# Detection of Suicidal Ideation Based on Social Media Using Machine Learning

Mahmuda Akter
ID: 19103124
And
Rayhan Mahmud
ID: 19103123

A Thesis in the Partial Fulfillment of the Requirements

for the Award of Bachelor of Computer Science and Engineering (BCSE)


Department of Computer Science and Engineering
College of Engineering and Technology
IUBAT – International University of Business Agriculture and Technology

Summer 2022

# Detection of Suicidal Ideation Based on Social Media Using Machine Learning

Mahmuda Akter
ID: 19103124
And
Rayhan Mahmud
ID: 19103123

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)

The thesis has been examined and approved,

_____

Prof. Dr. Utpal Kanti Das
Chairman and Professor
Dept. of Computer Science and Engineering

_____

Dr. Muhammad Hasibur Rashid Chayon
Associate Professor & Coordinator

_____

Ehsan Ahmed Niloy
Supervisor

Department of Computer Science and Engineering
College of Engineering and Technology
IUBAT – International University of Business Agriculture and Technology

Summer 2022

## Abstract

A significant public health issue is suicidal behavior. Suicidal ideas are frequently communicated through social media platforms. In order to inform suicide research and policy, there is a great deal of interest in tracking such communications for both population-wide and personal prevention goals. For this work looks towards the modeling and detection of suicidal in Natural Language Processing methods using unsupervised text ideation. In order to identify suicide intent in social media posts, we are collecting various types of datasets and implementing the dataset. Natural language documents have a variety of groups of data, tasks, and topics and are typically unstructured, which makes it difficult for Natural Language Processing to perform satisfactorily. As a result, we have to use machine learning techniques; one may collect semantic and syntactic information about words and look for additional words. Methods like Word2Vec, Glove are widely employed for machine learning model initialization due to their better efficiency when compared to random initialization. This study intends to develop a predictive algorithm and we are using two algorithms Word2Vec and Glove. In this section, we want to extract representations using the Word2Vec and Glove algorithms. Custom Word2Vec embedding will be pre-trained using the training dataset, and a readily available Glove embedding will be used. Suicide or not committing suicide is a binary variable in the issue statement that we are trying to predict. We used three models: CNN, BERT, and ELECTRA that include machine learning. Before being evaluated on the test dataset and the validation dataset, the models were improved on the train dataset. Accuracy, Precision, Recall, and F1 score will be used to evaluate the models performance on the test dataset. Using three models, the most important outcomes of the models are specially trained Word2Vec. We've decided to choose ELECTRA as our ultimate model because it scores high compared to the other two models. After that, we can put any kind of text on our system and the system give the result either it suicidal or non-suicidal and the probability. We aim to further highlight the significance of our suicide detection algorithm by integrating it into a practical Chabot designed to build close relationships with youngsters and guide them towards professional resources when signs of distress are observed. The help of our project, we were able to identify suicidal ideation from social media posts and develop models.

# Letter of Transmittal

3rd September 2022

The Chairman

Thesis Defense Committee

Department of Computer Science and Engineering

IUBAT– International University of Business Agriculture and Technology

4 Embankment Drive Road, Sector 10, Uttara Model Town

Dhaka 1230, Bangladesh

**Subject:** Letter of Transmittal.

Dear Sir,

With due respect, this is our pleasure to present our thesis report entitled— "Detection of Suicidal Ideation Based on Social Media Using Machine Learning", we have prepared this report as partial fulfillment of the thesis. We have tried our level best to prepare this project to the required standard. It was certainly a great opportunity to work on this paper to actualize our theoretical knowledge in the practical arena. Now, we are looking forward to your kind appraisal regarding this thesis report. We will remain deeply grateful to you if you kindly go through this report and evaluate our performance. We hope that you would find the report comprehensive and competent.

Yours sincerely,

_____

Mahmuda Akter

ID: 19103124

_____

Rayhan Mahmud

ID: 19103123

## Student's Declaration

I am Mahmuda Akter and my partner Rayhan Mahmud, we are student of IUBAT– International University of Business Agriculture and Technology, declaring that this thesis paper on the stated topic has only been prepared for the fulfillment of "Thesis" as the partial fulfillment of—Bachelor of Computer Science and Engineering degree.

It has not been prepared for any other purposes, rewards, or presentation.

Sincerely yours,

_____

Mahmuda Akter

ID: 19103124

_____

Rayhan Mahmud

ID: 19103123

## Supervisor's Certification

This is to certify that Thesis report on "Detection of Suicidal Ideation Based on Social Media Using Machine Learning" has been carried out by Mahmuda Akter bearing ID: 19103124 and Rayhan Mahmud ID: 19103123 of IUBAT-International University of Business Agriculture and Technology as a partial fulfillment of the requirement of Thesis course. The report has been prepared under my guidance and is a record of the bona-fide work carried out successfully. To the best of my knowledge and as per their declaration, no parts of this report have been submitted anywhere for any degree, diploma or certificate.

Now, they are permitted to submit the report. I wish them all success in their future endeavors.

Thesis Supervisor

_____

Ehsan Ahmed Niloy

Lecturer,

Department of Computer Science and Engineering

IUBAT- International University of Business Agriculture and Technology

## Acknowledgement

By the grace of Allah who is the most merciful and the most graceful, it's our pleasure to take this occasion to thank a few people, who have assisted, encouraged, directed and supported us throughout our Thesis program. Firstly, we want to thank our family, who has endowed their immeasurable-innumerable support and encouragement to attain this exquisite event of our life. We are very appreciative to Dr. Muhammad Hasibur Rashid Chayon, Associate Professor & Coordinator of Department of Computer Science and Engineering, IUBAT- International University of Business Agriculture and Technology for his unrelenting direction and sustain throughout the semester. We would like to pay our gratitude to our thesis advisor Eahsan Ahmed Niloy, Lecturer of Computer Science and Engineering Department, who has given us the opportunity to make such a report not only in this semester but also throughout our education life at IUBAT by giving his valuable suggestions and advice at any time, at any situation. We would be able to make this report effectively and properly only for his right direction.

Sincerely yours,

_____

Mahmuda Akter

ID: 19103124

_____

Rayhan Mahmud

ID: 19103123

# Table of Contents

# List of Figures

# List of Tables

# 1. Introduction

According to the Bangladesh Mental Health Study conducted by the National Institute of Mental Health (NIMH), depression is the most common mental disorder. The prevalence of mental disorders overall among people 18 years and older was 18.7%, according to a 2018 mental health survey jointly performed by the government and the National Institute of Mental Health (NIMH). The prevalence of diseases was 20.2% higher in people 60 years and older. The Bangladesh Bureau of Statistics estimates that 10,000 persons commit suicide in Bangladesh every year on average. Due to hormonal changes brought on by puberty or high levels of stress brought on by societal and academic expectations, young people are more likely to experience depression. The coronavirus disease 2019 (COVID-19) pandemic has wreaked havoc on the mental health of the Bangladeshi population. One study has found that the prevalence of depressive (57.9%), stress (59.7%) and anxiety (33.7%) symptoms in the adult population is now much higher than pre-pandemic rates. According to a report released last month, 14,436 persons in Bangladesh have committed suicide since March 2020, which is 70% more than the coronavirus was responsible for during same time. 49 percent of suicide victims, according to the report, were between the ages of 20 and 35.

Youths are more prone to talk about suicidal ideas online since they are frequent users of social media and digital technology. As a result, it is urgent to create big data systems that are effective in extracting suicidal intent from social media data using cutting-edge Natural Language Processing (NLP) methods.

In order to identify suicide intent in social media posts, this study intends to develop a predictive algorithm. We intend to incorporate our suicide detection model into a practical chatbot created to develop close relationships with teenagers and point them toward professional services when indicators of distress are noticed in order to emphasize the model's relevance even more.

# 2. Literature Review

- Using machine learning Algorithm to Detect suicide Risk Factors on Twitter.

  In this paper, they first applied three topic clustering algorithms in this study to discover topics discussed by users on Twitter, including Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF). Second, we explored cluster formations and user classifications into “HighRisk" or “AtRisk” using the identified topics. While LSA is a matrix factorization strategy that rotates topic vectors in the term space to best capture the diversity of terms, LDA is a probabilistic approach that maximizes the log likelihood of each term appearing in each topic. However, LSA generates themes that contain phrases that are inversely associated with those subjects; as a result, these results are difficult to understand. NMF, a nonnegative rank factorization approach, is used to accommodate for this drawback in order to make sure the themes are better at being understandable and well-separated. Two machine learning techniques were used to categorize users into "HighRisk" or "AtRisk" groups: Decision Tree classification and K-means clustering.

  This suggests that there may be information missing from the dataset utilized for this study that reflects the ideation of the other 5 suicide risk topics. Additionally, the dataset excludes the use of emoticons, hashtags, and other twitter meta-data that may signal suicidal intent. Additionally, there is a problem with the selection of tweets and people who are using sarcasm or making jokey statements in contrast to users who actually intend to commit suicide [1].

- An Unsupervised Learning Approach for Automatically to Categorize Potential Suicide Messages in Social Media.

  In this study, the authors suggest using unsupervised learning to automatically classify social media posts about suicide. Their approach is based on a conventional clustering algorithm that uses the semantic relations in the text of social media postings as input. The message can be categorized into several suicide alert levels using clustering results.

They collected data from four social networks such as Twitter, Instagram, and also from blogs and forums. Algorithms they used - Unsupervised data, Data acquisition and pre-processing, Semantic similarity measures, Clustering.

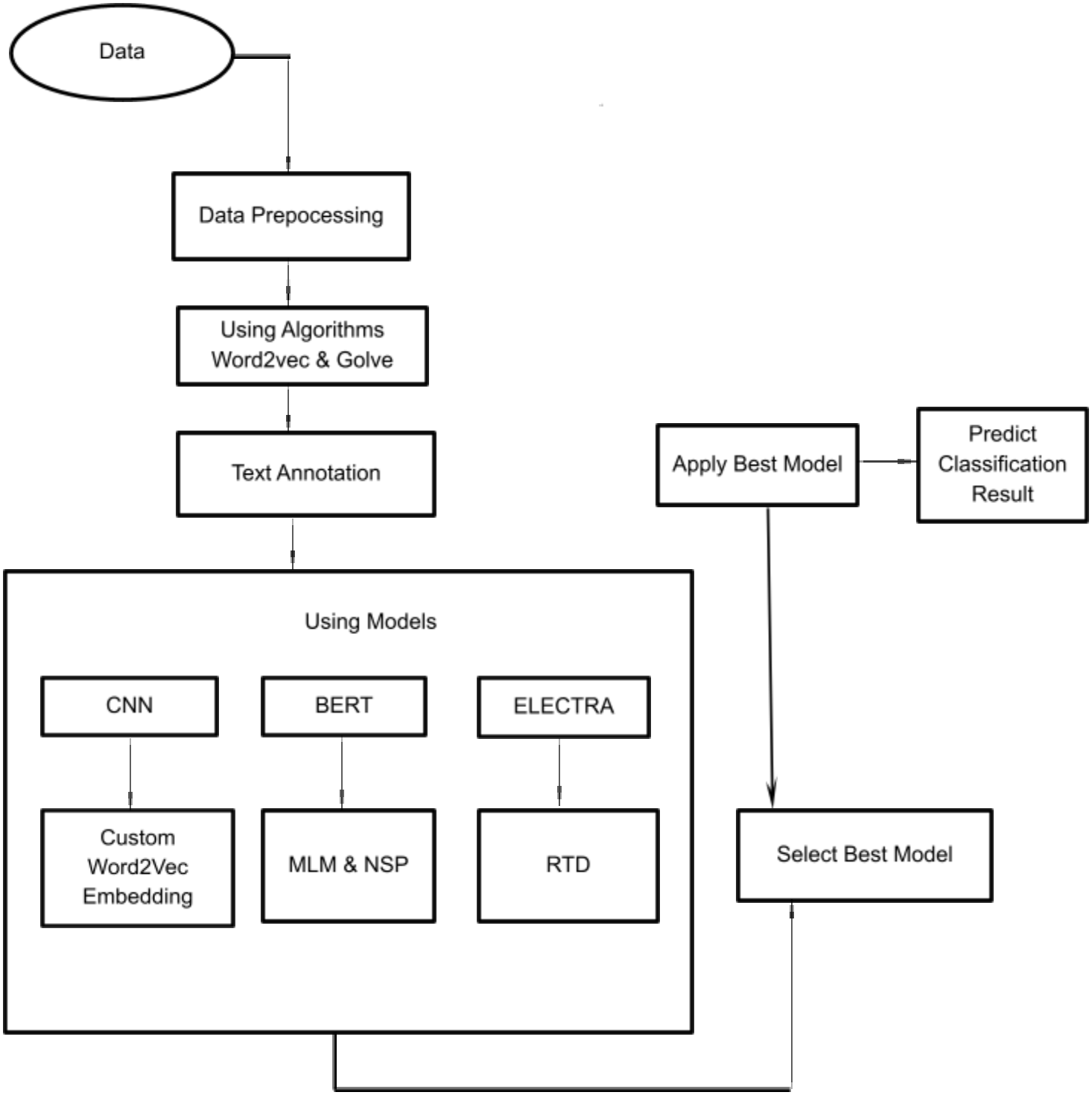
They have the failings of using more robust similarity metrics and clustering algorithms to improve match rates. They didn't use additional texts that are evenly distributed by category [2].

- Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks.

  In this paper, the four main parts of this work's methodology are as follows: the first is the requirements for manually creating the vocabulary associated with the suicide theme; the second is the gathering of Twitter data; the third is the requirements for automatically classifying data using machine learning algorithms implemented in Weka18; and the fourth is the requirements for a semantic analysis of these sentiments to enhance our results.

  They need to improve and develop their techniques for their method to be more accurate [3].

# 3. Research Methodology



# 4. Data Collection and Difference

**4.1 Data Collection:** The Suicide and Depression Detection dataset, which includes posts from the social media site Reddit, was acquired from Kaggle for use in this project. [25].

| index | text | class |
|---|---|---|
| 1 | Am I weird I don't get affected by compliments if it's coming from someone I know irl but I feel really good when internet strangers do it | non-suicide |
| 2 | Finally 2020 is almost over... So I can never hear "2020 has been a bad year" ever again. I swear to fucking God it's so annoying | non-suicide |
| 3 | i need helpjust help me im crying so hard | suicide |
| 4 | I'm so lostHello, my name is Adam (16) and I've been struggling for years and I'm afraid. Through these past years thoughts of suicide, fear, anxiety I'm so close to my limit . I've been quiet for so long and I'm too scared to come out to my family about these feelings. About 3 years ago losing my aunt triggered it all. Everyday feeling hopeless , lost, guilty, and remorseful over her and all the things I've done in my life,but thoughts like these with the little I've experienced in life? Only time I've revealed these feelings to my family is when I broke down where they saw my cuts. Watching them get so worried over something I portrayed as an average day made me feel absolutely dreadful. They later found out I was an attempt survivor from attempt OD(overdose from pills) and attempt hanging. All that happened was a blackout from the pills and I never went through with the noose because I'm still so afraid. During my first therapy I was diagnosed with severe depression, social anxiety, and a eating disorder. I was later transferred to a fucken group therapy for some reason which made me feel more anxious. Eventually before my last session with a 1 on 1 therapy she showed me my results from a daily check up on my feelings(which was a 2 - step survey for me and my mom/dad ) Come to find out as I've been putting feeling horrible and afraid/anxious everyday , my mom has been doing I've been doing absolutely amazing with me described as "happiest she's ever seen me, therapy has helped him" I eventually was put on Sertaline (anti anxiety or anti depression I'm sorry I forgot) but I never finished my first prescription nor ever found the right type of anti depressant because my mom thought I only wanted the drugs so she took me off my recommended pill schedule after ~3 week and stopped me from taking them. All this time I've been feeling worse afraid of the damage/ worry I've caused them even more. Now here with everything going on, I'm as afraid as I've ever been . I've relapsed on cutting and have developed severe insomnia . Day after day feeling more hopeless, worthless questioning why am I still here? What's my motivation to move out of bed and keep going? I ask these to myself nearly every night almost having a break down everytime. Please Please Please someone.. anyone help me. I'm so scared I might do something drastic, I've been shaped by fear and anxiety. Idk what to do anymore | suicide |

Figure 01: Sample Rows of Original Dataset

The dataset is divided into two columns, as shown in Figure 1 above, where "text" denotes the content of the posts and "class" denotes their labels. To create this dataset, 232,074 posts from the SuicideWatch and teenagers subreddits of Reddit were scraped. SuicideWatch posts were classified as such, whereas the posts gathered from teenagers were classified as non-suicidal.

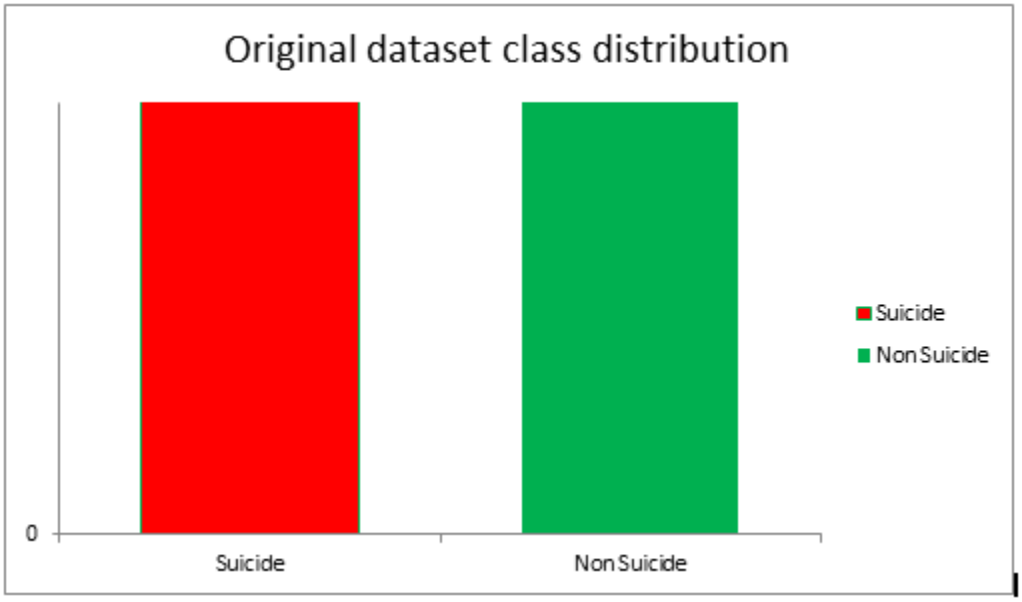



Figure 02: Original dataset class distribution

As shown in Figure 2 above, where each class contains 116,037 rows, or 50% of the dataset, the classes are equally distributed.

**4.2 Text Preprocessing:** To make the text data suitable formats for the ensuing model building, the text data needs to be preprocessed. Social media data typically requires more individualized preprocessing and cleaning procedures because it is more unstructured. So, using the steps listed below in the order shown in Figure 3 below, our data was cleaned.

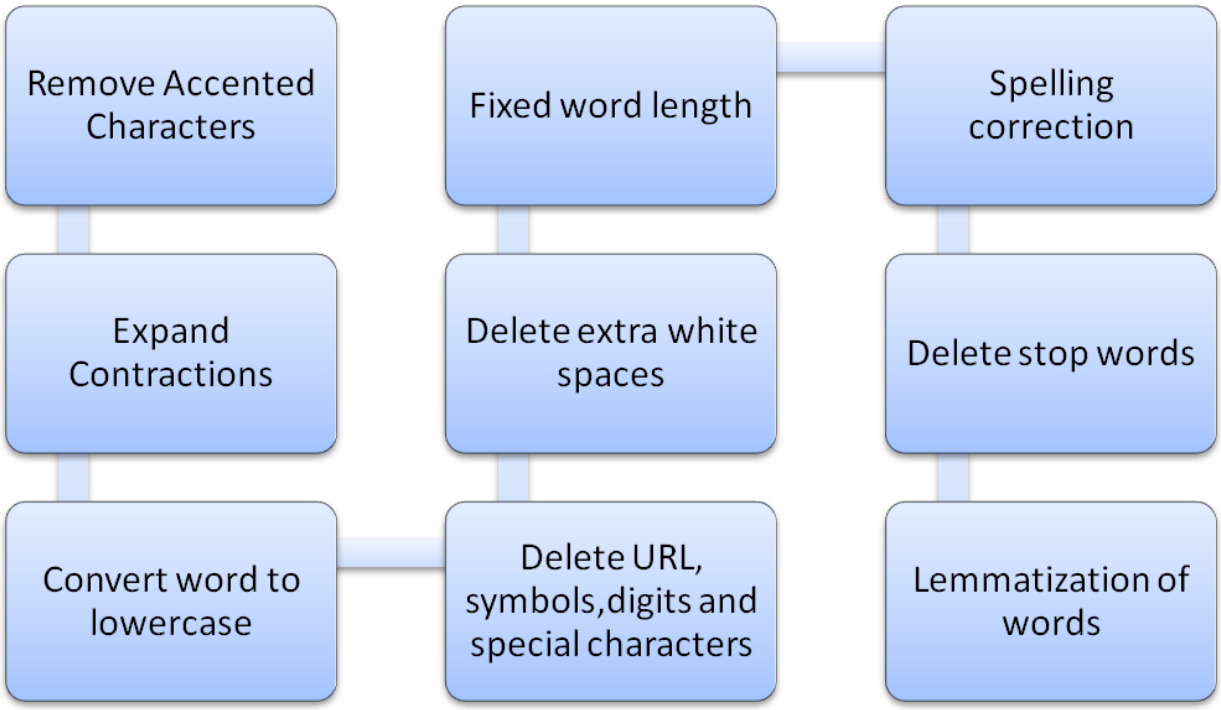



Figure 03: Data preprocessing steps

**Remove Accented Characters:** The meaning of words like "cafe" and "café" is the same. Accent characters are thus eliminated in order to portray related terms as one word.The vocabulary size of the dataset is decreased during this step, which also acts as a dimensionality reduction step.

**Expand contractions:** In contractions, some letters are left out of words to make them shorter. A contracted version of "I have" is the word "I've," for instance.

**Convert to lowercase:** Words with different cases, like "Boy" and "boy," should be represented as one and the same word because they have the same meaning. This also aids in lowering the dimensionality of the data by making all words lowercase.

**Delete URLs, symbols, digits, special characters, and extra whitespaces:** These characters have been deleted because they don't pertain to our use case and aren't relevant to our model. These components are eliminated in the following order: left to right.

**Fix word length:** Characters that are incorrectly repeated cause words to lengthen. For instance, the word "good" can be extended to "goooood." In order to prevent inaccurate information from being incorporated into our prediction model, additional unnecessary characters must be removed because English words can only contain a maximum of two repeated characters.

**Spelling correction:** Spelling is checked using the Symspell algorithm. It is based on the faster-than-normal spelling n packages Symmetric Delete Spelling Correction algorithm. This aids in separating two words that were accidentally combined because whitespace was missed. For instance, the word "help just" is changed to "help just" with a space to denote the word's boundaries.

**Delete stopwords:** Stopwords are words that are frequently used but have no real significance or meaning. Examples of this kind include "a," "that," and "the." Despite appearing on the list of frequently used stopwords, "no" and "not" have been kept because they imply a negative intonation that may be helpful when categorizing suicidal posts.

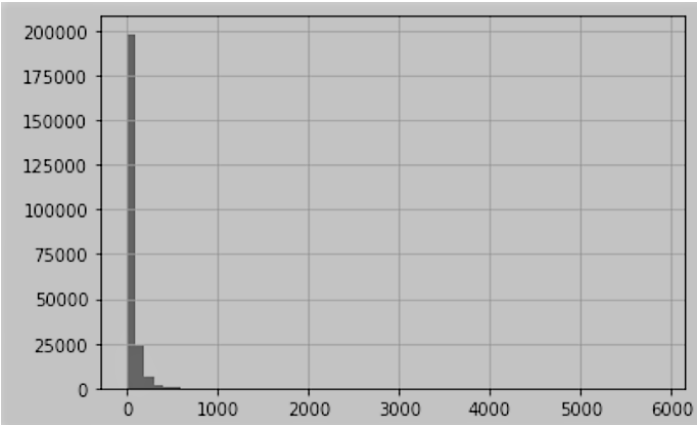
**Lemmatization of words:** The process of lemmatization involves reverting inflected words to their root form. For instance, when fed into models, the terms "eat," "eating," and "eaten" could all be represented by the same term because they all have the same meaning. Lemmatization, as opposed to stemming, examines the morphology of the words and takes into account their contextual meaning. Lemmatization is preferred over stemming in situations where, for instance, the lemma "better" is "good" cannot be recognized by stemming.

### 4.3 Data Cleaning:

**Remove Irrelevant Words**: We next performed some preliminary data analysis on the frequency of each word in the preprocessed posts. The word "filler," which had 55,442 occurrences among the top 30 most frequent words, caught our attention. Further investigation has led us to the conclusion that this token has no meaning and was probably just noise that was recorded during the data collection phase. We then went ahead and deleted this word from all posts.

**Remove Empty Rows after Preprocessing:** There were roughly 60 empty text fields with no words in them after text preprocessing and the removal of unnecessary words. These related rows were removed.

**Remove Outlier Rows with High Word Count:** The word count of the posts that had been preprocessed was then obtained. We noticed outliers in the word count with astronomically high numbers in Figure 4 above. The maximum word count of 5,850 differs significantly from the 75th percentile word count of 62. We have chosen to use the 75th percentile as a threshold for word count where processed posts with more than 62 words are removed in order to optimize for model training in later stages. We can focus on model training on shorter posts and shorten training time by subsetting the dataset.




| | |
|---|---|
| count | 232009.000000 |
| mean | 53.143029 |
| std | 89.614739 |
| min | 1.000000 |
| 25% | 12.000000 |
| 50% | 25.000000 |
| 75% | 62.000000 |
| max | 5850.000000 |

Figure 4. Histogram and Distribution of Word Count in Posts

**Final Dataset:**

| index | text | class | cleaned_text |
|---|---|---|---|
| 1 | Am I weird I don't get affected by compliments if it's coming from someone I know irl but I feel really good when internet strangers do it | non-suicide | weird not affect compliment come know girl feel good internet stranger |
| 2 | Finally 2020 is almost over... So I can never hear "2020 has been a bad year" ever again. I swear to fucking God it's so annoying | non-suicide | finally hear bad year swear fuck god annoying |
| 3 | i need helpjust help me im crying so hard | suicide | need help help cry hard |

Figure 5. Sample Rows of Cleaned Dataset

According to Figure 5 above, the final cleaned dataset has 174,436 rows, with the processed text fields being located in the "cleaned text" column.

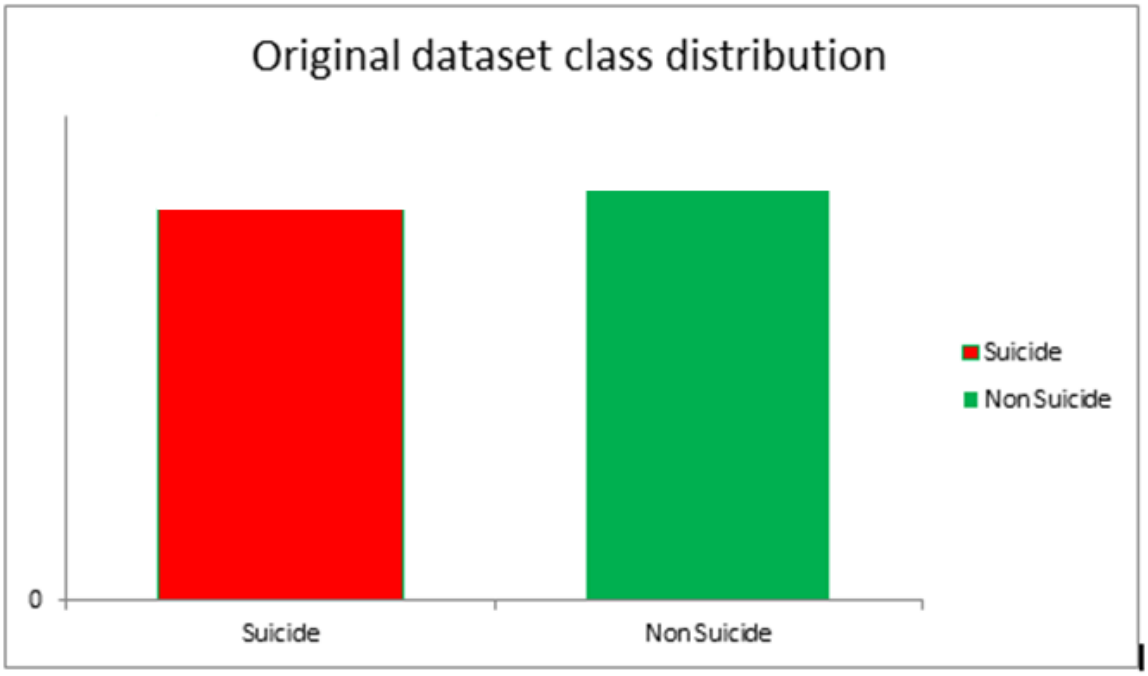



Figure 6. Cleaned Dataset Class Distribution

Figure 6 above displays the cleaned dataset's class distribution. The 5:5 ratio between suicidal and non-suicidal text has changed to a roughly 5:3 ratio when compared to the original dataset class distribution (see Figure 2 above). This finding suggests that more suicidal posts had longer word counts, which were then deleted as part of the data cleaning process. Although the class distribution is slightly unbalanced, this is still considered a typical classification issue, and model performance will not be impacted.

## 5. Data Exploration

With a ratio of 8:1:1, we first divided the dataset into the train, test, and validation sets. Then we further zoomed into each class while conducting data exploration on the train set.

**Word Cloud**

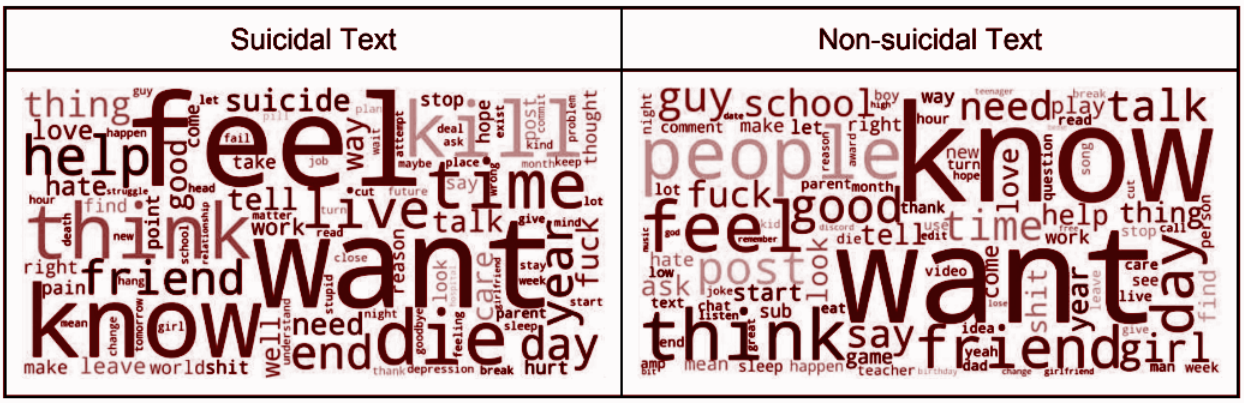



Figure 7. Word Cloud for Suicidal and Non-suicidal Text

First, we looked at the top 100 most used words, which are shown in Figure 7 above as word clouds. We noticed frequently occurring verbs like "feel," "know," and "want" in word clouds for both suicidal and non-suicidal texts. It's interesting to note how many suicidal texts contain words like "kill," "die," "hate," "end," and "help." Contrarily, non-suicidal texts contained a greater number of neutral and uplifting words, such as "good," "people," "friend," "girl," and "guy." According to this observation, suicidal texts typically express more negative emotions than non-suicidal texts.

# 6. Algorithm Selection

The term "representation learning" in the context of NLP refers to a collection of methods for converting unprocessed textual data into a form that can be used for machine learning tasks and is computationally effective. Natural language texts are frequently unstructured and have a variety of granularities, tasks, and domains, which makes it difficult for NLP to perform satisfactorily. As a result, word embeddings are frequently employed to represent words as a dense vector. They make it possible for machine learning classifiers to gather semantic and syntactic data about words and find other words that occur in similar contexts. Due to their higher effectiveness when compared to random initialization, models like Word2Vec, GloVe, and fastText are frequently used for deep learning model initialization. Our model will have to learn the word embeddings from scratch during training if pre-trained word embeddings are not loaded into the embedding layer, which could be difficult for the following reasons. First off, there's a chance that our training set will be deficient in rare terms. Due to a lack of knowledge, learned embeddings may not appropriately represent these words. Second, because there are so many trainable parameters, learning the embeddings from scratch will greatly lengthen our training procedure.

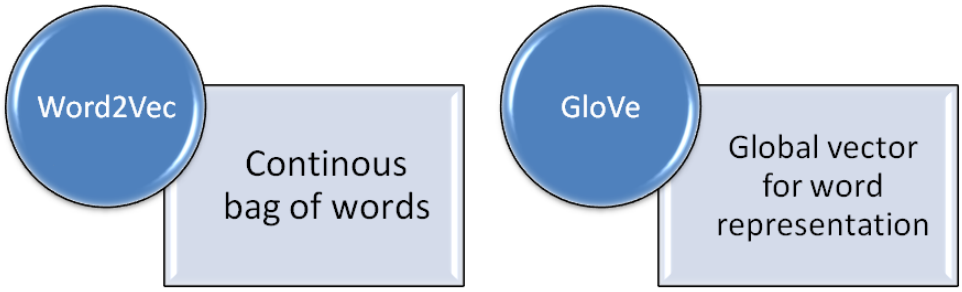



Figure 8. Representations Built

Using the Word2Vec and GloVe algorithms, we hope to extract representations in this section. Using the training dataset, custom Word2Vec embeddings will be pre-trained, and an easily accessible GloVe embedding will be employed (as seen in Figure 8 above).[4]

**6.1 Word2Vec:** With the use of Word2Vec embeddings, words can be represented densely in vector form while also revealing information about their semantic significance. Because it

groups together vectors of related words and infers the meaning of each word based on where it appears in the text, it is effective. Word connections with other terms in the corpus are provided by the estimations [6].

**Algorithm:** In particular, Word2Vec is a neural network-based prediction-based method that uses two algorithms: continuous bag-of-words (CBOW) and skip-gram (SG). While SG uses the target word as input and tries to predict the context words before and after it, CBOW seeks to predict a target word based on the list of context words. The three major components of Word2Vec are a vocabulary builder, a context builder, and a neural network. To create a corpus, the vocabulary builder uses the raw sentence data to extract special words. The context builder then turns the words to vectors by taking into account all of the words in the context window that include the target term. Last but not least, Word2Vec will train a neural network with a single hidden layer, where the number of neurons reflects the dimensions of the embedding[5].

**Implementation:** When a dataset is too small to provide useful bespoke embeddings, generic language use cases or pre-trained word off-the-shelf embeddings are often employed. Our train dataset's about 20,400 vocabulary and 139,500 data points should make for a favorable learning setting for a specially trained embedding layer.

**6.2 GloVe:** Another unsupervised learning approach to produce word vector representation is Global Vectors for Word Representation (GloVe). It combines both local and global context by using the word co-occurrence matrix, and unlike Word2Vec.

**Algorithm:** GloVe is a log-bilinear model that blends the logic of a count-based model with the linear structure used by techniques like Word2Vec. It is based on the ratios of probabilities from the co-occurrence matrix. Its weighted least-squares aim minimizes the difference between the vectors of two words' dot products and the logarithm of their chance of occurring together. The co-occurrence matrix is a V V matrix, where V is the vocabulary size. Given that the matrix will be enormous, we factorize it by minimizing a "reconstruction loss," which looks for lower-dimensional representations that can account for the majority of the variation in the high-dimensional data. We can separate relevant terms from irrelevant words using the matrix's ratio of probabilities, and we can further discriminate between the two relevant words.

**Implementation:** We have incorporated pre-trained GloVe word embeddings into our model to better compare various word embeddings. We have opted to employ the pre-trained 200-dimension Twitter embeddings, which have a vocabulary size of 1.2 million and have been trained on 2 billion tweets. These pre-trained word embeddings are an example of transfer learning because we are using the learned embeddings to complete a related task. We think that since our dataset also came from social media, some subtleties contained inside the embedding may enhance model performance. Since GloVe is better able to capture the semantic and syntactic meaning of a word with a bigger dataset used to train the embedding, we anticipate that GloVe will be able to improve model performance.

# 7. Model Building

We will construct various models in this part and assess how well they categorize suicidal text. In the problem statement, suicide or non-suicide is a binary variable that we are attempting to forecast. Transformers, machine learning, and deep learning are among the 3 models that were created. Among them are Convolutional Neural Network (CNN), Bidirectional Encoder Representations from Transformers (BERT), and Efficiently Learning on Encoder that Classifies Token Replacements Accurately (ELECTRA), which are shown in Figure 10 below.

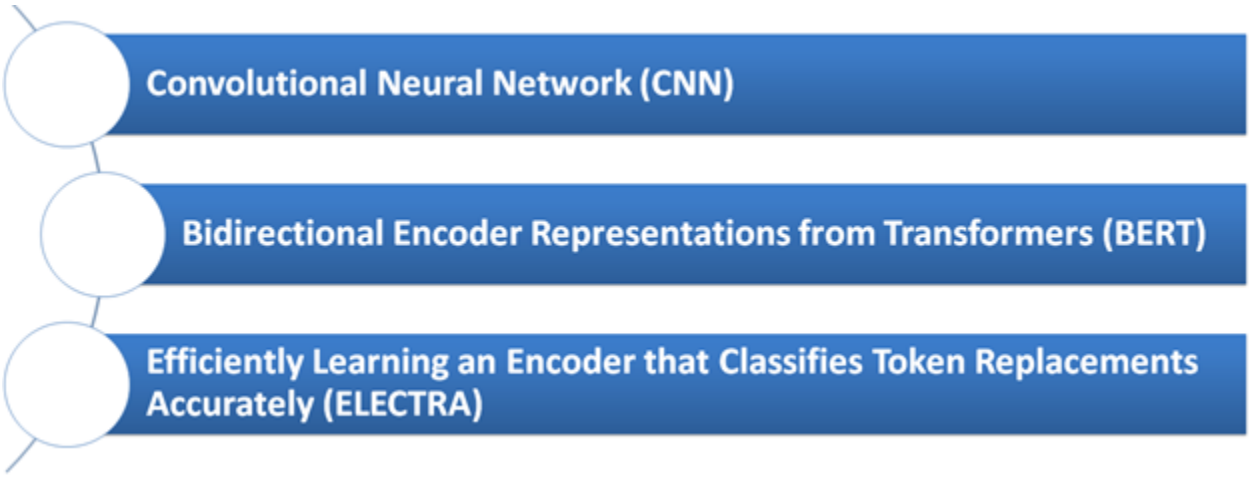



Figure 9. Models Built

On the train dataset, the models were refined, on the validation dataset, and on the test dataset, they were evaluated. Based on four evaluation metrics Accuracy, Precision, Recall, and F1 score we will assess the model's performance on the test dataset. To make sure that our model is applied more effectively in subsequent tasks, a stronger emphasis will be placed on the F1 score for our use case. False negatives are undesirable, but the F1 score gives a more accurate picture of the model's performance than recall does.

**7.1 Convolutional Neural Network (CNN):** As our project attempts to classify text data, the sequence of words plays a part in contributing to the connotation of a sentence. The Convolutional Neural Network (CNN) was proposed as an efficient way to classify text data as it is able to achieve decent prediction accuracy and consume lesser computational resources. As

such, we have experimented with the CNN model as one of the deep learning algorithms in our project.

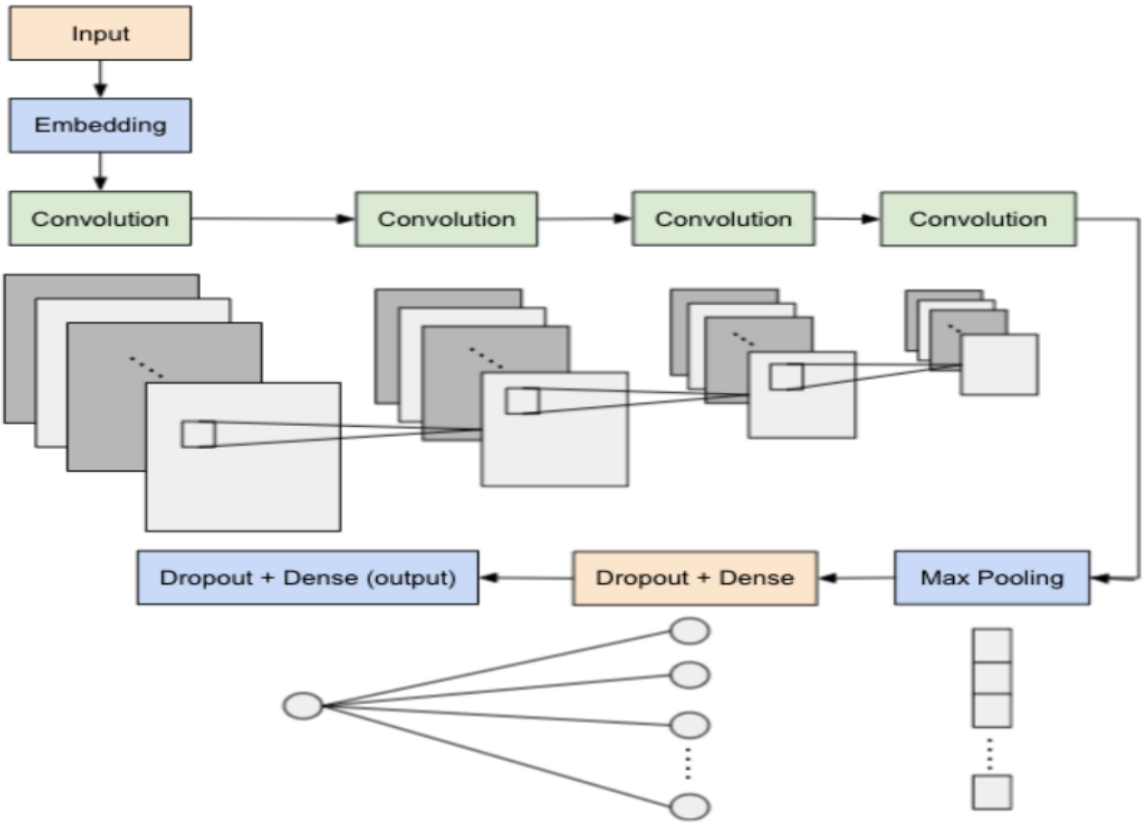



Figure 10. CNN Model Architecture

Our CNN model architecture consists of the following layers: an embedding layer, 4 convolutional layers, a pooling layer, a dropout layer, and a fully connected layer, as shown in Figure 10 above. We first performed tokenization on our text data to convert each word into an integer. The embedding layer is responsible for mapping each word of the sentence in our training data into a feature vector of a fixed embedding size [27].

**Model Variants:** We have experimented with the following model variants, with variations made to the Embedding layer:

• CNN Model 1: Random Initialisation (no pre-trained weights)

• CNN Model 2: Custom Word2Vec Embeddings (300-dimensions)

• CNN Model 3: Pre-trained GloVe Embeddings (200-dimensions)

**Model Performance:** The model performance for all CNN model variants can be seen in Table 1 below. The best model variant is Model 2 (Custom Word2Vec Embeddings) and it has performed the best across all metrics.

| CNN Model Variants | Accuracy | Recall | Precision | F1 Score |
|---|---|---|---|---|
| 1. Random Initialisation | 0.8985 | 0.8281 | 0.9010 | 0.8630 |
| **2. Custom Word2Vec Embedding** | **0.9285** | **0.9013** | **0.9125** | **0.9069** |
| 3. Pre-trained GloVe Embedding | 0.9001 | 0.8511 | 0.8858 | 0.8681 |

Table 1. CNN Models Performance Comparison

**7.2 BERT:** BERT, also known as Bidirectional Encoder Representations from Transformers, utilizes the encoder structure of a transformer for language modeling and was developed by Google in 2018 [23].

**Model Architecture:** BERT is pre-trained for two tasks: Next Sentence Prediction and Masked Language Model (MLM) & (NSP).

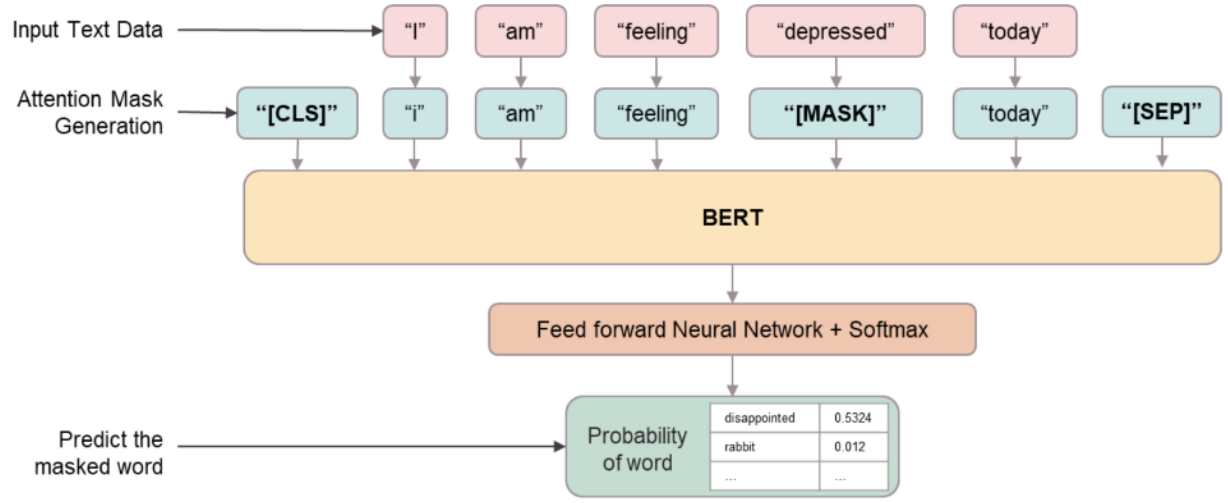



Figure 11. BERT Masked Language Model (MLM)

Consider the example in Figure 11 above. After masking, the input phrase "I am feeling depressed today" will become "I am feeling [MASK] today." The model is then taught to substitute the proper word for the masked tokens, enabling it to acquire more precise

representations via the attention process. For training reasons, MLM substitutes a "[MASK]" token for 15% of the words in the sequences.

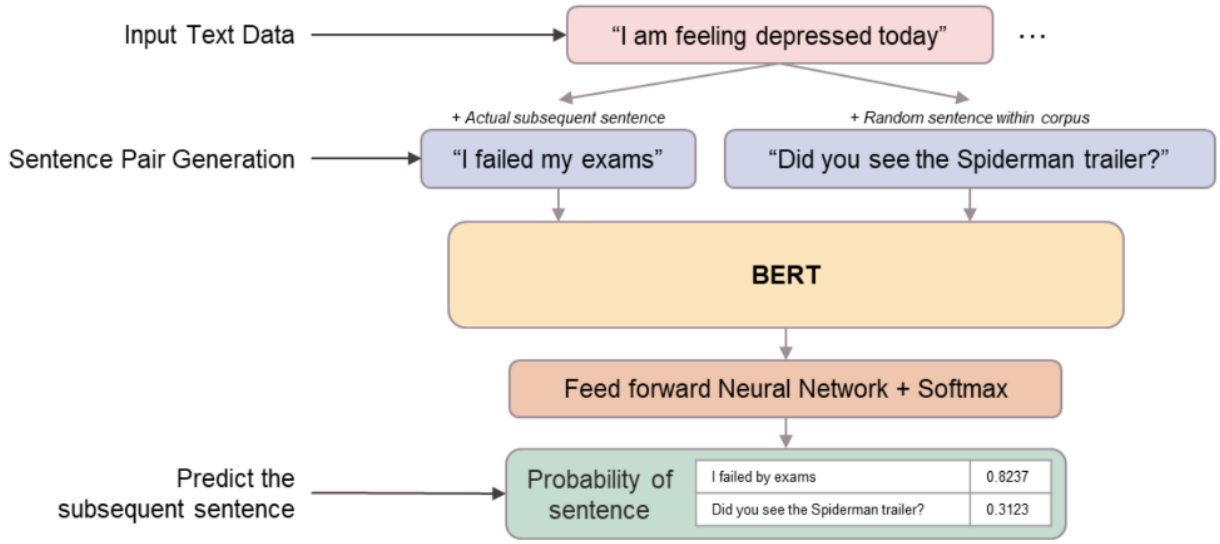



Figure 12. BERT Next Sentence Prediction (NSP)

While the NSP task is used to gather sentence-level information, MLM only captures sentence-level features at the token level. The model learns to predict whether the second sentence in a pair will come after the first throughout the training process by receiving pairs of sentences as input. As seen in Figure 12 above, the line "I failed my exams" actually follows the input text "I am feeling melancholy today," but the sentence "Did you see the Spiderman trailer?" is chosen at random from the corpus. The training data will consist of one actual input and one false input for each sentence. The additional tokens "[CLS]" and "[SEP]" are used to signify the beginning and ending of the sentence, respectively (Refer to Figure 11 above). BERT was made accessible in two variations, base and large and pre-trained using the 3.3 billion words of English Wikipedia and BooksCorpus data.

**Implementation:** Utilizing hugging face transformers, we used the BERT basic model for our implementation of the BERT model. On a test dataset, exhaustive experiments were performed to ascertain the model hyperparameters and implementation strategies. Since BERT was first trained on whole sentences, the same standard should be applied to the sentences we input, necessitating the employment of a different data preprocessing method. We sent the original data into the Hugging Face BERT tokenizer using the identical data points from earlier models. The

tokenizer generates unique tokens and attention masks as well as automatically translates tokens from the pretrained vocabulary to their associated token ids in a format that is conducive to training. There are three basic methods that are frequently employed to fine-tune a model: (1) Train the full pre-trained architecture. (2) Train a few layers from the pre-trained architecture. (3) Freeze the complete pre-trained architecture and train additional layers.

**Model Variants:** We have experimented with the following model variants, with variations in fine-tuning:

• BERT Model 1: Pre-trained

• BERT Model 2: Fine-tuned

**Model Performance:** Table 4 below shows the model performance for each BERT model variant. Model 2 (Fine-tuned BERT) is the best model variation, greatly outperforming Model 1 (Pre-trained BERT) on all criteria. Pre-trained models won't likely function as well as a model that has been specially trained for our particular use case, much like the various feature representation strategies we have tested haven't worked as well. Contrary to popular belief, pre-trained BERT forecasts the majority of inputs as positive despite having a poor F1 score. Fortunately, by focusing on the F1 score, a more accurate assessment of the model's performance can be made, making the pre-trained BERT unsatisfactory.

| BERT Model Variants | Accuracy | Recall | Precision | F1 Score |
|---|---|---|---|---|
| 1. Pre-trained BERT | 0.4681 | 0.9295 | 0.4156 | 0.5744 |
| **2. Fine-tuned BERT** | **0.9757** | **0.9669** | **0.9701** | **0.9685** |

Table 2. BERT Models Performance Comparison

**7.3 ELECTRA:** In 2020, Google announced ELECTRA, also known as Efficiently Learning an Encoder that Classifies Token Replacements Accurately, one of their most recent pre-trained transformer models.. ELECTRA, on the other hand, outperforms the aforementioned models while consuming only a fourth of the necessary processing resources on a select benchmark datasets.

**Model Architecture:** Replaced Token Detection (RTD), a brand-new pre-training task from ELECTRA, addresses the drawbacks of the MLM technique from BERT, which corrupts the input with masked tokens. Instead of learning to anticipate every single input token, BERT only learns to predict a limited selection of them (the masked tokens), which minimizes the amount it learns from each phrase. In general, BERT requires a lot of compute yet still gives good results when used to downstream NLP tasks

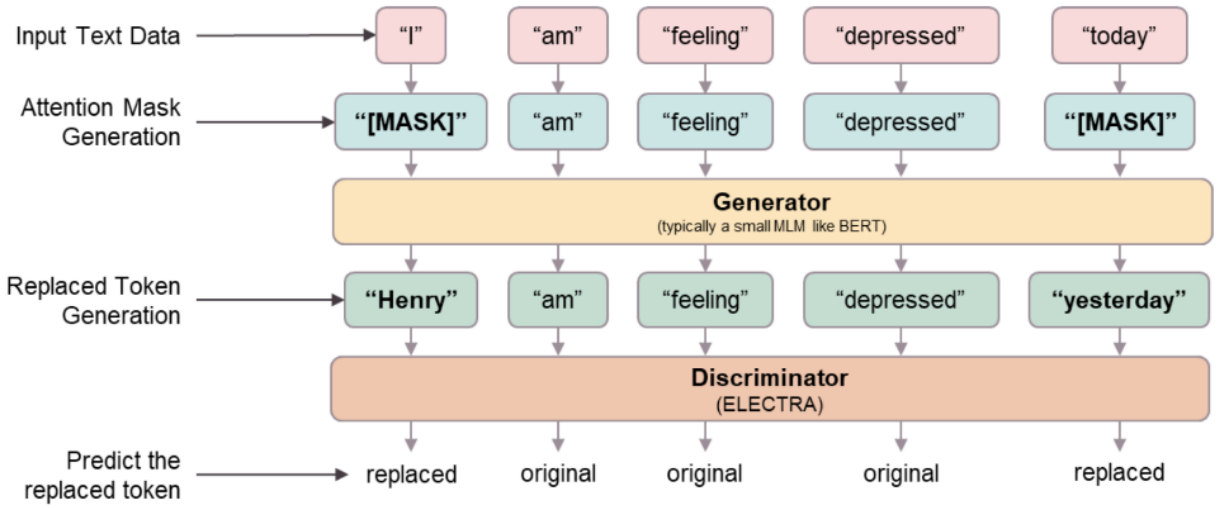



Figure 13. ELECTRA Replaced Token Detection (RTD)

In contrast, ELECTRA employs RTD to train a bidirectional model in which "[MASK]" tokens are substituted for wrong but plausible fakes. The input phrase "I am feeling depressed today" will be changed into "Henry am feeling depressed yesterday," as seen in Figure 13 above, using the identical example from the BERT section earlier. The replacement sentence makes more sense compared to sentences replaced with "[MASK]" tokens, although it still does not entirely suit the context. The discriminator then tries to determine which tokens have been replaced, drawing inspiration from generative adversarial networks (GANs) to tell the difference between authentic and fraudulent input data. The model must learn an appropriate data representation in order to complete this binary classification job, which supports learning from all input positions. A generator, often a small MLM like BERT, produces the replacement tokens. The discriminator and generator are then trained together using the same input word embeddings.

**Implementation:** We are using Hugging Face Transformers, a technique akin to BERT, to create the ELECTRA basic model. In order to compare the performance of the two transformer models more effectively, we have chosen the base model, which assumes a network of identical size. The Hugging Face ELECTRA tokenizer, which does the necessary preprocessing, received the source data in a similar manner.

**Model Variants:** We have experimented with the following model variants, with variations in fine-tuning:

• ELECTRA Model 1: Pre-trained

• ELECTRA Model 2: Fine-tuned

**Model Performance:** Table 3 below shows the model performance for each ELECTRA model variant. Model 2 (Fine-tuned ELECTRA) is the best model version; it performs much better than Model 1 (Pre-trained ELECTRA) in terms of accuracy, precision, and F1 score. Contrarily, Model 1 (Pre-trained ELECTRA) had a greater recall score that was nearly 1, but since our focus is on F1 score, it has no bearing on the model we choose [14].

| ELECTRA Model Variants | Accuracy | Recall | Precision | F1 Score |
|---|---|---|---|---|
| 1. Pre-trained ELECTRA | 0.4025 | **0.9908** | 0.3918 | 0.5615 |
| **2. Fine-tuned ELECTRA** | **0.9792** | 0.9788 | **0.9677** | **0.9732** |

Table 3. ELECTRA Models Performance Comparison

## 8. Model Selection

The most important outcomes of the models run are summarized in Table 4 below. Using the specially trained Word2Vec.

| Best Models | Accuracy | Recall | Precision | F1 Score |
|---|---|---|---|---|
| CNN | 0.9285 | 0.9013 | 0.9125 | 0.9069 |
| **BERT** | 0.9757 | 0.9669 | **0.9701** | 0.9685 |
| **ELECTRA** | **0.9792** | **0.9788** | 0.9677 | **0.9732** |

Table 4. Models Performance Comparison

All measurements show that the BERT & ELECTRA models perform better than the others. The highest F1, accuracy, and recall scores were obtained by ELECTRA. The model with the highest precision score, BERT, is closely behind this. The two models' performances are comparable, however ELECTRA has been shown to perform more quickly and effectively than BERT thanks to the proposed RTD, which directly addresses the drawback of combining MLM with BERT. We've decided to choose ELECTRA as our ultimate model.

# 9. Prediction Result

By using this we have found this kind of output for  text prediction. For example we can monitor our friends' posts as well as relatives.

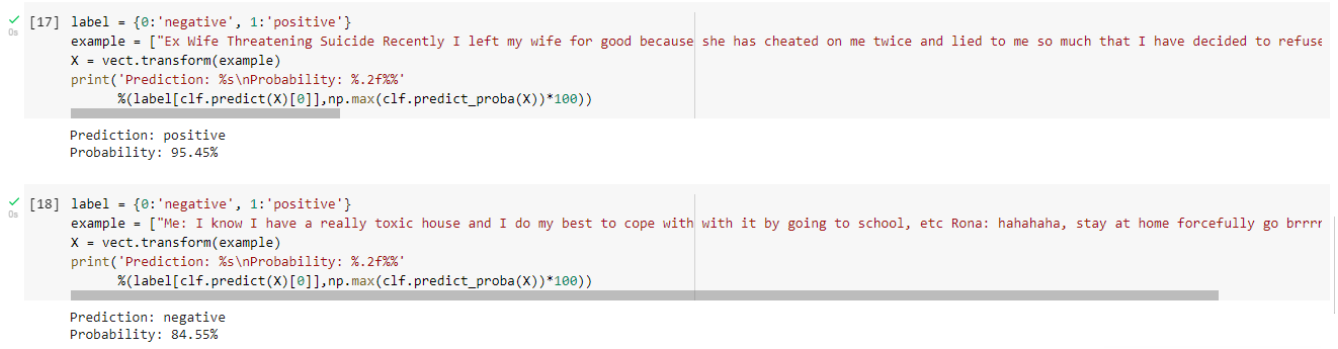



Figure:14. Predictive Result for Detection

Then if we think that his/her post are seems to suicidal we put their post on it and get the best probable output.

# 10. Mental Health Chatbot

In bangladesh, youth suicide rates are on the rise, thus more needs to be done to help and support them. Youths may be reluctant to seek assistance due to the continued societal stigma associated with mental illness and may prefer anonymous means to avoid being judged. As a result of their discomfort confiding in a professional context and their lack of trust in strangers to reveal sensitive information and experiences pertaining to their mental health condition, youngsters also resist getting help. Chatbots can aid in the early identification of suicide behaviors and offer support to people who are vulnerable.

**Model Architecture:** The goal of neural response generation (NRG) models, which include chatbots, is to predict a response based on text input. The two main categories of chatbots are (1) retrieval-based approaches, which utilize heuristics to choose an answer from a list of predefined answers, and (2) generative methods, which employ machine learning to create answers from scratch. To accomplish our goal, we will combine the aforementioned techniques to create a mental health chatbot, employing retrieval-based techniques to deliver appropriate responses to suicidal text messages and generative techniques to generate conversational responses.

**Implementation:** Developed by Microsoft in 2019, DialoGPT is a large-scale pretrained dialogue response generation transformer-based model for multi-turn dialogues that is used in the chatbot's generative component. 147 million multi-turn conversations from Reddit discussion threads were used to train DialoGPT, which was derived from the GPT-2 8 model. Utilized was the initial, pre-trained DialoGPT model from the Hugging Face Transformers library. On the other hand, a single-turn conversation Turing test revealed that DialoGPT responses were of human response quality. The pre-trained DialoGPT chatbot, however, is unable to respond appropriately to suicidal messages because it was not designed for that purpose. As a result, we altered a retrieval-based chatbot component to fit our use case. Along with a directory of local helplines that users can call for quick support, we have gathered a collection of consoling words from other websites dedicated to preventing suicide.

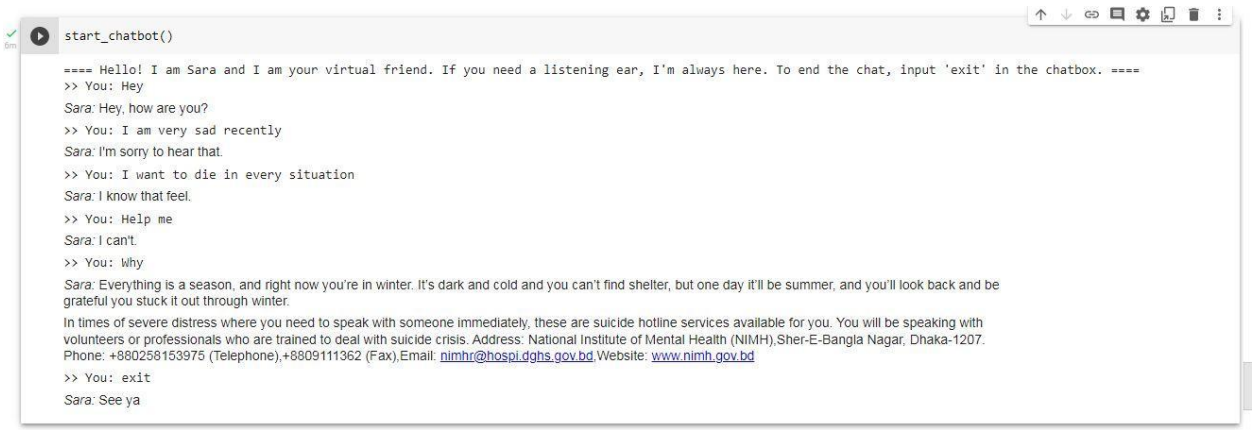



Figure 15. Mental Health Chatbot

**Limitations:** The popularity of chatbots has grown in recent years across a variety of industries, including consumer and financial services. However, there are still barriers that prevent chatbots from offering sufficient sympathetic responses in a real-world conversation, which is essential for those seeking assistance. Due to this restriction, chatbots may give improper responses at crucial moments or even incorrectly forecast the danger of suicide. Chatbots can't completely replace medical treatment and counseling, but new strategies are needed to combat the soaring suicide rates among young people. We firmly believe that mental health chatbots have the potential to be a part of the answer, but more research is needed in this area before they are widely adopted [15].

# 11. Future Improvements

## 11.1 Business Improvements:

**11.1.1 Building a Multilingual Chatbot:** Participants in a study evaluating the usability of mental health care chatbots recommended that a helpful feature of a chatbot would be to accept inputs of different languages, which is thought to facilitate more interaction with users. We can reach a larger international population if we add other languages to our chatbot, which presently only supports interactions in English. This might be accomplished by using conversational data that can be gathered by scraping multiple social media platforms to train and improve our models in different languages.

**11.1.2 Integration of Chatbot onto Social Media Platforms:** Social media use is one of the most well-liked online hobbies, and Statista Research Department (2021) projects that there will be 4.41 billion users using it worldwide in 2025. Researchers have suggested using social media to monitor mental health symptoms due to the rise in its use. Therefore, we might be able to add our chatbot to a well-known social networking site to improve its exposure and use. As an example of a well-known social media platform, Facebook has 2.91 billion monthly active members as of this writing (Statista Research Department, 2021). Facebook has made an effort to highlight content that may be related to suicide in addition to having a vast user base. There have been complaints about privacy standards since Facebook neglected to obtain users' permission before disclosing such information to a third party. Therefore, we suggested integrating our chatbot with Facebook to deliver prompt support and lessen the privacy controversy.

## 11.2 Technical Improvements:

### 11.2.1 Semi-supervised Learning to Improve Data Quality:

The data labels could not accurately represent the products. We can use semi-supervised learning approaches to do pseudo labeling in order to enhance the quality of the data through more precise labels. It can be used to train deep neural networks in a supervised manner using a modest quantity of data that has been labeled by humans. After that, the example can be used to create labels for data that isn't labeled. We can train our models to predict suicide behavior using both

human and pseudo-labelled data. We can make the model built more reliable by enhancing the quality of the data fed into it. The fundamental benefit of pseudo tagged data is that it can greatly reduce the amount of manual labor necessary, even if it may be slightly less accurate than human annotated data.

### 11.2.2 Larger Transformers Models to Improve Model Performance:

The ELECTRA base model modified for our dataset is currently the model that performs the best. Due to computational limitations, the ELECTRA big model could not be integrated. Larger transformer models can be used to enhance the performance of the model, but it should be noted that they take longer to train and may overfit the training dataset.

**11.2.3 Reinforcement Learning to Improve Chatbot Response:** We may apply reinforcement learning to steadily increase the skills and answers of our chatbots. Reinforcement learning enables the chatbot to learn by interacting with its surroundings, the end users, much to how humans learn by interacting with their environment. Then, it gathers rewards that can be compared to user feedback gathered following a chat with the bot. The bot's learning goal is to maximize rewards overall, which enables it to continuously enhance its capacity to produce more pertinent responses. In this way, the chatbot learns from user feedback and creates its own internal control system, making it increasingly effective at replying to end users. However, the learning process necessitates a large number of user interactions, which makes it take a while before it performs well.

# 12. Conclusion

Youth depression in Bangladesh is still a significant social problem, and prompt intervention depends on early detection. With the help of our project, we were able to identify suicidal wording in social media posts and develop models that received the best F1 score possible with ELECTRA (0.9732). By using our system, we can get a result either it positive in suicidal or not and also get the probality. We can take any kind of text or post of our relatives and friends form any social media platform and put them on our system then we get predictive result which are actually suicidal or non suicidal. The detection technique was subsequently included into a useful chatbot, enabling us to connect with people who are in need. We intend to improve the functionality of our chatbot and detection algorithm going forward and increase the number of people we can help.

# 13. References


1. Fodeh, S., Li, T., Menczynski, K., Burgette, T., Harris, A., Ilita, G., ... & Raicu, D. (2019, November). Using machine learning algorithms to detect suicide risk factors on twitter. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 941-948). IEEE
2. Parraga-Alava, J., Caicedo, R. A., Gómez, J. M., & Inostroza-Ponta, M. (2019, November). An unsupervised learning approach for automatically to categorize potential suicide messages in social media. In 2019 38th International Conference of the Chilean Computer Science Society (SCCC) (pp. 1-8). IEEE.
3. Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, *113*, 65-72
4. Anala, S. (2020, October 26). A Guide to Word Embedding. Towards Data Science.
5. Andrea C. (2019, September 11). How to train the word2vec model. Towards Data Science.
6. Bhanawat, V. (2019, June 28). The Architecture of Word2Vec. Medium.
7. Biswas, D. (2020, September 15). Self-improving chatbots based on Deep Reinforcement Learning. Medium.
8. Brownlee, J. (2020, January 14). A gentle introduction to imbalanced classification. Machine Learning Mastery.
9. Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armor, C., McTear, M. (2019). Assessing the Usability of a Chatbot for Mental Health Care.
10. Ce, P., Tie, B. (2020, November 10). An Analysis Method for Interpretability of CNN Text Classification Model.
11. Cheng, R. (2020, July 22). BERT Text Classification Using Pytorch. Medium.
12. Chollet, F. (2020, May 12). Keras Documentation: Transfer Learning & Fine-tuning. Keras.
13. Clark, K., & Luong, T. (2020, March 10). More efficient NLP model pre-training with electra. Google AI Blog.
14. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. International Conference on Learning Representations.
15. Cooper, N.. (n.d.). Finetuned DialoGPT model on Spanish Conversations. Hugging Face.
16. Culurciello, E. (2019, January 10). The fall of RNN / LSTM. Medium.
17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
18. Garbe, W. (2017, May 7). 1000x faster Spelling Correction. Towards Data Science.
19. Goggin, B. (2019, January 7). Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts.

20. Goldberg, Y. (2015, October 6). A Primer on Neural Network Models for Natural Language Processing.
21. Goled, S. (2021, March 16). Why transformers are increasingly becoming as important as RNN and CNN? Analytics India Magazine.
22. Horev, R. (2018, November 17). Bert explained: State of the art language model for NLP.Medium.
23. Joshi, P. (2020, July 26). Transfer learning NLP: Fine tune bert for text classification. Analytics Vidhya.
24. Jurafsky, D., Martin, J. (2021, September 21). Logistic Regression. Speech and Language Processing.
25. Kaggle. (2021). Suicide and Depression Detection. Kaggle.
26. Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., Rudzicz, F. (2019). A survey of word embeddings for clinical text.
27. Kim, H., Jeong, Y. S. (2019 April 29). Sentiment Classification Using Convolutional Neural Networks.
28. Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2020, October A Survey on Text Classification: From Shallow to Deep Learning.
29. Martínez-Miranda J. (2017). Embodied Conversational Agents for the Detection and Prevention of Suicidal Behaviour: Current Applications and Open Challenges.
30. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.
31. Minarr, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. (2021). Deep Learning Based Text Classification.
32. Mnasri, M. (2019). Recent advances in conversational NLP: Towards the standardization of Chatbot building.
33. Pai, A. (2020, March 16). An Essential Guide to Pretrained Word Embeddings for NLP Practitioners. Analytics Vidhya.