Sentiment Classification on Suicide Notes using BERT, Bi-LSTM, and Multi-head Attention

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE

Submitted by

ARMAAN AGRAWAL (2K23/AFI/04)

Under the supervision of

DR. ROHIT BENIWAL

Department of Computer Science and Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **ARMAAN AGRAWAL**, Roll No's – **2K23/AFI/04** students of M.Tech (Artificial Intelligence), hereby declare that the thesis titled **"Sentiment Classification on Suicide Notes using BERT, Bi-LSTM, and Multi-head Attention"** which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi — Armaan Agrawal

Date: — (2K23/AFI/04)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the thesis titled "**Sentiment Classification on Suicide Notes using BERT, Bi-LSTM, and Multi-head Attention**" which is submitted by **Armaan Agrawal**, Roll No's – **2K23/AFI/04**, Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the research work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

Dr. Rohit Beniwal

(Supervisor)

Department of Computer Science and Engineering

Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Dr. Rohit Beniwal for his continuous guidance and mentorship that he provided me during the research work. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this research as well as its industrial relevance. He was always ready to help me and clear my doubts regarding any hurdles in this research. Without his constant support and motivation, this research would not have been successful.

Place: Delhi — Armaan Agrawal

Date: — (2K23/AFI/04)

Abstract

Nowadays, there is a lot of focus on mental health and mental illnesses. Suicide is a topic of serious concern and a lot of effort is being put in by researchers and mental health professionals to combat this issue. "Sentiment Analysis on Suicide Notes" is a lesser-explored area. Understanding the emotions of the people committing the act of suicide is crucial to rage a fight against suicide and to develop areas of suicide prevention. This study focuses on works that relate to the task of "Sentiment Analysis on Suicide Notes".

Despite the importance of this task, there aren't enough datasets available to train models to identify different emotions in suicide notes. A variety of emotions are being represented in the dataset I am using. They are "happy", "sad", "neutral", "love", "hate", and "proud".

In this study, I will train the model in three phases. First, I will extract features from the input sentences using Bidirectional Encoder Representations from Transformers (BERT). This will be followed by sequence modeling performed using the Bidirectional Long Short-Term Memory (Bi-LSTM) network. In the third phase, to emphasize more attention on the important segments of the sentence, I'll use a Multi-head Attention mechanism. This would help increase the overall classification efficiency as it combines the strength of three different components. Resulting with Precision, Recall, and F1-Score all at 81%, this approach proves to be an effective one. This will help study the emotional content of the message written in suicide notes and will surely help in promoting initiative toward mental health in both personal and professional contexts.

Contents

List of Tables

List of Figures

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory
NLP	Natural Language Processing
I2B2	Informatics for Integrating Biology and the Bedside
CEASE	Corpus of Emotion Annotated Suicide notes in English
SIS	Suicidal Intent Sentiment

Chapter 1

INTRODUCTION

1.1 Overview

Data Mining deals with drawing out meaningful patterns and insights from huge amounts of unstructured text data. Natural Language Processing (NLP) has shown a significant advancement in the field of Data Mining [1] [2], by helping us see the patterns and trends that we might otherwise miss. NLP techniques applications can be seen in multiple domains, including machine translation, text summarization, information retrieval, chatbots and virtual assistants, and sentiment analysis [3].

Sentiment Analysis (or Opinion Mining), is a Natural Language Processing approach that helps understand the emotional tone of a given text. This has been a real breakthrough for the field [4]. By understanding human emotions, sentiment analysis can be helpful in multiple domains like customer feedback analysis, social media analysis, and psychological research.

Major studies and research in this field are done on data collected from Twitter and other E-commerce platforms like Amazon [5]. Such studies provide valuable information to governments and businesses, by understanding public interests, opinions, and attitudes.

However, there is a clear research gap in terms of analyzing suicide notes, being particularly a significant topic, which may help to study the emotions of an individual and provide any aid on time [6].

The task of analyzing and studying suicide notes dates back to 1957. A lot of work in this area is more concerned with recognizing Suicide Ideation, i.e. people with depressing thoughts [7] [8]. The introduction of various Natural Language Processing (NLP) techniques alongside the surge of Deep Learning Networks and Language Models has accelerated the progress in this task [9] [10] [11]. A typical Suicide Ideation task requires text data i.e. tweets, notes, posts, and blogs by different people that are labeled as suicidal or non-suicidal. Another work that complements this task of recognizing suicide ideation is the work of sentiment analysis in suicidal nodes. This work involves a labeled dataset with text data i.e. notes, tweets, blogs, and posts which are written moments before the act of suicide by the committer. This data is then broken down into individual sentences and each sentence is then labeled with an emotion from the annotator. The task of recognizing human sentiment from suicide notes allows for a better understanding of the emotional setup of the people who commit suicide.

One can perform sentiment analysis on different levels like Document, Sentence, and Phrase Levels. [12], among which I have chosen Sentence-level sentiment analysis, which helps to study the emotions expressed in each sentence of the notes.

This study first focuses on the works of Sentiment Analysis in Suicide Notes. There are three main datasets for this task. The first is the "I2B2 dataset for Sentiment Analysis in Suicide Notes", the second one is the "CEASE dataset", and the third is "Suicidal Intent Sentiment Dataset".

The first dataset, i.e. "I2B2 Dataset for Sentiment Analysis" was released in 2011 in the "I2B2/VA/Cincinnati Medical Natural Language Processing Challenge". A task of the challenge was the sentiment classification of suicide notes. The dataset used for the task was made up of 900 suicide notes. Each suicide note was broken down to a sentence level and then each sentence was labeled into one of the 15 emotions. These emotions were 'abuse', 'anger', 'blame', 'fear', 'guilt', 'hopelessness', 'sorrow', 'forgiveness', 'happiness_peacefulness', 'hopefulness', 'love', 'pride', 'thankfulness', 'instructions', and 'information'. A total of 24 teams took part in the challenge. This study covers the results provided by eight of those teams. One thing to observe is that although the dataset consists of enough suicide notes to train a model, the actual number of samples for both training and evaluation is not sufficient. For 10 of the 15 emotion classes, the number of samples in the test dataset doesn't exceed 50, and for 5 of those, the number is below 20. With so few samples the training and evaluation are both sub-optimal. A better alternative would be to increase the data or lower the number of emotion classes so that each category has enough data, for effective model training and evaluation. The "I2B2 dataset for Sentiment Analysis in Suicide Notes" is no longer publicly available as it has been discontinued.

The second dataset, i.e. "CEASE dataset", was introduced in 2020 as a substitute for the "I2B2 dataset for Sentiment Analysis in Suicide Notes" since the latter was no longer publicly available. It is a standard corpus of suicide notes which is used for emotion detection in patients. It contains data from 205 suicide notes. This dataset contains 2393 data points (sentences) alongside their labels. Each label is one of the fifteen emotion labels: 'forgiveness', 'happiness_peacefulness', 'love', 'pride', 'hopefulness', 'thankfulness', 'blame', 'anger', 'fear', 'abuse', 'sorrow', 'hopelessness', 'guilt', 'information', 'instructions'. However, this dataset, like the "I2B2 Dataset for Sentiment Analysis in Suicide Notes' suffers from the problem of insufficient data for optimal training and evaluation. 13 out of the 15 classes contain less than 50 samples for the testing set, and 7 of those 13 have the number below 10.

The third dataset, "Suicidal Intent Sentiment Dataset" was created by collecting 248 suicide notes from various sources, including Reddit, Quora, Storypick, and several blogs from Info-media sites. These notes were meticulously tokenized into sentences, resulting in 2632 individual sentences. Each sentence was manually annotated and categorized into one of the six sentiment categories: 'happy', 'sad', 'neutral', 'love', 'hate', and 'proud'. Each category have nearly same percentage of data.

In this study, I will combine three different state-of-the-art NLP techniques for three

different tasks; feature extraction, sequence modeling, and attention mechanism to focus more on the significant sections of the text. For these tasks I will be utilizing Bidirectional Encoder Representations from Transformers (BERT) [13], Bidirectional Long Short-Term Memory (Bi-LSTM) [14], and Multi-Head Attention Mechanism [15] respectively. These models are chosen because BERT allows us to capture the details hidden in the text and Bi-LSTM works well with sequential data. To enhance the model's performance, the Multi-Head Attention mechanism is used which allows the model to focus on the most important parts of the input sequence [16].

Using the above combination, I have seen a promising increase in the model's capability to correctly identify the emotions in a sentence, evidenced by its high Precision, Recall, and F1-Score values.

1.2 Motivation

Increasing mental health issues leading to the ideation of suicide is a growing global concern. Despite so much advancement in the field of artificial intelligence and natural language processing, one domain which is still relatively left unexplored is the "Sentiment Classification on Suicide Notes". Major reason for this gap is due to lack of availability of publicly accessible dataset, because of its sensitive nature and also the privacy concern for sharing the data with proper consent.

Not too late though, there seems to be an urgent and essential research required in this area. Suicide notes often contains subtle emotional cues, expressions of despair, and being the final message before one ends his life, if analyzed timely and correctly, can offer helpful insights regarding the psychological state of an individual.

The broader motivation behind this study is this work towards humanity. Identifying the emotions one holds before they quit their life on timely manner and integrating the system on social media platforms and counseling services may help provide timely aid which can lead to potentially save a life. This study bridges the gap between emotional understanding and the growing computational intelligence leading to social good and not limiting only to technical excellence.

Chapter 2

LITERATURE REVIEW

2.1 Overview of Existing Research

This section briefly summarizes twelve different papers related to the task of Sentiment Analysis in Suicide Notes. The first eight proposed works are based on the "I2B2 Dataset for Sentiment Analysis", the next two are based on the "CEASE dataset", and the last two are based on the "SIS dataset".

Yang et al. [17] proposed a hybrid model for sentiment analysis. Their system is made up of five major components. The first module is the text preprocessing module which performs the standard preprocessing operations for classification. The second module is the negation detection module which identifies negation signals in a second. Identifying the negation signals and marking them allows for a better understanding of the content of the text by the model. The third module is the emotion instance identification module which uses three machine learning algorithms to classify the sentence into one of the given classes, it also uses a conditional random field-based model that works on a token level for the same classification task as the three machine learning models. The fourth module is the result integration module which combines the results of the four models to achieve a final result. The fifth module is the post-processing module which works towards identifying potential sentences that can belong to the 'instructions' emotion class. They achieved an overall F1-Score of 61.39% and, for 7 of the 15 classes, an F1-Score of more than 50%.

Desmet and Hoste [18] used Support Vector Machine (SVM) techniques to create a system for fine-grained sentiment analysis for the above-defined Medical Natural Language Processing Challenge. Besides using the simple SVM technique, they also used the Bootstrap Resampling Technique with SVM to maximize the F1 score by optimizing the threshold value for the SVM Binary Classifier. Their work involves taking into account the fact that the 15 emotions in the dataset may not be mutually exclusive, therefore, they create a classifier model for each of the fifteen emotions. They achieved F1 scores of more than 40% for six of the seven most common emotions. They achieved the best results with an overall F1 Score of 53.87%. This overall score was achieved by averaging the values of the most successful classifiers for each emotion.

Wang et al. [19] presented a hybrid approach based on Machine Learning and rule-based classifiers for the above-defined Medical Natural Language Processing Challenge. They worked on using the N-gram-based, syntax-based, context-based, and knowledge-based information in the text along with certain class-specific features, as the feature set

to train an SVM model. The RBF (Radial Basis Function) is selected as the kernel for the model. For the rule-based learning classifier, several patterns were identified first and only the best patterns were kept for each class. Their results show that uni-gram and bi-gram features are more useful than tri-gram features for classification. Also, class-specific features are more useful than adding context features for classification. Individually, the best results achieved by the SVM model and the Rule Based Classifier had an F1 Score of 48.83%, and 45.36% respectively. The best results they achieved had an F1 Score of 50.38% obtained via combining both models to create a hybrid classifier.

Luyckx et al. [20] presented a thresholding approach using the support vector machine for the above-defined Medical Natural Language Processing Challenge. They convert the sentence with more than one label into multiple sentences with a single label by breaking down each sentence into fragments. They extract context-based and lexicon-based features from the sentences. They used the RBF function as the kernel for the SVM model. They experimented with the classification scheme and performed the classification task with three different schemes: 'emotion/no-emotion with thresholding', 'emotion-only', and 'emotion/no-emotion without thresholding'. Out of all the classification schemes, they achieved an F1-Score of 50.18%, for 'emotion/no-emotion with thresholding'. For the original classification task, they scored a Precision of 67.32%, a Recall of 34.42%, and an F1-Score of 45.36%.

Sohn et al. [21] presented three different models for the classification task. These are the multinomial Naive Bayes model, a rule-based classifier, and a hybrid of both. Data preprocessing is performed using NER (Named Entity Recognition), re-annotation, and Token Normalization. The best results they achieved were obtained by the hybrid model with a Precision of 57.09%, Recall of 55.74%, and an F1-Score of 56.40%.

Xu et al. [22] presented an SVM model using data augmentation. To deal with the problem of having smaller training data, they add manually annotated sentences to the training dataset. They combined the SVM model with three different preprocessing techniques: N-gram feature selection, Bag-of-N-gram feature selection, and pattern matching. For eight categories ('fear', 'guilt', 'hopelessness', 'love', 'sorrow', 'hopefulness', 'thankfulness', 'happiness_peacefulness') which are both subjective and have sufficient data, they rely on the linear SVM model with Bag-of-N-gram feature selection. The two objective categories ('information', and 'instruction') having sufficient training data, rely on the linear SVM model with N-gram feature selection. The remaining five subjective categories ('abuse', 'anger', 'blame', 'forgiveness', and 'pride') don't have sufficient training data and rely on the pattern matching technique. The outputs of the three models are combined to achieve a unified answer to a sentence. They resulted in a Precision of 56%, a Recall of 62%, and an F1-Score of 59%.

Cherry et al. [23] created an SVM model for the above similar task of classification coupled with various auxiliary features which allows the model to correct the imbalance in the number of data samples in each class of the dataset. The auxiliary features added to the data are bias feature; sentence features like number of words, names, tense, etc.; count of the synonyms for each label's name in the sentence; feature to describe the characters and feature which provides document level information. They achieved a Precision of 55.72%, Recall of 54.72%, and F1-Score of 55.22%.

McCart et al. [24] created three different models and their ensembles for classification. The three different models were the Rule-based model, the Statistical Text Mining model which couples data preprocessing and various machine learning models (Decision Trees, Support Vector Machine, K-Nearest Neighbor), the Weight Scheme-based model which assigns weight features and various auxiliary features to each data point (sentence), coupled with multiple machine learning models. They created an ensemble of these models and techniques and applied a voting scheme that gives equal weightage to each model and returns the answer with the most votes, only if that answer passes a set threshold of weight required to be obtained. Their best ensemble model achieved a Precision of 49.92%; a Recall of 50.55%; and an F1 score of 50.23

Ghosh et al. [25] used three deep learning techniques for the task of sentiment analysis on suicide notes using the "CEASE" dataset. The three different techniques were Convolutional Neural Network (CNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM). A standalone model from each of the three techniques was created and two hybrid models were also designed using an ensemble of these three techniques. Out of the two hybrid models, one was implemented by using all three models individually and then applying a voting strategy afterward, the other hybrid model was created by combining the three models and creating a Multi-Layer Neural Network Model. Among those three individual models, they achieved the best results using the LSTM model with a Precision, Recall, and F1-Score of 59%. The best results they achieved were obtained via the Multi-Layer Neural Network Model with a Precision Score of 59%, a Recall Score of 60%, and an F1 Score of 59%. However, it is also to be noted that the results obtained via the Multi-Layer Neural Network Model are just slightly better than the LSTM model and the training time to accuracy trade-off does not seem worth it.

Ghosh et al. [26] extend the "CEASE Dataset" by 120 Suicide Notes. They use this extended dataset for three different but related tasks. These tasks are Detection of Depressing Thoughts in an Individual, Single-Label Sentiment Analysis, and Multi-Label Sentiment Analysis. They created various individual and multi-models for these tasks using multiple Deep Learning Techniques. Out of all the models the Multitask Framework, which performed best, used the pre-trained GloVe (Global Vectors) embeddings, a Bi-GRU (Bidirectional GRU) based encoding layer, an Attention layer, and an External Knowledge module. Alongside these common layers, it also had a task-specific deep neural network module for each task. The best results they achieved for the Single-Label Sentiment Analysis Task had an F1-Score of 51.90%. The best individual system for the Single-Label Sentiment Analysis Task achieved an F1-score of 50.8% being worse than the Multitask Framework. They suspect that the better performance from the multitask framework may be due to the shared knowledge among various tasks.

Beniwal and Dhobal [27] used Bidirectional Long Short-Term Memory (Bi-LSTM) on Suicidal Intent Sentiment Dataset (SIS Dataset) resulting with Accuracy, Recall, and F1-Score of 72% and 73% of Precision.

Bansal and Beniwal [28] worked on the same dataset to train a hybrid model having capabilities of Generative Pre-trained Transformers (GPT), Bi-LSTM and Convolutional Neural Networks (CNN) to achieve Precision and F1-Score of 77% and Recall of 78%.

2.2 Comparative Analysis of Previous Works

Table 2.1 and Table 2.2 below provide a comparative study of the different works in the field of "Sentiment Classification on Suicide Notes".

Table 2.1 provides a comparison on the approaches taken by different researchers for this task. Majorly there are only three datasets used namely the "I2B2 Dataset", the "CEASE Dataset" and the "SIS Dataset". For the Preprocessing and Feature Selection part, a lot of different methods have been used. It is observed that four of the works perform the N-gram feature selection to create the feature set. Also, it is to be noted that the dataset is not spell checked and contains a lot of mistakes in grammar and spelling. A couple of the works have performed spell checking during the preprocessing with motivation to have more consistency in the dataset. However, the other works seem to have the general motivation that the way the word has been spelled might provide some information regarding the emotional setup of the writer i.e it may have been spelled incorrectly or exaggerated due to anger or sadness, and therefore most of the works have not performed the spell checking.

In regards to the Models being used, we find that the dataset is not big enough to allow for a language model or even a deep learning model to be useful and therefore most works rely on the machine learning based models for the classification. This is a problem since various deep learning and language models could have been used to effectively perform the classification part but the smaller size of the dataset prevents that. Either increasing the data or reducing the number of classes can allow us to use the deep learning based and language based models for the classification.

Few of the recent works have made us of deep learning models, and a significant improvement can be seen.

Table 2.1: Comparison of different works based on dataset used, processing and feature selection methods used, and models used

Paper	**Dataset**	**Pre-Processing + Feature Selection Method**	**Models**
Yang et al.	I2B2	POS Tagging + Word Features + Context Features + Syntactic Features + Semantic Features	CRF + NB, ME, SVM + Vote - based merging strategies + Post Preprocessing for instructions class
Desmet and Hoste	I2B2	Spelling Correction + Lemmatization + POS Tagging + N-gram Features + Subjectivity Features	SVM + Bootstrap Resampling
Wang et al.	I2B2	N-gram Features + Syntax Based Features + Context Based Features + Class Based Features + Pattern Based Features	SVM + RBF
Luyckx et al.	I2B2	Reannotation + N-gram Features + Lexicon Based Features + Context Based Features + Lexical Based Features	SVM + Bootstrap Resampling
Sohn et al.	I2B2	NER+ POS Tagging + Reannotation + Token Normalization	Multinomial NB + RBF
Xu et al.	I2B2	Data Augmentation + N-Gram Features + Bag-of-N-Gram Features + Pattern Based Features	Linear SVM + Pattern Matching
Cherry et al.	I2B2	Sentence Based Features + Bias Feature + Character Based Features + Document Level Based Features	SVM
McCart et al.	I2B2	Spelling Correction + Reannotation + Stemming + Sentence Based Features	RBF + DT,SVM,KNN + Voting Scheme
Ghosh et al.	CEASE	GloVe	LSTM
Ghosh et al.	CEASE	GloVe	Bi-GRU Encoding Layer + Attention Layer + External Knowledge Module
Beniwal and Dhobal	SIS	Tokenization + Zero Padding + One Hot Encoding	Bi-LSTM
Bansal and Beniwal	SIS	Tokenization + Zero Padding + Label Encoding	GPT + Bi-LSTM + CNN

Table 2.2: Result comparison of different papers based on Precision, Recall and F1-Score value

Paper	**Precision**	**Recall**	**F1-Score**
Yang et al.	-	-	61.39%
Desmet and Hoste	-	-	53.87%
Wang et al.	-	-	50.38%
Luyckx et al.	67.32%	34.42%	45.36%
Sohn et al.	57.09%	55.74%	56.40%
Xu et al.	56%	62%	59%
Cherry et al.	55.72%	54.72%	55.22%
McCart et al.	49.92%	50.55%	50.23%
Ghosh et al.	59%	60%	59%
Ghosh et al.	-	-	51.90%
Beniwal and Dhobal	73%	72%	72%
Bansal and Beniwal	77%	78%	77%

Chapter 3

METHODOLOGY

3.1 Data Acquisition

I have used the dataset built by Dobhal and Beniwal (https://github.com/armaan99/Suicidal-Intent-Sentiment-Dataset) where they have collected 248 suicide notes from various sources, including Reddit, Quora, Storypick, and several blogs from Info-media sites. These notes were meticulously tokenized into sentences, resulting in 2632 individual sentences. Each sentence was manually annotated and categorized into one of the six sentiment categories: 'happy', 'sad', 'neutral', 'love', 'hate', and 'proud'. Some randomly chosen samples from the dataset are illustrated in Table 3.1 along with their sentiment categories. Additionally, Fig. 3.1 and Fig. 3.2 depict the frequency and percentage distribution of sentences across each category respectively, providing a clear overview of the dataset's composition.

Table 3.1: Random Samples from Dataset

Text	**Label**
I will be so happy in heaven	happy
I've had a good life	proud
Please don't ever forget me	sad
You clown police	hate
Keep me in your hearts	love
Someday you'll be proud of me	proud
Thanks for making my life special	happy
Let us for a moment be sensible	neutral
I hate the way things turned out	hate
I have done everything in my power	neutral
It is terrifying	sad
I love you all	love

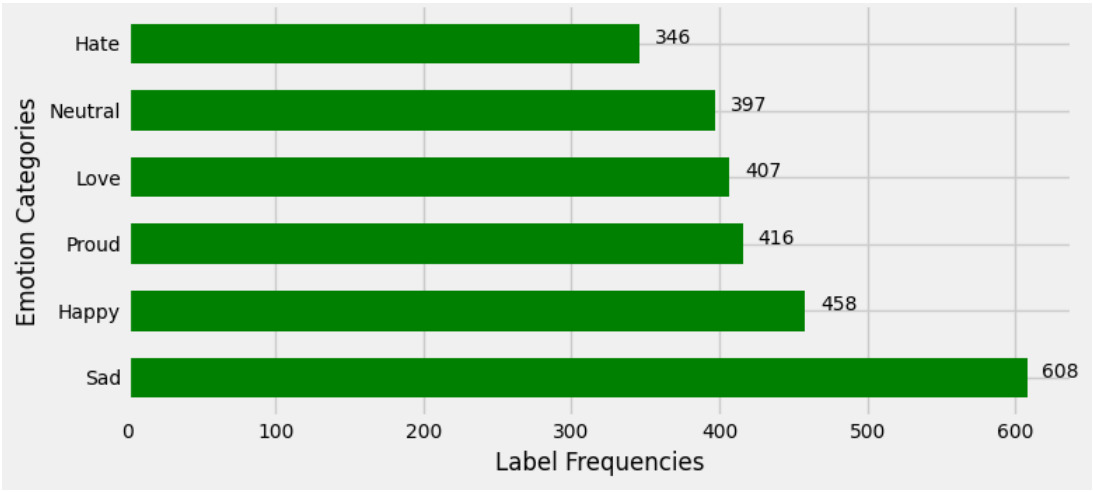

Figure 3.1: Frequency distribution of Sentences across each Category

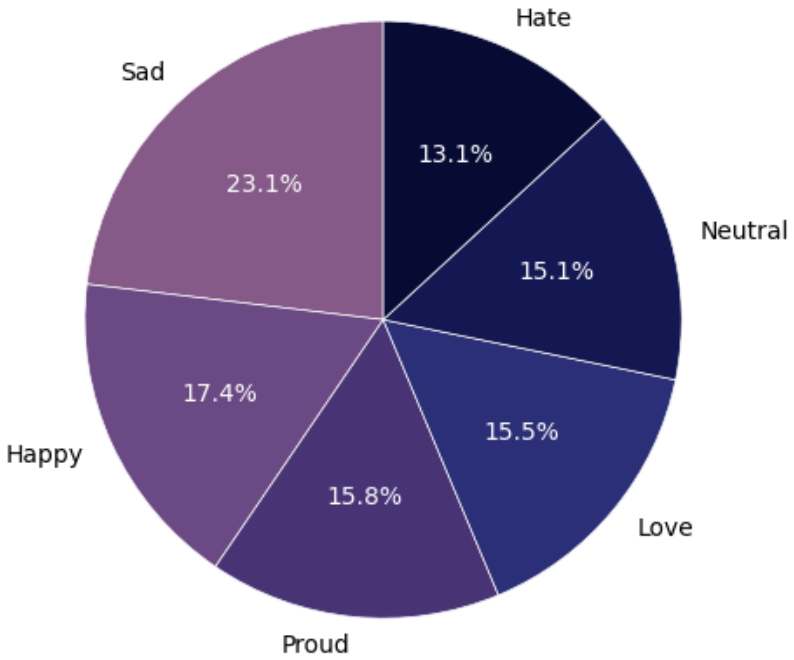

Figure 3.2: Percentage distribution of Sentences across each Category

3.2 Data Pre-Processing

In this study, the data preprocessing pipeline involves several key stages, ensuring that the source data gets transformed into a suitable format for input into the classification model.

Initially, the dataset is divided into features and target variables. The features comprise the first column of the dataset, which includes the sentences from the suicide notes. The target variable comprises the class labels in the second column, representing the sentiment categories: 'happy', 'sad', 'neutral', 'love', 'hate', and 'proud'.

At first, I have used a Label Encoder to encode the classified target labels into a numerical format to streamline the procedure, where a distinct integer will be assigned to each sentiment type.

Also, the textual data needed to be transformed into numerical vectors because this is an NLP task. To do this, I have used the BERT tokenizer. Sentences will get tokenized into words and sub-words by the BERT tokenizer, which will capture the subtleties of the text's context. I have provided the tokenized sequences zero padding to ensure each sentence is the same length after tokenization. This is necessary to ensure that the dimension of input being fed to the model remains the same as varying input length might interfere with the model's training.

I have divided the dataset in the ratio of 80:20 into the training and the test sets. This split gives 2105 records for training the model and 527 remaining records will be used for the model's evaluation.

3.3 Model Architecture

I have used a series of advanced methodologies like feature extraction, sequence modeling, and attention mechanism altogether, to get benefitted from the combined strengths of individual components. These techniques become the three primary components of the model architecture, playing an important role in transforming raw text into meaningful representations. Using these meaningful outcomes, the model can make much more accurate predictions.

The first phase begins with feature extraction. This is done using a pre-trained BERT model, freezing all its layers to save computational time by preventing it from getting updated at training time. This allows us to leverage the knowledge of pre-trained BERT effectively. BERT will help us collect contextual information from input sequences which is essential for understanding the nuances of language. Before feature extraction, the BERT Tokenizer will first tokenize the input sequences. Using truncation and padding techniques, the BERT tokenizer assures that the text is tokenized into a fixed-length format. This is an essential step in managing different sentence lengths. BERT is then applied to the tokenized text, to extract rich feature representations from these sequences. This generates contextual embeddings for every token.

In the second phase of sequence modeling, I have used a Bi-LSTM network, chosen for its ability to capture long-term dependencies in the input sequences, which works by

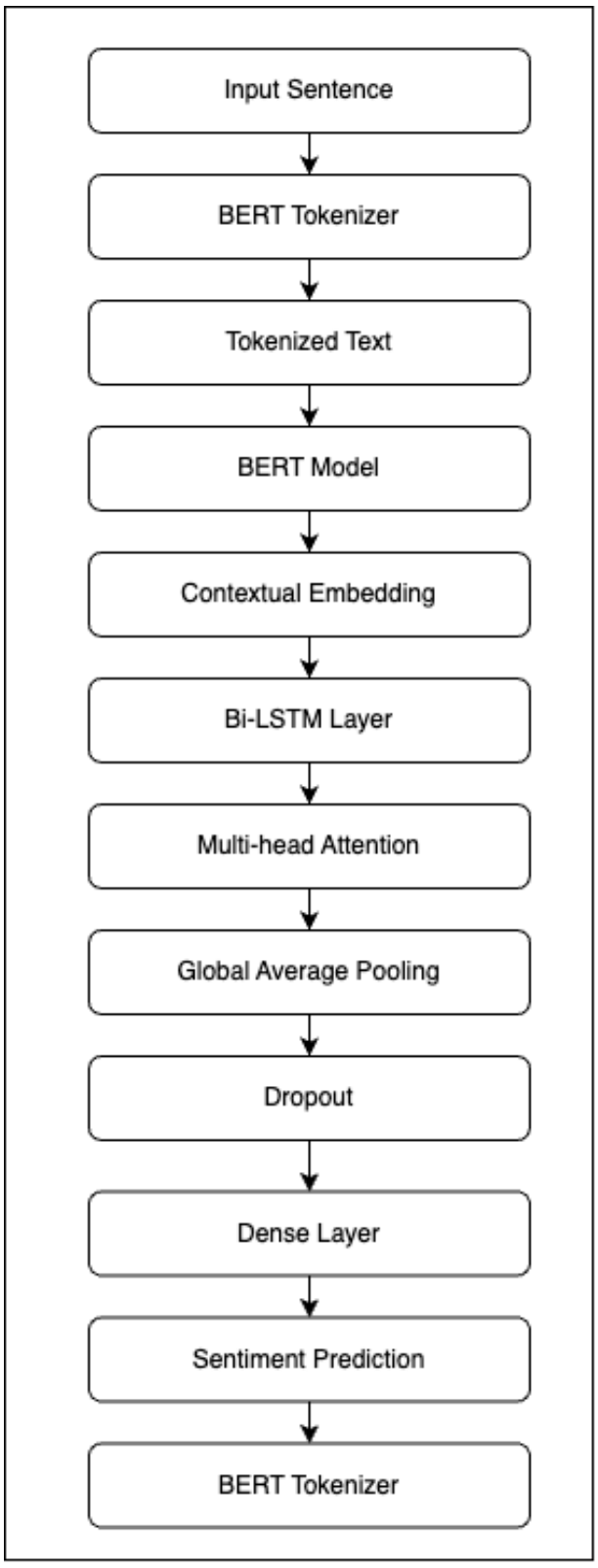

Figure 3.3: Data flow through various layers of the model architecture

Table 3.2: A comprehensive overview of the hyper-parameters used in different layers of our model architecture

Component	**Parameter**	**Value**
Data Split	Training Set	80% (2105 records)
Data Split	Testing Set	20% (527 records)
BERT Tokenizer	Max Length	60
BERT Model	Layers Frozen	All layers
Bi-LSTM	Units	64
Bi-LSTM	Return Sequences	True
Multi-Head Attention	Number of Heads	4
Multi-Head Attention	Key Dimension	64
Dropout	Rate	0.5
Dense Layer	Activation Function	Softmax
Model Compilation	Loss Function	Sparse Categorical Crossentropy
Model Compilation	Optimizer	Adam
Adam Optimizer	Learning Rate	0.001
Training Configuration	Batch Size	32
Training Configuration	Number of Epochs	20

processing the text in both forward and backward directions. This bidirectional approach ensures that the model learns the context completely. The output of BERT is fed as an input into the Bi-LSTM network. This network captures the sequential information from the text and generates a hidden state that gets passed on to the attention mechanism.

To increase the model's ability to classify emotions correctly, I have added a Multi-Head Attention mechanism, that allows the model to concentrate more on the significant segments of the text. This will make the model prioritize essential phrases in the sequences that play a key role in determining the sentiment associated with them. In the multi-head attention mechanism, the attention heads capture different aspects of input sequences, giving the model multiple perspectives on the data. This enhances the model's capacity to identify minute sentiment distinctions.

The above-described elements are combined to create the final model architecture. The input layer will receive the tokenized sequences, which are then processed through the Bi-LSTM network and Multi-Head Attention layer. Output received from the attention mechanism is now pooled using a Global Average Pooling (GAP) layer which aggregates the data throughout the sequence. Following this, we will use a Dropout layer which helps prevent over-fitting. Lastly, is a Dense layer, and the activation function used is Softmax, to give the final sentiment classification.

Chapter 4

RESULTS and DISCUSSION

I have trained the model with Python programming language which has good library support for NLP tasks, and the entire development and experimentation were performed on Google Colab because of its powerful and collaborative environment.

The dataset is divided into a ratio of 80:20 in training and test sets. This split gave 2105 records for training the model. The left-over 527 records were used to evaluate the model's performance. This stratified approach ensures a representative distribution of sentiment categories across both sets, enhancing the reliability of our evaluation.

To compile the model, we will employ the 'Categorical Crossentropy' loss function, which is appropriate for multi-class classification problems. The optimization process was carried out using the 'Adam' optimizer with a 0.001 learning rate, striking a balance between convergence speed and stability. Fig. 4.1 represents the Confusion Matrix, illustrating the detailed breakdown of correct and incorrect classifications.

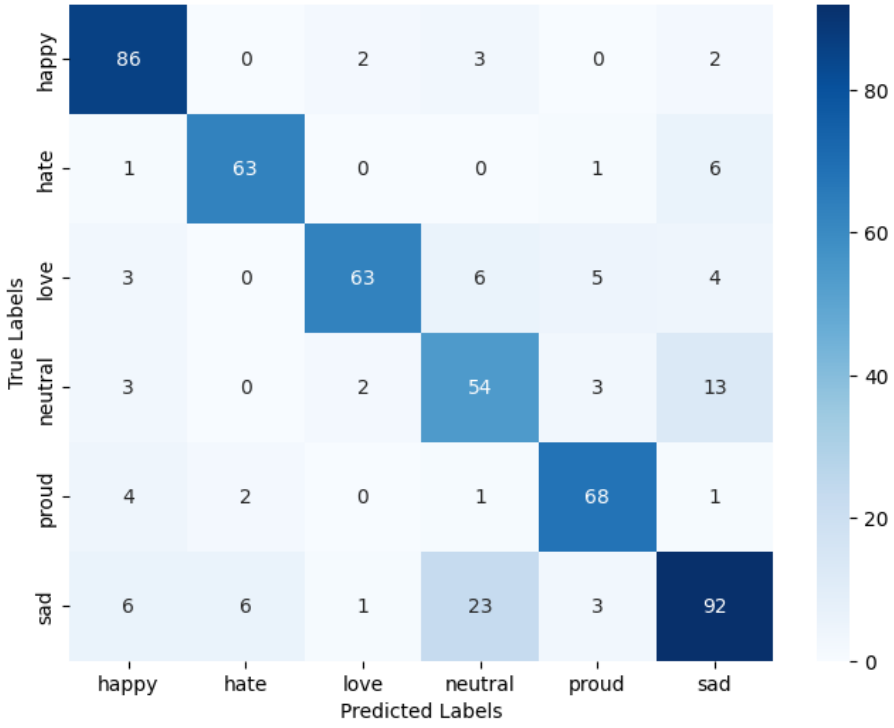

Figure 4.1: Confusion Matrix - Testing Data

Also as depicted in the classification report, Fig. 4.2, the model I have trained had achieved an impressive accuracy, recall, precision, and F1-score all at 81%. These results underscore the effectiveness of my approach in accurately categorizing the sentiments in the suicide notes.

Classification Report - Testing Data:

	precision	recall	f1-score	support
happy	0.83	0.92	0.88	93
hate	0.89	0.89	0.89	71
love	0.93	0.78	0.85	81
neutral	0.62	0.72	0.67	75
proud	0.85	0.89	0.87	76
sad	0.78	0.70	0.74	131
accuracy			0.81	527
macro avg	0.82	0.82	0.81	527
weighted avg	0.81	0.81	0.81	527

Figure 4.2: Classification Report of Predicted Labels

When compared to the previous work of Dhobal and Beniwal, Table 4.1, who reported the value of accuracy, recall, precision, and F1-score as 72%, 72%, 73%, and 72% respectively, and also with the work of Bansal and Beniwal, where they have reported the value for recall, precision, and F1-score as 77%, 78% and 77% respectively, our model demonstrates a significant improvement. The enhancement can be attributed to the hybrid approach I have used, which integrates advanced techniques for tokenization, sequence modeling, and attention mechanisms to better focus on relevant text parts.

Table 4.1: Comparison of Bi-LSTM with BERT + Bi-LSTM + Multi-head Attention

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Bi-LSTM	72%	73%	72%	72%
GPT + Bi-LSTM + CNN	78%	77%	78%	77%
BERT + Bi-LSTM + Multi-head Attention	81%	81%	81%	81%

In summary, the model I have trained not only outperforms existing benchmarks but also showcases the potential of hybrid approaches in enhancing the reliability and accuracy of sentiment analysis in sensitive contexts such as suicide notes.

Chapter 5

CONCLUSION AND FUTURE SCOPE

The current work shows how well modern NLP techniques may be used to analyze suicide letters' sentiments. Our model produced notable gains in the precision, recall, and F1-score over earlier work by using BERT for feature extraction, Bi-LSTM for sequence modelling, and a Multi-head Attention mechanism to enhance concentration on relevant regions of the input sequence. In particular, our model resulted in a recall, precision, and f1-score all at 81%, significantly outperforming the performance metrics published in previous studies.

However, future investigation can be done to further amplify the applicability and robustness of sentiment classification in this context. This can be done by increasing the dataset size by collecting a diverse range of suicide notes from different cultural and linguistic backgrounds that can help enhance the model's generalizability. One can explore other deep learning architectures or integrate multiple architectures like Convolutional Neural Networks (CNNS), and other variants of Transformers. One can also tailor a more sophisticated attention mechanism to make a more refined extraction of details hidden in the suicide notes.

To conclude our study that contributes to the expanding area of research on sentiment classification in sensitive contexts, future work might promise to bridge the gap between technological advancements and their utilization in mental health, which may help provide better support to an individual at risk.

Bibliography

[1] Flayeh, Azhar Kassem, Yaser Issam Hamodi, and Nashwan Dheyaa Zaki. "Text Analysis Based on Natural Language Processing (NLP)." 2022 2nd International Conference on Advances in Engineering Science and Technology (AEST). IEEE, 2022.

[2] Chen, Linjie, et al. "A Method for Extracting Information from Long Documents that Combines Large Language Models with Natural Language Understanding Techniques." 2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD + AI). IEEE, 2023.

[3] Yadollahi, A., Shahraki, A.G. and Zaiane, O.R., 2017. Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 50(2), pp.1-33.

[4] Yadav, Payal, and Dhatri Pandya. "SentiReview: Sentiment analysis based on text and emoticons." 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 2017.

[5] Gopinath, A., et al. "Perceptual Based Sentiment Analysis of Consumer Reviews using Rational Machine Learning Techniques for E-Commerce Applications." 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT). IEEE, 2023.

[6] Desmet, Bart, and Véronique Hoste. "Emotion detection in suicide notes." Expert Systems with Applications 40.16 (2013): 6351-6358.

[7] Mishra, Ajey Shakti, et al. "Cyberbullying and Suicide Ideation Detection via Hybrid Machine Learning Model." 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2023.

[8] Sakthi, U., Thomas M. Chen, and Mithileysh Sathiyanarayanan. "CyberHelp: Sentiment Analysis on Social Media Data Using Deep Belief Network to Predict Suicidal Ideation of Students." 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). IEEE, 2023.

[9] Katoch, Sapna, Balwinder Kaur Dhaliwal, and Gurpreet Singh. "Deep Learning and Natural Language Processing-Based Model for the Prediction of Suicidal Ideation in Military Personnel." 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2023.

[10] Lim, Yan Qian, Ming Jie Lee, and Yim Ling Loo. "Towards a machine learning framework for suicide ideation detection in Twitter." 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS). IEEE, 2022.

[11] Rappai, Sherin, and Gobi Ramasamy. "Computational Methods to Predict Suicide Ideation among Adolescents." 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, 2022.

[12] Balaji, Penubaka, O. Nagaraju, and D. Haritha. "Levels of sentiment analysis and its challenges: A literature review." 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). IEEE, 2017.

[13] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[14] Hu, Jingxuan. "Sentiment Classification Model of Online Reviews Based on Word Features and Bi-LSTM." 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA). IEEE, 2022.

[15] Fang, Fang, et al. "Text Classification Model Based on Multi-head self-attention mechanism and BiGRU." 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS). IEEE, 2021.

[16] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[17] Yang, Hui, et al. "A hybrid model for automatic emotion recognition in suicide notes." Biomedical informatics insights 5 (2012): BII-S8948.

[18] Desmet, Bart, and Véronique Hoste. "Emotion detection in suicide notes." Expert Systems with Applications 40.16 (2013): 6351-6358.

[19] Wang, Wenbo, et al. "Discovering fine-grained sentiment in suicide notes." Biomedical informatics insights 5 (2012): BII-S8963.

[20] Luyckx, Kim, et al. "Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification." Biomedical informatics insights 5 (2012): BII-S8966.

[21] Sohn, Sunghwan, et al. "A hybrid approach to sentiment sentence classification in suicide notes." Biomedical informatics insights 5 (2012): BII-S8961.

[22] Xu, Yan, et al. "Suicide note sentiment classification: a supervised approach augmented by web data." Biomedical informatics insights 5 (2012): BII-S8956.

[23] Cherry, Colin, Saif M. Mohammad, and Berry De Bruijn. "Binary classifiers and latent sequence models for emotion detection in suicide notes." Biomedical informatics insights 5 (2012): BII-S8933.

[24] McCart, James A., et al. "Using ensemble models to classify the sentiment expressed in suicide notes." Biomedical informatics insights 5 (2012): BII-S8931.

[25] Ghosh, Soumitra, Asif Ekbal, and Pushpak Bhattacharyya. "Cease, a corpus of emotion annotated suicide notes in English." Proceedings of the twelfth language resources and evaluation conference. 2020.

[26] Ghosh, Soumitra, Asif Ekbal, and Pushpak Bhattacharyya. "A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes." Cognitive Computation 14.1 (2022): 110-129.

[27] Beniwal, Rohit, and Abhishek Dobhal. "Sentiment Classification on Suicide Notes Using Bi-LSTM Model." International Conference on Data Analytics & Management. Singapore: Springer Nature Singapore, 2023.

[28] Bansal, A. and Beniwal, R., 2024, July. Sentiment Classification on Suicide Notes Using GPT, Bi-LSTM and CNN. In 2024 Asia Pacific Conference on Innovation in Technology (APCIT) (pp. 1-5). IEEE.

List of Publications

Conferences

1. A. Armaan and B. Rohit (2025). "Sentiment Classification on Suicide Notes using BERT, Bi-LSTM, and Multi-head Attention", *In international Conference on AI and Robotics (AIR). Springer.*
2. A. Armaan and B. Rohit (2025). "Sentiment Classification Approaches on Suicide Notes: A Review" [Communicated].

NAZARBAYEV UNIVERSITY

International Conference on AI and Robotics (AIR) 2025

organized by

The Center of Excellence in Medical Robotics and Research, Nazarbayev University Kazakhstan

CERTIFICATE OF PRESENTATION

This is to certify that

Armaan Agrawal

has presented the paper titled **"Sentiment Classification on Suicide Notes using BERT, Bi-LSTM, and Multi-head Attention"** authored by **Armaan Agrawal, Rohit Beniwal** in the ***International Conference on AI and Robotics (AIR) 2025*** held at The Center of Excellence in Medical Robotics and Research, Nazarbayev University Kazakhstan during **May 09-11, 2025**.

Prof. Prashant Jambwal
General Chair

Prof. Shahid Hussain
General Chair

Springer

SCRS\AIR2025\PC\319

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis_____

_____

Total Pages _____ Name of the Scholar_____

Supervisor (s)

(1)_____

(2)_____

(3)_____

Department_____

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: _____ Similarity Index: _____, Total Word Count: _____

Date: _____

Candidate's Signature **Signature of Supervisor(s)**

Armaan Agrawal Rohit Beniwal

Sentiment Classification on Suicide Notes.pdf

Delhi Technological University

Document Details

Submission ID
trn:oid:::27535:100979820

Submission Date
Jun 15, 2025, 4:00 PM GMT+5:30

Download Date
Jun 15, 2025, 4:03 PM GMT+5:30

File Name
Sentiment Classification on Suicide Notes.pdf

File Size
321.3 KB

22 Pages

6,151 Words

33,840 Characters

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

48 Not Cited or Quoted 9%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
6% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **48** Not Cited or Quoted 9%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
6% Publications
4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	"Proceedings of Data Analytics and Management", Springer Science and Business... 2%
2	Internet	dokumen.pub <1%
3	Internet	medinform.jmir.org <1%
4	Internet	www.mdpi.com <1%
5	Internet	journals.sagepub.com <1%
6	Publication	Jaiteg Singh, S B Goyal, Rajesh Kumar Kaushal, Naveen Kumar, Sukhjit Singh Sehr... <1%
7	Internet	arxiv.org <1%
8	Publication	Bogale, Amen Woldesenbet. "An Anomaly-Based Machine Learning Approach for ... <1%
9	Internet	academic-accelerator.com <1%
10	Internet	rave.ohiolink.edu <1%

11	Submitted works	
VIT University on 2025-03-19		<1%
12	Internet	
annals-csis.org		<1%
13	Internet	
bdbanalytics.ir		<1%
14	Internet	
digibuo.uniovi.es		<1%
15	Internet	
irjet.net		<1%
16	Publication	
Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artific...		<1%
17	Publication	
Petrus-Nihi, Eluwumi. "Vector Field Embedded Into Images, as Well as Into Gradie...		<1%
18	Publication	
Shalli Rani, Ayush Dogra, Ashu Taneja. "Smart Computing and Communication fo...		<1%
19	Submitted works	
University of Essex on 2023-08-25		<1%
20	Submitted works	
University of Wolverhampton on 2025-05-19		<1%
21	Publication	
Xinyu Cao, Hou Liangwen, Haitao Wang, Lei Liu. "Microblog-oriented Multi-scale ...		<1%
22	Publication	
Yingying Ni, Wei Ni. "A multi-label text sentiment analysis model based on sentim...		<1%
23	Internet	
philpapers.org		<1%
24	Internet	
www.coursehero.com		<1%

25 Internet

www.politesi.polimi.it <1%

26 Publication

Jingxuan Hu. "Sentiment Classification Model of Online Reviews Based on Word F... <1%

27 Submitted works

Liverpool John Moores University on 2022-09-05 <1%

28 Submitted works

Liverpool John Moores University on 2025-02-27 <1%

29 Submitted works

University of Birmingham on 2015-12-22 <1%

30 Submitted works

University of Lancaster on 2022-09-14 <1%

31 Publication

Yong Ren, Jinfeng Han, Yingcheng Lin, Xiujiu Mei, Ling Zhang. "An Ontology-Base... <1%

32 Internet

encyclopedia.pub <1%

33 Internet

ieiespc.org <1%

34 Internet

pmc.ncbi.nlm.nih.gov <1%

35 Internet

spectrum.library.concordia.ca <1%

36 Internet

theses.lib.polyu.edu.hk <1%

Armaan Agrawal Rohit Beniwal

Sentiment Classification on Suicide Notes.pdf

Delhi Technological University

Document Details

Submission ID
trn:oid:::27535:100979820

Submission Date
Jun 15, 2025, 4:00 PM GMT+5:30

Download Date
Jun 15, 2025, 4:03 PM GMT+5:30

File Name
Sentiment Classification on Suicide Notes.pdf

File Size
321.3 KB

22 Pages

6,151 Words

33,840 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.