

A comparative analysis of transformer-based architectures for enhancing suicidal ideation and detection using pre-trained language model

Project & Thesis-II

CSE 4250

A thesis Report

Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Submitted by

Mainul Islam	190104016
Samia Afrin	190104132
Md Kamrul Hasan	190104144
Abdullah Al Mahmud	190104146

Supervised by

Dr.Taslim Taher



Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

November 21,2023

CANDIDATES' DECLARATION

Enrolled in the Bachelor of Science in Computer Science and Engineering program at Ahsanullah University of Science and Technology, Dhaka, Bangladesh, hereby declare that the thesis entitled "**A Comparative Analysis of Transformer-Based Architectures for Enhancing Suicidal Ideation and Detection using Pre-trained Language Model**" presented in this report is the result of our investigation conducted under the supervision of Dr. Taslim Taher from the Department of Computer Science and Engineering. This work spanned over two final-year courses, CSE4100: Project and Thesis I, and CSE4250: Project and Thesis II, in accordance with the curriculum of the Department.

We further declare that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma, or other qualifications.

Mainul Islam
190104016

Samia Afrin
1901040132

Md Kamrul Hasan
190104144

Abdullah Al Mahmud
190104146

CERTIFICATION

This is to certify that the thesis entitled "A Comparative Analysis of Transformer-Based Architectures for Enhancing Suicidal Ideation and Detection using Pre-trained Language Model," presented by us in this report, is the culmination of our investigation conducted under the supervision of Dr. Taslim Taher from the Department of Computer Science and Engineering at Ahsanullah University of Science and Technology.

We affirm that the work presented in this thesis is original and has not been submitted elsewhere for the award of any degree.

Group Members

Mainul Islam	190104016
Samia Afrin	190104132
Md Kamrul Hasan	190104144
Abdullah Al Mahmud	190104146

Dr. Talim Taher
Assistant Professor & Supervisor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dr. Md. Shahriar Mahbub
Professor & Department Head
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

ACKNOWLEDGEMENT

At the outset, we extend our heartfelt gratitude to the Almighty for bestowing upon us the fortitude and resolve to accomplish this research endeavor.

Our sincere appreciation goes to our esteemed supervisor, Dr. Taslim Taher, whose unwavering guidance and support have been instrumental in navigating the intricacies of this research project. His profound expertise and invaluable insights have significantly enriched our understanding, for which we are truly indebted.

In closing, we express our deep thanks to our families and friends whose unwavering love and support have been a constant source of motivation. Their encouragement has been pivotal in overcoming the challenges encountered during the course of this research.

This thesis is gratefully supported, in part, by Ahsanullah University of Science and Technology.

ABSTRACT

Suicide, a global public health challenge, necessitates effective detection of suicidal ideation for timely prevention. In the era of social media, individuals openly express thoughts, providing an opportunity for automated detection. This research leverages natural language processing and machine learning, utilizing a dataset from "Suicide Watch" and "depression" subreddits. Collected through the Push-shift API, the dataset spans from the inception of "Suicide Watch" to January 2, 2021. The proposed model comprises three layers: Data Collection, Embedding, and Classification. Various transformer models (BERT, ALBERT, RoBERTa, DistilBERT) demonstrate distinct performance characteristics. BERT and ALBERT exhibit superior accuracy (98%), while RoBERTa maintains commendable accuracy (95%). DistilBERT, with streamlined efficiency, achieves competitive results (95.51%). Future work focuses on Multimodal Enrichment, Performance Analysis, Temporal Suicidal Ideation Detection, and enhancing accuracy and speed. This research contributes to advancing suicide prevention through technology and machine learning.

Contents

<i>CANDIDATES' DECLARATION</i>	i
<i>CERTIFICATION</i>	ii
<i>ACKNOWLEDGEMENT</i>	iii
<i>ABSTRACT</i>	iv
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Overview	1
1.2 Motivation	1
1.3 Dataset Collection and Description:	2
1.3.1 Data Description	3
1.4 Problem Statement	3
2 Literature Review	5
2.1 A Transformer Based Approach To Detect Suicidal Ideation Using Pre- Trained Language Models	5
2.1.1 Paper Description	5
2.1.2 Model & Accuracy	6
2.2 Detection of Suicide Ideation in Social Media Forums Using Deep Learning	6
2.2.1 Paper Description	6
2.2.2 Model & Accuracy	7
2.3 Detection of Depression and Suicide Risk Based on Text From Clinical Interviews Using Machine Learning	7
2.3.1 Paper Description	8
2.3.2 Model & Accuracy	8
2.4 Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications	8
2.4.1 Paper Description	9

2.5	A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning	9
2.5.1	Paper Description	10
2.5.2	Model & Accuracy	10
3	Background Studies	11
3.1	Transformers	11
3.1.1	Transformer Architecture:	12
3.2	Text Pretrained models	14
3.2.1	BERT	14
3.2.2	RoBERTa	16
3.2.3	ALBERT	17
3.2.4	DistilBERT	18
3.3	Text embedding models	19
3.3.1	GloVe	19
4	Work Process	21
4.1	Research Design	21
4.1.1	Data Layer	23
4.1.2	Embedding Layer	23
4.1.3	Classification Layer	23
4.2	Data Acquisition	24
4.3	Data Preprocessing	24
4.3.1	Eliminating Extra Spaces and Lower casing	25
4.3.2	Text Refinement	25
4.3.3	Removing Duplicate data and Null values	26
4.3.4	Removing Stop Words and Lemmatization	27
4.4	Data Visualization	29
4.4.1	Dataset Balancing	29
4.4.2	Average Word length in each text	30
4.4.3	Data Unigram Analysis	31
4.4.4	Data bigram analysis:	31
4.4.5	Data Trigram Analysis:	32
4.4.6	Stop Word and Punctuation Analysis	32
4.4.7	Textual Structure Analysis	33
4.4.8	Sentential Diversity Analysis	34
4.5	Label Encoding	35
4.6	Test Model Training	35
4.6.1	BERT	35
4.6.2	ALBERT	36

4.6.3	ROBERTA	37
4.6.4	DistilBERT	37
4.6.5	GloVe Embeddings	38
5	Result and Performance Analysis	39
6	Contribution & Research Gap	41
6.1	Contribution	41
6.2	Research Gap :	41
6.2.1	Language-Specific Limitation	41
6.2.2	Research Gap and Need for Inclusivity	42
6.2.3	Potential for Multilingual Replication	42
6.2.4	Consideration of Socioeconomic Factors	42
6.2.5	Exploration of Different Transformer Models	42
6.2.6	Incorporation of Emojis for Enhanced Understanding	42
7	Conclusion and Future work	44
7.1	Future work	44
7.2	Conclusion	45
	References	46

List of Figures

1.1	Dataset shape	3
1.2	Sample of data	3
3.1	The architecture of a transformer model	13
3.2	The architecture of The Masked Language Modeling	15
3.3	Architecture of Next Sentence Prediction.	16
3.4	The Architecture of Roberta model.	17
3.5	The Architecture of Albert model.	18
3.6	The Architecture of DistilBERT model.	19
3.7	The Architecture of GloVe model.	20
4.1	Proposed Model Diagram	22
4.2	Eliminating Extra Spaces and Lower casing	25
4.3	Text Refinement: Emojis, Links, HTML, and Punctuation Removal.	26
4.4	Removing Duplicate Data and Null Values	27
4.5	Removing Stop Words and Lemmatization	28
4.6	Data before cleaning	28
4.7	Data after cleaning	29
4.8	Dataset Balance	30
4.9	Average Word length in each text	30
4.10	Data Unigram Analysis	31
4.11	Data Bigram Analysis	31
4.12	Data Trigram analysis	32
4.13	Stop Word and Punctuation Analysis	33
4.14	Textual Structure Analysis	34
4.15	Sentential Diversity Analysis	34
4.16	Label Encoding	35

List of Tables

2.1	Model Performance	6
2.2	Model Performance	7
2.3	Model Performance	8
2.4	Model Performance	10
4.1	Classification Report	36
4.2	Classification Report	37
4.3	Classification Report	37
4.4	Classification Report	37
4.5	Classification Report	38

Chapter 1

Introduction

1.1 Overview

Suicide continues to devastate countless lives, with statistics showing roughly 80,000 lost annually in the U.S. alone. [1]. While prevention efforts are widespread, pinpointing at-risk individuals presents a considerable hurdle. This paper investigates how state-of-the-art natural language processing can significantly aid in such detection via social media. We propose leveraging NLP's strength to recognize worrying language, sentiments, and patterns within posts that may indicate suicidal thoughts or plans. Shorter, simpler sentences will be interspersed with more complex constructions and varied phrasing to explore these objectives. Specifically, this study aims to create a model proficient at automatically filtering online communications for linguistic signs of endangerment, serving as an early warning system for intervention.

While previous efforts in suicide prevention have proven valuable, a divide remains in fully understanding how technical advances may aid those at risk. This work hopes to illuminate the ways natural language processing could support more proactive identification of suicidal ideation, recognizing subtle language patterns in a compassionate manner. By exploring the capability of models to contribute to this important domain, this research aims to enhance current initiatives to address suicide and safeguard human life, thereby benefitting individuals and communities worldwide.

1.2 Motivation

Suicide is a contributor to mortality rates with the World Health Organization reporting around one million deaths annually from this cause. This underscores the importance of

implementing suicide prevention measures, which can be greatly supported by advancements, in technology. Suicide is a cause of death, on a scale. The World Health Organization reports that around one million individuals lose their lives to suicide annually. This emphasizes the importance of implementing strategies for preventing suicide, which can be greatly supported by advancements in technology.

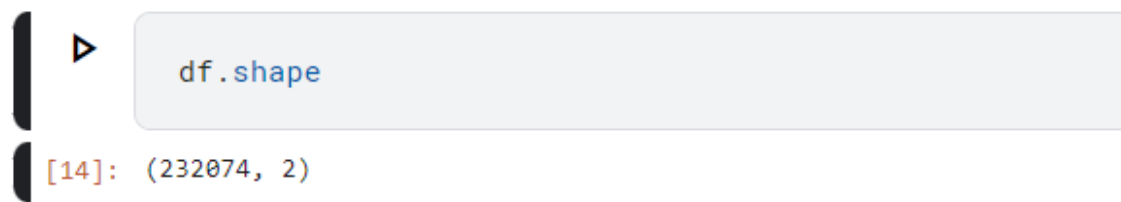
Henceforth, forging an automated contrivance that can precisely perceive and categorize self-destructive contemplation from communal media statistics harbors the potential to safeguard lives and foster psychological well-being. By employing sophisticated apparatuses of machine learning like transformers, scholars can construct potent frameworks that can aid in discerning endangered persons and furnishing them with the requisite backing and means. This constitutes the impetus driving our inquiry into the detection of self-destructive ideation, as we endeavor to make a valuable contribution towards the formulation of more efficacious schemes for averting suicides.

1.3 Dataset Collection and Description:

The dataset employed in this research is a publicly available collection of text data extracted from the "Suicide Watch" and "depression" subreddits on the Reddit platform. The data was acquired using the Push-shift API, a reliable tool for gathering social media content. For "Suicide Watch," posts spanning from December 16, 2008 (its inception) to January 2, 2021, were meticulously collected. In the case of "depression" posts, data was captured from January 1, 2009, to January 2, 2021 www.kaggle.com.

This dataset encompasses a substantial volume of content, containing 233,372 rows and 2 columns. The two columns likely represent text data and associated metadata, which may include information like timestamps and user details. Notably, this dataset has been previously utilized in various research studies centered around suicide prevention and intervention. Its extensive coverage, particularly in the "Suicide Watch" subreddit, offers valuable insights into the linguistic expressions and themes related to mental health and suicide on online platforms.

The utilization of this dataset is founded on its relevance to the research objectives, as it encapsulates real-world conversations and discussions on these sensitive topics, making it a valuable resource for training and evaluating suicide risk detection models.



```
df.shape
```

```
[14]: (232074, 2)
```

Figure 1.1: Dataset shape

1.3.1 Data Description

The dataset consists of social media posts and the title of the posts [2]. For protecting the user's privacy the username was omitted from the dataset. The dataset only contains the "Post Title" which is the title of the post that was set by the user and The "Post Text" which is the body of the user post where the user describes their problems and thoughts. Users can give a title that has a limit of 300- characters. Most suicidal related posts can have a title that indicates suicidal thoughts or a cry for help in the title. For this reason, the title of the post is important for detecting suicidal related posts. A user can express their thoughts properly in the post section as there is no character limit in a post for the user The dataset consists of two attributes.

1. **Text:** This attribute tells us the thought of a suicidal person. Generally, it indicates whether the person is willing to suicide or not.
2. **Suicidal ideation:** It indicates that the respective text is suicidal or non-suicidal.

	text	suicide_ideation
0	ex wife threatening suicidercently left wife ...	suicide
1	weird dont get affected compliments coming som...	non-suicide
2	finally 2020 almost never hear 2020 bad year e...	non-suicide
3	need helpjust help im crying hard	suicide
4	im losthello name adam 16 ive struggling years...	suicide

Figure 1.2: Sample of data

1.4 Problem Statement

The identification and prevention of suicidal ideation represent a pressing concern within the realm of mental healthcare. With the prevalence of social media platforms

as a communication medium, there emerges a potential avenue for early detection and intervention in cases of mental health crises. However, the overwhelming volume of user-generated content on these platforms renders manual monitoring an unfeasible task. This scenario underscores the increasing demand for automated systems that possess the capability to accurately discern individuals who may be exhibiting signs of suicidal ideation.

Recent advancements in the field of natural language processing have showcased the superiority of transformer-based architectures, including well-known models like BERT and its variants, when compared to traditional deep learning approaches. While these transformer-based models have exhibited remarkable performance in various NLP tasks, their application in the context of suicidal ideation detection remains largely uncharted.

Hence, this research confronts the challenge of conducting a comprehensive comparative analysis of transformer-based architectures to ascertain which specific approach yields the most effective results in enhancing the detection of suicidal ideation within social media posts. By addressing this issue, we aim to contribute to the development of more accurate and efficient automated systems that can play a vital role in suicide prevention and mental health support.

Chapter 2

Literature Review

Literature review includes all of the main themes and subthemes found within the general topic chosen for the study. These themes and subthemes are usually interwoven with the methods or findings of the prior research. Also, a literature review sets the stage for and offers readers justifications for the purpose and methods of the original research being reported in a manuscript [3].

2.1 A Transformer Based Approach To Detect Suicidal Ideation Using Pre-Trained Language Models

[4] **Author:** Farsheed Haque,Ragib Un Nur,Shaeekh Al Jahan,Zarar Mahmud and Faisal Muhammad Shah.

2.1.1 Paper Description

In this paper, they investigate Suicidal Ideation detection from social media posts of users. they have used user posts from SuicideWatch.SuicideWatch is a subReddit where users anonymously post about their sufferings, traumatic incidents, or their fight with mental illness. These posts often have words or intentions that indicate suicidal thoughts in their head. they have taken these posts with full user privacy and used Natural Language Processing (NLP) to create a model that will detect Suicidal Ideation in these posts. this paper mainly focuses on the possibility of using Transformers, a state-of-the-art Deep Learning model to detect Suicidal Ideation. When it comes to NLP, handling sequential data is very important. And as transformers do not require processing sequential data in order, it performs better than other recurrent neural networks like Bi-LSTM. Since transformer models allow parallelism it made possible training

on larger datasets, and that is how the pre-trained systems like BERT and its variants ALBERT, Roberta, and XLNET were developed. Authors made a classification model that shows Transformer based models such as BERT, ALBERT, ROBERTa, and XLNET perform significantly better than old recurrence-based neural networks like Bi-LSTM in the area of sentiment analysis, suicidal ideation detection to be precise. Therefore, several algorithms they choose to serve the purpose such as

- (a) Bi-LSTM
- (b) BERT
- (c) BERT
- (d) ALBERT
- (e) ROBERTa
- (f) XLNET

2.1.2 Model & Accuracy

Table 2.1: Model Performance

Model name	Accuracy
Bi-LSTM	84.39
BERT	88.3
ALBERT	87.77
ROBERTa	95.21
XLNET	89.01

2.2 Detection of Suicide Ideation in Social Media Forums Using Deep Learning

[5] **Author:** Michael Mesfin Tadesse , Hongfei Lin , Bo Xu and Liang Yang.

2.2.1 Paper Description

This text describes a study that aimed to detect signs of suicide ideation in social media forums using deep learning methods. The study used different data representation techniques to reformulate text into a format that the system could recognize, and focused on the potential of a hybrid LSTM-CNN model. The study found that the hybrid model considerably improved the accuracy of text classification, and demonstrated the

strength and potential of CNN. However, the study acknowledges the limitations of data deficiency and annotation bias. The study concludes that it can contribute to future machine learning research for building an easily accessible and highly effective suicide detection and reporting system implemented in social media networks as an efficient intervention point between at-risk individuals and mental health services. Therefore, several algorithms they choose to serve the purpose such as .

- (a) Random forest
- (b) Support vector machine
- (c) Navies bayes
- (d) XGBOOST
- (e) LSTM-CNN

2.2.2 Model & Accuracy

Table 2.2: Model Performance

Model name	Accuracy
Random forest	85.6
Support vector machine	83.5
Naive Bayes	82.5
XGBOOST	88.3
LSTM-CNN	93.8

They perform their results in two main phases. begining by examining the data analysis results in the entire labeled corpus of Reddit posts.

- (a) First, they analyze the most frequent n-grams in suicide-indicative posts linked with suicidal intents, and compare them with the n-grams in non-suicidal posts.
- (b) Next, to measure the signs of suicidal thoughts, they use their proposed set of features and compare the performance of deep learning classifier with the baselines in terms of evaluation metrics.

2.3 Detection of Depression and Suicide Risk Based on Text From Clinical Interviews Using Machine Learning

[6] **Author:** Daun Shin, Kyungdo Kim, Seung-Bo Lee, Changwoo Lee, Ye Seul Bae, Won Ik Cho, Min Ji Kim, C. Hyung Keun Park, Eui Kyu Chie, Nam Soo Kim and Yong Min Ahn.

2.3.1 Paper Description

The paper is a research study that aimed to diagnose depression and predict suicide risk based on the words spoken by participants in a semi-structured interview. The study recruited 83 healthy and 83 depressed patients, and the recording was transcribed into text after only the words uttered by the participant were extracted. The study used machine learning techniques, specifically the Naive Bayes classifier algorithm, to analyze the text data and predict the risk of suicide among patients with depression. The study found that depression can be diagnosed through machine learning based on the Naive Bayes classifier technique, and the accuracy was confirmed by constructing an ensemble model that predicts the risk of suicide among patients with depression. The study concluded that participants' words during an interview show significant potential as an objective and diagnostic marker through machine learning for predicting suicide risk. The paper provides valuable insights into the potential of using machine learning techniques to diagnose depression and predict suicide risk based on the words spoken by participants in a semi-structured interview.

2.3.2 Model & Accuracy

Table 2.3: Model Performance

Model name	Accuracy
Demographic	0.542
Text	0.602
Ensemble	0.747
Sensitivity	0.816
Specificity	0.647
AUC	0.800

2.4 Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications

[7] **Author:** Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang

2.4.1 Paper Description

This paper offers a comprehensive overview of machine learning methods and their applications in the detection of suicidal ideation. It delves into various approaches, spanning clinical techniques, content analysis, and deep learning models, categorizing them based on their data sources and highlighting specific tasks and datasets within the field. The paper underscores the importance of integrating external knowledge and the potential of using machine learning to identify individuals at risk of suicide, enabling timely interventions. Various techniques, such as psychological lexicon dictionaries, Support Vector Machines (SVM), topic models, and deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Bidirectional Encoder Representations from Transformers (BERT), are explored. Additionally, the paper emphasizes the incorporation of affective attributes in suicide detection.

The tasks involved in suicidal ideation detection are outlined, covering suicide text categorization, interpretation of suicidal messages, identification of suicide attempters, response generation, and the identification of mental health disorders and self-harm risks. The paper highlights the use of common datasets from platforms like Reddit and Twitter for researching suicidal ideation.

Despite its comprehensive review, the paper lacks specific accuracy values for the discussed models and primarily focuses on reviewing machine learning methods and their applications in suicidal ideation detection. It emphasizes techniques, data sources, and research limitations, including data scarcity and annotation biases. The paper concludes by proposing future research directions, such as leveraging emerging learning techniques, enhancing intention understanding, considering temporal information, and enabling proactive conversational interventions. It also highlights the need to bridge the gap between clinical mental health detection and automated machine-based detection. In summary, this paper offers a valuable overview of the landscape of machine learning for detecting suicidal ideation, with insights into its potential and future research directions, but without detailed accuracy analyses.

2.5 A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning

[8] **Author:** Rezaul Haque, Naimul Islam, Maidul Islam and Md Manjurul Ahsan.

2.5.1 Paper Description

In this study, the researchers employed the Tweepy API to collect a substantial dataset of 49,178 real-time tweets related to suicidal ideation. To ensure the dataset’s quality, each tweet underwent meticulous annotation, categorizing them as either ‘suicidal’ or ‘non-suicidal’ after a comprehensive analysis. Given the informal and noisy language commonly found in tweets, the study applied Natural Language Toolkit (NLTK) packages for text pre-processing, significantly improving the clarity of the textual data. Feature extraction techniques, including CountVectorizer and word embedding, were specifically tailored to enhance detection accuracy for both Machine Learning (ML) and Deep Learning (DL) models.

The study went on to rigorously train various DL models, such as LSTM, BiLSTM, GRU, BiGRU, and C-LSTM, using Keras, a high-level API based on TensorFlow. These DL models were subjected to a thorough evaluation, considering a range of performance metrics, including accuracy, AUC, precision, recall, and F1-score. Furthermore, the research included a comprehensive comparative analysis, systematically contrasting the performance of DL models with traditional ML methods, including Random Forest, Support Vector Classifier, Stochastic Gradient Descent Classifier, Logistic Regression, and Multinomial Naive Bayes.

2.5.2 Model & Accuracy

The deep learning models, especially BiLSTM, outperform machine learning models across all metrics, indicating their superior performance in detecting suicidal ideation on Twitter. BiLSTM achieved the highest accuracy, precision, recall, and F1-Score (93.6% and 0.93) among all models. In contrast, the top-performing machine learning model, Random Forest, achieved an accuracy of 93.0% and a precision, recall, and F1-Score of 0.92. This demonstrates the superior effectiveness of deep learning models, particularly BiLSTM, in identifying expressions related to suicidal ideation on Twitter.

Table 2.4: Model Performance

Model	Accuracy	Precision	Recall	F1-Score	AUC
BiLSTM	93.6%	0.93	0.93	0.93	0.93
LSTM	93.5%	0.93	0.93	0.93	0.93
BiGRU	93.4%	0.93	0.93	0.93	0.93
CLSTM	93.2%	0.91	0.93	0.93	0.93
Random Forest (RF)	93.0%	0.92	0.92	0.92	0.92
Support Vector Classifier (SVC)	91.9%	0.91	0.91	0.91	0.91
Stochastic Gradient Descent (SGD)	91.7%	0.91	0.91	0.91	0.91
Logistic Regression (LR)	91.2%	0.91	0.91	0.91	0.91
Multinomial Naive Bayes (MNB)	84.6%	0.84	0.84	0.84	0.84

Chapter 3

Background Studies

Background studies play a pivotal role in research endeavors across various disciplines. They serve as the foundation upon which new investigations are built. Understanding the existing body of knowledge in a specific field is crucial as it provides researchers with the necessary context to delve into their own inquiries. Through comprehensive background studies, researchers can identify gaps and areas in need of further exploration, ensuring that their work is both relevant and novel. By examining prior research, they gain insights into methodologies, theories, and findings that guide their own investigations. Importantly, background studies prevent redundancy, allowing researchers to avoid revisiting questions that have already been answered. They also facilitate interdisciplinary insights, encouraging the integration of ideas and methods from diverse fields. Ethical considerations, theoretical frameworks, and literature reviews are all strengthened by a solid understanding of prior work. Ultimately, background studies are essential for the responsible, informed, and credible advancement of scientific knowledge and the development of meaningful research contributions.

3.1 Transformers

A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence. Transformer models apply an evolving set of mathematical techniques, called attention or self-attention, to detect subtle ways even distant data elements in a series influence and depend on each other [9]. The Transformer model first proposed in the paper “Attention Is All You Need” and is now a state-of-the-art technique in the field of NLP in 2017 by Vaswani et al., 2017 and achieved state-of-the-art results in language translation using only a fraction of the previous training times. It is an encoder–decoder type of model (see Figure 4A for the general idea of an encoder–decoder model), where the encoder

maps the vector representations of the tokens from an input text (the input embeddings) to an internal representation. The decoder then uses the internal representation and maps it to the output sequences (the target language, for instance) [10].

3.1.1 Transformer Architecture:

Understanding the Transformer model's architecture might seem overwhelming at first due to its complexity. However, breaking it down into its essential parts reveals a more manageable framework. The Transformer, known for its excellence in handling sequences, consists of four main components.

- (a) **Tokenization** : Tokenization is a fundamental step in natural language processing, breaking down text into units like words and punctuation. It facilitates structured representation, feature extraction, and normalization, serving as a crucial preprocessing stage for various language-related tasks.
- (b) **Embedding** : Embedding is a crucial process in natural language processing (NLP) where words or pieces of text are transformed into numerical vectors. Each word is mapped to a vector, and similar words have similar vector representations, capturing semantic relationships. This numerical representation enables machine learning models to understand and process textual data, making it suitable for tasks like sentiment analysis, language translation, and document classification. The embedding process facilitates the extraction of meaningful features and enhances the model's ability to capture contextual information from the input text.
- (c) **Positional encoding** : Positional encoding is a critical step in natural language processing that addresses the challenge of representing the order or position of words within a sentence. While standard methods like vector addition to represent sentences suffer from commutativity, positional encoding introduces a sequence of predefined vectors added to word embeddings. This ensures that each word's vector carries information about its position in the sentence. Consequently, sentences with the same words in different orders receive distinct vectors, enabling the model to capture unique positional information. This approach, exemplified in the modification of vectors for words like "Write," "a," "story," and ".", prevents ambiguity and enhances the model's ability to differentiate between sentences with similar word compositions.
- (d) **Transformer block** : The Transformer block is a pivotal architectural element that elevates the capability of neural networks in natural language processing. Introduced in the groundbreaking paper "Attention is All You Need," the Transformer

block integrates the attention mechanism, a fundamental innovation contributing to the remarkable success of transformer models. Components of a Transformer Block are :

1.Attention Component : At the heart of the Transformer block lies the attention mechanism. This component allows the model to imbue each word in a sequence with context, enabling it to consider the importance of every word concerning others. By incorporating context into each word, the attention mechanism captures intricate relationships and dependencies, significantly enhancing the model's understanding of sequential data.

2.Feedforward Component : Complementing the attention mechanism is the feedforward component. This neural network layer processes the contextualized information obtained from the attention mechanism, facilitating the extraction of high-level features and patterns within the input sequence.

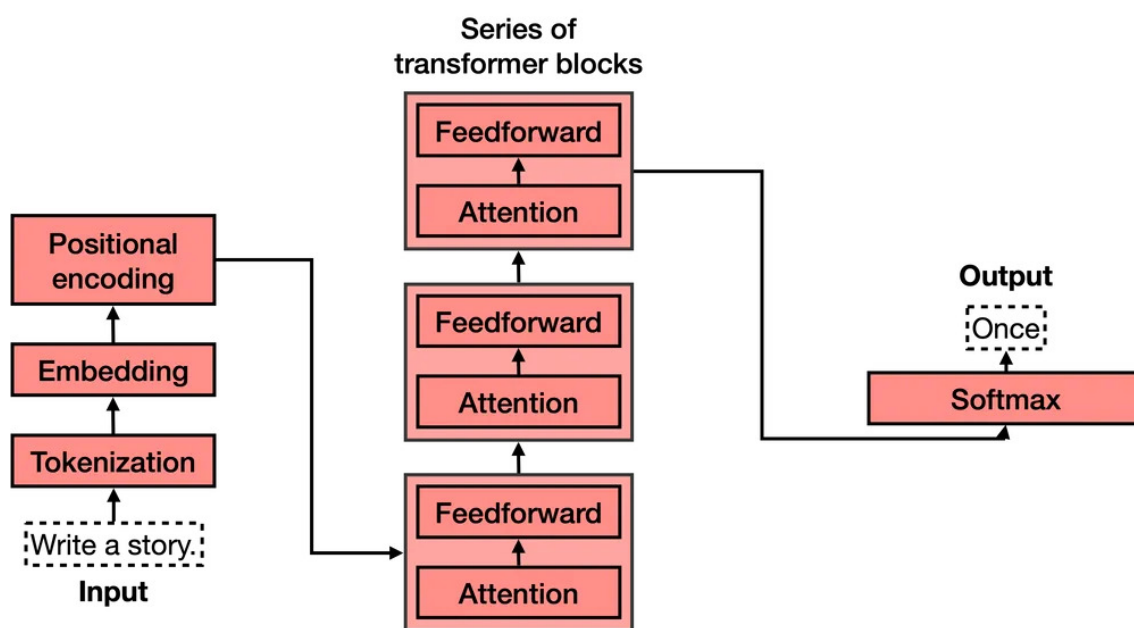


Figure 3.1: The architecture of a transformer model

3.2 Text Pretrained models

A pretrained model is a model that has been trained on a large dataset and can be used as a starting point for other tasks. Pretrained models have already learned the general patterns and features of the data they were trained on, so they can be fine-tuned for other tasks with relatively little additional training data.

In natural language processing (NLP), pre-trained models are often used as the starting point for a wide range of NLP tasks, such as language translation, sentiment analysis, and text summarization. By using a pre-trained model, NLP practitioners can save time and resources, as they don't have to train a model from scratch on a large dataset [11].

3.2.1 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. BERT models help machines understand and interpret the meaning of the text. The BERT model is pre-trained with two learning objectives that force the model to learn semantic information within and between sentences. The masked language modeling (MLM) task forces the BERT model to embed each word based on the surrounding words. The next sentence prediction (NSP) task, on the other hand, forces the model to learn semantic coherence between sentences [12].

Masked LM (MLM) : Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a token [12]. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. In technical terms, the prediction of the output words requires:

- (a) Adding a classification layer on top of the encoder output.
- (b) Multiplying the output vectors by the embedding matrix, transforming them into the vocabulary dimension.
- (c) Calculating the probability of each word in the vocabulary with softmax.

The BERT loss function takes into consideration only the prediction of the masked values and ignores the prediction of the non-masked words.

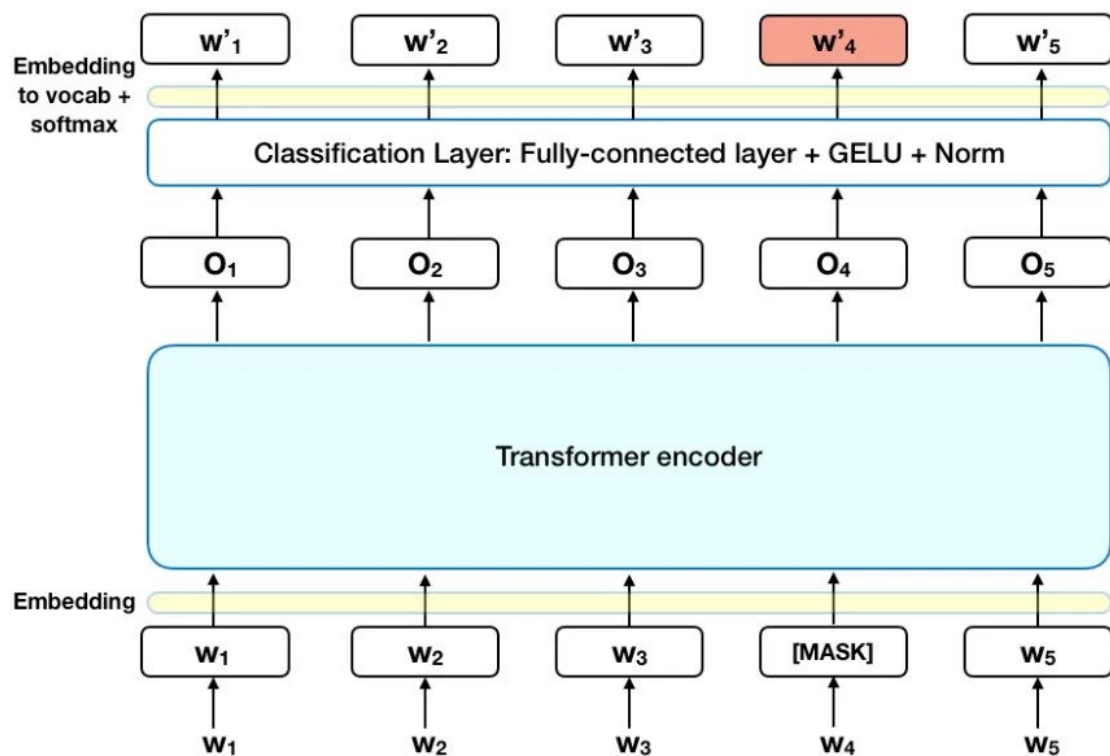


Figure 3.2: The architecture of The Masked Language Modeling

Next Sentence Prediction (NSP) : In the BERT training process, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document. During training, 50% of the inputs are a pair in which the second sentence is the subsequent sentence in the original document, while in the other 50% a random sentence from the corpus is chosen as the second sentence. The assumption is that the random sentence will be disconnected from the first sentence [13]. To help the model distinguish between the two sentences in training, the input is processed in the following way before entering the model:

- (a) **Token embeddings:** A [CLS] token is added to the input word tokens at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence.
- (b) **Segment embeddings:** A marker indicating Sentence A or Sentence B is added to each token. This allows the encoder to distinguish between sentences.

- (c) **Positional embeddings:** A positional embedding is added to each token to indicate its position in the sentence.

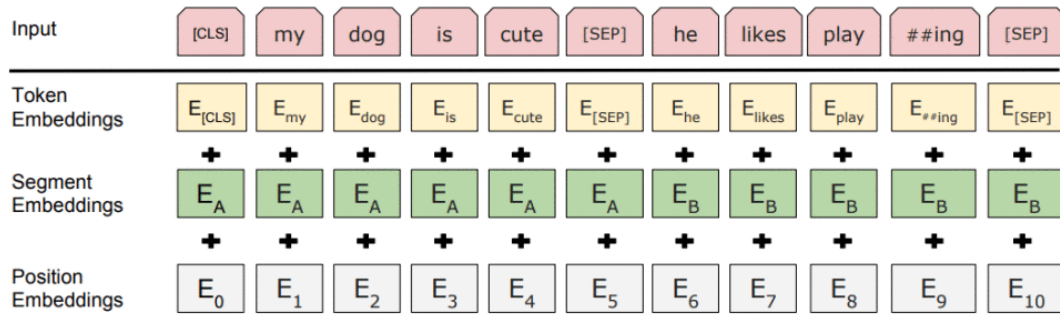


Figure 3.3: Architecture of Next Sentence Prediction.

3.2.2 RoBERTa

RoBERTa, short for Robustly Optimized BERT Pretraining Approach, is a transformer-based language model that builds upon the architecture and principles of the BERT (Bidirectional Encoder Representations from Transformers) model [14]. One key difference between RoBERTa and BERT is that RoBERTa was trained on a much larger dataset and using a more effective training procedure. In particular, RoBERTa was trained on a dataset of 160GB of text, which is more than 10 times larger than the dataset used to train BERT. Additionally, RoBERTa uses a dynamic masking technique during training that helps the model learn more robust and generalizable representations of words [15]

Modifications to BERT: RoBERTa closely resembles BERT in architecture, but to enhance performance on the BERT framework, the authors introduced simple design changes in both the model's structure and training approach [14].

- (a) **Removal of NSP Objective :** First, it removes the Next Sentence Prediction (NSP) objective, altering the training focus from predicting document segment relationships. The authors experimented with various NSP configurations, finding that its removal either matches or slightly improves downstream task performance.
- (b) **Training with Larger Batches & Sequences :** Second, RoBERTa adopts a training approach with larger batch sizes and longer sequences compared to BERT's original regimen. Training involves 125 steps with 2K sequences and 31K steps with 8K sequences, offering advantages in terms of improved perplexity on masked language modeling and enhanced end-task accuracy. Larger batches also facilitate parallelization in distributed training.

- (c) Dynamic Masking Pattern : Lastly, RoBERTa introduces dynamic masking by duplicating training data and applying different masks ten times over 40 epochs. This contrasts with BERT's static masking during data preprocessing. The dynamic masking strategy enhances adaptability, allowing the model to learn from varied mask patterns, thereby refining its language understanding capabilities.

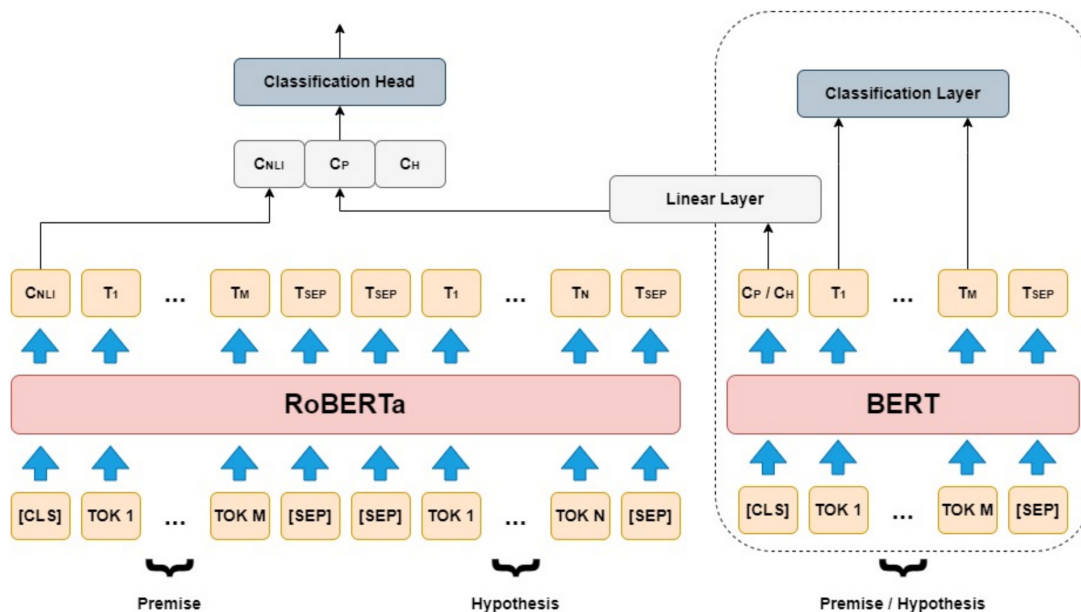


Figure 3.4: The Architecture of Roberta model.

3.2.3 ALBERT

ALBERT, short for "A Lite BERT," is a lightweight variation of the BERT model. It is designed to address BERT's computational inefficiencies while maintaining or even surpassing its performance on various natural language processing (NLP) tasks. ALBERT achieves this by introducing model parameter reduction techniques, enabling efficient training on large-scale datasets. The ALBERT model introduces innovative mechanisms to address the challenges associated with increasing model size in pretraining natural language representations. With a focus on mitigating GPU/TPU memory limitations and reducing training times, ALBERT incorporates two parameter-reduction techniques. These techniques enhance memory efficiency and training speed, allowing the model to scale more effectively than the original BERT. Additionally, ALBERT utilizes a self-supervised loss that prioritizes modeling inter-sentence coherence, consistently improving performance on downstream tasks involving multi-sentence inputs. Notably, the best-performing ALBERT model achieves state-of-the-art results on benchmark datasets such as GLUE, RACE, and SQuAD, while boasting fewer parameters compared

to BERT-large. This underscores the model's efficiency and effectiveness in large-scale natural language processing applications. [16]

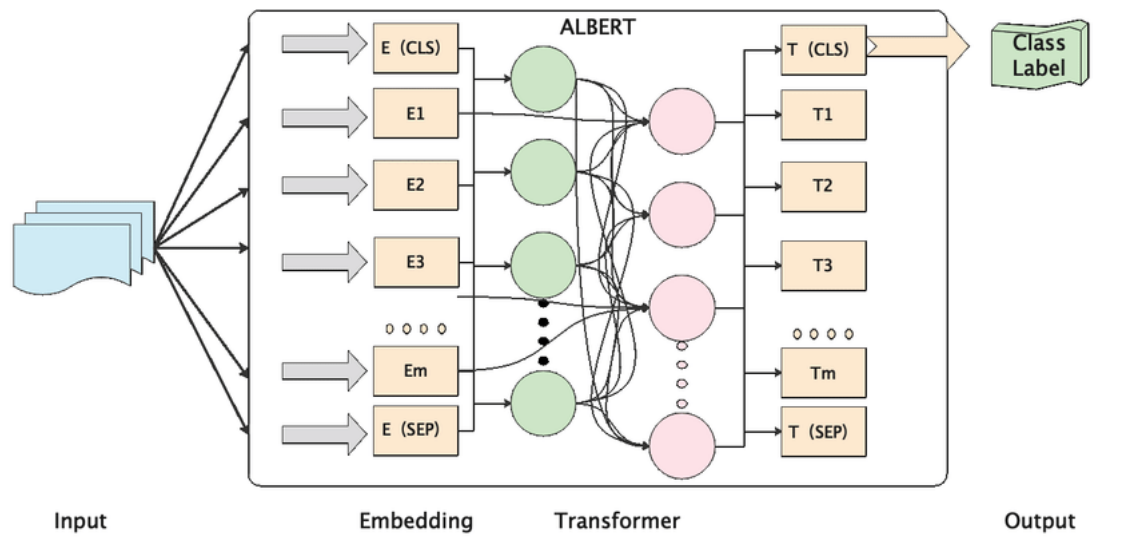


Figure 3.5: The Architecture of Albert model.

3.2.4 DistilBERT

DistilBERT, short for Distilled BERT, is a compressed and smaller version of the BERT (Bidirectional Encoder Representations from Transformers) model. It was introduced by Sanh et al. in the paper "DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter" in 2019 [17]. The primary goal of DistilBERT is to reduce the size and computational requirements of the original BERT model while maintaining its overall effectiveness. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark.

Here are the key details and architecture of DistilBERT:

- Transformer Architecture:** Like BERT, DistilBERT is based on the transformer architecture, which allows it to capture contextual relationships and dependencies between words in a sentence.
- Distillation Process:** DistilBERT is created through a process called knowledge distillation, where a larger model (BERT) is used to train a smaller model (DistilBERT) to mimic its behavior. This process involves transferring the knowledge from the teacher model (BERT) to the student model (DistilBERT) while making it more computationally efficient.

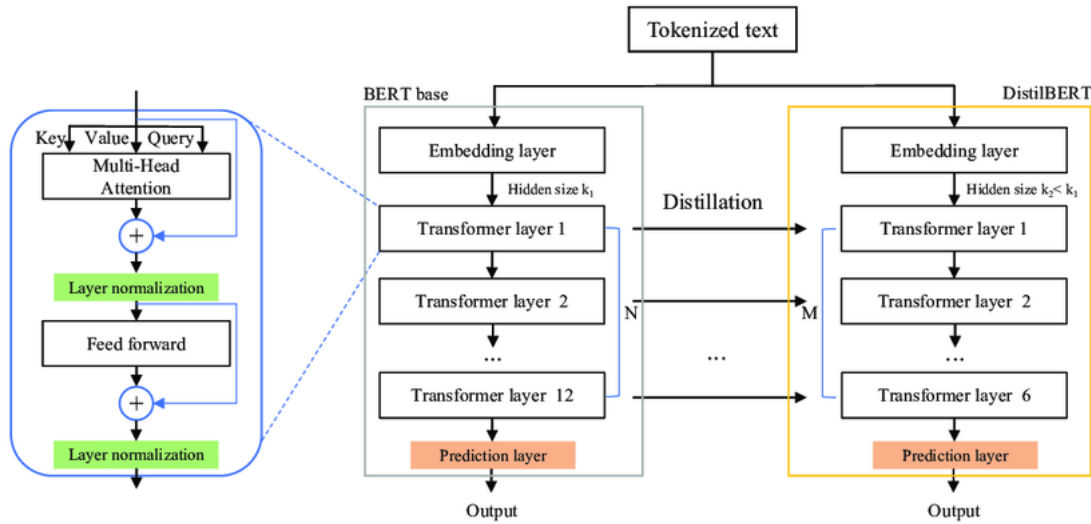


Figure 3.6: The Architecture of DistilBERT model.

3.3 Text embedding models

A text embedding model is a type of model that converts textual data into numerical vectors, often in a continuous vector space. These vectors capture semantic relationships and contextual information about the words or phrases in the text, enabling algorithms to work with and understand textual information in a more meaningful way.

3.3.1 GloVe

GloVe, or Global Vectors for Word Representation, is a widely used word embedding model designed to capture semantic relationships and contextual information of words in a continuous vector space. Developed by researchers at Stanford University, GloVe stands out for its ability to generate meaningful word representations by leveraging global statistical information from the entire corpus. Unlike some other word embedding models, GloVe does not rely solely on local context but incorporates a holistic view of word co-occurrence patterns. The primary steps in GloVe's model architecture are as follows [18].

- (a) **Constructing the Word Co-Occurrence Matrix:** GloVe starts by building a word co-occurrence matrix based on the frequency of word pairs appearing together in a given context window. This matrix represents the global statistical patterns of word associations in the corpus.
- (b) **Factorization of the Co-Occurrence Matrix:** The co-occurrence matrix is factorized to obtain word vectors. The factorization process captures the underlying semantic relationships between words, emphasizing their co-occurrence patterns.

- (c) Learning Word Representations: The model learns continuous vector representations for each word based on the factorized co-occurrence matrix. These vector representations, or embeddings, encode semantic information, allowing words with similar meanings to have similar vector representations in the continuous vector space.
- (d) Training and Optimization: GloVe uses optimization techniques to fine-tune the word embeddings, ensuring that they accurately reflect the global statistical properties of the corpus. This process refines the vectors, enhancing their ability to capture semantic nuances and relationships.

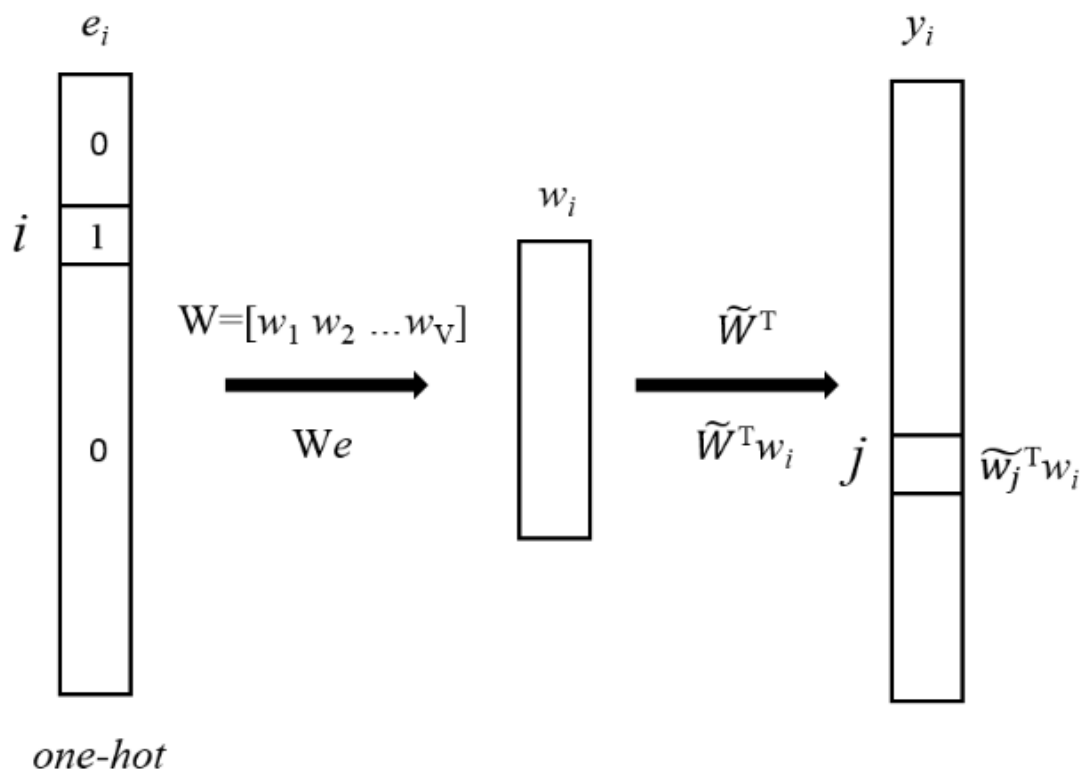


Figure 3.7: The Architecture of GloVe model.

GloVe's simplicity, efficiency, and ability to capture global semantic relationships have made it a popular choice for a variety of natural language processing tasks, including text classification, sentiment analysis, and machine translation.

Chapter 4

Work Process

This research endeavors to augment the precision of text classification for the identification of suicidal ideation within social media posts by conducting a comparative analysis of various natural language processing (NLP) models. The envisaged model comprises three distinctive layers—data, embedding, and classification—each tailored to fulfill specific functions within the classification pipeline. The study not only seeks to advance the field of NLP-based text classification but also holds the potential to make substantial strides in understanding and addressing mental health challenges in the digital realm.

4.1 Research Design

The proposed model is designed in the form of three layers

- (a) Data Collection Layer
- (b) Embedding Layer
- (c) Classification Layer

Each layer has its specific functions. The data collection layer deals with the data collection from the social platform, data cleaning and data preprocessing. The Embedding layer deals with word embedding. Various conventional models (GloVe) and transformers models (BERT, ALBERT, ROBERTA, DistilBERT) will give their footmarks in this sector. Then in the classification layer, the data will be classified using binary classification such as Bi-LSTM, Simple Neural Network.

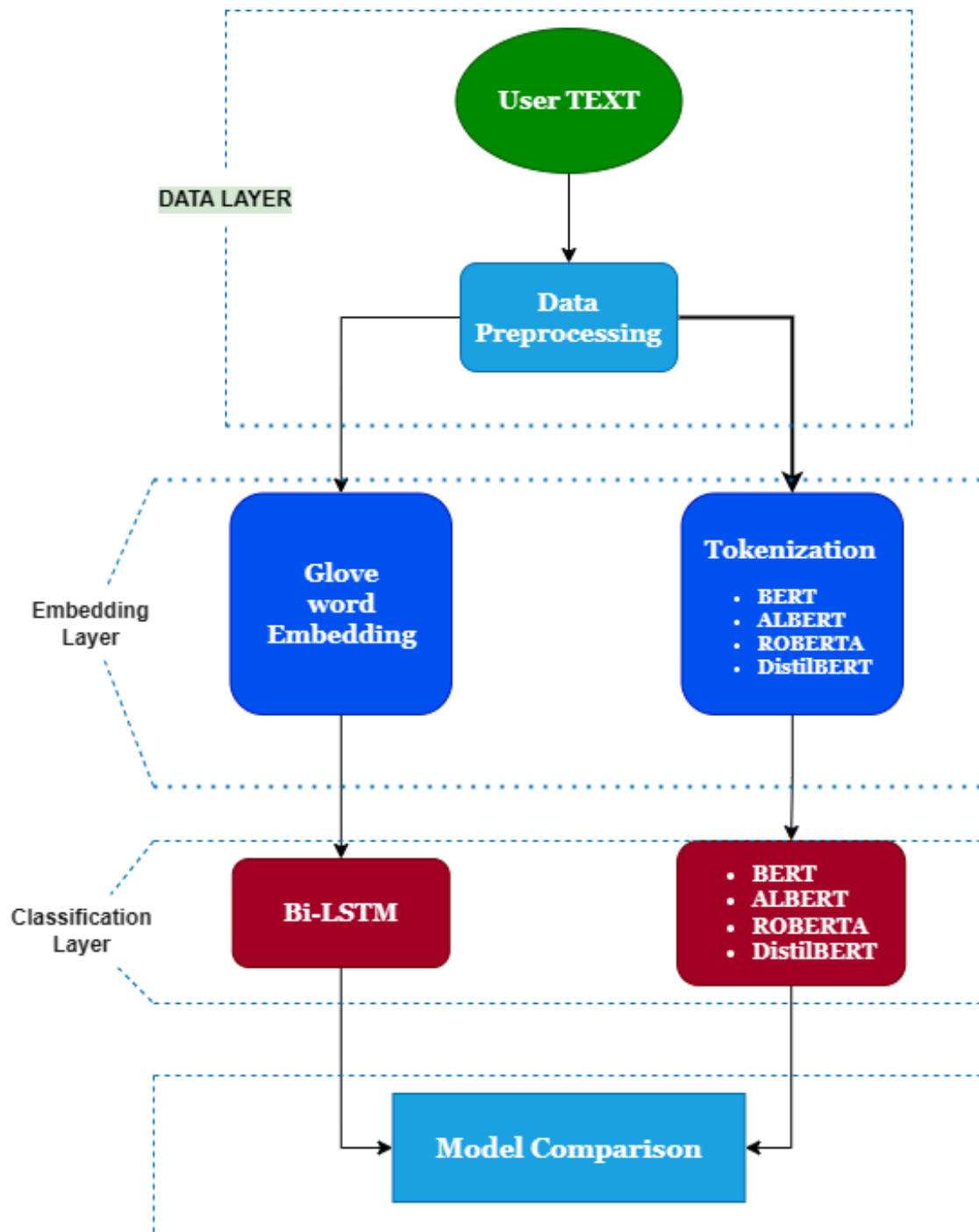


Figure 4.1: Proposed Model Diagram

4.1.1 Data Layer

The data layer part consists of preprocessing the suicidal suggestive texts. For that various social platform has been used such as “SuicidalWarch”, mentalhelp, and "depression" subreddits of the Reddit platform. And we have selected this platform as our source of data. The resulting user text needs to be passed into filters to improve the accuracy of our proposed model. As raw user posts have abbreviated words and it will be difficult for any automated machine to get the main context of the post. For this, we have expanded the abbreviated words to their correct form. We also removed special characters, any URL addresses, and redundant whitespaces with a single white space. Again We removed any emoticon which could give us an inconsistent result. This preprocessing will make the user raw texts to a format that can be understood by the word embedding in this layer, the collected data is looked alike.

4.1.2 Embedding Layer

The embedding layer serves as a pivotal component in our neural network architecture, responsible for converting discrete words or tokens into continuous vector representations. In this layer, we leverage two sophisticated word embedding models and a transformer-based model.

- (a) **Word embedding models :** Firstly, the GloVe (Global Vectors for Word Representation) model captures semantic relationships and contextual information inherent in words. This pre-trained model provides vector representations that encapsulate the meanings of words, contributing to a nuanced understanding of suicide-related content.
- (b) **Transformer-based model :** we integrate a transformer-based model (BERT, ALBERT, ROBERTA, DistilBERT) into the embedding layer. This model excels in capturing intricate contextual dependencies and patterns within the input text. By combining the strengths of both GloVe and the transformer-based model, our embedding layer ensures a comprehensive and contextually rich representation of the preprocessed textual data. This strategic fusion enhances the model's capacity to discern and interpret the underlying meanings of words, fostering improved performance in subsequent layers of the neural network.

4.1.3 Classification Layer

The classification layer is a crucial component in NLP models for text classification. Despite the extensive research conducted so far, there is still room for improvement

in terms of accuracy, efficiency, and generalization. Our future research plans aim to address these challenges by investigating various approaches for improving the performance of the classification layer. We believe that our research will contribute to the development of more robust and scalable NLP models for text classification.

4.2 Data Acquisition

The data collection process involved leveraging the Pushshift API to systematically retrieve all posts within specific time ranges from two distinct subreddits, namely "suicidewatch" and "Depression." The data encompasses posts made on the "suicidewatch" subreddit spanning from December 16, 2008, to January 2, 2021. For the "Depression" subreddit, the collection period spans from January 1, 2009, to January 2, 2021. This approach ensures a comprehensive dataset for analysis, providing insights into the content and discussions within these online communities over the specified time frames. [19]

4.3 Data Preprocessing

Data preprocessing, a pivotal phase in the analytical pipeline, aligns with the adage that "clever algorithms benefit from more data, but superior data surpasses sheer volume" [20]. Given the intricacies inherent in datasets, it becomes imperative to address potential issues such as inappropriate entries, null values, and irregularities before subjecting the data to machine learning algorithms. The cleaning process involves meticulous handling of missing values, as they contribute little to the overarching analytical objectives. Furthermore, structuring the raw data by imposing order, such as sorting columns in an ascending fashion, proves instrumental. This deliberate preparation not only ensures the dataset's integrity but also optimizes it for subsequent advanced analytics, ultimately enhancing the efficacy of machine learning algorithms.

4.3.1 Eliminating Extra Spaces and Lower casing

In the first step of data preprocessing, the dataset is transformed to lowercase. This critical transformation ensures uniformity in text data by standardizing all characters to their lowercase form. The significance lies in promoting consistency during subsequent analyses, as variations arising from the use of uppercase and lowercase letters are mitigated. Additionally, the effectiveness of natural language processing tasks, such as sentiment analysis and text classification, is enhanced by treating words in a case-insensitive manner, capturing their semantic meaning more accurately.

Extra spaces within the dataset are eliminated as part of the data preprocessing procedure. This involves detecting and removing redundant spaces between words and sentences. The importance of this step lies in enhancing the quality and cleanliness of the text data. Extraneous spaces can introduce noise and disrupt the natural flow of text, potentially impacting the performance of downstream natural language processing tasks. By systematically eliminating extra spaces, the dataset is refined, ensuring a more accurate and reliable foundation for subsequent analyses and model training.

```
df['text']=df['text'].str.lower()

# remove the extra space in heading & tailing
df_text['cleaned_text(space Remove)'] = df_text['cleaned_text'].str.strip()

# take string & splits it into a list of substrings based on
# whitespace (spaces, tabs, or newline characters)
df_text['cleaned_text(space Remove)'] = df_text['cleaned_text(space Remove)'].apply(lambda x: ' '.join(x.split()))
```

Figure 4.2: Eliminating Extra Spaces and Lower casing

4.3.2 Text Refinement

In the initial stages of our data preprocessing pipeline, meticulous attention is directed towards refining the dataset by systematically eliminating non-essential elements. This comprehensive approach involves the removal of emojis, links, HTML tags, and punctuation from the text data. The rationale behind this process is multifold. Primarily, it ensures that the dataset is honed to contain only the pertinent textual content, free from distracting elements. Furthermore, the systematic purification contributes to the creation of a more consistent and standardized input for the subsequent stages of analysis. This nuanced preparation not only simplifies the dataset but also establishes a solid foundation for the subsequent analytical steps. By purifying the dataset from unnecessary elements, we lay the groundwork for a more accurate and efficient model. This strategic cleansing enables the model to focus exclusively on the core linguistic patterns, enhancing its ability to discern meaningful insights without interference from unrelated symbols or structures.

```

emoji_pattern = re.compile("[
    u"\U0001F600-\U0001F64F" # emoticons
    u"\U0001F300-\U0001F5FF" # symbols & pictographs
    u"\U0001F680-\U0001F6FF" # transport & map symbols
    u"\U0001F1E0-\U0001F1FF" # flags (iOS)
    u"\U00002500-\U00002BEF" # chinese char
    u"\U00002702-\U000027B0" # Dingbats
    u"\U000024C2-\U0001F251" # Miscellaneous symbols and pictographs
    u"\U0001F926-\U0001F937" # Additional emojis
    u"\U00010000-\U0010ffff" # Extended emojis
    u"\u2640-\u2642" # Gender symbols
    u"\u2600-\u2B55" # Miscellaneous symbols, geometric shapes, and dingbats
    u"\u200d" # Zero-width joiner
    u"\u23cf" # Eject button
    u"\u23e9" # Eject button (variation)
    u"\u231a" # Watch
    u"\ufe0f" # Variation selector (used to specify emoji variation)
    u"\u3030" # Wave dash
    u"]+", re.UNICODE])
emoji_cleaned_text = emoji_pattern.sub(' ', text)

# URLs pattern
url_pattern = r'https?://\S+|www\.\S+'
url_cleaned_text = re.sub(url_pattern, ' ', emoji_cleaned_text)

#remove contractions
contraction_reomved_text=contractions.fix(url_cleaned_text)

#punctuation Patter
punctuation_pattern = r'^\w\s'
punctuation_cleaned_text = re.sub(punctuation_pattern, ' ', contraction_reomved_text)

# '_' character removed from suffix or prefix or both of a word
pattern = r'\b_+|\b_'
underscored_removed_text=re.sub(pattern, '', punctuation_cleaned_text)

```

Figure 4.3: Text Refinement: Emojis, Links, HTML, and Punctuation Removal.

4.3.3 Removing Duplicate data and Null values

Our dataset containing posts from many forums, where experiences and thoughts related to suicidal ideation are shared by users, the understanding of the model can be distorted by duplicate posts from the same user or identical content shared across multiple threads. The identification and removal of duplicate entries ensure that each user's perspective uniquely contributes to the analysis, preventing biases in the transformer-based model.

Handling Null values through imputation or removal is crucial for creating a comprehensive dataset. This step ensures that the transformer-based model receives complete and informative input, enhancing its ability to understand and respond to varying degrees of suicidal ideation.

```
# drop duplicate
# (232074, 2)
df.duplicated().sum()
#0
df.drop_duplicates(inplace=True)
df.shape
# (232074, 2)

# find missing values
df.isnull().sum()
#0
```

Figure 4.4: Removing Duplicate Data and Null Values

4.3.4 Removing Stop Words and Lemmatization

In this critical phase of data preprocessing, two fundamental techniques are applied: the removal of stop words and lemmatization. Stop words, ubiquitous yet semantically limited, are deliberately excluded to simplify and streamline the dataset. Additionally, lemmatization is employed to standardize words to their base or root form, fostering consistency and reducing lexical variations. Through this combined approach, the primary objective is to elevate the quality and coherence of the dataset. By discarding stop words, the impact of frequently occurring but non-informative terms is mitigated, allowing the model to concentrate on substantive content. Concurrently, lemmatization ensures words are represented in their most neutral form, minimizing redundancy and facilitating more precise analyses. This meticulous process establishes the foundation for a refined dataset that seamlessly aligns with subsequent analysis stages, facilitating a more nuanced and accurate comprehension of the underlying linguistic patterns.

```

#nltk
import nltk
from nltk.tokenize import word_tokenize

# stopwords
nltk.download('stopwords')
nltk.download('wordnet')
from nltk.corpus import stopwords
nltk_stopwords = set(stopwords.words('english'))

# lemmatizer
lemmatizer = nltk.WordNetLemmatizer()
# =====
# Stop Words
# =====
# NLTK Stop Words
nltk_stopwords = set(stopwords.words('english'))

```

Figure 4.5: Removing Stop Words and Lemmatization

After undergoing a comprehensive data preprocessing pipeline, our final dataset emerges as a meticulously curated and refined corpus. The dataset is devoid of extraneous elements, such as emojis, links, HTML tags, and punctuation, ensuring that only the essential textual content remains. This focused dataset has been subjected to lowercasing, the removal of duplicate entries, and the elimination of stop words. Furthermore, lemmatization has been applied to standardize words to their base form, promoting uniformity and reducing lexical variations. The result is a streamlined, consistent, and linguistically refined dataset, ready for subsequent analysis. This pre-processing not only enhances the dataset's quality but also optimizes it for natural language processing tasks, allowing for a more accurate understanding of the underlying linguistic patterns related to suicidal ideation and detection.

Data before cleaning

```

Am I weird I don't get affected by compliments if it's coming from so
meone I know irl but I feel really good when internet strangers do it

```

Figure 4.6: Data before cleaning

Data after cleaning

```
weird get affected compliments coming someone know irl feel really go  
od internet strangers
```

Figure 4.7: Data after cleaning

4.4 Data Visualization

Data visualization is a pivotal component of a thesis paper, offering a visual representation of complex findings and patterns within the dataset. It serves as a powerful tool to communicate research outcomes effectively to both expert and non-expert audiences. Visualization enhances the interpretability of results, making intricate relationships and trends accessible at a glance. This aids in conveying the significance of findings, reinforcing the narrative, and facilitating a deeper understanding of the research outcomes. Furthermore, visualizations serve as compelling evidence, strengthening the credibility of the research by providing a transparent and comprehensive portrayal of the analyzed data. In essence, data visualization is indispensable for transforming raw data into insightful narratives, making research more accessible, compelling, and impactful.

4.4.1 Dataset Balancing

With a dataset comprising 232,074 entries, a critical consideration is achieving balance between the two classes—suicidal and non-suicidal. The binary nature of the dataset, with 116,037 instances representing suicide-related content and an equivalent number designated as non-suicidal, underscores the importance of achieving parity. This balance is pivotal for preventing biases and ensuring that the machine learning model is exposed to an equitable representation of both classes. By equally distributing instances across the two segments, the dataset becomes a well-structured binary labeled dataset, fostering fair and unbiased model training. This strategic balancing acts as a foundation for robust model performance, enhancing its ability to discern patterns and make accurate predictions across both classes.

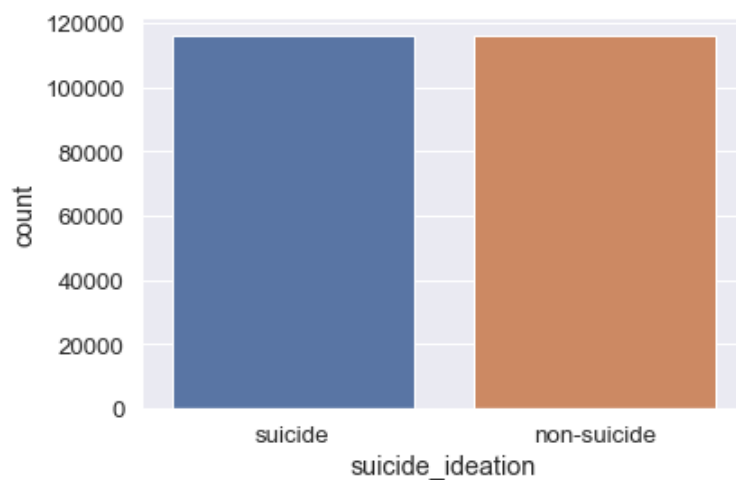


Figure 4.8: Dataset Balance

4.4.2 Average Word length in each text

In analyzing average word lengths, the plot reveals a notable distinction: suicidal text exhibits shorter word lengths compared to non-suicidal text. The average word length for suicidal content is 0.004, while non-suicidal content boasts a higher average of 0.008. This disparity in word lengths highlights a potential linguistic feature that may contribute to distinguishing between the two classes in the dataset.

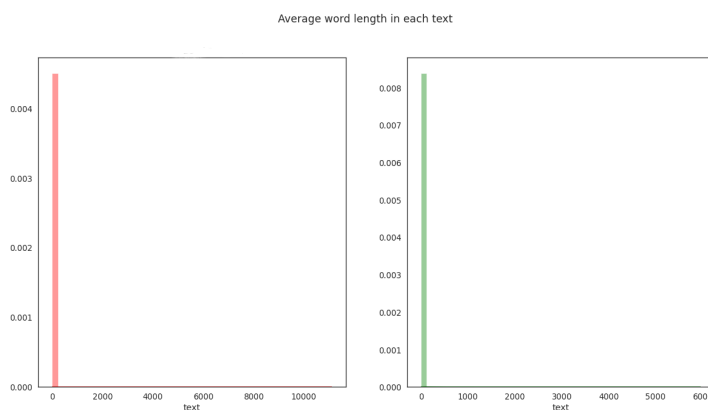


Figure 4.9: Average Word length in each text

4.4.3 Data Unigram Analysis

Unigram analysis can provide insights into the vocabulary, common terms, and linguistic characteristics of a text. It is often a foundational step in text processing tasks, including sentiment analysis, topic modeling, and text classification. By understanding the prevalence and significance of individual words, researchers and practitioners can gain valuable information about the structure and content of textual data.

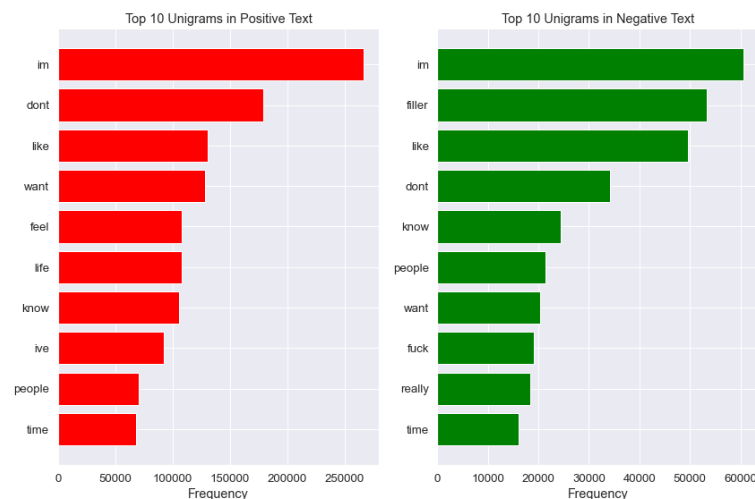


Figure 4.10: Data Unigram Analysis

4.4.4 Data bigram analysis:

Bigram analysis is commonly used in natural language processing and text mining to uncover patterns, associations, and contextual information within a body of text. It provides insights into how words interact with their neighboring words, capturing more nuanced information about the structure and semantics of the language.

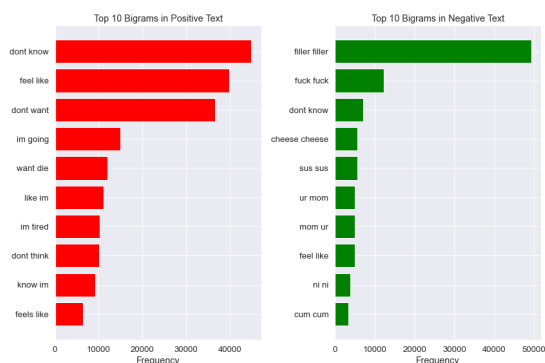


Figure 4.11: Data Bigram Analysis

Bigram models are often employed in various NLP tasks, such as language modeling, part-of-speech tagging, and machine translation. The utilization of bigram models enhances the understanding of language by considering the relationships between adjacent words, enabling the extraction of meaningful features and improving the accuracy of various natural language processing applications.

4.4.5 Data Trigram Analysis:

Trigram analysis involves studying sequences of three consecutive words in a text. It provides a more detailed perspective on language structure and semantics, capturing patterns and associations within a body of text. Trigrams contribute to various natural language processing tasks, revealing intricate linguistic patterns and enhancing comprehension of meaning and context in textual data.

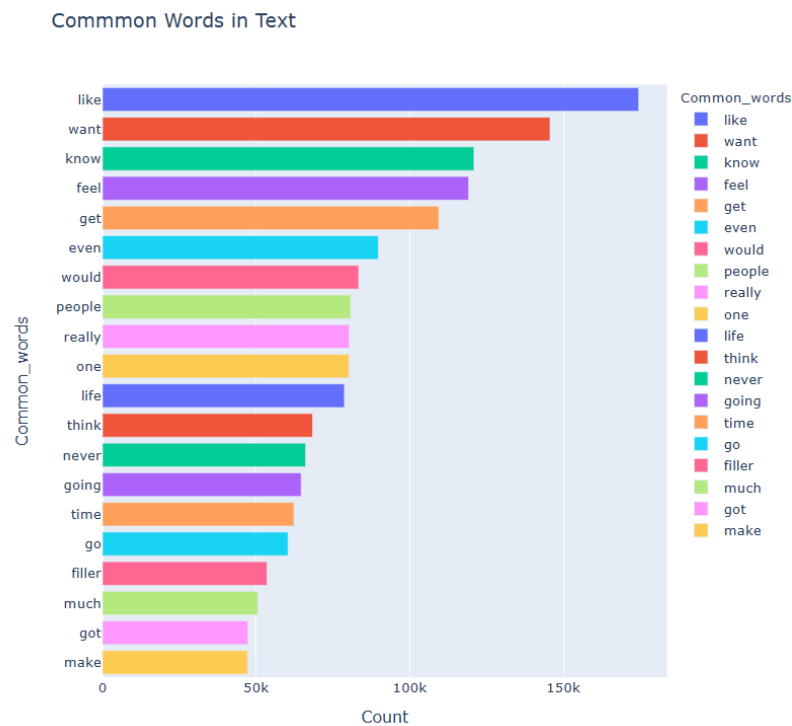


Figure 4.12: Data Trigram analysis

4.4.6 Stop Word and Punctuation Analysis

Stop word count and punctuation count analysis in textual content provides essential insights into linguistic features. Stop words, frequently occurring but semantically insignificant, contribute to overall text structure, influencing readability. Examining stop word count helps discern writing style and informational content. Punctuation count analysis reveals expressive elements and emotional tone, enhancing understanding of structural nuances. Contrasting these metrics enriches interpretation, offering nuanced insights into linguistic composition.

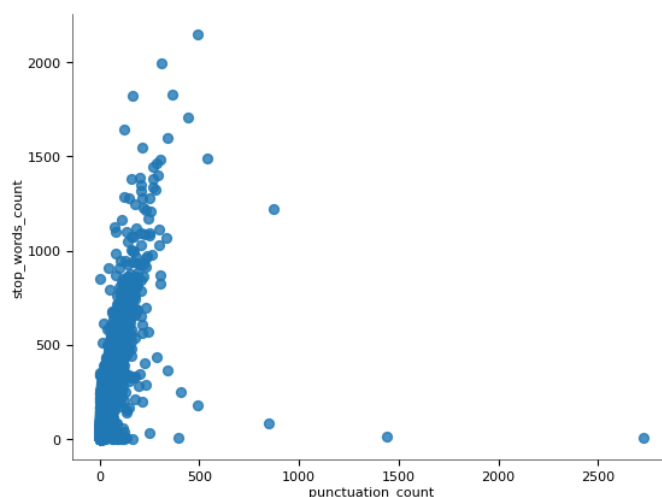


Figure 4.13: Stop Word and Punctuation Analysis

4.4.7 Textual Structure Analysis

Analyzing word and sentence counts provides valuable insights into the structure of textual data. Examining the distribution of words per sentence allows us to understand the average complexity and informativeness of the language. A higher word count per sentence may indicate more elaborate expressions. Conversely, tracking the sentence count aids in understanding the document's overall length and complexity. This analysis contributes to a comprehensive understanding of the dataset's linguistic features, guiding subsequent steps in natural language processing and machine learning.

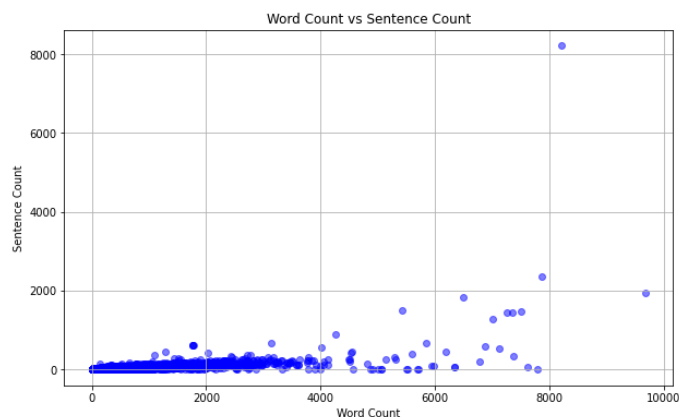


Figure 4.14: Textual Structure Analysis

4.4.8 Sentential Diversity Analysis

In evaluating sentential diversity, we compare the total number of sentences to the count of unique sentences within the dataset. This analysis provides insights into the overall uniqueness and variety of expressions across the corpus. A higher ratio of total sentences to unique sentences may indicate redundancy or repetition in the content. Conversely, a lower ratio suggests greater diversity and distinctiveness in the language used. This metric serves as a valuable indicator of the richness and variety of textual expressions within the dataset, contributing to a nuanced understanding of linguistic patterns and potential repetitive content

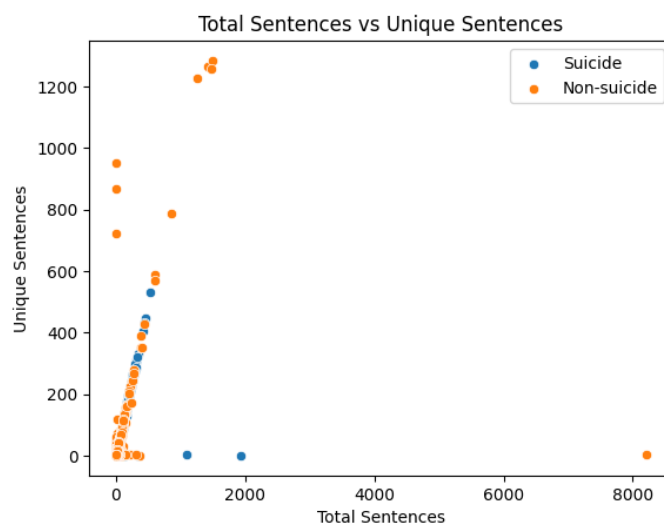


Figure 4.15: Sentential Diversity Analysis

4.5 Label Encoding

In our dataset, the 'suicide ideation' attribute reflects whether a text pertains to suicide or non-suicide content, denoted by binary values (1 and 0). Label Encoding extracts this information and creates a new 'label' attribute, simplifying the dataset for binary classification tasks. This transformation is vital as it equips the model to effectively differentiate between suicidal and non-suicidal texts, enhancing its ability to make accurate predictions in the context of suicide ideation.

[6]:

	text	suicide_ideation
0	ex wife threatening suiciderecently left wife ...	1
1	weird dont get affected compliments coming som...	0
2	finally 2020 almost never hear 2020 bad year e...	0
3	need helpjust help im crying hard	1
4	im losthello name adam 16 ive struggling years...	1
...
232069	dont like rock going get anything go	0
232070	tell many friends lonely everything deprived p...	0
232071	nee nrohably tastes like saltv tea someone dra...	0

Figure 4.16: Label Encoding

4.6 Test Model Training

Initially for our research, we choose BERT and Glove models. GloVe is a count-based model that generates word embeddings based on co-occurrence statistics, while BERT is a context-based model that generates contextualized word embeddings using a transformer-based architecture. The gloVe is fast and efficient but lacks contextual understanding, while BERT excels at capturing contextual relationships and achieves state-of-the-art performance on various NLP tasks. The choice between GloVe and BERT depends on the specific task and available resources.

4.6.1 BERT

BERT relies on a Transformer (the attention mechanism that learns contextual relationships between words in a text) [21]. Bidirectional Encoder Representations from Transformers (BERT) uses the "masked language model" (MLM) that helps representing the left and the right context in the model. This gives the ability to train a bidirectional

transformer model. The masked language model works by masking token from the text and to predict the token based on the context. The BERT can be denoted by the number of layers. One of them is the BERTLarge which has 24 layers and BERTBase which has 12 layers. We have worked with the BERTBase as it takes less computational time and it gives almost the same results in sentiment analysis as BERTLarge. The BERT is pre-trained over the different pretraining task with unlabeled data. For classification, the pre-trained BERT takes the user text token as input with the [CLS] tag. We fine-tune the BERT end to end hyperparameters as necessary for our task.

Model Summary:

Table 4.1: Classification Report

	Precision	Recall	F1-Score
Non-suicide	0.98	0.98	0.98
Suicide	0.97	0.97	0.97
Accuracy			0.98
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98

4.6.2 ALBERT

While doing different tasks of NLP, we have seen that the increasing size of the pre-training model can give us a better result. But in some cases, it becomes impossible to increase the model size because of the hardware limitation. To resolve the problem lite BERT(ALBERT) architecture was introduced [15]. First ALBERT introduces two different matrices for the word embedding matrix. By separating this the model can now incorporate different sizes of the hidden layer that does not have any effect in the word embedding layer. The architecture of ALBERT introduces cross-layer parameter sharing. This prevents the increase of parameters with the depth of the network. It also tackles the next sentence prediction loss that the BERT model architecture has. ALBERT has presented self-supervised loss that is based on consistency between the sentences. By using pre-trained ALBERT in our suggested model we get faster training time then BERT.

Model Summary:

Table 4.2: Classification Report

	Precision	Recall	F1-Score
Non-suicide	0.98	0.99	0.99
Suicide	0.99	0.97	0.98
Accuracy			0.98
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98

4.6.3 ROBERTA

Demonstrate that expanding the dataset leads to accuracy, in tasks. To achieve this researchers working on the language model utilized a dataset of over 160GB without any reduction in sequences, during pre training. This approach enhances the models accuracy by eliminating the Next Sentence Prediction (NSP) loss.

Model Summary:

Table 4.3: Classification Report

	Precision	Recall	F1-Score
Non-suicide	0.97	0.99	0.99
Suicide	0.98	0.97	0.98
Accuracy			0.98
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98

4.6.4 DistilBERT

The DistilBERT classification layer is a crucial component that transforms the contextualized embeddings generated by the transformer base into meaningful predictions for text classification tasks, making the model adaptable and effective for a wide range of natural language processing applications.

Model Summary:

Table 4.4: Classification Report

	Precision	Recall	F1-Score
Non-suicide	0.98	0.99	0.99
Suicide	0.99	0.97	0.98
Accuracy			0.98
Macro Avg	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98

4.6.5 GloVe Embeddings

GloVe (Global Vectors for Word Representation):

GloVe is a type of learning algorithm that doesn't require supervision. It aims to create vector representations of words by analyzing the information in a text. By examining how words appear together in a collection of texts GloVe captures the meanings and similarities, between them. In this approach, each word is represented as a vector, in a dimensional space.

Bidirectional Long Short-Term Memory (BiLSTM):

A Bidirectional LSTM refers to a kind of neural network (RNN) design that can analyze input sequences in both the forward and backward directions. This two way processing enables the model to better grasp connections and patterns, within the input sequence thereby improving its comprehension of context and long term relationships. This combination frequently leads to performance, in tasks that demand a comprehension of language meaning and context.

Model Summary:

Table 4.5: Classification Report

	Precision	Recall	F1-Score
Non-suicidal	0.91	0.93	0.92
Suicidal	0.95	0.93	0.94
Accuracy			0.93
Macro Avg	0.93	0.93	0.93
Weighted Avg	0.93	0.93	0.93

Chapter 5

Result and Performance Analysis

This section presents the outcomes of training and testing the dataset on various models, with the data distribution divided into training (70%), validation (10%), and testing (20%) sets. The subsequent analysis highlights the performance metrics of precision, recall, F1-scores, and overall accuracy for each model.

Model	precision	recall	f1-scores	Accuracy
GloVe	0.98	0.98	0.98	98%
BERT	0.98	0.98	0.98	98%
ALBERT	0.98	0.98	0.98	98.34%
RoBERTa	0.95	0.95	0.95	95%
DistilBERT	0.97	0.97	0.97	97.51%

The performance evaluation of the various transformer-based models, including GloVe, BERT, ALBERT, RoBERTa, and DistilBERT, revealed distinct characteristics that contribute to their effectiveness in identifying suicidal ideation in social media posts.

The GloVe embedding with Bi-LSTM demonstrates a commendable accuracy of 98%, accompanied by precision, recall, and F1-score metrics also standing at 0.98. This traditional approach showcases a robust capability in effectively distinguishing between suicidal and non-suicidal texts.

On the other hand, transformer models, particularly BERT and ALBERT, exhibit superior precision, recall, and F1-scores, with an impressive accuracy of 98%. The nuanced differences in performance between BERT and ALBERT (98% vs. 98.34%) highlight the incremental advantages conferred by ALBERT’s architectural modifications.

RoBERTa maintains a noteworthy accuracy of 95%, albeit with slightly lower precision, recall, and F1-scores compared to BERT and ALBERT. DistilBERT, with its streamlined

architecture, achieves competitive results with a 95.51% accuracy, demonstrating efficiency in resource utilization.

In conclusion, while the GloVe embedding with Bi-LSTM offers robust performance, the transformer models, particularly BERT and ALBERT, outshine in precision and recall metrics, emphasizing their effectiveness in discerning suicidal content. The choice among these approaches ultimately hinges on the specific use-case requirements and available computational resources.

Chapter 6

Contribution & Research Gap

6.1 Contribution

This research makes several noteworthy contributions to the field of mental health awareness and suicide prevention through advanced technology and machine learning methodologies.

- (a) **Innovative Detection Model** : Introducing a novel machine learning approach for detecting suicidal ideation in social media, our three-layer model (Data Collection, Embedding, Classification) enhances accuracy through advanced natural language processing techniques.
- (b) **Transformer Model Evaluation** : Thoroughly assessing popular transformer models (BERT, ALBERT, RoBERTa, DistilBERT), our research provides valuable insights into their performance nuances, aiding researchers and practitioners in informed model selection for similar language processing tasks.
- (c) **Global Mental Health Impact** : Addressing the critical global challenge of suicide prevention, our study contributes to advancing mental health awareness. The proposed model serves as an early intervention tool, supporting global mental health initiatives and suicide prevention efforts.

6.2 Research Gap :

6.2.1 Language-Specific Limitation

The study is currently tailored exclusively for detecting suicidal ideation in English text, creating a notable gap in language specificity. Future investigations and enhancements

are needed to extend the model's applicability to various languages, addressing this limitation and ensuring a more globally effective approach to suicide prevention.

6.2.2 Research Gap and Need for Inclusivity

The language-specific focus highlights a research gap, emphasizing the necessity for studies that explore and enhance the model's effectiveness across diverse linguistic contexts. This inclusivity is crucial for a more comprehensive and globally applicable suicide prevention strategy.

6.2.3 Potential for Multilingual Replication

The experiment, conducted solely in English, suggests the potential for replication in other languages, such as Bangla. This expansion could provide insights into language-specific nuances of suicidal ideation, contributing to a more culturally informed understanding of the phenomenon.

6.2.4 Consideration of Socioeconomic Factors

Given that the dataset was collected from a suicide API and is currently convenient in English, there is an opportunity to explore the impact of Bangladeshi socioeconomic factors on suicidal ideation. This would contribute to a more nuanced understanding of the complex interplay between language, culture, and mental health.

6.2.5 Exploration of Different Transformer Models

Considering the availability of other transformer models, there is a suggestion to explore their performance in detecting suicidal ideation. This exploration could potentially lead to models with higher accuracy and effectiveness compared to the current study.

6.2.6 Incorporation of Emojis for Enhanced Understanding

To capture a more comprehensive picture, future research could explore the impact of emojis on suicidal ideation. Emojis play a significant role in online communication and may offer valuable insights into emotional states that traditional text analysis might miss.

In summary, the study identifies language-specific limitations and suggests avenues for future research, including multilingual replication, consideration of socioeconomic

factors, exploration of different transformer models, and the incorporation of emojis for a more nuanced understanding of suicidal ideation.

Chapter 7

Conclusion and Future work

7.1 Future work

- (a) **Multimodal Enrichment** : We aim to enrich our dataset by integrating audio and video content, enhancing our model's ability to comprehend context and emotional nuances in social media posts concerning mental health.
- (b) **Performance Analysis** : Conducting a longitudinal analysis to explore how the model's performance evolves over time, considering the dynamic nature of language use on social media platforms
- (c) **Temporal Suicidal Ideation Detection** : Investigating real-time detection of suicidal ideation in data streams with a focus on temporal information. Identifying stages of suicide risk, including stress, thoughts, and plans, facilitates timely monitoring for changes in mental status.
- (d) **Improve accuracy and speed** : The machine learning models that were created are quite good at recognising and understanding suicidal content, but there is still room for improvement. Future work can focus on improving the accuracy and speed of the models by using more advanced machine learning methods and looking into other data sources.

7.2 Conclusion

In conclusion, our research tackles the pressing global issue of suicide with an innovative approach to identify suicidal ideation. By tapping into the rich content of social media, specifically the "Suicide Watch" and "depression" subreddits, we crafted a sophisticated three-layered model—Data Collection, Embedding, and Classification. Introducing various transformer models like BERT, ALBERT, RoBERTa, and DistilBERT, we conducted a robust analysis.

The standout performers, BERT and ALBERT, achieved an impressive 98% accuracy, showcasing their prowess in accurately categorizing suicidal content. RoBERTa maintained a solid 95% accuracy, offering insights into the impact of training nuances. DistilBERT, with its streamlined efficiency, secured competitive results at 95.51%, making it a feasible option for resource-constrained situations.

Looking ahead, our future work aims to enrich the dataset through Multimodal Integration, conduct longitudinal analyses for Performance Evaluation, explore Temporal Suicidal Ideation Detection, and enhance model accuracy and speed. This research contributes to the global effort in suicide prevention by leveraging technology and machine learning. By providing a nuanced understanding of suicidal ideation, our work strives to enable more effective and timely interventions, fostering mental health support worldwide.

References

- [1] W. H. Organization *et al.*, “National suicide prevention strategies: Progress, examples and indicators,” 2018.
- [2] A. Name, “Suicidal thought detection dataset.” <https://www.kaggle.com/code/abhijitsingh001/suicidal-thought-detection>, Year.
- [3] A. S. Denney and R. Tewksbury, “How to write a literature review,” *Journal of criminal justice education*, vol. 24, no. 2, pp. 218–234, 2013.
- [4] F. Haque, R. U. Nur, S. A. Jahan, Z. Mahmud, and F. M. Shah, “A transformer based approach to detect suicidal ideation using pre-trained language models,” in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, 2020.
- [5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of suicide ideation in social media forums using deep learning,” no. 1-19, p. Page range, 2020.
- [6] D. Shin, K. Kim, S.-B. Lee, C. Lee, Y. S. Bae, W. I. Cho, M. J. Kim, C. H. K. Park, E. K. Chie, N. S. Kim, and Y. M. Ahn, “Detection of depression and suicide risk based on text from clinical interviews using machine learning: Possibility of a new objective diagnostic marker,” *Frontiers in Psychiatry*, vol. 13, 2022.
- [7] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, “Suicidal ideation detection: A review of machine learning methods and applications,” *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, 2020.
- [8] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, “A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning,” *Technologies*, vol. 10, no. 3, p. 57, 2022.
- [9] NVIDIA, “What is a transformer model?,” 2022.
- [10] A. Chandra, L. Tünnermann, T. Löfstedt, and R. Gratz, “Transformer-based deep learning for predicting protein properties in the life sciences,” *Elife*, vol. 12, p. e82819, 2023.
- [11] “Top 5 pre-trained model in nlp.” <https://shorturl.at/wxyJ9>.

- [12] W. De Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, “Bertje: A dutch bert model,” *arXiv preprint arXiv:1912.09582*, 2019.
- [13] “Bert explained: State of the art language model for nlp.” <https://shorturl.at/bcnY8>.
- [14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [15] “Overview of roberta model.” <https://www.geeksforgeeks.org/overview-of-roberta-model/>.
- [16] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019.
- [17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.
- [18] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” pp. 1532–1543, 2014.
- [19] N. Komati, “Suicide watch dataset.” <https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch>, Year.
- [20] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, “Multi-modal emotion recognition using deep learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 02, pp. 52–58, 2021.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Open sourcing BERT: State-of-the-art pre-training for natural language processing.” <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>, 2018.

Generated using Undergraduate Thesis L^AT_EX Template, Version 1.4. Department of
Computer Science and Engineering, Ahsanullah University of Science and Technology,
Dhaka, Bangladesh.

This thesis was generated on Tuesday 21st November, 2023 at 7:37am.