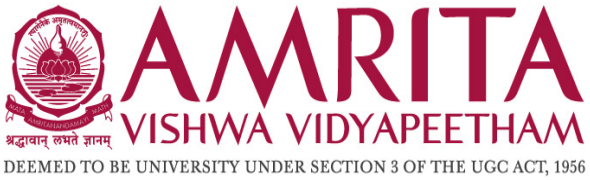
Suicide Detection Using Deep Learning

A PROJECT REPORT

Submitted by

Adithyan Sukumar - (CB.EN.U4AIE19004)
Anirudh Vadekedath - (CB.EN.U4AIE19008)
Arjun Anil - (CB.EN.U4AIE19012)
Rajath Rajesh - (CB.EN.U4AIE19051)
Vasudevan KM - (CB.EN.U4AIE19067)

in partial fulfillment for the award of the degree of
BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE)



Center for Computational Engineering and Networking

AMRITA SCHOOL OF COMPUTING

AMRITA VISHWA VIDYAPEETHAM

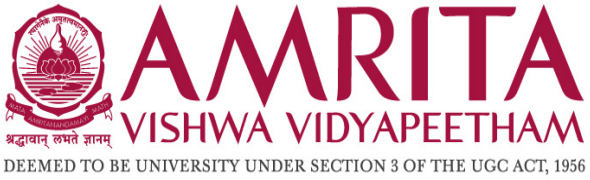
COIMBATORE - 641112 (INDIA)

MAY - 2023

AMRITA SCHOOL OF COMPUTING

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112



BONAFIDE CERTIFICATE

This is to certify that the thesis entitled **Suicide Detection Using Deep Learning** submitted by **Adithyan Sukumar (CB.EN.U4AIE19004), Anirudh Vadakedath (CB.EN.U4AIE19008), Arjun Anil (CB.EN.U4AIE19012), Rajath Rajesh (CB.EN.U4AIE19051) and Vasudevan KM (CB.EN.U4AIE19067)** for the award of the **Degree of Bachelor of Technology** in the **"Computer Science and Engineering (Artificial Intelligence)"** is a bonafide record of the work carried out by them/him/her under my guidance and supervision at Amrita School of Computing, Coimbatore.

Guide name **Dr. Premjith B** & **Dr. Jyothish Lal G**

Dr. K.P.Soman
Professor and Head
CEN

Submitted for the university examination held on - 04/05/2023

INTERNAL EXAMINER **EXTERNAL EXAMINER**

Contents

Acknowledgement

We would like to acknowledge Dr. Premjith B and Dr. Jyothish Lal G for their support in completing this project. We would also like to thank all our

List of Figures

List of Abbreviations

NLP	Natural Language Processing
ML	Machine Learning
EHR	Electronic Health Record
DL	Deep Learning
TF-IDF	Term Frequency - Inverse Document Frequency
SVM	Support Vector Machine
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
STM	Single Task Model
MTM	Multi-Task Model
SMHD	Self-reported Mental Health Diagnoses
HAN	Hierarchical Attention Network
KNN	K-Nearest Neighbours
POS	Part Of Speech
LSTM	Long Short-Term Memory
DNN	Deep Neural Network
PHQ	Patient Health Questionnaire
LRP	Layer-wise Relevance Propagation
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
BERT	Bidirectional Encoder Representations from Transformers
ALBERT	A Lite Bidirectional Encoder Representations from Transformers
NSP	Next Sentence Prediction
GUSE	Google Universal Sentance Encoder

Abstract

Early detection of suicidal ideation in depressed individuals can allow for adequate medical attention and support, which can be life-saving. Recent NLP research focuses on classifying, from the given text, if an individual is suicidal or clinically healthy. However, there have been no major attempts to differentiate between depression and suicidal ideation, which is a separate and important clinical challenge. Due to the scarce availability of EHR data, suicide notes, or other verified sources, web query data has emerged as a promising alternative. Online sources, such as Reddit, allow for anonymity, prompting honest disclosure of symptoms, making it a plausible source even in a clinical setting. However, online datasets also result in inherent noise in web-scraped labels, which necessitates a noise-removal process to improve performance. Recent literature mainly focuses on the use of Deep Learning models for classification, which is why we propose a transformer-based classifier to classify suicide vs depression ideation. Our extensive experimentation with transformer-based classifiers displays their strong performance of it as a new clinical application for a challenging problem.

Chapter 1

Introduction

One of the biggest issues plaguing the current generation is depression and various other mental illnesses, which if untreated may frequently lead to suicidal thoughts or attempts. At both the individual and group level, diagnosing depression and determining whether it poses a danger of suicide attempt are significant issues. The growth of social media has helped us to identify and detect people who have depression and suicide tendencies through their tweets and posts [23], as the Internet and specifically social media has grown, online forums have developed into popular resources for struggling individuals to seek guidance and assistance. Conversely, as the Internet and specifically social media have grown, online forums have developed into popular resources for struggling individuals to seek guidance and assistance. These forums have potential to be scraped to create datasets for automated systems of mental health diagnosis, as they are extensive and free to access. Especially for neural network-based approaches that require large datasets to be trained efficiently, a growing number of studies are using this data for diagnostic purposes.

Reddit has emerged as an important data source for diagnosing mental health dis-

orders. Reddit is an online social media forum in which users form communities with defined purposes referred to as subreddits. Certain subreddits discuss dealing with mental health and openly explain their situations (r/depression and r/SuicideWatch). Reddit specifically allows users to create alternate and discardable accounts to ensure privacy and anonymity, which promotes disclosure and allows those with little support systems in real life to receive support online. The wide user base, honesty of these online settings, and moderated screening of these posts to ensure legitimacy provides an unprecedented opportunity for computationally analyzing mental health issues on a large scale. Despite the extensive research into classifying between healthy and mentally unstable patients through text, there remains little work focused on detecting when individuals with underlying mental health struggles such as depression are at risk of attempting suicide.

This represents an important clinical challenge, both for the advancement of how depression is treated and for implementing interventions. Distinguishing between suicidality and depression is a more fine-grained task than distinguishing between suicidal and healthy behavior, explaining the lack of current solutions. Several research was done on this topic to compare different DL techniques for early detection of depression on Social Media platforms using twitter data. Some proposed method includes, a combined model of SVM and Naive Bayes algorithm for good accuracy. Online data has traditionally been difficult to use in such fine-grained situations, because labels for such data are often unreliable given their informal nature and lack of verification. Labeling data based on subreddit relies on self- reporting, since each user chooses which subred-

dit they feel best reflects their mental state; thus, they may over or under report their diagnosis. Not all instances call for transparency at the prediction level; in certain cases, DL models must become explainable throughout the learning phases. For example, DL models that draw on selected datasets may produce bias that is difficult to spot and necessitates a higher explainability level. Understanding the predictions for the entire category and being able to interpret the hidden representations of these networks would indicate whether certain protected attributes are biasing the predictions in some way.

The work comprises of two datasets which are similar in contents but has a huge difference in size. Dataset two which is very large compared to the dataset one has been used for the training along with the four BERT models that is ALBERT, RoBERTa, DistilBERT, BERT. BERT models have been used for both word embedding and classification tasks with three different testing methods. The following guidelines were followed when the tests were carried out in Pytorch. The models were trained using Adam as the optimizer and a learning rate of 2e-5 over the course of two epochs. Since two GPUs were used for the training, a per-device batch size of 16 was chosen. An identical experimental strategy was applied to all BERT classifiers. The performance of the models is measured by accuracy, precision, recall, and the f1-score.

Out of the four models DistilBERT performed better in all three testing methods with accuracy around 94. On behalf of this DistilBERT is used for model explainability. Python library called Captum is used for this purpose where it is clear that the algorithm is learning the phrases for depression and suicide prediction and that the label is not being predicted at random.

1.1 Literature Survey

Dr. Jennifer et. al. [24] compared different DL techniques for early detection of depression on Social Media platforms using twitter data. proposed a combined model of SVM and Naive Bayes algorithm for a good accuracy. Shaoxiong Ji et. al. [11] showed an illustrative review on the different methods such as Attention based, RNNs and CNNs for Suicide Ideation Detection. Yaakov Ophir et al. [15] used Single Task Model (STM), to predict suicide risk from Facebook postings directly and a Multi-Task Model (MTM), which included hierarchical, multilayered sets of theory-driven risk factors. Nhan Cach Dang et al. [3] provided the complexities in NLP and solves them using deep-learning models by using TF-IDF and word embeddings to a series of datasets. R Ramakrishnan et. al. [19] provides an embedding method that, uses VMD to get better embeddings. Ashima Yadav et. al. [6] reviews DL techniques on unstructured data. A taxonomy of sentiment analysis is presented and discussed, along with the implications of popular deep learning architectures. The survey summarizes the popular datasets, along with its key features, DL model applied, and accuracy obtained from them. Edwin D. Boudreaux et. al. [16] provides an overview of machine learning applied to suicide prediction, summarizes exemplar published studies for illustration, and explores future directions for research. Ivan Sekulic and Michael Strube [22] proposed method for detecting the mental health on social media dataset named as SMHD dataset using Logistic regression, SVM, Supervised FastText, and Hierarchical

Attention Network (HAN). HAN was able to produce better prediction in most cases, they also mentioned the limitation of HAN model in smaller datasets. Prof. S. J. Pachouly [17] proposed machine learning approaches like SVM, KNN, Decision tree and Naïve Bayes along with feature extraction technique like Bag of Words, TF-IDF, POS Tagging to classify depression in a scale of 0-100% from tweets collected from twitter. Bahman Zohuri and Siamak Zadeh [27] discuss the uses of DL and ML in suicide risk management. Topics such as Facial Emotion Recognition and Emotion Detection along with Optical Character Recognition were discussed. Lang He et. al. [9] worked on providing a review on the detection of depression in individuals using audiovisual cues using different DL methods such as LSTM, DNN, CNN, etc. Emily Schriver M.S, et. al. [21] applied univariate and multivariable logistic regressions to determine significant risk factors associated with suicide ideation responses from the Patient Health Questionnaire (PHQ-9). They compared how different factors such as age, race, availability of medical insurance among others affected suicide ideation. Shafie Gholizadeh et al. [7] compared several popular ML model explainability techniques, focusing in particular on those related to Natural Language Processing (NLP) models. One such approach that performed better than the gradient-only based and permutation-based explainability was layer-wise relevance propagation (LRP). We can get a lot insight into the black-box NLP models by applying this NLP explainability study, and we can lower the risk of selecting an incorrect or inappropriate model. It was also easier to comprehend why the model gave incorrect predictions when LRP scores indicated examples of false positive and false negative outcomes. Hila Chefer et al. [4] proposed

the first method to explain prediction by any Transformer-based architecture, including bi-modal Transformers and Transformers with co- attentions. They offer general solutions and apply them to the three architectures that are most frequently used, including encoder- decoder attention, pure self-attention, and self-attention paired with co-attention. Hui Liu et al. [13] propose a novel generative explanation framework that learns to make classification decisions and generate fine-grained explanations at the same time. The explainability factor and the minimum risk training approach are introduced, which teach people how to come up with more reasonable explanations. They created two novel datasets that include summaries, rating scores, and fine-grained reasons. They then conducted experiments on both datasets and compared the results with a number of powerful neural network baseline systems. On both datasets, they perform better than other baselines and can simultaneously produce concise explanations. Ian C. Covert et al. [5] describe a new unified class of methods, removal-based explanations, that are based on the principle of simulating feature removal to quantify each feature's influence. This framework unifies 26 existing methods, including several of the most widely used approaches: SHAP, LIME, Meaningful Perturbations, and permutation tests. Loukas Ilias et al. [10] are using several transformer-based models, with BERT achieving the highest accuracy accounting for 87.50%. Concurrently, they propose an interpretable method to detect AD patients based on siamese networks reaching accuracy up to 83.75%. Next, they introduce two multi-task learning models, where the main task refers to the identification of dementia (binary classification), while the auxiliary one corresponds to the identification of the severity of dementia

(multiclass classification). Their model obtains accuracy equal to 86.25% on the detection of AD patients in the multi-task learning setting. Finally, they present some new methods to identify the linguistic patterns used by AD patients and non-AD ones, including text statistics, vocabulary uniqueness, word usage, correlations via a detailed linguistic analysis, and explainability techniques (LIME). Andrew Poulton et al. [18] proposes a framework which provide the best accuracy and most intuitive explanations amongst the many available models and explainability techniques and assesses several popular transformer-based models with various explainability methods on the widely used benchmark dataset from Semeval-2013. Krithik Ramesh et al. [20] investigate two different explainability frameworks. Specifically, Label Attribution and Optimal Transport of Vision-Language semantic spaces with the VisualBERT multimodal transformer model provide an interpretability process towards understanding attention interactions in multimodal transformers. They provide a case study of the Visual Genome and Question Answer 2 Datasets trained using VisualBERT. Vasu Agarwala et al. [26] explained a stacked model which fine tunes the informational insight gained from the data at each step and then tries to make a prediction more perfect. The work proposed by Julia El Zini and Mariette Awad [1] presents the first comprehensive survey on explainability methods in the NLP field that combines ExAI methods on the input-, processing- and output levels. According to the assumptions made on explained models, it will help to make the distinction between model-agnostic and model-specific method. Leilani H. Gilpin et al. [2] proposed a paper named Explaining Explanations: An Overview of Interpretability of Machine Learning, which suggests that for machine learning sys-

tems to gain widespread acceptance among a sceptical public, they must be able to present or permit sufficient explanations for their conclusions. So far, development has been promising, with initiatives in deep network processing explanation, deep network representation explanation, and system-level explanation generation yielding positive results. In this paper authors discuss about different methods of explaining the results produced by model. Narine Kokhlikyan et al. [12] introduced a novel, unified, open-source model interpretability library. The library includes generic implementations of several gradient and perturbation-based attribution algorithms, often known as feature, neuron, and layer significance algorithms, as well as a collection of assessment metrics. It is suitable for both classification and non-classification models, as well as graph-structured models based on Neural Networks (NN). Tanvirul Alam et al. [25] used transformers to classify bangla text and used Captum to show which all words contribute in learning of the text. The work also shows that finetuning from transformer models can yield better performance compared to traditional methods that make use of hand-crafted features, and also deep learning models like CNN and LSTM trained on distributed word representation. Efficient shapley values estimation by amortization for text classification is a paper proposed by Chenghao Yang et al. [8] shows that obtaining stable explanation scores on long text inputs is difficult. We propose to estimate the explanation scores efficiently using an amortised model trained to suit pre- collected reference explanation scores, inspired by the notion that distinct cases can share similarly relevant properties. Paper proposed by Mitchell Naylor et al. [14] investigated practical aspects of interpretable and explainable ML methods in a clinical

text classification case study, demonstrating some of the interpretability tools available to healthcare NLP practitioners, discussing existing definitions for explainability and interpretability, and introducing a framework that can evaluate the quality of explanations across text classification models, including the infidelity and relative sensitivity of the presuppositions.

1.2 Problem statement

Suicide is a leading cause of death at a worldwide level. It is also observed that depression also leads to suicide. Since most people share their thoughts on social media it can be used to identify people going through a depression phase which might lead to suicide. Therefore in this project, we are planning to incorporate different word embedding methods and to use attention-based deep-learning models such as transformers for accurately classifying suicide versus depression.

1.3 Objectives

The main objectives of the work are as follows:

- Apply the text data to classify between suicide and depression using transformer based classifiers.
- Rigorous analysis of the classifiers on different datasets.
- Understanding the misclassified data.

1.4 Organization of the thesis

The paper is organized as follows: Section 1.1 presents the related surveys to Transformer based text embedding and classification and the explainability of these models, section 1.2 provides our problem statement and section 1.3 provides the objectives of this work. Section 2.1 presents a description of the dataset and section 2.2 provides an understanding of the different BERT models. Section 3 presents the methodology followed by an explanation of the different methodologies used in this thesis, section 3.5 presents an understanding on the evaluation metrics used in this thesis. Section 4 explains the results obtained and Section 5 provides a conclusion to this paper followed by the future work that con be done on this thesis.

Chapter 2

Background

Now it is the time to articulate the research work with ideas gathered in above steps by adopting any of below suitable approaches:

2.1 Dataset Description

2.1.1 Dataset 1

The dataset contains 1,895 total posts which is a collection of reddit posts from the r/SuicideWatch and r/Depression subreddits. This was collected by Ayaan Haque and Viraaj Reddi by scrapping the above mentioned subreddits. The datasets consist of 11 features and one label column. The processing done on the data were stop words removal, punctuation removal and emoji removal. After processing the data and doing some data analysis it was found that average sentence length for self-text feature was 171 and the average sentence length for cleaned selftext feature was 81 and average comments per post was 4, which can also be observed in Figure 2.1. The train test split was taken as 80:20.

After some analysis, this is what was observed for this dataset:

	Least num of comments	Most num of comments	Average num of comments
Number of comments	0	202	4

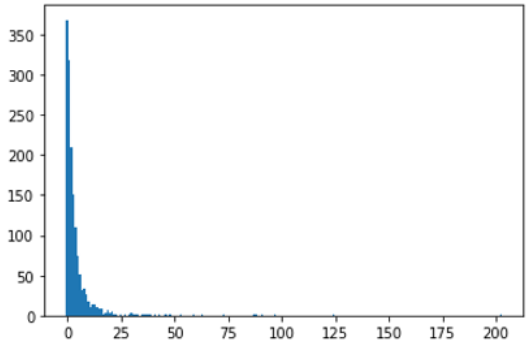


Figure 2.1: Comments vs comments per post

2.1.2 Dataset 2

The Dataset contains around 3,50,000 posts which is a collection of reddit posts from the r/SuicideWatch and r/Depression subreddits from Jan 1, 2009, to Jan 2, 2021. They are classified into two different classes. The Depression class consist of 2,32,987 data points and class Suicide consist of 1,17,017 data point. The dataset is a collection of posts from "SuicideWatch" and "depression" subreddits of the Reddit platform. The posts are collected using Pushshift API. All posts that were made to "SuicideWatch" from Dec 16, 2008(creation) till Jan 2, 2021, were collected while "depression" posts were collected from Jan 1, 2009, to Jan 2, 2021. The processing done on the data were stop words removal, punctuation removal and emoji removal. After processing the data and doing some data analysis it was found that average sentence length was 171. The train test split was 80:20. On further analysis we found some common words for each

class.

Class	Common Words
Depression	'myself', 'filler', 'depression', 'depressed', 'anymore', 'feel', 'shit', 'suicide', 'wo', 'suicidal'
Suicide	'end it', 'bye', 'n't want', 'to end', 'kill myself', 'to die', 'ca n't', 'to kill', 'killing myself', 'kill'

2.2 BERT Models

2.2.1 BERT

Bidirectional Encoder Representations from Transformers BERT is essentially a transformer design with an Encoder stack. An encoder-decoder network using self-attention on the encoder side and attention on the decoder side is known as a transformer design. The Encoder stack in BERTLARGE has 24 levels compared to BERTBASE's 12 layers. These go beyond the Transformer architecture as it was originally outlined in the paper (6 encoder layers). In addition, BERT (LARGE and BASE) architectures have more attention heads (12 and 16 respectively) and larger feedforward networks (768 and 1024 hidden units) than the Transformer architecture proposed in the original article. It has 8 attention heads and 512 hidden components. While BERTLARGE has 340M values, BERTBASE only has 110M. On numerous activities involving natural language processing and language modelling, BERT was able to increase accuracy (or F1-score). The primary innovation offered by BERT is the ability to use semi-supervised learning for a variety of NLP tasks, enabling transfer learning in NLP.

2.2.2 DistilBERT

DistilBERT is a distilled version of BERT; smaller, faster, cheaper and lighter than other models. Compared to bert-base-uncased, it executes 60% faster and maintains performance to over 95% with 40% fewer parameters. It is a method to pre-train a smaller general-purpose language representation model, which can then be fine-tuned with good performances on a wide range of tasks. It does not have token_type_ids, and can be trained to predict the same probabilities as the larger model. It has been trained to forecast the same probabilities as the larger model by distilling the pre-trained BERT model. The objective is to predict the masked tokens correctly and find a cosine similarity between the hidden states of the student and the teacher model.

2.2.3 ALBERT

A Lite BERT (ALBERT) ALBERT eliminates the main barriers to scaling pre-trained models by combining two parameter reduction methods. The first one is a factorized embedding parameterization. We distinguish between the size of the hidden layers and the size of the vocabulary embedding by breaking the big vocabulary embedding matrix into two smaller matrices. The hidden size can grow more readily thanks to this division without substantially growing the vocabulary embeddings' parameter size. Cross-layer parameter sharing is the second method. With this method, the number won't increase along with the network's depth. Both methods significantly reduce the number of parameters for BERT while preserving performance, thereby enhancing parameter efficiency. With 18x fewer parameters and a training time of about 1.7x

faster, an ALBERT configuration comparable to BERT-large exists. The methods for parameter reduction also function as a type of regularization that stabilizes the training and aids in generalization. Additionally, they present a self-supervised loss for sentence-order prediction (SOP) to enhance ALBERT's performance even more. In order to address the shortcomings of the next sentence prediction (NSP) loss suggested in the original BERT, SOP primarily concentrates on inter-sentence coherence.

2.2.4 RoBERTa

RoBERTa: A Robustly Optimized BERT Pretraining Approach discovered that BERT was significantly undertrained and suggests RoBERTa, an improved training strategy that can perform as well as or better than all post-BERT methods. They make a few straightforward changes, such as training the model longer, in larger batches, with more data, removing the goal of predicting the next sentence, training on longer sequences, and dynamically altering the masking pattern used on the training data in order to create RoBERTa.

Chapter 3

Methodology

The methodology introduced in this project aims at performing sentiment analysis for detecting Suicide Depression on Reddit data.

3.1 Methodology 1

In the first set of experiments, Dataset 1 with 1895 posts has been used for the process. In this method, we are using different BERT models mentioned above for word embeddings. The embedded word vector from the models will be the input for the DNN for classifying text into Suicide versus Depression. Figure 3.1 explains how the old dataset [1] is used in training and testing.

3.2 Methodology 2

The second set of experiments is also conducted on dataset one. Here BERT, DistilBERT, ALBERT, and RoBERTa have been used for both word embedding and classification. Once the models were trained, evaluation metrics such as Accuracy and

[1] Dataset 1

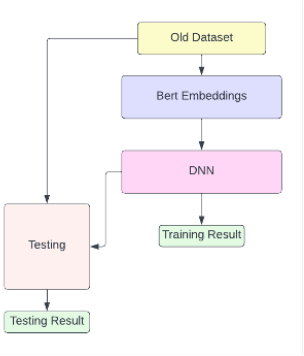


Figure 3.1: Methodology 1

F1-score are calculated on the test data. Figure 3.2 how the different BERT models are trained on the given dataset.

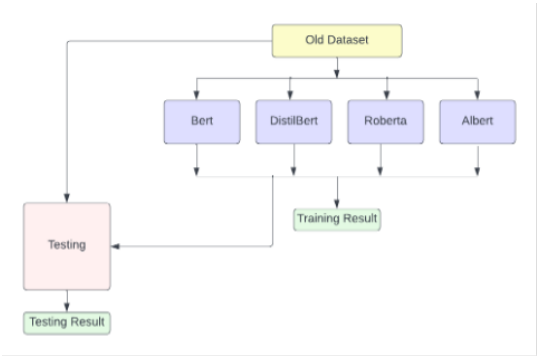


Figure 3.2: Methodology 2

3.3 Methodology 3

For the third method, Dataset two which is very large compared to the dataset one has been used for the training along with the four BERT models mentioned above. BERT models have been used for both word embedding and classification tasks. For testing purposes, we implemented three different testing methods, which include:

- Trained using Dataset two and tested using Dataset one.

- Trained using Dataset two and tested using Dataset two with suicide and depression classes.

- Trained using Dataset two and tested using Datast two with suicide and non-suicidal classes.

From Figure 3.3 we can understand how the New Dataset [2] and the Old Dataset [3] us used for training and testing.

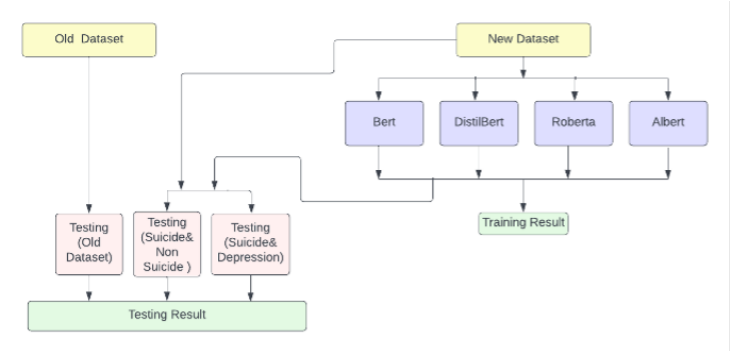


Figure 3.3: Methodology 3

3.4 Model Explainability

Following trials, the model needed to be interpreted, and for this, we used Captum. It is a utility that PyTorch users use to make models more interpretable. Model interpretability techniques have gained importance as a consequence of the rise in model complexity and the resulting lack of transparency. Model understanding is a focus for practical machine learning applications across industries as well as an active field of

[2]Dataset 2
[3]Dataset 1

study. Modern algorithms from Captum, such as Integrated Gradients, give us a simple way to comprehend which characteristics are influencing a model's output. Captum makes it simpler for ML scholars to use interpretability algorithms that can communicate with PyTorch models. Additionally, it enables researchers to compare their work to other existing algorithms that are readily accessible in the library.In model explainability, we have shown the importance given by the model on each word and its contribution to the overall label prediction. Through this we hope to show that the model is able to focus on the important words to predict rather than randomly predicting the right labels.

3.5 Evaluation Metric

The evaluation metric we used to evaluate the performance of the model include Accuracy and F1-score.

3.5.1 Accuracy

The accuracy in general measures the proportion of correct predictions to the total number of instances evaluate . The formula used for calculating accuracy of a model is given in equation 3.1.

$$Accuracy, A = \frac{tp + tn}{tp + tn + fp + fn} \tag{3.1}$$

where,

tp = True positive, Correctly predicted positive classes

tn = True negative, Correctly predicted negative classes

fp = False positive, Incorrectly predicted positive classes

fn = False negative, Incorrectly predicted negative classes

3.5.2 Precision

Precision explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where False Positive is a higher concern than False Negatives.

$$precision = \frac{tp}{tp + fp} \tag{3.2}$$

3.5.3 Recall

Recall explains how many of the actual positive cases we were able to predict correctly with our model. It is a useful metric in cases where False Negative is of higher concern than False Positive.

$$recall = \frac{tp}{tp + fn} \tag{3.3}$$

3.5.4 F1-score

It gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall. The formula used for calculating the F1-score of a model is given in equation 3.4.

$$f1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.4}$$

Chapter 4

Results & Discussion

The experiments were carried out under the following circumstances in Pytorch. The models were trained for 2 epochs, the optimizer used was Adam and the learning rate was set to 2e-5. Since the training was conducted using two GPUs, the per-device batch size was set to 16. The experiment setup was kept the same for all BERT classifiers. The metrics used to evaluate the performance of the models are accuracy, precision, recall and F1-score.

4.1 Methodology 1

As explained in the previous section we use DNNs for classification and the transformers for embedding purposes. These are the results obtained: From this it can be observed

Embedding	**BERT**	**ALBERT**	**RoBERTa**	**DistilBERT**
Train	0.992	0.5124	0.7086	0.987
Validation	0.815	0.5461	0.71	0.696
Testing	0.711	0.51	0.692	0.688

that the DNNs were able to better classify with BERT embeddings than with any other transformer embeddings.

4.2 Methodology2

As explained this methodology uses transformer models for embedding as well as classification.

4.2.1 BERT Large Uncased

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.68	0.73	0.70	0.70
Suicide	0.72	0.66	0.69	

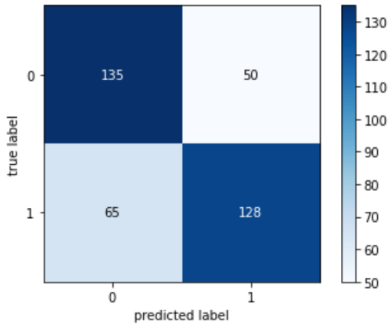


Figure 4.1: Confusion matrix of BERT Large Uncased model

The low accuracy shows that some records are misclassified and an example is shown below:

Text: Everyone is constantly interrupting me and talking over me. Whenever I try talking about my problems no one seems to care and just counteracts it by saying how they have it harder. I mean, I guess they do but I wish I could just talk to someone at school without them saying how they have like so much more work and they got two hours of sleep last night. That's why I hate being nice and supportive sometimes. I listen to other people, I help other people and I feel like it almost never gets returned.

I'm there for other people but it feels like no one is there for me. And it's not only at school. At home, I'm the youngest and it feels like they just ignore what I say, and don't listen to anything I have to say. They are so proud of my sister, and I am too, but they don't care as much about my accomplishments and awards. Maybe it's because they aren't as impressive as hers?

Predicted: *depression*

Actual: *suicide*

4.2.2 DistilBERT Base Uncased

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.64	0.61	0.62	0.64
Suicide	0.64	0.67	0.66	

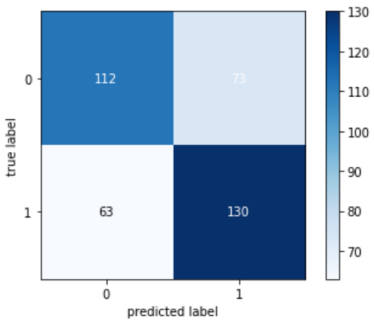


Figure 4.2: Confusion matrix of DistilBERT Base Uncased model

Examples of the misclassified data:

- Text: feel low feel like never good enough hopeless everything self medicating wine well bloody merry go round feel constantly saying fine nothings wrong wear happy smiley mask hide really tablets therapy past hasn worked wonderful man

healthy kids grown seem happy

Predicted: *depression*

Actual: *suicide*

- Text: guess hooking member gang get quicker not want anyone stop offer advice tired giving back fake thank appreciate concern messages

  Predicted: *depression*

  Actual: *suicide*

4.2.3 ALBERT Base

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.72	0.55	0.62	0.67
Suicide	0.65	0.79	0.71	

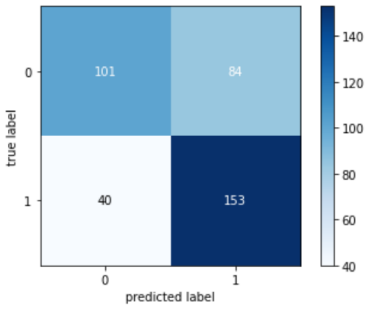


Figure 4.3: Confusion matrix of ALBERT Base model

Examples of the misclassified data:

- Text: Ever since I attempted and failed I have been going downhill in terms of depression. I was feeling more numb than depressed when I attempted. I would've gone to the hospital but I dont have the finances for that. I now have

been falling back into a deep dark depression where I cant get up and cant shower and I'm eating less. I think about death and dying still very frequently, especially if anything goes wrong.

Predicted: *depression*

Actual: *suicide*

- Text: There is utter absence of meaning in my day to day life. I would love nothing more than to feel or do anything other than artificial emotions and automated actions, but alas must take that pill or else it will all crumble, oh and that pill, and maybe these 3 too. It's a fine line between authentic peace and self destruction that I've been on for what feels like a life time, but unfortunately it's only been a year and that life time is all I have to look forward to.

  Predicted: *suicide*

  Actual: *depression*

4.2.4 RoBERTa Large

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.72	0.65	0.71	0.72
Suicide	0.72	0.75	0.73	

An example of misclassified data:

Text: Give me one good reason I should not kill myself right now. The whole world is going to shit and no one can change that. The only people that could help refuse to. The world will never be a good place, it will always be on the brink. Everyone

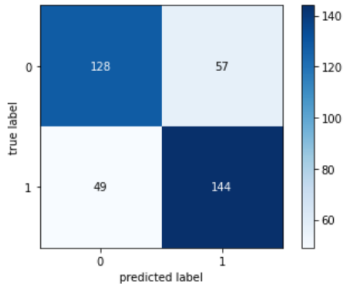


Figure 4.4: Confusion matrix of RoBERTa Large model

always says to just look on the bright side, but what is the point? For every good thing you could say, I could name off 3 bad things. Why shouldn't I just end it right now and save myself the trouble. Everyone dies eventually, there is no point in being afraid of the inevitable. I house I can't afford will be underwater anyways, if were not all dead by the time I need one. Everyone always says that I shouldn't kill myself because people would be sad. Welp, I'm not a good person. I'm selfish, rude, hopeless, and just annoying in general. As is everyone. To me there is no point in prolonging the inevitable. So please if you have a good actual answer that is not just "some people would be inconvenienced" than please tell me.

Predicted: *suicide*

Actual: *depression*

4.3 Methodology 3

This methodology tested the model with 3 sets of data. Initially, the model was trained on the larger SuicideWatch dataset which has around 300,000 values. Once the models

were trained, 3 methods were used as shown below. The 1st method used the initial dataset, the dataset with inadequate data, to test the model on. Next, the models were tested on the test set of this data itself. Finally, the models were tested on another dataset which consisted of Suicide and Non-Suicide data as classes.

4.3.1 BERT Base Uncased

4.3.1.1 Method 1

This method uses Dataset 1 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.84	0.76	0.80	0.81
Suicide	0.79	0.86	0.812	

4.3.1.2 Method 2

This method uses Dataset 2 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.94	0.93	0.94	0.92
Suicide	0.87	0.89	0.88	

4.3.1.3 Method 3

This method uses Dataset 1 to classify between Non-Suicide and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Non-Suicide	0.90	0.95	0.99	0.94
Suicide	0.99	0.89	0.94	

4.3.2 DistilBERT Base Uncased

4.3.2.1 Method 1

This method uses Dataset 1 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.79	0.75	0.77	0.78
Suicide	0.77	0.81	0.79	

4.3.2.2 Method 2

This method uses Dataset 2 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.89	0.99	0.94	0.94
Suicide	0.99	0.88	0.93	

4.3.2.3 Method 3

This method uses Dataset 1 to classify between Non-Suicide and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Non-Suicide	0.90	0.99	0.94	0.94
Suicide	0.99	0.88	0.93	

4.3.3 ALBERT Base Uncased

4.3.3.1 Method 1

This method uses Dataset 1 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.70	0.82	0.76	0.74
Suicide	0.79	0.66	0.72	

4.3.3.2 Method 2

This method uses Dataset 2 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.93	0.92	0.92	0.90
Suicide	0.85	0.87	0.86	

4.3.3.3 Method 3

This method uses Dataset 2 to classify between Non-Suicide and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Non-Suicide	0.89	1.00	0.94	0.93
Suicide	1.00	0.87	0.93	

4.3.4 RoBERTa Base Uncased

4.3.4.1 Method 1

This method uses Dataset 1 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	0.81	0.80	0.80	0.81
Suicide	0.81	0.82	0.81	

4.3.4.2 Method 2

This method uses Dataset 2 to classify between Depression and Suicide

Classes	Precision	Recall	F1-Score	Accuracy
Depression	1.00	0.50	0.67	0.67
Suicide	0.50	1.00	0.67	

Classes	Precision	Recall	F1-Score	Accuracy
Non-Suicide	1.00	1.00	1.00	1.00
Suicide	1.00	1.00	1.00	

4.3.4.3 Method 3

This method uses Dataset 2 to classify between Non-Suicide and Suicide

From the results, it can be seen that in most cases the Suicide vs non-Suicide data is able to get a higher accuracy than Suicide vs Depression. This is because the text in Suicide vs Non-Suicide is visually far more distinguishable than the text in the Suicide vs Depression dataset. The below figure 4.5 is a comparison of the accuracies of all the models used in comparison to all the datasets used.

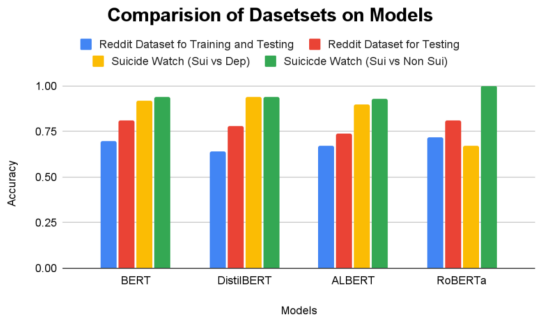


Figure 4.5: A comparison of the accuracies of the different datasets on multiple models

Figure 4.5 shows that RoBERTa is able to properly classify Suicide from Non-Suicide, but it struggles with Suicide and Depression classes, while BERT and DistilBERT models equally perform on Suicide vs Depression and Suicide vs Non-Suicide datasets. Of BERT and DistilBERT, since DistilBERT took a lower amount of time to train the model, model explainability was done on DistilBERT and a few examples of the results can be seen below.

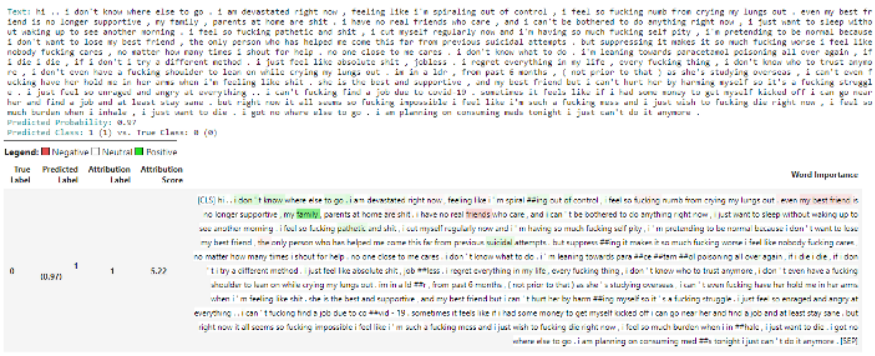


Figure 4.6: Model Explainability Result 1

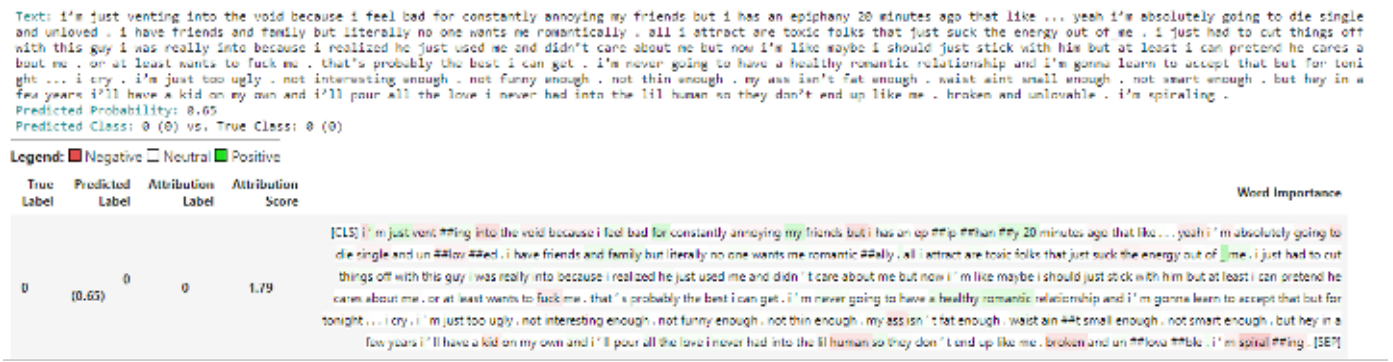


Figure 4.7: Model Explainability Result 2

Of the many texts tested on, these are a few results. It can be seen that the words contributing to Suicide and Depression are highlighted.

Chapter 5

Conclusion and Future Work

In this project, we address the performances of transformer-based models such as BERT, and DistilBERT along with others on different datasets. We were able to find that by providing more data, models are able to perform much better, as shown in the results. We can also conclude that models were much better at classifying Suicide vs Non-Suicide than Suicide vs Depression. This is mainly because the text in Suicide vs Non-Suicide is far more distinguishable than Suicide vs Depression. Based on the results of the model explainability experiments, we were also able to confirm why the model made the predictions that it did. In the future, different word embeddings can be experimented with in order to inflate the accuracies of different models. Larger models can also be used as the computing infrastructure for our current experiments could not work with large models. A pipeline where embeddings such as GUSE and other embeddings' output can be passed into a transformer model such as BERT to be classified could also be introduced.

Bibliography

[1] Vasu Agarwal et al. "Analysis of classifiers for fake news detection". In: *Procedia Computer Science* 165 (2019), pp. 377–383.

[2] Tanvirul Alam, Akib Khan, and Firoj Alam. "Bangla text classification using transformers". In: *arXiv preprint arXiv:2011.04446* (2020).

[3] Edwin D Boudreaux et al. "Applying machine learning approaches to suicide prediction using healthcare data: overview and future directions". In: *Frontiers in psychiatry* 12 (2021), p. 707916.

[4] Hila Chefer, Shir Gur, and Lior Wolf. "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 397–406.

[5] Ian C Covert, Scott Lundberg, and Su-In Lee. "Explaining by removing: A unified framework for model explanation". In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 9477–9566.

[6] Nhan Cach Dang, Marıa N Moreno-Garcıa, and Fernando De la Prieta. "Sentiment analysis based on deep learning: A comparative study". In: *Electronics* 9.3 (2020), p. 483.

[7] Shafie Gholizadeh and Nengfeng Zhou. "Model explainability in deep learning based natural language processing". In: *arXiv preprint arXiv:2106.07410* (2021).

[8] Leilani H Gilpin et al. "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[9] Lang He et al. "Deep learning for depression recognition with audiovisual cues: A review". In: *Information Fusion* 80 (2022), pp. 56–86.

[10] Loukas Ilias and Dimitris Askounis. "Explainable identification of dementia from transcripts using transformer networks". In: *IEEE Journal of Biomedical and Health Informatics* 26.8 (2022), pp. 4153–4164.

[11] Shaoxiong Ji et al. "Suicidal ideation detection: A review of machine learning methods and applications". In: *IEEE Transactions on Computational Social Systems* 8.1 (2020), pp. 214–226.

[12] Narine Kokhlikyan et al. “Captum: A unified and generic model interpretability library for pytorch”. In: *arXiv preprint arXiv:2009.07896* (2020).

[13] Hui Liu, Qingyu Yin, and William Yang Wang. “Towards explainable NLP: A generative explanation framework for text classification”. In: *arXiv preprint arXiv:1811.00196* (2018).

[14] Mitchell Naylor et al. “Quantifying explainability in nlp and analyzing algorithms for performance-explainability tradeoff”. In: *arXiv preprint arXiv:2107.05693* (2021).

[15] Yaakov Ophir et al. “Deep neural networks detect suicide risk from textual facebook posts”. In: *Scientific reports* 10.1 (2020), p. 16685.

[16] Ahmed Husseini Orabi et al. “Deep learning for depression detection of twitter users”. In: *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. 2018, pp. 88–97.

[17] SJ Pachouly et al. “Depression Detection on Social Media Network (Twitter) using Sentiment Analysis”. In: *Int. Res. J. Eng. Technol* 8 (2021), pp. 1834–1839.

[18] Andrew Poulton and Sebas Eliens. “Explaining transformer-based models for automatic short answer grading”. In: *Proceedings of the 5th International Conference on Digital Technology in Education*. 2021, pp. 110–116.

[19] Rohith Ramakrishnan et al. “Analysis of Text-Semantics via Efficient Word Embedding using Variational Mode Decomposition.” In: *PACLIC*. 2021, pp. 711–720.

[20] Krithik Ramesh and Yun Sing Koh. “Investigation of Explainability Techniques for Multimodal Transformers”. In: *Data Mining: 20th Australasian Conference, AusDM 2022, Western Sydney, Australia, December 12–15, 2022, Proceedings*. Springer. 2022, pp. 90–98.

[21] Emily Schriver et al. “Identifying risk factors for suicidal ideation across a large community healthcare system”. In: *Journal of affective disorders* 276 (2020), pp. 1038–1045.

[22] Ivan Sekulić and Michael Strube. “Adapting deep learning methods for mental health prediction on social media”. In: *arXiv preprint arXiv:2003.07634* (2020).

[23] Ruba Skaik and Diana Inkpen. “Using social media for mental health surveillance: a review”. In: *ACM Computing Surveys (CSUR)* 53.6 (2020), pp. 1–31.

[24] S Smys and Jennifer S Raj. “Analysis of deep learning techniques for early detection of depression on social media network-a comparative study”. In: *Journal of trends in Computer Science and Smart technology (TCSST)* 3.01 (2021), pp. 24–39.

[25] Chenghao Yang et al. “Efficient Shapley Values Estimation by Amortization for Text Classification”. In: ().

[26] Julia El Zini and Mariette Awad. "On the explainability of natural language processing deep models". In: *ACM Computing Surveys* 55.5 (2022), pp. 1–31.

[27] Bahman Zohuri and Siamak Zadeh. "The utility of artificial intelligence for mood analysis, depression detection, and suicide risk management". In: *Journal of Health Science* 8.2 (2020), pp. 67–73.