suicide Ideation Detection on Social Media. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2415–2428. doi: 10.18653/v1/2021.eacl-main.205.
13. **Saxena, R.R. (2024).** Applications of Natural Language Processing in the Domain of Mental Health. TechRxiv. doi: 10.36227/techrxiv.173014748.80471770/v1.
14. **Soumya, K., Garg, V.K. (2022).** Named Entity Emotion Intensity Tagging for Suicidal Ideation Detection from Social Media Texts During MT. International Journal of Intelligent

Engineering and Systems, Vol. 15, No. 6. doi: 10.22266/ijies2022.1231.22.

15. **Stella, M., Swanson, T.J., Li, Y., Hills, T.T., Teixeira, A.S. (2022).** Cognitive Networks Detect Structural Patterns and Emotional Complexity in Suicide Notes. Frontiers in Psychology, Vol. 13, pp. 917630. doi: 10.3389/fpsyg.2022.917630.
16. **Teixeira, A.S., Talaga, S., Swanson, T.J., Stella, M. (2021).** Revealing Semantic and Emotional Structure of Suicide Notes with Cognitive Network Science. Scientific Reports, Vol. 11, No. 1, pp. 19423. doi: 10.1038/s41598-021-98147-w.
17. **Uban, A.-S., Chulvi, B., Rosso, P. (2021).** An Emotion and Cognitive Based Analysis of Mental Health Disorders from Social Media Data. Future Generation Computer Systems, Vol. 124, pp. 480–494. doi: 10.1016/j.future.2021.05.032.
18. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J. et al. (2017).** Attention Is All You Need. Advances in Neural Information Processing Systems, pp. 5998–6008. doi: 10.48550/arXiv.1706.03762.
19. **World Health Organization. (2021).** Suicide Worldwide in 2019: Global Health Estimates. World Health Organization.
20. **Yöyen, E., Keleş, M. (2024).** First- and Second-Generation Psychological Theories of Suicidal Behaviour. Behavioral Sciences, Vol. 14, No. 8, pp. 710. doi: 10.3390/bs14080710.
21. **Zhang, T., Yang, K., Ji, S., Ananiadou, S. (2023).** Emotion Fusion for Mental Illness Detection from Social Media: A Survey. Information Fusion, Vol. 92, pp. 231–246. doi: 10.1016/j.inffus.2022.11.031.

Article received on 22/06/2025; accepted on 13/11/2025.
**Corresponding author is Yulia Ledeneva.*