National College of Ireland

A Deep Learning Approach for Suicide Risk Assessment using Reddit

MSc Research Project
Data Analytics

Shrinidhi Chandrashekhar Shetty
Student ID: x18199780

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shrinidhi Chandrashekhar Shetty
Student ID:	18199780
Programme:	MSc. in Data Analytics **Year:** 2019-20
Module:	MSc. Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	17 August 2020
Project Title:	A Deep Learning Approach for Suicide Risk Assessment using Reddit
Word Count:	**7182 Page Count: 20**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shrinidhi C Shetty

Date: 16-August-2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deep Learning Approach for Suicide Risk Assessment using Reddit

Shrinidhi Chandrashekhar Shetty
x18199780@student.ncirl.ie

Abstract

The advancement in technology has ironically connected the world and separated from each other. The world is drastically progressing in the technical aspect and exponentially dropped in terms of preserving composite mental health. The mental stress from the external environment leads to developing self-harm thoughts. Suicide is the most alarming trend in today's life. It is necessary to build proactive systems to stop people from taking such drastic decisions to end their life. The research focuses on analyzing the pattern followed in stages of Suicide Risk- starting from Suicide Ideation, Suicide Indicator, Suicide Behaviour, and Suicide Attempt. The research begins with a comprehensive analysis of building a large labeled dataset from the sparse dataset available with a semi-supervised learning approach to overcome the scarcity of data. The dataset is critically evaluated to model a multi-class text classifier and understand the factors and sentiments related to each of the target classes. The research builds models based on content analysis and deep neural networks to incorporate an assessment tool in online sites to monitor the suicide risk patterns. The results can be a compiled report for mental health care workers to understand the underlying pattern.

1 Introduction

Suicide is at an alarming rate in the world. World Health Organisation (WHO) says that around 8,00,000 people end their life every life because of suicidal thoughts[1]. That means one person is dying for every 40 seconds. According to the Centres for Disease Control and Prevention (CDC), there is an addition of people who try to attempt suicide or have suicide ideas than committing it. Suicide is a major public health problem that is affecting people of all age groups. Committing suicide is subjective in each person's perspective. But the researchers say that it can be related to factors like violence, health issues, economic concerns. WHO says suicide is the leading cause of death in 15-19-year-olds which can be due to bullying, child abuse, or violence? These external abuses can affect a person's mental stability and cause depression. The psychological strains can be a leading cause to commit suicide. Although the reasons for suicide are uncertain to figure out, there lie many protective factors to prevent them.

There are various preventive measures and protective strategies undertaken by various organizations to control the death tolls due to suicide. Social Media Platforms are the easiest to access and to communicate. It is a medium where people express their thoughts and their vulnerabilities. It has also become a platform where cyberbullying coexists with friendliness. Social Media has become the center for cyberbullying and trolling. The platform has become

[1] https://www.who.int/news-room/fact-sheets/detail/suicide

the epitome of these cyber violence and friendliness as well. Reddit is a social forum where people can ask queries, answer, or share content. Reddit has over 330 million active users around the world and is growing as huge as Twitter. It always posts texts, images, videos, links, or create a poll. Each of the posts contains a title and the content. The character limit for title is 300 whereas for the post it is 40000. The content on Reddit is categorized and is called subreddit. 'SuicideWatch' is one subreddit where Redditors share their suicidal thoughts or post supportive comments, respectively. 'SuicideWatch' was created in 2008 and has over 214000 users. This community supports their peers by posts or comments to the ones struggling with suicidal thoughts. Reddit does not take charge or police the individuals from not committing suicide but rather provides a platform to speak up anonymously so that the peers can understand their thoughts and support or empower each other. Hence there is a need to actively track the posts on 'SuicideWatch' subreddit and collect data that can help in analyzing patterns from suicidal thoughts phase to suicide attempt phase in the individuals and prevent such unlikely deaths promptly.

Earlier research has focused on segregating the users in social media platforms like Twitter and Reddit as people at High and Low risk of suicidal behavior or classifying the users based on suicidal or non-suicidal poets. Various machine learning algorithms and deep learning algorithms have been employed to explore the patterns in the Reddit posts. The research will be focusing on quantitatively answering the research question - What are the factors that influence the categories of suicide risk? and How well can the Deep Learning algorithms and Natural language Processing classify the Suicide Risk categories?

The rest of the document is organized as follows. In section 2, related work on natural language processing and machine in suicide risk assessment is discussed. Section 3 provides a detailed methodology of the research starting from data collection to the performance evaluation. Section 4 concludes the document with discussion and future work.

2 Literature Review

A considerable amount of literature has been published on the identification of suicide risks in the online platform. The role of Artificial Intelligence (AI) in assessing mental health issues is quantitatively powerful. With the abundant data available through social media, clinical health records and personal digital devices have led to the emergence of AI methods in incorporating Machine Learning (ML) with Natural Language Processing (NLP) for the prevention or detection of suicidal thoughts. A relationship exists between the human mind expressing its thoughts through language and mental health insights. Investigating suicidal risk is continuing to a question of great interest in the field of Data Analytics. This section describes the numerous studies conducted in assessing suicide risks. (D'Alfonso, 2020)

2.1 A study on the content analysis and feature engineering in NLP.

Over the last decade, most research on NLP has emphasized on the importance of the mining sentiments from the text data that can reveal opinions and inferences on the subject or topic. NLP has proven to be a promising technique in extracting features like polarity, sentiment, opinions from the text that can reflect an individual's thoughts. The paper emphasizes the

importance of the features extracted from the textual data from the online platform. Various ML algorithms rely on the data scraped from social media platforms or discussion forms for the classification process. (Cobos et al., 2019)

Most studies on suicide risk detection have focused on exploring online content. The author says that the advent of online social media and technology has taken a toll on the mental health aspect. Platforms like Facebook, Twitter, and Reddit have a rich source of text data that has led to innovative researches related to suicide and mental health issues. These microblogging sites have millions of users whose thoughts can be analyzed through the content and represent sample data for the whole population. The insights from the sample can be mapped to the population. The study also focuses on the relationship between depression and suicide in the text data. (Tadesse et al., 2019)

The paper suggests that social networks and discussion forums have become part of daily life. The usage has resulted in Social Network Mental Disorders (SNMD) like cyberbullying, net compulsion. The author suggests that this has an advent effect on mental health and on developing suicidal thoughts. Thus, early intervention of such scenarios is necessary to prevent any fatal deaths. The study considers data related to social media and clinical records to determine mental health disorders. (Shuai et al., 2018)

2.2 The need for Semi-Supervised Learning approach in NLP

A large number of works of literature in problems involving sentiment analysis suffers from the availability of large datasets. Mostly in suicide risk assessment problems, questionnaires or psychometric tests are considered as the baseline for collecting data. Using this approach, researchers were able to differentiate the characteristics of people suffering from suicidal thoughts and those with not. The author implements autoencoders to label the sentiments in the documents concentrating on two-word phrases. The research could successfully tag multiple labels associated with documents with vector space representations of words in the documents. The author suggests using a vector space rather than any pre-defined lexicons or polarity shift rules in the documents. Pre-defined lexicons and polarity rules are not suitable in sentiment analysis tasks as they do not consider the context and subjectivity of the lexicons in the documents. The new datasets formed from the semi-supervised encoders overcome the problem of generalization. (Socher et al., n.d.).

The author discusses the effects of the sample set size on text mining tasks. The collection of clinical data from hospitals for suicide risk predictions can be difficult because of consent issues and preparing a database of questionaries can be expensive in labor wise and following the protocols. It is important to identify the effect of sample size on the type of problems, approach technique, and the effectiveness of the learning algorithms on them. The paper argues that the stability and reliability of results from any machine learning algorithm depend on the size of the dataset. (Matykiewicz and Pestian, n.d.; Venek et al., 2017)

2.3 Comparison of Machine Learning and Deep Learning approaches in Natural Language Processing

The author proposes a depression detection tool with a context-based Deep Neural Network. The research considers various factors related to an individual to do a contextual analysis and formulate a regression equation to predict the suicide risk or not. The author judges the risk of an individual's suicidal thoughts based on the extent of good, risky, bad. The context DNN model does not consider any statistical measure to differentiate the various categories of risk levels. The usage of the regression equation in a context analysis cannot understand the complex hyperplanes separating the different categories of suicide risks. (Baek and Chung, 2020)

The author proposes a content analysis method to detect any suicidal lexicons in Twitter posts. The content analysis extracts the usage of each word in each document. These features apply to machine learning algorithms in understanding the simple hidden patterns in the documents. They do not work well with the deep neural networks as they look for complex structures in the corpus. The hyperplanes in the text corpus target classes that statistically significant with each other can mine for a lot of hidden patterns to understand documents based on context and structure. (Kumar and Rao, 2019; Tadesse et al., 2019)

The author proposes that the language used in individuals with depression or mental distress happens to be different from regular people. The research evaluates the early depression assessment systems with the Early Risk Detection metric. The metric critically analysis the risk factors in the online posts. It provides time for acting on the risk involved or in overcoming the fatalities. The research was conducted on various pre-trained word embeddings like GloVe and fastText. Although the metric is better in intervention terms. It demands comparatively less false-positive cases. (Trotzek et al., 2020)

The rate of suicide in youths and adolescents is high compared to other age categories. And there is a strong correlation among today's youth and technology. Most of them have access to steady internet and chat messages. The author proposes in building a chatbot that can proactively patrol the data streams and aid in assessing any future fatalities. The research tries to handle the situation of an individual early rather than at the peak of a crisis. This is a good approach to sensitive topics like suicide. (Elliott et al., 2019)

2.4 Need for further research

After an extensive review of literature in suicide detection and prevention, there are unexplored parts about the scarcity of data and the availability of a balanced dataset. The literature also lacks in categorizing the online posts into the four main categories of risk. The research not only classifies the data samples into their appropriate categories but also deals with the imbalanced dataset and a self-learning approach to better understand the data and their hidden complex patterns.

3 Research Methodology

The research follows the Knowledge Discovery in Databases (KDD) approach. The project is Descriptive research which aims at determining the categories of risk in suicide posts from Reddit. The authors define the KDD process as a process where data mining techniques are applied for extracting patterns. More than any advanced data mining approach,

understanding the problem statement and data is important. KDD comprises seven steps that start with data understanding to data cleaning, integration of data sources, selection of appropriate data, transformation, data mining, evaluation of results, and presenting the knowledge. With massive bytes of data being generated every day, the need to build models and theories on the increasing volumes of data can be computationally challenging. With massive bytes of data being generated every day, the need to build models and theories on the increasing volumes of data can be computationally challenging. In this new generation, humans will need tools to assist in extracting and storing knowledge from the data. Therefore, KDD addresses the problem of data overload. (Usama Fayyad et al., 1996). It will be advantageous to store the factors related to each of the suicide risk categories and refer to them for future suicide posts on social media platforms for early interventions before any fatalities.

3.1 Problem Understanding

The aim of the project to assess the online posts which are in the form of English language textual data to suicide risk level (Ideation, Indicator, Behaviour, Attempt). Natural Language Processing is the frontier of Artificial Intelligence and Data Analytics. NLP bridges the gap between the text data and machine learning models. The purpose of the research is to understand the stages of suicide risk which can be utilized as an early intervention tool for physicians or patrolling the content on the social network for avoiding catastrophe. In addition to that, the lack of textual data for research purposes is solved by a semi-supervised learning approach that can aid in future projects.

3.2 Data Selection

The input and output data are the English language sentences and the corresponding suicide risk labels. Reddit consists of a subreddit called 'r/**SuicideWatch'** which is a group that supports users hassling with suicide thoughts. Data Mining Models rely on massive volumes of labeled data which was challenging for NLP researches. The manual labeling of data is expensive and time-consuming. To increase the amount and quality of data semi-supervised learning approach is utilized to prepare the dataset. Therefore, labeled, and the unlabelled dataset is selected. The initial dataset[2] consists of 4786 users' text data. These data are categorized into 6 target classes: suicide ideation, suicide indicator, suicide behavior, suicide attempt, supportive and non-informative. A real-world dataset was scraped Reddit-'r/**SuicideWatch'** that consists of 40k random user's text data samples without output labels. The Reddit dataset which is used in the research is listed in Table 1.

Table 1: Dataset Attributes

Attribute	Description	Target Class Label
Reddit Post (labeled)		**Ideation**

[2] https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit

	English language textual data	**Indicator** **Behavior** **Attempt** **Supportive** **Non-Informative**
Reddit Post (unlabelled)		--not labeled--

The characteristics of the labeled dataset utilized in the research are listed in Table 2. The dataset is highly imbalanced with 46.1% of supportive texts and with least 2.9% of suicide attempt texts. The unlabelled dataset consists of 40469 documents scraped from Twitter using shift.io API[3].

Table 2: Dataset Characteristics

Class	**n**	**%**
Supportive	1452	46.1
Indicator	513	16.3
Ideation	502	15.9
Non-Informative	445	14.1
Behavior	150	4.8
Attempt	91	2.9
Total	**4786**	**100**

3.3 Data Pre-processing

Of the initially labeled dataset collected, the test samples with missing output labels are rejected. The area of research focuses on categorizing the text data among the four categories of Suicide Indicator, Suicide Ideation, Suicide Behaviour, and Suicide Attempt. The text data of other target classes are neglected. Since the data is in textual format, it needs to be cleaned for the removal of unnecessary characters. The cleaning process for the dataset should be in line with the word embeddings used in the later stages of the research. The dataset should get rid of HTML encodings, URLs, punctuations, special characters because pre-trained word embedding does not contain vectors for them, and they do not add any meaning to the research. The apostrophe must be replaced, and the contractions of the English language should be expanded, and all the alphabets converted to lower case. Around 20% of online posts consist of STOP WORDS that should be removed. All the cleaning is conducted with a 'spaCy' library. The cleaned dataset is split into train and test set using scikit-learn library. Since the dataset is imbalanced and to avoid skewness in the train and test portion, the stratification of datasets is introduced. The samples are distributed in equal proportions for the target classes in both the train and test set. The dataset is split into train and test set before any data transformation to keep the vector information of both sets mutually exclusive while learning from train data and predicting on test data. The train and test set are in 80:20 proportion.

[3] https://pushshift.io/api-parameters/

3.4 Data Transformation

The input to Deep Learning Models cannot be textual data. Hence the text data should be represented in number type. The data transformation part comprises of three steps to convert clean text data to number representation.

Tokenizer: The cleaned text sentences are split into words called tokens. Tokens that are frequently available in the text corpus are retained and less frequent words are neglected. This helps in reducing the noise in the dataset. The tokenizer by default rejects few non-essential words. A utility function from Keras library called 'keras.preprocessing.text.tokenizer' is used. The tokenizer generates a dictionary of the unique words and the corresponding unique number which is sorted in the order of frequency of words. It is stored in 'tokenizer.word_index'. The tokens for the sample text data is stored in 'tokenizer.text_to_sequences'.

Pad the Token Sequence: The deep learning models expects all the input sample to have an equal length. Hence the tokenized text sequences need to be padded with 0s to make all the text sequences to a pre-defined length.

Encoding the Target Class Labels: The target class label is of string type and needs to be encoded to number for the Deep Neural Networks to understanding the output class. A label encoder from the sci-kit-learn library is used.

Creating Embedding Matrix: Word Embedding representations are used to represent words with similar meanings. They overcome the generalization issue by representing the rare words input to the model in terms of most similar meaning words. Pre-trained Word Embeddings are the large corpus of word representations trained with massive volumes of datasets that can be utilized for various similar tasks. Global Vectors or GloVe embeddings is one such pre-trained word embedding utilized in the project. It is advantageous because it works in unsupervised learning fashion, it covers global statistics for the word corpus and considers the probabilities of two words occurring together in the dataset. The GloVe embeddings are downloaded from the official website. An Embedding Matrix is created for the dataset for each sample's word sequence is referenced from the GloVe embeddings.

3.5 Data Mining

The classical Machine Learning algorithms are vague in learning from complex unstructured datasets. Another advantage of Deep Learning is their adaption of Transfer learning. Pre-Trained models on data that are in large volumes can be utilized for various similar tasks. The ability to use the pre-trained models for the model training can save huge computational power and bulk amount of data. The research uses two deep learning algorithms for text classification.

Text convolution Neural Network (Text CNN): The idea of using a Convolution Neural Network (CNN) for Text Classification was first introduced in 2014. CNN is widely efficient in image classification and object detection tasks. CNN is used in text classification tasks with the idea of representing documents in the form of images that the Neural Network can understand. The similarities in representations of images and documents in the form of a matrix are the central concept of logic. (Kim, 2014). In addition to that CNN considers the co-occurrences of words while learning.

Bi-direction Long Short Term Memory (BiLSTM): Recurrent Neural Networks (RNN) can store the information from the previous task in hidden states and apply them to the current tasks. It enables the network to learn the dependency of words on each other in a sentence along with context. Long Short Term Memory (LSTM) is a subcategory of RNN that can remember information. BiLSTM is a special case that remembers the contextual information on both sides of the network. Hence the name.(Dong et al., 2020)

3.6 Evaluation and Interpretation

The quality of the research is dependent on the quantitative metrics that measure the performance. It is not viable to use standard evaluation metrics for a multi-class classification problem. The standard metrics cannot be reliable or even misleading when the dataset consists of a skewed distribution of classes. The confusion matrix can aid in understanding the classification accuracy for each class label. The standard metrics like Average, Precision, Recall, F1 score are calculated for each class by taking an average of one class against the whole classifier. Ranking metrics like the Train/Validation Loss are diagnosed. The Macro Average and Weighted average of metrics provide a better understanding of the results for multi-class classification problems. (Basha et al., 2020)

4 Design Specification

The approach to answering the research question in the project is an experimental type. The factors that distinguish the suicide risk levels in individuals are answered with quantitative research methods. The research starts with data collection from the Reddit website that is categorized into four levels of suicide risk namely- Ideation, Indicator, Behaviour, and Attempt. Reddit is scraped with shift.io API. Data is collected overtime periods from the subreddit called 'SuicideWatch'. The scraped data is an unlabelled set of data. A small dataset that is labeled is obtained from a data source. A major problem in text mining is the lack of availability of labeled data. To overcome that problem, a semi-supervised learning approach called pseudo labeling is undertaken. In addition to that, a balanced dataset is created by data augmentation techniques. Pre-trained word2vec called 'GloVe' embeddings are used to represent textual data in a sequence of numbers. The final dataset after the pre-processing and transformation is around 40k. Deep Neural Networks are selected for their ability to determine the non-linear relationship in complex data sources. The dataset is entered into Text CNN and BiLSTM for modeling on train data proportion. The trained models are tuned by tuning the hyperparameters to obtain reliable models. The accuracy of the trained model is verified with test data. Evaluation metrics are calculated to measure the quality of the analysis. The results are interpreted and discovered for knowledge that will answer the research question proposed in the project. The framework of methods and techniques undertaken in the research is depicted in Figure 1.

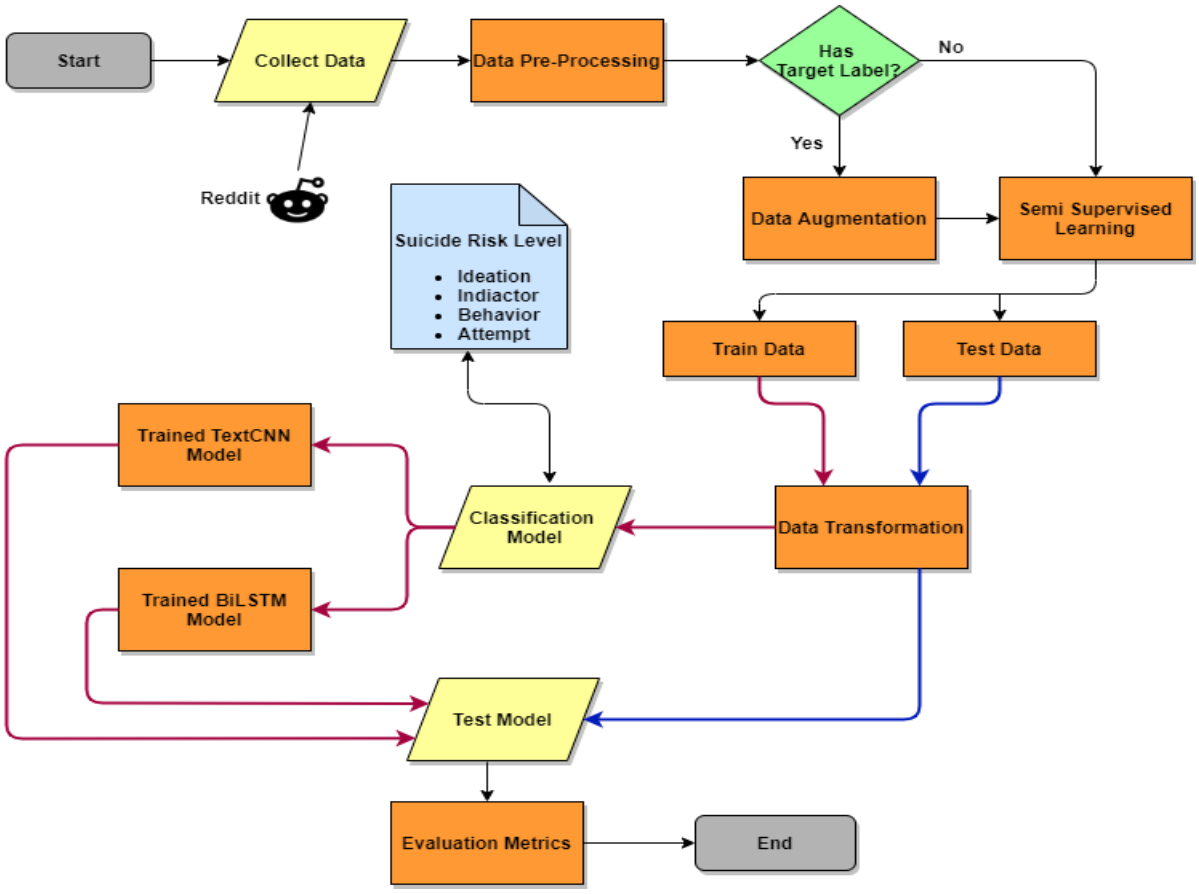

Figure 1: Research Flowchart

5 Technical Implementation

The Machine learning and Deep Learning algorithms used in the research are explained in this section. There are three main stages of modeling in the research. They are as follows:

5.1 Data Augmentation Model

In multi-class text classification, imbalanced target classes can cause an effect on the accuracy of the model. Hence it is necessary to have a balanced dataset in the research process. The balanced dataset can be used in the learning process to label the large volume of the unlabelled dataset. Thus, solving the problem of an imbalanced dataset and lack of supervised data(Abdurrahman and Purwarianti, 2019). Text data samples in the target classes of Suicide Behaviour and Suicide Attempt are 0.2% of the majority class data (Suicide Ideation and Suicide Indicator). In this scenario, to avoid overfitting of data and to improve the generalization of the model, data augmentation techniques are used. It is a technique to create a diverse training set with an available dataset. Textual Data Augmentation is conducted in four simple steps. A random set of n words are selected from the dataset excluding STOP WORDS and searched for their synonyms. The random words are replaced with corresponding synonyms in any position of a sentence and iterated for n number of times. Two words in each document interchange their position for n iterations. A word with probability p is deleted from

each document. The addition of augmented text data can add some noise to the dataset. But it also makes the model robust and avoids overfitting of data. The process is conducted until a balanced dataset is obtained. The steps in data augmentation are explained in Figure 2. Data Augmentation is applied to minority target classes namely-Suicide Behaviour and Suicide Attempt. The pre-defined values are 'alpha'-number of words that will change in a document, probability 'p' – threshold probability to delete a word in a document, and number of augmented sentences 'n' for each original sentence. The output of this method will be a balanced dataset.

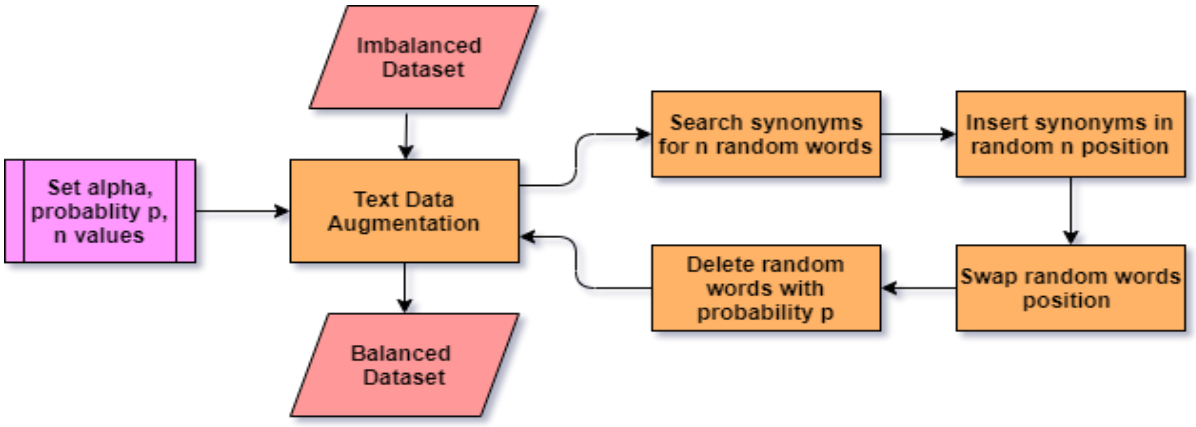

Figure 2: Data Augmentation Flowchart

5.2 Semi-Supervised Labelling Model

The availability of data is crucial for any data mining process. In the field of Natural Language Processing, there is abundant text data but not supervised text data. To overcome this problem, semi-supervised learning is helpful. Self-Learning is one of the techniques in semi-supervised learning. (Dasgupta and Ng, 2009)The research project uses this technique. The flow of the Semi-Supervised approach is illustrated in Figure 3. Initially, LSTM is modeled using the small portion of the labeled dataset for 1000 epochs. The neural network learns from the labeled dataset. Next, the labels for the unlabelled dataset are predicted and the probabilities are set to a high confidence level of *(>=0 .80).* The documents for which the probabilities are higher than the threshold are added to the labeled dataset and removed from the unlabelled set. The next step is to iterate over the same steps from the beginning until all the samples in the unlabelled sets are removed or do not meet the confidence level criteria. This technique leverages the data volume and enhances the learning capacity for the model. Initially, a labeled dataset of 4000 was available for the research and after applying this technique additional 30000 documents are available to conduct the research.

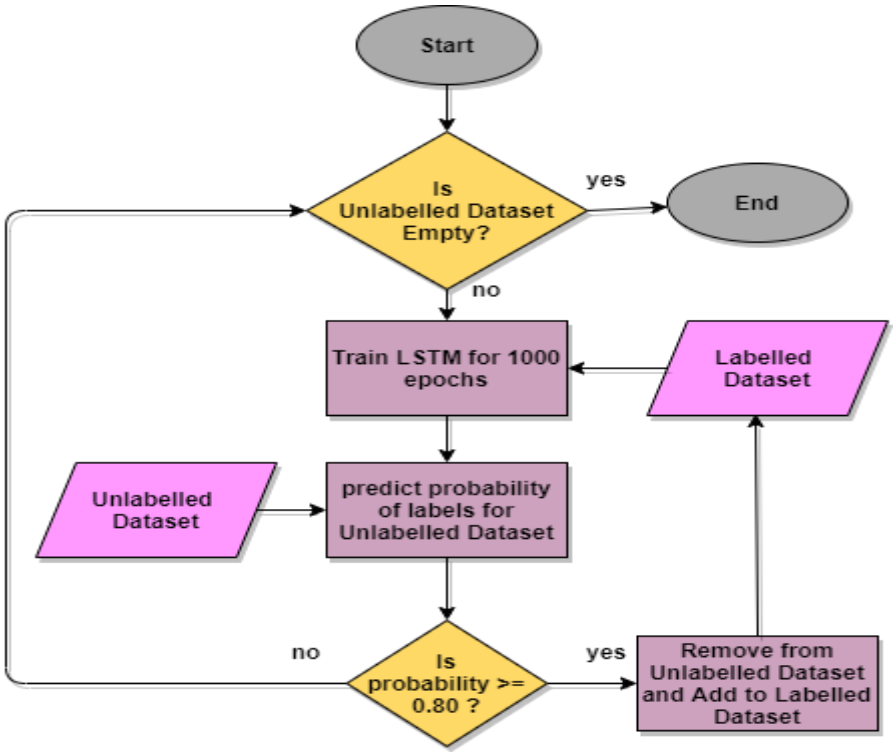

Figure 3: Semi-Supervised Labelling Flowchart

5.3 Text CNN Model

The input data to a CNN model must be in a matrix format. CNN is well known for its application in image recognition and classification problems. Images are represented in terms of a matrix with pixel values. In the same way, text data is represented in terms of digits in the form of an embedding matrix by referencing the GloVe embeddings of the pre-trained model. The length of the longest sentence in the dataset is assigned to maxlen. The embedding size is set to 300. An embedded matrix of size 300*maxlen is created that represents the text dataset. TextCNN considers the co-occurrences of words in the documents.
The parameters set in the model are as follows:
Library: Keras, Pytorch, scikit-learn.
Filters: [1,2,3,5] - check for unigrams, bigrams, trigrams.
Activation Function: ReLU (Rectified Linear Unit)
Loss Function: Cross Entropy Loss
Optimizer: Adam (Adaptive Moment Optimization)
Epochs: 15
GPU library: CUDA

5.4 BiLSTM Model

The Text CNN algorithm considers the co-occurrences of words in the documents but does not understand the context of the whole document. Long Short Term Memory can store information from previous states. They understand the sequence and context of the words. BiLSTM stores the information related to the context of the document on both the directions of the Network which is beneficial to a text classification problem.

The hyperparameters set in the model are as follows:
Library: Keras, Pytorch, scikit-learn.
Hidden Size: 64
Activation Function: ReLU (Rectified Linear Unit)
Loss Function: Cross Entropy Loss
Optimizer: Adam (Adaptive Moment Optimization)
Epochs: 15
GPU library: CUDA

6 Experimental Evaluation

6.1 Impact of Data Augmentation on Dataset Size

The initial dataset was not adding any value to the research because of the small size. It is also a known fact that the data for the research belongs to a sensitive area. It is proof that not much data related to people committing suicide was available. But this technique of Data Augmentation eased the process. Text Data Augmentation enhances the usability of the dataset with available data size. The document count for each class category before and after applying simple data augmentation techniques is shown in Figure 4. The result is a well-balanced dataset. The effect of augmented data on the performance of the models is seen in Table 3.

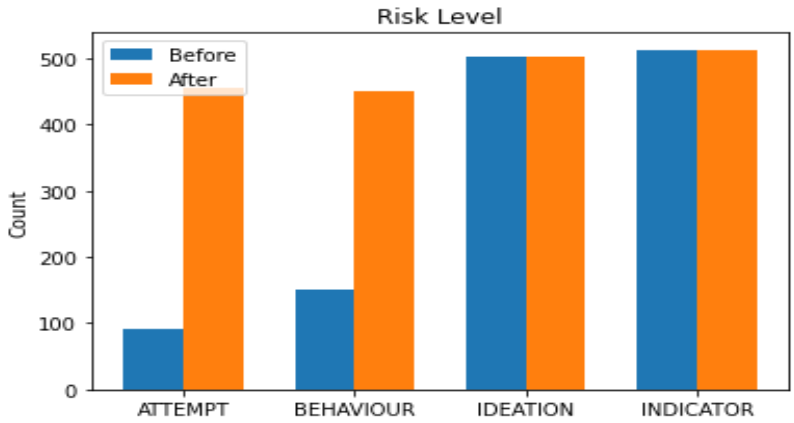

Figure 4: Semi-Supervised Labelling Flowchart

6.2 Boosting the diversity of Dataset with Self-Learning

Semi-supervised learning boosts the diversity of the dataset. It is evident from Figure 6 that the dataset is in large volume and balanced. The Neural Networks crave for large volumes of data. In the self -learning process, the model became robust to noisy data. The unlabelled dataset consisted of recent data from Reddit, it was shocking that the diversity of data was shocking looking at the balanced dataset. In addition to posts related to suicide, there were a ton of posts that showed compassion and positivity towards the people facing suicidal thoughts. Semi-supervised learning can potentially solve the issues of data scarcity that is hindering most of the data mining problems. Data Augmentation step is revisited to create a final balanced dataset. The size of the dataset for each target class is shown in Figure 5.

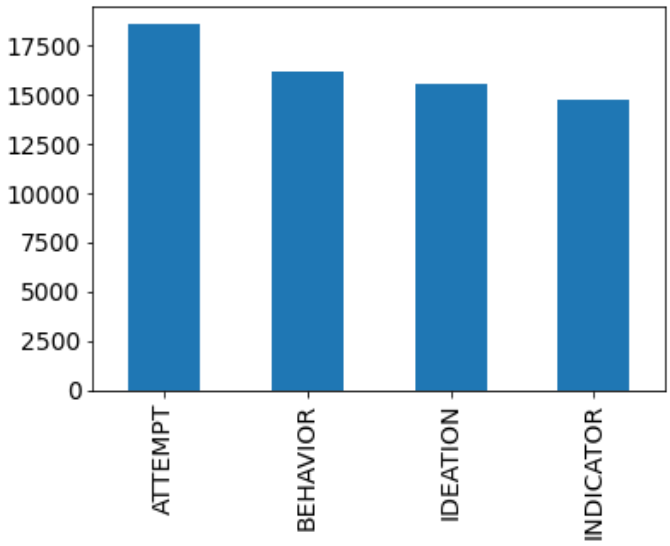

Figure 5: Dataset after Semi-Supervised Learning

6.3 Exploratory Data Analysis on the Text Corpus

A Sentiment analysis technique called Valence Aware Dictionary for Sentiment Reasoning (VADER) is used. The VADER score is a normalized composed of both polarity and intensity of the documents. It is available in the NLTK package. Vader score is used to understand the underlying emotion in the documents. The scores for the four categories are shown in Figure 6. The average Vader score for all the categories lies in the negative region, with Suicide Attempt posts being the most negative one for obvious reasons.

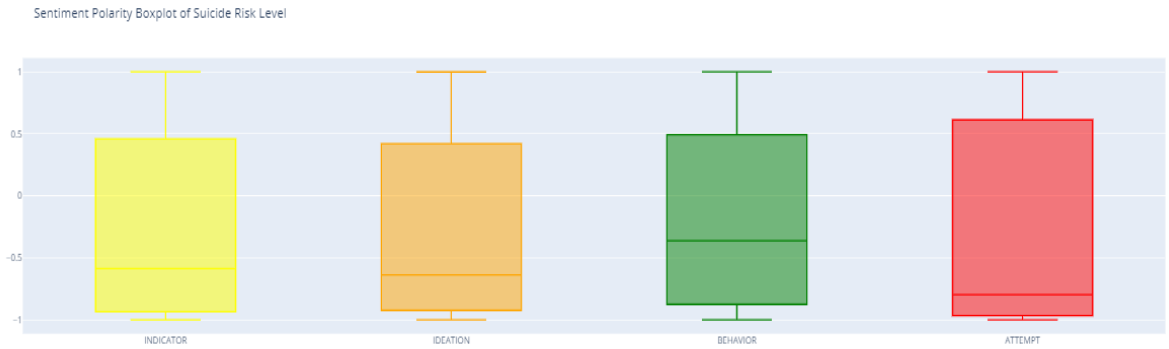

Figure 6: VADER Sentiment Analysis

Chi-Square Test is used to determine the precedence of the importance of each word towards the target variable. For text classification problems, the chi-square test identifies the words in the documents that are highly correlated to its corresponding target class. In Figure 7, it is seen that the occurrences of attempt, kill, involuntary, ambulance, fail, gun, skype is the most used words in the documents. It states that the sentences that contain these words are likely to come under any one of the categories of suicide risk. Chi Square test is conducted to recognise the features that contribute the most in the correlation towards the target class. It is a visualization of the words that holds more weightage in belonging to any of the target classes. In other words, if these words are present in some posts then there is high chance of that post belonging to one among the four target classes of suicide severity.

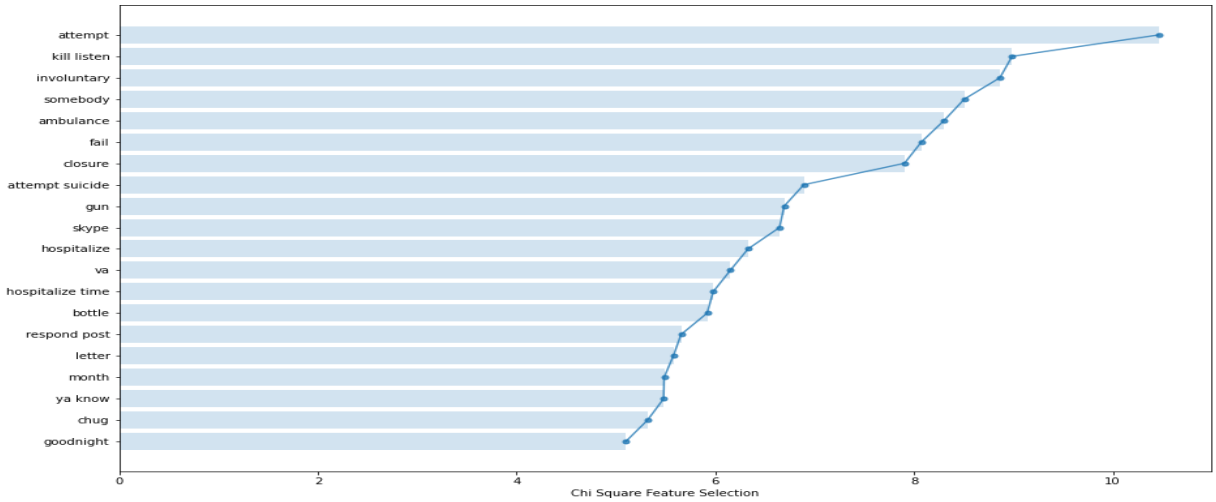

Figure 7: Chi-Square Feature Selection of Documents.

6.4 Performance Evaluation of Text CNN and BiLSTM Models

The loss curve is used to debug the learning style of the neural network. The loss curve of both training data and validation data provides the loss incurred in the network for each epoch. The train/validation loss curve for both models is a good learning rate. Confusion metrics account for the documents that are predicted appropriately. The details of the graph for each model is explained below:

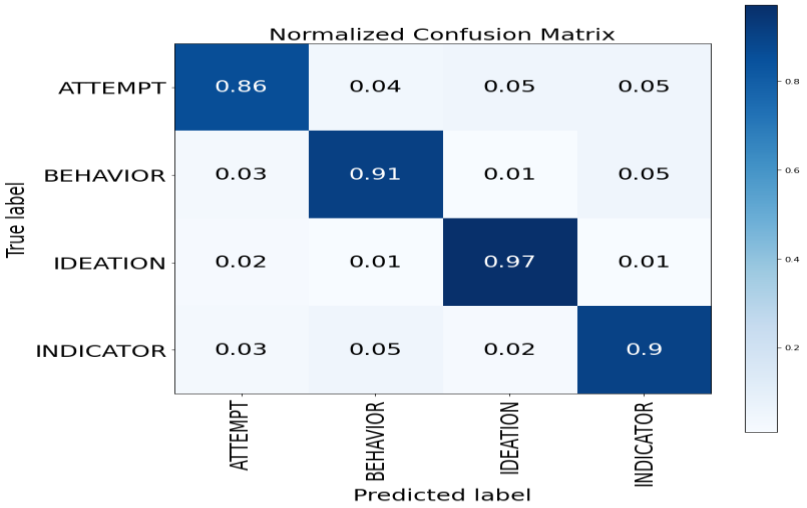

Figure 8: Confusion Matrix for Text CNN Model

The Loss Graph of text CNN shows that training and validation are correlated to each other during the 15 epochs off training the data. They seem to have a strong positive correlation with each other and are learning at a constant rate. The training phase of the model is reliable to apply to external datasets. The confusion matrix shows that all the labels are predicted to appropriate labels. Around 80% of the labels are correctly predicted. Especially, high-risk categories like Suicide Attempt and Suicide Behavior are have classified accurately which is a good sign. Both the Loss Graph and Confusion Matrix are shown in the above Figure 8 and Figure 9. The Text CNN model has given good results.

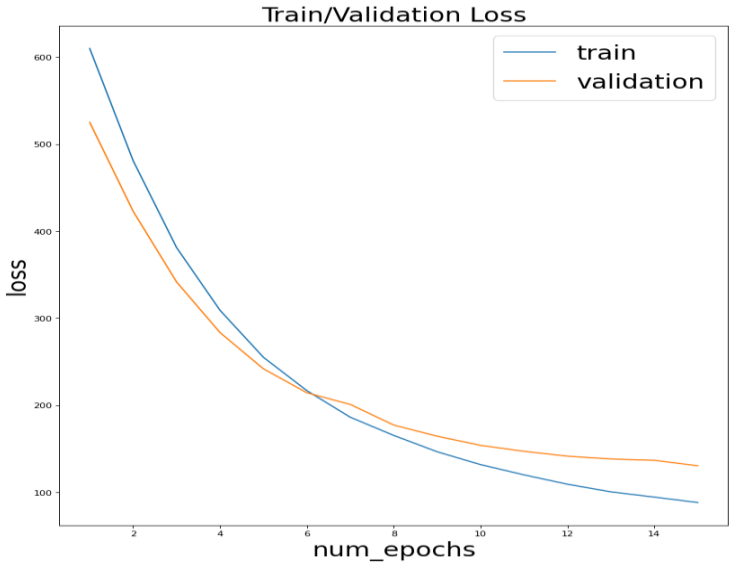

Figure 9: Train/validation Loss Graph - TextCNN

The Loss Graph of BiLSTM and its Confusion matrix is shown in the below Figure 10 and Figure 11. The train and validation curve seems to be strongly correlated until the 6th epoch. After which there is a glitch in the correlation between them. The neural network would not be working well with the data from the previous states. Although the results are positive and good, the trained model may not be robust and reliable with external new documents. Confusion Matrix of the resulting matrix predicts the labels accurately. Again, the highly important classes are classified into their respective class accurately.

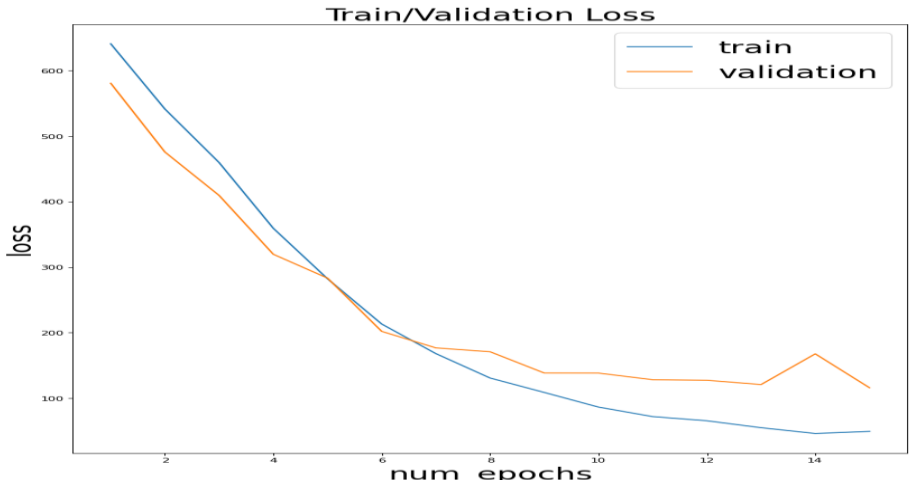

Figure 10: Train/validation Loss Graph - BiLSTM

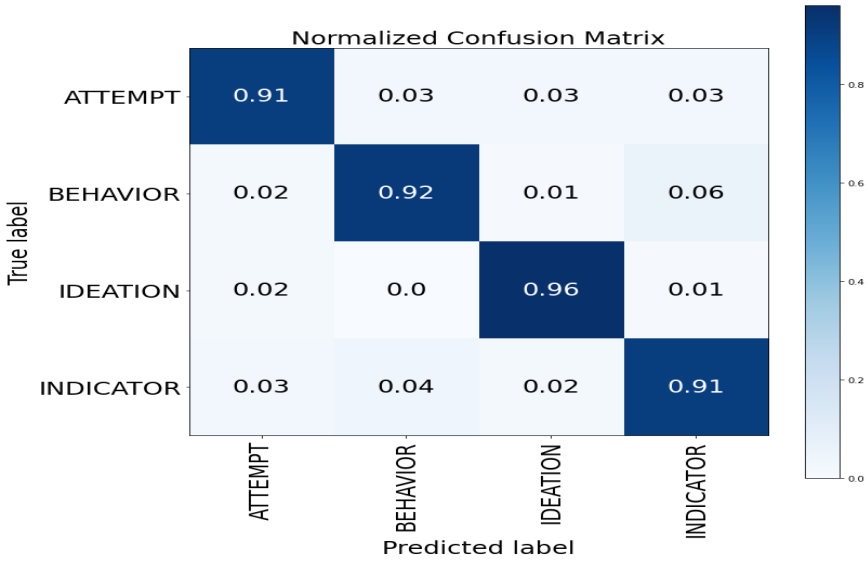

Figure 11: Confusion Matrix - BiLSTM

6.5 Empirical Evaluation of Deep Learning Models

The evaluation metrics for all the Models are mentioned in Table 3. The accuracy of the model increased along with the increase in dataset size, dataset balance, usage of pre-trained word embeddings, and the adoption of Neural Networks.

In multi-class classification problems, the classical evaluation metrics like accuracy, precision, recall, F1 score are calculated for each target class considering one classifier over the whole classifier. Then the average of all the metrics for each of the target classes gives the overall numbers for each model. All the numbers are pulled from the confusion matrix and calculated for results.

The accuracy raised with each updated model. Although precision and recall, values remain consistent throughout all the models. Since the task is a multi-class classifier the importance lies in classifying the appropriate categories of data rather than concentrating on one target class. The F1 score is a measure to find a harmonic mean between precision and recall. All the three values- precision, recall, and F1 score are consistent with each other

Table 3: Model Comparison.

Model	Word Embeddings	Accuracy	Precision	Recall	F1 score
		Macro Average Metrics			
LSTM	Word2Vec	45%	-	-	-
LSTM with Data Augmentation	Word2Vec	71%	93%	92%	92%
Text CNN	GloVe	90%	91%	91%	91%
BiLSTM	GloVe	92%	92%	92%	92%

6.6 Discussion

The overall performance of the model improved with the dataset size, balanced dataset, and neural network implementation. The sentiment analysis in the different categories of risk of Suicide was also recognized. The classifier was able to categorize the risks into the target classes appropriately. The research concentrated on four categories of risk namely- Suicide Ideation: Individuals considering or thinking about suicide, Suicide Indicators: Individuals with any mental or health issues, Suicide Behavior: Individuals actions that show their suicidal thoughts and Suicide Attempt: Individuals with pre-defined plan to give up on life. In the research, all the four categories mainly contained common basic bi-grams and tri-grams of words. These words are shown in Table 4. The interesting revelation was that each category was different from each other when the least used or less frequent words in the documents were analyzed. They showed that Ideators mostly had issues with aging, school life, drugs, personality disorders, etc. Indicators showed words like jobless, money issues, no friends, no food. The behavior category showed the emotions of talking to someone for help, making new friends, monotonous work, looking for something new, etc. Attempters used words to thank the people they know, failure, hopelessness, how are they planning to commit suicide, etc. These patterns made it a little different among the target classes. But the Ideation and indicator group seems to have almost a thin line of difference between them. They can be a single category of study for further research. To sum up, Suicide Indicators and Ideators shows the initial development of suicidal thoughts, Suicide Behavior showed the actions that meant suicide. Suicide Attempters are the people who have reached the final stage and well planned on how to end their life.

Table 4: Top bigrams and trigrams

Top bigrams	Top trigrams
Feel like	Feel like hyperactive behavior
Want die	Feel like shit
Year old	Want die want
Hyperactive behavior	Know feel like
Good friend	Play video game
High school	Feel like life
Year ago	Life feel like

People tend to express their thoughts and emotions more pro-actively on social networking sites. Social Media Networks can contribute towards social issues like suicide, depression, mental disorders in a large perspective. The data involved in investigating sensitive matters like Suicide can be overwhelming and difficult. The research helps to fill up the scarcity of data by developing a self-learning model. The ensemble of self-learning and evaluating can be deployed as a centralized monitoring tool on networking sites. The dataset achieved by self-learning impersonates a small sample of population data. The dataset shares similar characteristics and correlation with the real-world data. The stages of committing suicide to start with ideas of suicide, moving on to developing a thought process, then showing the characterizes in one's actions to the final stage of committing suicide is clustering with the dataset in the research. The research figures out the thin line of difference between the four stages. Similarly, the research details can be a behavioral analysis report for the health care workers to understand the pattern in the process of committing suicide.

7 Conclusion and Future Work

With the advancements in social media and machine learning algorithms, there is an abundance of data available to perform text mining. The content on the social media platform is where one can address their problems and support. That is the trend that is being followed now. A lot of text data related to suicidal thoughts and behavior is also available. Hence the research focuses on addressing the suicide risk categories present in the online posts on social media. The research involved tasks like augmenting text data to create a balanced dataset, modeling a self-learner to boost the labeled dataset size, and finally implemented the Text CNN and BiLSTM to render better results. The models showed progressive improvement in the accuracy as the data got balanced, increased in volume, and understood complex features through neural network architecture. The evaluation metrics had an improvement linearly from 40% to 90%.

The field of Deep Learning is emerging with advanced techniques to solve any complex data mining problems. It would be interesting to use Graph Neural networks in analyzing the patterns in text data for future work. There are many more semi-supervised learning techniques like transfer learning and adversarial learning which can be implemented to make the model more robust to noisy data. There are many subreddits in Reddit on sensitive categories like depression, mental disorders. The data from these domains can be transferred and cross-referenced to look for patterns in them.

References

Abdurrahman, Purwarianti, A., 2019. Effective Use of Augmentation Degree and Language Model for Synonym-based Text Augmentation on Indonesian Text Classification, in: 2019 International Conference on Advanced Computer Science and Information Systems (ICACSIS). Presented at the 2019 International Conference on Advanced Computer Science and information Systems (ICACSIS), IEEE, Bali, Indonesia, pp. 217–222. https://doi.org/10.1109/ICACSIS47736.2019.8979733

Baek, J.-W., Chung, K., 2020. Context Deep Neural Network Model for Predicting Depression Risk Using Multiple Regression. IEEE Access 8, 18171–18181. https://doi.org/10.1109/ACCESS.2020.2968393

Basha, N., Ziyan Sheriff, M., Kravaris, C., Nounou, H., Nounou, M., 2020. Multiclass data classification using fault detection-based techniques. Comput. Chem. Eng. 136, 106786. https://doi.org/10.1016/j.compchemeng.2020.106786

Cobos, R., Jurado, F., Blazquez-Herranz, A., 2019. A Content Analysis System That Supports Sentiment Analysis for Subjectivity and Polarity Detection in Online Courses. IEEE Rev. Iberoam. Tecnol. Aprendiz. 14, 177–187. https://doi.org/10.1109/RITA.2019.2952298

D'Alfonso, S., 2020. AI in mental health. Curr. Opin. Psychol. 36, 112–117. https://doi.org/10.1016/j.copsyc.2020.04.005

Dasgupta, S., Ng, V., 2009. Discriminative models for semi-supervised natural language learning, in: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning

for Natural Language Processing - SemiSupLearn '09. Presented at the the NAACL HLT 2009 Workshop, Association for Computational Linguistics, Boulder, Colorado, pp. 84–85. https://doi.org/10.3115/1621829.1621840

Dong, Yongfeng, Fu, Y., Wang, L., Chen, Y., Dong, Yao, Li, J., 2020. A Sentiment Analysis Method of Capsule Network Based on BiLSTM. IEEE Access 8, 37014–37020. https://doi.org/10.1109/ACCESS.2020.2973711

Elliott, B., Warren, J., Darragh, M., Goodyear-Smith, F., 2019. Towards a Youth Mental Health Screening Analytics Tool, in: Proceedings of the Australasian Computer Science Week Multiconference on - ACSW 2019. Presented at the the Australasian Computer Science Week Multiconference, ACM Press, Sydney, NSW, Australia, pp. 1–9. https://doi.org/10.1145/3290688.3290717

Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp. 1746–1751. https://doi.org/10.3115/v1/D14-1181

Kumar, E.R., Rao, A.K.V.S.N.R., 2019. Suicide Prediction in Twitter Data using Mining Techniques: A Survey, in: 2019 International Conference on Intelligent Sustainable Systems (ICISS). Presented at the 2019 International Conference on Intelligent Sustainable Systems (ICISS), IEEE, Palladam, Tamilnadu, India, pp. 122–131. https://doi.org/10.1109/ISS1.2019.8907987

Matykiewicz, P., Pestian, J., n.d. Effect of small sample size on text categorization with support vector machines 9.

Shuai, H.-H., Shen, C.-Y., Yang, D.-N., Lan, Y.-F.C., Lee, W.-C., Yu, P.S., Chen, M.-S., 2018. A Comprehensive Study on Social Network Mental Disorders Detection via Online Social Media Mining. IEEE Trans. Knowl. Data Eng. 30, 1212–1225. https://doi.org/10.1109/TKDE.2017.2786695

Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D., n.d. Semi-supervised recursive autoencoders for predicting sentiment distributions 11.

Tadesse, M.M., Lin, H., Xu, B., Yang, L., 2019. Detection of Depression-Related Posts in Reddit Social Media Forum. IEEE Access 7, 44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180

Trotzek, M., Koitka, S., Friedrich, C.M., 2020. Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. IEEE Trans. Knowl. Data Eng. 32, 588–601. https://doi.org/10.1109/TKDE.2018.2885515

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, 1996. From Data Mining to Knowledge Discovery in Databases. AI Mag. 17. https://doi.org/10.1609/aimag.v17i3.1230

Venek, V., Scherer, S., Morency, L.-P., Rizzo, A.S., Pestian, J., 2017. Adolescent Suicidal Risk Assessment in Clinician-Patient Interaction. IEEE Trans. Affect. Comput. 8, 204–215. https://doi.org/10.1109/TAFFC.2016.2518665