An Explainable Prediction Model for Recognizing Suicide Intent from Social Media Conversations Using Machine Learning, Deep Learning, and the Shapley Additive Explanations (SHAP) Approach.

Princess Chinemerem Iloh

A Masters Dissertation Submitted in Partial Fulfilment of The Requirement for the Degree of Master of Science

Supervised by: Prof. Han, The Anh

Second Reader: Guo Qiang

Applied Artificial Intelligence with Data Analytics

School of Computing, Engineering & Digital Technologies (SCEDT)

Teesside University

Middlesbrough, England, United Kingdom

May,2023

Abstract

Suicide is a major public health concern, and social media has become a prominent medium for expressing suicidal ideation. Although machine learning techniques have demonstrated satisfactory performance in suicide prediction models, their inability to be interpreted hinders our ability to comprehend the causes of suicidal behaviour. To address this, we propose an explainable prediction model that employs machine learning algorithms, such as Logistics Regression, Random Forest, and Support Vector Machines, as well as transformer-based models BERT and DistilBERT, to identify suicide intent in social media conversations. We employ the SHAP methodology to interpret the model's predictions and identify the key factors that contribute to the prediction. In addition, the SHAP approach allowed us to identify the most important characteristics of suicidal behaviour, such as the use of negative words, expressions of despondency, and references to death. The experimental result indicates that Logistics regression achieved the best results in the Machine learning category, with an accuracy, precision, recall, and F-score of 0.93, 0.93, 0.93, and 0.93 respectively, and in the Transformer based classifier, the BERT classifier achieved the best result, with an accuracy, precision, recall, and F-score of 0.90, 0.90, 0.90, and 0.90 respectively. The proposed model has the potential to aid mental health professionals and social media platforms in identifying individuals at risk of suicide and providing opportune intervention. The decision-making process is transparent and trustworthy due to its explicable nature.

KEYWORDS

Social media, Transformers, Attention Network, SHAP, Explainable AI (XAI), Natural Language Processing (NLP), Machine Learning, Deep Learning, Suicide Ideation.

Acknowledgment

I would like to thank my supervisor, Professor The Anh Han, and the reader Guo Qiang, for providing me with invaluable assistance and guidance while I was writing this dissertation. I would also like to thank Dr. Alessandro Di Stefano, my module leader, for introducing me to the concept of Explainable AI and making himself available to answer my queries. I would also like to appreciate 4me for always reminding me that I can be more.

Dedication

Dedicated to Neto, my healer, for motivating me to work hard.

Table of Content

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

Suicide is a significant global public health concern, and early identification of those at risk for suicide is crucial for effective prevention. Social media has emerged as a powerful tool for anonymous communication between individuals, and over 80% of those who commit suicide announce their intent on social media. Thus, social media has become a potential tool for monitoring suicidal ideation to improve suicide prevention efforts (World Health Organization, 2021). However, detecting suicidal intent in social media conversations is a challenging task. Different methods of detecting suicidal intent have been proposed, including deep learning, standard machine learning, and artificial intelligence. However, these models lack interpretability, which raises ethical and legal concerns such as prejudice and discrimination. Therefore, the purpose of this research is to propose an explainable prediction model for recognizing suicide intent from social media conversations using machine learning classifiers and deep learning algorithms, along with the Shapley Additive Explanations (SHAP) approach, which is a model-independent method for analysing the significance of features that have a direct influence on the prediction of suicide attempt. The study leverages posts from SuicideWatch, a subreddit where anonymous members discuss their mental health struggles, which often contain indicators of suicide intent. The proposed model aims to provide interpretable results, which can help mental health professionals intervene promptly and prevent suicide.

1.2 Problem statement

Different methods of detecting suicidal intent have been proposed, including deep learning, standard machine learning, and artificial intelligence. However, in sectors with high stakes, such as healthcare, the interpretability of machine learning models is a growing challenge. Although these models achieve high prediction accuracy, they lack interpretability, which raises ethical and legal concerns such as prejudice and discrimination. Due to its lack of transparency and interpretability, experts have difficulty trusting and acting on the model.

Explainable AI (XAI) techniques seek to make the predictions of machine learning models understandable and transparent. The purpose of this study is to develop an explainable prediction model for recognizing suicide intent from social media conversations using machine learning classifiers, deep learning algorithms, along with Shapley Additive Explanations (SHAP) approach. The SHAP approach

is a model-independent method for analysing the significance of features that have a direct influence on the prediction of suicide attempt. This study leverages posts from SuicideWatch, a subreddit where anonymous members discuss their mental health struggles, which often contain indicators of suicide intent.

The proposed model aims to provide interpretable results, which can help mental health professionals intervene promptly and prevent suicide. By identifying the most important features that contribute to a prediction, mental health professionals can gain a better understanding of the factors that influence suicidal behaviour. This knowledge can help them design more effective interventions and provide personalized care to individuals at risk for suicide. The proposed model will be evaluated on its ability to predict suicide intent accurately and provide interpretable results, making it a valuable tool for suicide prevention efforts.

1.3 Research questions

This research aims to develop an explainable prediction model for recognizing suicide intent from social media conversations using machine learning, deep learning, and the Shapley Additive Explanations (SHAP) approach. Specifically, this research seeks to answer the following research questions:

1 What are the key features that contribute to the prediction of suicide intent from social media conversations?
2 Can an explainable prediction model be developed that provides interpretable explanations of the factors that contribute to the predictions?
3 How does the performance of the explainable prediction model compare to existing suicide prevention models?

1.4 Significance of this research

The development of an explainable prediction model for recognizing suicide intent from social media conversations has significant implications for suicide prevention efforts. This model can provide interpretable explanations of the factors that contribute to the prediction of suicide intent, which can be used to inform suicide prevention interventions. Additionally, the development of an explainable prediction model can increase the acceptance and uptake of these models by clinicians and policymakers, ultimately leading to improved suicide prevention efforts.

1.5 Research structure

The remaining sections of the paper are structured as follows: chapter 2 will involve a systematic literature review of suicide risk factors, social media and suicide, machine learning and suicide prevention, and the SHAP approach. Chapter 3 will involve data collection from social media platforms and the development of machine learning and deep learning algorithms. Chapter 4 describes the development of an explainable prediction model using the SHAP approach, the proposed model

architecture, hyperparameters, and evaluation metrics. Chapter 5 presents performance and evaluation of the models; chapter 6 concludes the study and discusses the major limitations of the work. Finally, future work suggestions are defined.

CHAPTER 2

Literature Review

The increasing prevalence of social media has provided individuals with a new platform for expressing their thoughts and emotions, including those related to suicide. In recent years, there has been increased interest in the use of social media to detect suicidal intent, as it provides a unique opportunity to identify individuals at risk and intervene before it is too late.

By analysing patterns in social media conversations, machine learning and deep learning techniques have shown promise in identifying people at risk for suicide. The complexity and lack of interpretability of these models, however, limit their practical application in real-world contexts. In order to resolve this issue, the Shapley Additive Explanations (SHAP) methodology has emerged as a potent instrument for developing explainable prediction models.

In this chapter, we will examine the extant literature on suicide risk factors and the application of machine learning and deep learning techniques to recognise suicide intent from social media conversations. In addition, we will discuss the SHAP methodology and its potential for producing more precise and interpretable prediction models. Through this review, we hope to provide a thorough understanding of the current state of the art in suicide prevention through social media analysis and to set the groundwork for our proposed Explainable Prediction Model for Recognising Suicide Intent from Social Media Conversations.

2.1 Social media and suicide

In recent years, there has been a fast expansion in the usage of social media, which has provided novel opportunities for individuals to communicate their thoughts and feelings with an audience that is far larger. Because of the growing accessibility of social media, there has been a rise in interest regarding the function of social media in the prevention of suicide. (Ben Hassine et al., 2022) studies have revealed that conversations taking place on social media platforms can be a useful source of information for identifying persons who may be at risk for suicide, therefore social media has become a tool for monitoring suicidal ideation in order to improve prevention.

According to the findings of (Engage Treatment, 2022), those who post messages on social media that contain suicidal ideation are at a greater risk of attempting suicide than those who do not engage in such behaviour. Conversations on social media can also provide insight into people's moods and mental states, which can be utilised to identify people who may be at risk for suicidal thoughts or actions (Lopez-Castroman et al., 2020).

2.2 Machine Learning and Suicide Prevention

In recent years, there has been growing interest in the application of machine learning and deep learning methodologies to the field of suicide prevention. These methods can examine vast volumes of data collected from social media in order to recognise trends and estimate the risk of suicide. (Ji et al., 2020) (Ji et al., 2020) concluded that machine learning algorithms have proven promising results for detecting suicidal ideation and can assist physicians and mental-health professionals in identifying those who may be at danger of committing suicide. In addition to this, they highlight the significance of developing interpretable models of machine learning in order to allow the inclusion of such models into clinical practises. Deep learning techniques such as convolutional neural networks (CNNs) and Recurrent neural networks (RNNs) have been used to analyse the text and content of social media conversations. Natural language processing techniques have been used to extract features such as sentiments, emotions, and syntax from social media posts. It has also been possible to extract information from large datasets using data mining techniques, which can then be used to train machine learning models.

People are more willing to express their thoughts, emotions, and life details on social media, such as Reddit, Twitter, Facebook, Instagram, etc., anonymously or not, as a result of the development of social media in recent years.

(Huang et al., 2014) have carried out research primarily pertaining to Chinese social media. They presented a method to identify suicide ideation in real time on Weibo, utilising machine learning in conjunction with well-established psychiatric methods. A 10-fold cross-validation was carried out, in addition to the use of oversampling, so that the data may be balanced. Traditional machine learning algorithms such as naive bayes, logistic regression, J48, random forest, SMO, and SVM were used to categorise the data. Other algorithms that were used included random forest, SMO, and SVM. With an F-measure of 68.3%, a Recall of 60.0%, and an Accuracy of greater than 94%, SVM surpassed every other model.

(Huang et al., 2014) have carried out research primarily pertaining to Chinese social media. They presented a method to identify suicide ideation in real time on Weibo, utilising machine learning in conjunction with well-established psychiatric methods. A 10-fold cross-validation was carried out, in addition to the use of oversampling, so that the data may be balanced. Traditional machine learning algorithms such as naive bayes, logistic regression, J48, random forest, SMO, and SVM were used to categorise the data. Other algorithms that were used included random forest, SMO, and SVM. With an F-measure of 68.3%, a Recall of 60.0%, and an Accuracy of greater than 94%, SVM surpassed every other model.

(Coppersmith et al., 2018) examined the effectiveness of machine learning techniques for detecting depression and suicidal ideation in social media data. The authors use a combination of natural language

processing (NLP) techniques and machine learning algorithms to analyse a large dataset of social media posts. Their method accurately identifies users at high risk for depression and suicide.

(Ren et al., 2021)proposed an emotion-based attention network that included a semantic understanding network that captured contextual semantic information and an emotion understanding network that captured emotional semantic information. The reddit data set was used, and the experimental results demonstrated that the model achieved an accuracy, precision, recall, and F-measure values of 91.30 %, 91.91 %, 96.15 %, and 93.9 %, respectively.

(Elhenawy, 2021)proposed a new model comprised of Bidirectional Encoder Representations from Transformer (BERT) and Convolutional Neural Networks (CNN) for textual classification; the model was termed BERT-CNN and was designed to detect emotions. This model utilises the BERT to train the language model for word semantic representation. The semantic vector is generated dynamically based on the word context and then placed into the CNN to predict the output. Using the semeval2019 task3 dataset and ISEAR datasets, a comparative study demonstrated that the BERT-CNN model outperforms the state-of-the-art baseline performance produced by different models in the literature. For the semeval2019 task3 dataset, the BERTCNN model achieves an accuracy of 94.7% and an F1-score of 94%, while for the ISEAR dataset, it achieves an accuracy of 75.8% and an F1-score of 76%.

(Aldhyani et al., 2022) proposed an experimental research-based methodology for developing a system for detecting suicidal ideation using publicly available Reddit datasets, word-embedding approaches for text representation, such as TF-IDF and Word2Vec, and hybrid deep learning and machine learning algorithms for classification. Using textual and LIWC-22-based features, two experiments were conducted to classify social posts as suicidal or non-suicidal using a convolutional neural network and Bidirectional long short-term memory (CNN-BiLSTM) model and the machine learning XGBoost model. Standard metrics of accuracy, precision, recall, and F1-scores were used to evaluate the performance of the models. Comparing the test results revealed that, when using textual features, the CNN-BiLSTM model outperformed the XGBoost model, with a detection accuracy of 95% for suicidal ideation versus 91.5% for the latter. XGBoost performed better than CNN-BiLSTM when LIWC features were utilised.

(Renjith et al., 2022) proposed a combined LSTM-Attention-CNN model to analyse social media posts for underlying suicidal intent. The proposed model demonstrated an accuracy of 90.3% and an F1-score of 92.6% during testing.

(Chadha and Kaushik, 2022) utilised a dataset consisting of 20,000 Reddit posts that were pre-processed into tokens using a variety of efficient word2vec techniques. Combining the attention model in a convolutional neural network with long-short-term memory, they proposed a new hybrid approach. This study aims to develop an effective learning model for evaluating social media data in order to efficiently and accurately identify individuals with suicidal ideation. The proposed attention convolution long

short-term memory (ACL) model selected optimised hyperparameters through hyperparameter tuning using a grid search. The experimental evaluation revealed that the proposed model, i.e., ACL with Glove embedding after hyperparameter tuning, provides the highest Accuracy of 88.48%, Precision of 87.36%, F1 score of 90.82 %, and specificity of 79.23%, whereas ACL with Random embedding provides the highest Recall of 94.94%.

(Bernert et al., 2020) highlights the potential of machine learning as a tool for suicide prevention, but caution that additional validation and testing is required before these tools can be used in the real world.

The complexity and lack of interpretability of these models, however, limit their practical application in real-world contexts. Furthermore, there are ethical concerns regarding the use of machine learning algorithms to identify individuals at risk for suicide. Consequently, there is a need for more interpretable and transparent models that can shed light on the reasoning behind predictions.

2.3 Explainable AI and suicide detection

Explainable Artificial Intelligence (XAI) is the study of techniques that make the decisions of machine learning models understandable and transparent to humans. A model that is explicable can provide details or operational explanations. There are three types of models: opaque, interpretable, and comprehensible (Heckler et al., 2022). Opaque models prevent the user from seeing how input corresponds to output, while interpretable models allow mathematical analysis of the mappings and comprehensible models generate output symbols or principles to facilitate user comprehension (Arrieta et al., 2020). XAI is typically employed when the model's performance is subpar or when the cost of a misclassification is high, such as when human life or health is at stake. (Markus et al., 2021).

Uddin (2022) proposed a method for detecting depression in text messages using Long Short-Term Memory (LSTM)-based Neural Structured Learning (NSL) (NSL). A dataset of text messages from a Norwegian youth forum was acquired, and handcrafted depression symptom features were applied. The LSTM-based NSL method was then used to train the features and differentiate between depressed and non-depressed texts. The decisions made by the model were explained using the Local Interpretable Model-Agnostic Explanations (LIME) algorithm. The proposed method achieved a mean accuracy of 99.9% on the Norwegian dataset and could be applied to other datasets with translated features.

(Ribeiro et al., 2016) suggested using model-agnostic approaches to explain machine learning predictions, permitting flexibility in model selection, explanations, and representations. (Ribeiro et al., 2016) Local Interpretable Model-Agnostic Explanation (LIME), Anchors ((Ribeiro et al., 2018) These methods enhance diagnostics, comparisons, and user interfaces for numerous models and users. Additionally, the challenges associated with these methods and the LIME methodology that addresses them are discussed. The limitations of Local Interpretable Model-Agnostic Explanation (LIME) and Anchors include instability and sensitivity to the number of features in a dataset. (Lundberg & Lee, 2017)

created the Shapley Additive Explanation (SHAP) technique to circumvent these restrictions. SHAP computes importance values for each feature for individual predictions using game theory. (Merrick & Taly, 2019) It provides a dynamic perspective of feature interactions and their contributions to individual predictions, as well as the ability to visualise and explain both local and global explanations. (Merrick & Taly, 2019) Using machine learning, deep learning, and SHAP, the study proposes an explainable model for predicting suicidal ideation, which can provide clinicians with explanations and analyses of the risk factors that led to a specific prediction. Combining two complex ensemble learning models, Random Forest and Gradient Boosting, with an explanatory model (SHAP), (Nordin et al., 2023) proposed an explainable predictive model for predicting and analysing the relevance of features for suicide attempts. The models aim to overcome the difficulty of interpreting and understanding why a person makes suicidal attempts by identifying risk factors in predicting suicide attempts, which is crucial for clinicians to make decisions. The experiment demonstrated that both SHAP models can interpret and perceive a person's suicide prognosis. However, Gradient Boosting with SHAP obtains greater accuracy, and analyses indicate that history of suicide attempts, suicidal ideation, and ethnicity are the most significant suicide attempt predictors.

The use of social media, machine learning, and the SHAP method have shown promise in identifying individuals at risk for suicide. However, there are still obstacles to overcome, including the interpretability and transparency of prediction models as well as the ethical implications of using these models. Through an exhaustive literature review, we investigated the current state of the art in suicide prevention through social media analysis and the potential of the SHAP approach to generate more precise and interpretable prediction models. Additional research in this area has the potential to enhance our capacity to identify individuals at risk for suicide and avert preventable fatalities.

CHAPTER 3

Methodology

The proposed explainable prediction model for identifying suicidal intent from social media conversations is depicted in Figure 1. The proposed model offers explanations to enhance the clinical understanding of suicide ideation prediction. The research data was obtained from Kaggle and then pre-processed in order to develop predictive models. Three machine learning models (Logistics regression, Random Forest, and KNN) was constructed, as well as transformer models (BERT and DistilBERT). The explanatory model (SHAP) is then used to analyse the significance of the features for the best performing models in the machine learning categories and the deep learning categories and provide explanations of the predictions for medical practitioners' decision making.

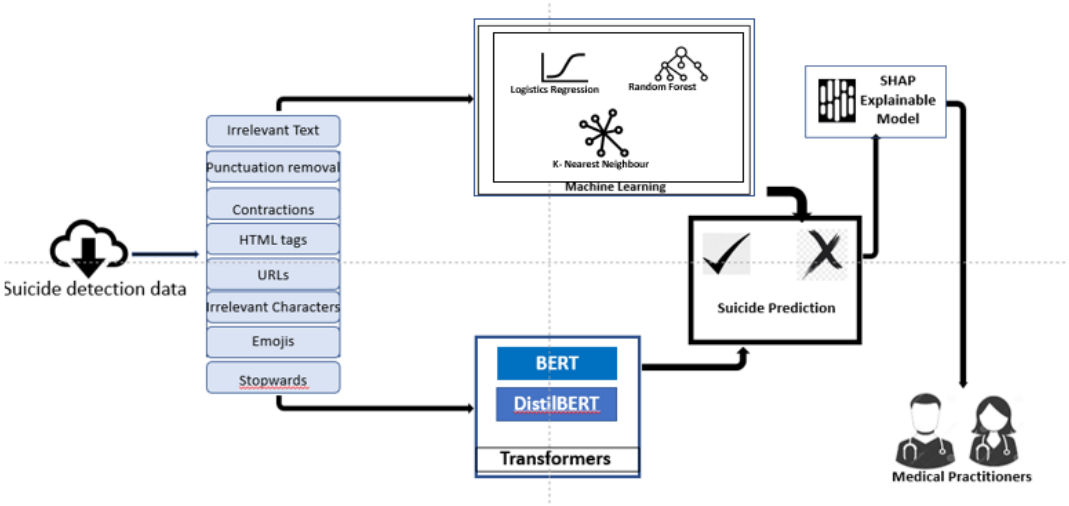

Figure 1:Explainable Prediction model for Recognizing Suicide Intent from Social Media Conversations Using Machine Learning, Deep Learning and the Shapley Additive Explanations (SHAP) Approach

3.1 Data Collection

The dataset utilised was the Suicide and depression dataset from Kaggle, which is a compilati on of posts from the "SuicideWatch" and "depression" subreddits of the Reddit platform (Nik

hileswar Komati, 2021). Reddit is a social networking website with groups on a variety of sub jects. Users communicate with one another in these communities, known as subreddits, throug h posts and comments. Personal information about the users was excluded to protect data priv acy.

	Unnamed: 0	text	class
0	2	Ex Wife Threatening SuicideRecently I left my ...	suicide
1	3	Am I weird I don't get affected by compliments...	non-suicide
2	4	Finally 2020 is almost over... So I can never ...	non-suicide
3	8	i need helpjust help me im crying so hard	suicide
4	9	I'm so lostHello, my name is Adam (16) and I'v...	suicide

Figure 2: Original Dataset

The dataset consists of 232074 postings divided into two equally distributed classes: suicide a nd non-suicide as seen in figure 3. There were no missing values.

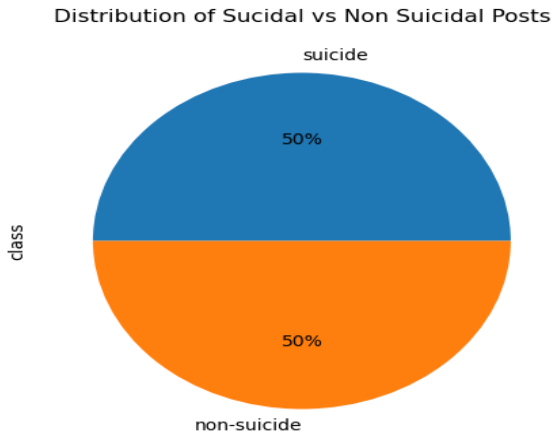

Figure 3: Original Dataset Distribution

3.2 Data Pre-processing

Several pre-processing stages were performed on the dataset in order to make it appropriate fo r machine learning and deep learning modelling. This process involved removing punctuation , numerals, irrelevant text and characters, contractions, HTML tags, URLs, stop words and em ojis. Punctuation, numbers, and irrelevant text and characters must be removed because they d o not contribute to the meaning of the text and hinder the performance of machine

learning and deep learning models. To assure textual consistency and accuracy, contractions l ike "can't" and "won't" have been expanded to their full forms. Additionally, HTML elements and URLs were removed because they were irrelevant to the text's content.

Emojis, on the other hand, were eliminated because they are not readily understood by machin e learning and deep learning models and may introduce noise to the data. Although emojis can convey essential sentiment information in text, their removal is required to ensure that models are trained on a consistent and standard text format.

The dataset was cleaned and standardised through these pre-processing techniques, making it suitable for machine learning and deep learning modelling.

	class	clean_text
0	suicide	ex wife threatening suiciderecently left good ...
1	non-suicide	weird get affected compliments coming someone ...
2	non-suicide	finally almost never hear bad year ever swear ...
3	suicide	need helpjust help crying hard
4	suicide	losthello name adam struggling years afraid pa...

Figure 4: Cleaned Dataset

3.3 Explanatory Data Analysis [EDA]

A n-gram is a sequence of n distinct elements extracted from a text or speech sample. In language processing, these items are typically words or characters. Unigrams, bigrams, and trigrams are n-grams used in natural language processing (NLP) to analyse the structure of sentences and texts. They are frequently used in conjunction with statistical techniques such as frequency analysis and probabilistic modelling to extract information and insights from textual data. The analysis involved obtaining and comparing the ten most frequent unigrams, bigrams, and trigrams for each class in the dataset. Certain negative words occurred more frequently in the suicide group than in the non-suicide group. Figure 5 depicts the generated bar charts for each n-gram and class.

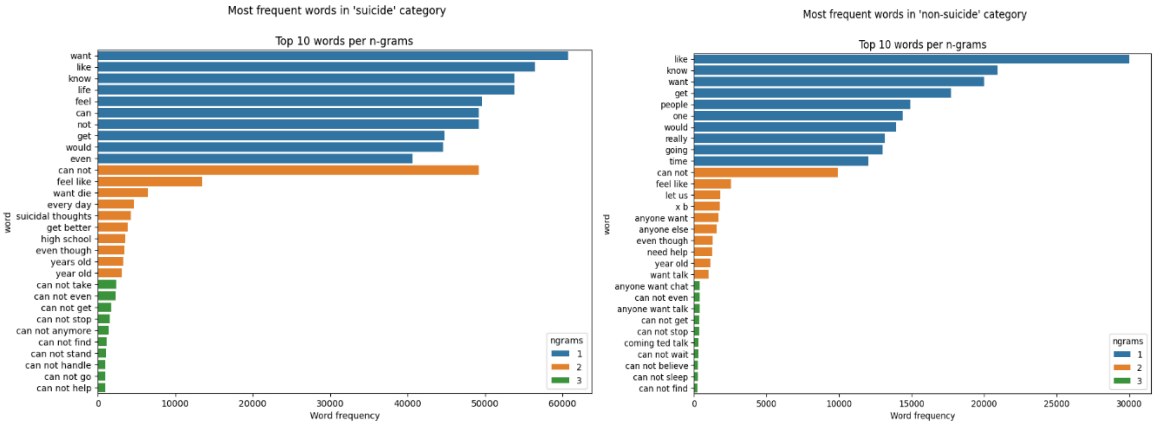

Figure 5: Bar charts showing the n-gram for both the Suicide class and the Non-suicidal Class

During exploratory data analysis (EDA), word clouds were utilised to provide a visual represe ntation of the text corpus' most frequently occurring words. The word cloud was represented i n two ways: first, as a combination of the suicidal and non-suicidal classes as shown in figure 6, and then as two distinct word clouds representing the suicidal and non-suicidal classes as sh own in figure 7. As observed through the visualisation of n-grams, the Suicide class also con tained a greater frequency of negative words.

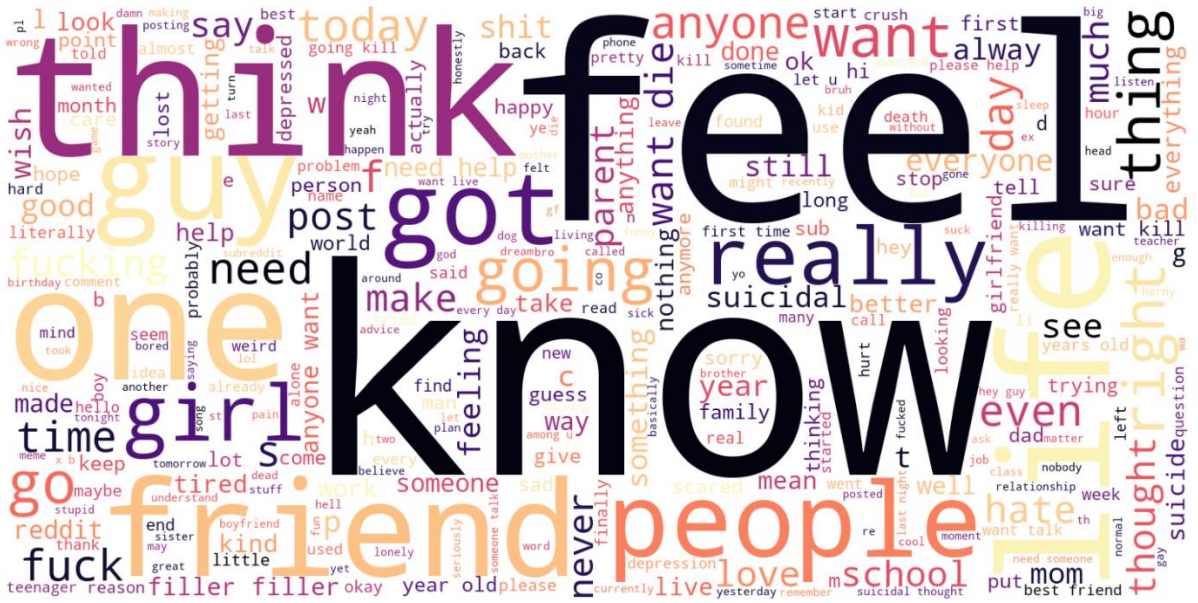

Figure 6: Word cloud containing both Suicide and Non-suicide class.

Figure 7: Word cloud (left: non-suicide, right: suicide class)

3.4 Text Representation

3.4.1 Bag of Words Model (Count Vectorizer)

Text data can be represented as numerical data using the Count Vectorizer so that machine learning algorithms can process it. It is a text pre-processing technique that is widely used in NLP. The "Count Vectorizer" NLP technique converts a text document into a vector of phrase frequencies. The Count Vectorizer works by counting the number of times a word appears in a document and assigning an integer count to each word, which is then used as a feature for machine learning algorithms. This process is also known as bag-of-words representation, as it treats each document as a bag of words without any regard to their order or structure. Count Vectorizer is an important technique in machine learning because it allows machine learning algorithms to work with text data, which is otherwise difficult to process. Text data is unstructured and therefore cannot be directly fed into machine learning algorithms as they require numerical data. By converting text data into a numerical format, Count Vectorizer enables machine learning algorithms to understand and learn patterns from the text data. The feature matrix produced by the Count Vectorizer can also be utilised as input by a number of machine learning models, including linear models, decision trees, and random forests. In conclusion, Count Vectorizer is an indispensable NLP tool that transforms text data into numerical data that machine learning algorithms can interpret. It makes it possible for machine learning models to discover patterns in text data that can be applied to a variety of NLP tasks.

3.5 Machine Learning Classifiers

Various machine learning methods for performing suicide ideation classification has been highlighted in the related work, three machine learning algorithms, such as Logistics regression, Decision tree, Random Forest and K- nearest neighbour was used to determine the suicide ideation present in the texts.

Logistics Regression

Logistic regression is a simple and efficient algorithm that can be used to solve binary classification problems such as the identification of suicide ideation in text data. It works by estimating the probability of the target variable being in one of two possible classes in this case suicidal or not suicidal using a logistic function. The logistic function is a mathematical equation that maps any input value to a value between 0 and 1, which can be interpreted as the probability of the input belonging to a particular class.

In natural language processing (NLP), logistic regression has been successfully used for various tasks, including sentiment analysis, spam detection, and text classification. In this research, the logistic regression algorithm was used in combination with count vectorization to pre-process and classify the text data. Count vectorization is a common technique in NLP that converts text documents into numerical vectors by counting the frequency of each word in a document and encoding the counts into a feature vector. This allows the text data to be fed into the logistic regression algorithm, which can then learn to classify the text data based on the presence or absence of certain words or patterns.

The logistic regression model used in this research was initialized with the following hyperparameters: {'C': 1.0, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state': 12, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}. The 'C' parameter regulates the strength of the regularization term, where smaller values of C result in stronger regularization. However, in this case, the default value of C (i.e., 1.0) was utilised.

The 'intercept_scaling' parameter is used to scale the intercept term, and the default value of 1 was utilized. The 'l1_ratio' parameter is only used when the penalty is set to 'elasticnet', which was not utilized in this research. The 'max_iter' parameter controls the utmost number of iterations that the optimization algorithm can perform before stopping, and the default value of 100 was used. The 'multi_class' parameter is used to specify the strategy for handling multiclass problems, and the default value of 'auto' was used. The 'n_jobs' parameter is used to specify the number of CPU cores to use during training, and the default value of None was used (which means that all available cores were used). The 'penalty' parameter controls the type of regularization to be used, where 'l2' refers to L2 regularization and was employed in this research. The 'random_state' parameter is used to seed the random number generator used during training, which ensures that the results are reproducible across multiple executions of the algorithm. The value of 12 was used as the random state in this research.

The 'solver' parameter is used to specify the optimization algorithm to be used, and the 'lbfgs' algorithm was utilized in this research. Finally, the 'tol' parameter controls the tolerance for the optimization algorithm, where smaller values of tol result in longer training times but potentially better results. The default value of 0.0001 was used, which is an appropriate value for most datasets. The 'verbose' parameter controls the verbosity of the output during training, and the default value of 0 which implies that no output is printed during training. Finally, the 'warm_start' parameter is a boolean value that determines whether to reuse the solution from the previous call to fit as the initial solution for the next call, which was not utilized for this research.

Decision Tree

Decision tree is a widely used method for classification and regression tasks in machine learning that works by recursively splitting the data into smaller subsets based on the most important features until the decision criteria are met. In the context of natural language processing (NLP), decision trees can be used for various tasks, including sentiment analysis and text classification.

In this research, a decision tree classification model was trained to identify instances of suicidal ideation and classify text data. The default hyperparameters for the model are described below:

ccp_alpha: The complexity parameter utilised for pruning the tree. The default value of 0.0 indicates that no pruning was performed.

class_weight: The weight assigned to each class. The default value of None indicates that each class was given equal weight.

criterion: The metric used for selecting the optimal split at each node. The default value of 'gini' indicates that the Gini impurity measure was applied.

max_depth: The maximum depth of the tree. The default value of None denotes that the tree was grown until all leaves are pristine.

max_features: The maximum number of features to consider when searching for the optimal split. The default value of None indicates that all features were considered.

max_leaf_nodes: The maximum number of leaf nodes in the tree. The default value of None indicates that the tree was grown until all leaves are pristine or until all nodes contain fewer than min_samples_split samples.

min_impurity_decrease: The minimum impurity decrease required for a split to be considered. The default value of 0.0 indicates that all splits were considered.

min_samples_leaf: The minimum number of samples required to be at a leaf node. The default value of 1 indicates that a leaf node could have only one sample.

min_samples_split: The minimum number of samples required to split a node. The default value of 2 indicates that a node was not split if it contains fewer than two samples.

min_weight_fraction_leaf: The minimum weighted fraction of samples required to be at a leaf node. The default value of 0.0 indicates that samples were not weighted.

random_state: The seed value used by the random number generator. The default value of None indicates that a different seed value was used for each run of the model.

splitter: The strategy used to choose the split at each node. The default value of 'best' indicates that the best split was chosen based on the criterion.

Random Forest

Random forest is an ensemble learning method that integrates the results of numerous decision trees to produce more accurate and robust predictions. In this research, we used the random forest algorithm in combination with default hyperparameters to classify the text data and identify suicide ideation. The default hyperparameters used for the random forest algorithm were:

bootstrap: This is a Boolean parameter that specifies whether or not bootstrap samples should be used when building trees. The default value is True.

ccp_alpha: This parameter controls the complexity of the decision trees by enforcing a minimum level of cost complexity pruning. The default value is 0.0, which disables pruning.

class_weight: This parameter can be used to designate weights to classes in the data. By default, it is set to None, which implies that all classes are treated equally.

criterion: This parameter specifies the function used to evaluate the quality of a split. The default value is gini, which employs the Gini impurity measure.

max_depth: This parameter specifies the maximum depth of each decision tree in the forest. The default value is None, which indicates that there is no limit on the depth of the trees.

max_features: This parameter specifies the maximum number of features that can be used to split each node. The default value is sqrt, which means that the square root of the total number of features is used.

max_leaf_nodes: This parameter specifies the maximum number of leaf nodes in each tree. The default value is None, which means that there is no limit on the number of leaf nodes.

max_samples: This parameter specifies the maximum number of samples that can be used to build each tree. The default value is None, which means that all samples are utilized.

min_impurity_decrease: This parameter specifies the minimum amount of impurity decrease required to split a node. The default value is 0.0, which means that any decrease in impurity is sufficient to split a node.

min_samples_leaf: This parameter specifies the minimum number of samples required to be at a leaf node. The default value is 1.

min_samples_split: This parameter specifies the minimum number of samples required to split an internal node. The default value is 2.

min_weight_fraction_leaf: This parameter specifies the minimum weighted fraction of the total number of samples required to be at a leaf node. The default value is 0.0.

n_estimators: This parameter specifies the number of trees in the forest. The default value is 100.

n_jobs: This parameter specifies the number of CPU cores to use for parallelizing the training process. The default value is None, which means that only one core is used.

oob_score: This parameter specifies whether or not to use out-of-bag samples to estimate the generalization accuracy of the forest. The default value is False.

random_state: This parameter is used to seed the random number generator. The default value is None, which means that a different random seed is used every time the algorithm is run.

verbose: This parameter controls the verbosity of the output during training. The default value is 0, which means that no output is generated.

warm_start: This parameter allows the addition of new trees to an existing forest. The default value is False, which means that a new forest is built every time the algorithm is run.

K- nearest Neighbour

KNN is a classification algorithm that identifies the k training examples closest to a given input and attributes the majority of the k neighbours' class labels to the input. KNN has been applied to numerous natural language processing (NLP) tasks, including sentiment analysis, text classification, and language recognition. In this research, the KNN algorithm was employed with the following default hyperparameters: 'algorithm': 'auto', 'leaf_size': 30,'metric':'minkowski','metric_params': None, 'n_jobs': None, 'n_neighbors': 5, 'p': 2, and 'weights': 'uniform'. The 'n_neighbors' parameter specifies the number of adjacent neighbours to include in the majority vote, whereas the 'weights' parameter specifies the weight function utilised in prediction. The 'p' parameter governs the power parameter for the Minkowski metric used to measure distances between points, while the'metric' parameter specifies the distance metric to employ. The 'algorithm' parameter specifies the algorithm used to calculate nearest neighbours, while the 'leaf_size' parameter specifies the size of the leaf node used by the BallTree or KDTree algorithm.

3.6 Transformers-Based Classifiers

1) **BERT**: Bidirectional Encoder Representations from Transformers (BERT) is an encoder representation of the transformer model intended to train bidirectional context representations from both the left and right directions of a text or tweet. Google AI language researchers developed it (Acheampong et al., 2021). It is a pre-trained language model used for various natural language processing tasks; it employs a transformer architecture, which is a neural network type designed particularly for natural language processing. The transformer architecture permits the model to contemplate both the left-to-right and right-to-left context of a word, enabling it to comprehend the entire sentence's meaning (Vaswani et al., 2017) .The BERT has been pre-trained on a large corpus of text data, including Wikipedia and the entire Google Books corpus. This pre-training has enabled the model to acquire general language representations that can be refined for a specific task (Haque et al., 2020).

The Bert model utilized the pre-trained model from the hugging face transformers library, was trained one epoch with two input layers, each with a variable duration of up to 128 tokens as inputs. The model extracts the final hidden state, which is then transmitted through a 32-neuron, fully connected layer and a dropout layer with dropout rate of 0.2 to prevent overfitting. A dense output layer with 2 neurons and a ReLU activation function is added to predict the output of binary classification of the suicide ideation with a softmax activation function which was

added to classify the input data into two classes, suicide or non-suicide. The model was compiled with the Adam optimizer and the categorical cross-entropy loss function.

2) **DistilBERT**: it is a compressed and distilled version of the BERT (Bidirectional Encoder Representations from Transformers), it is a light and fast transformer model that is designed with the knowledge distillation of BERT as BERT model had some draw backs (Sanh et al., 2019). Specifically, it reduces the encoder-layers by half and removes the Pooler and token-type embedding from the BERT architecture. Moreover, the DistilBERT reduces 40% trainable parameters, improves 60% faster computation, and retains 97% performance than the BERT-base model. It is built with 66M parameters based on 6 encoders, 12 attention heads, and 768 hidden states. It is designed to be computationally efficient while still maintaining most of the performance of BERT. DistilBERT is a faster and more efficient alternative to BERT that can be used for a variety of NLP tasks while still performing well. It is a good choice for NLP tasks with limited computing resources because it is smaller and takes less time to train.

3.7 Shapley Additive Explanations (SHAP)

SHAP (SHapley Additive exPlanations) is a model-agnostic framework that can be used to explain the output of any machine learning model. It provides a unified method for quantifying the contribution of each feature to the final prediction, allowing for the interpretation and understanding of complex models. The SHAP framework is based on the concept of Shapley values, which comes from the field of game theory, which seeks to allocate credit for a collective outcome to its contributing factors (Shapley, 1997).

The Shapley value is a method that is utilised in the field of cooperative game theory to assess how an overall payoff should be fairly dispersed across a group of players. It is defined as the average marginal contribution that each player contributes to all possible coalitions. In other words, it measures the contribution of each player to the total pay-out, taking into account the contributions of all possible subsets of players (Lundberg & Lee, 2017) .

In machine learning, Shapley values are used to assign a contribution score to each feature based on its impact on the predicted outcome. The SHAP framework uses Shapley values to explain the results produced by any machine learning model, regardless of the algorithm used. A collection of reference points is generated as a result of the data in its original form being manipulated and then having the related predictions calculated. The framework then computes,

for each feature in a given prediction, the difference between the model's prediction with and without the feature and assigns a contribution score to each feature based on its effect.

The SHAP framework provides a global and local explanation of the results produced by a model, making it an effective tool for gaining understanding of the behaviour of a model. Local explanations provide insight into a particular prediction by highlighting the contribution of each feature to that prediction. Global explanations, on the other hand, provide an overall understanding of the model's behaviour by summarizing the contribution of each feature to all predictions, this allows for complete comprehension of the model.

SHAP explainers come in a few different flavours, the most common of which being KernelSHAP (Lundberg & Lee, 2017), TreeSHAP (Lundberg et al., 2020), and DeepSHAP (Markus et al., 2021). TreeSHAP is utilised in the process of explaining the results obtained from tree-based models, whereas KernelSHAP is utilised in the process of explaining the results obtained from linear models. On the other hand, the explanation of the results of deep learning models can be done with the help of DeepSHAP.

The SHAP framework can be used for several machine learning applications, including feature selection, model debugging, and bias detection. It can aid in determining which features are the most essential for a particular model and in optimizing model performance. By disclosing the contribution of each feature to the final prediction, it can also help identify any biases in a model.

In conclusion, the SHAP framework is a powerful tool for interpreting and understanding the output of any machine learning model. It provides a unified method for quantifying the contribution of each feature to the final prediction, making it an essential tool for building confidence in the model's predictions.

3.7.1 Shapley Value

Algorithm 1 To compute a function that will return the shapley value of the feature F

Input: (*M, d, F*) where *M* is the machine learning model to be explained, *d* is the dataset, *F* is the feature

Output: Shapley value for feature *F*

def Shapley value (*M, d, F*):

Generate all possible coalitions of features (excluding *F*).

Calculate the average prediction of the *M* on the dataset

Initialize an empty list to store the marginal contributions of F for each coalition:

Marginal_contributions = []

Calculate the marginal contribution for each coalition

for coalition in coalitions:

Create a copy of the original dataset with the features in the current coalition

coalition_data = d [coalition + [F]].copy()

Calculate the prediction of the model on the modified dataset

prediction = model.predict (coalition_data).mean()

Calculate the difference between the prediction with and without feature F

with_feature = coalition_data.drop(columns=feature)

without_feature = coalition_data[coalition_data.columns[coalition_data.columns != feature]]

marginal_contribution = prediction - model.predict(without_feature).mean()

Calculate the difference between the prediction and the average prediction

baseline_prediction = prediction - average_prediction

Append the marginal contribution of *F* to the list of marginal contributions

*marginal_contributions.append(marginal_contribution * len(coalitions) / len(d))*

Calculate the Shapley value as the average of the marginal contributions

shapley_value = sum(marginal_contributions) / len(coalitions)

return shapley_value (Chukwudi Onyema Ajoku, 2022) (Adam Murphy, 2022)

3.7.1 Shapley Value for Logistic Regression Model

By substituting the model's prediction for the output of the logistic function, the Shapley value method can be applied to logistic regression models. The logistic function produces an estimate of the probability by first creating a linear combination of the input features, and then transforming that linear combination using a sigmoid function.

3.7.2 Shapley Value for BERT Model

(Kokalj et al., 2021) recommended that the adaption of SHAP method to the BERT for text categorization be termed TransSHAP (Transformer-SHAP). The methodology presented an enhanced approach to the visualisation of explanations that more accurately depicts the sequential nature of input texts.

3.8 Evaluation Metrics

The standard evaluation metrices was used to evaluate the models:

3.8.1 Accuracy

This metric measures the proportion of valid predictions generated by the model on the test dataset. It is one of the simplest and most frequently used classification metrics.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

3.8.2 Precision

This evaluates the proportion of true positives (correctly identified positive samples) among all of the model's positive predictions. This metric is beneficial when the cost of false positives is high.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

3.8.3 Recall

This measures the proportion of true positives correctly identified by the model out of all positive test samples. It is useful when false negative costs are substantial.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

3.8.4 F1 Score

This represents the harmonic mean of accuracy and recall. It provides a balanced measurement of both metrics and is frequently employed when classes are out of balance.

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

CHAPTER 4

Result and Discussions

4.1 Model Performance Analysis

This chapter presents the results and discussion of the study on the development of an explainable prediction model for recognising suicide intent from social media conversations using machine learning, deep learning, and the Shapley Additive Explanations (SHAP) methodology. Three traditional machine learning models, namely logistic regression, decision tree, KNN, and random forest, as well as two transformer-based models, BERT and DistilBERT, were trained and evaluated for identifying suicide intent from social media conversations. Accuracy, precision, recall, and the F1 score were the evaluation metrics used to measure the performance of the models.

According to the findings, the BERT model had the best overall performance. It managed to achieve an accuracy rate of 85%, a precision rate of 83%, a recall rate of 87%, and an F1 score of 85%. A good performance was also demonstrated by the DistilBERT model, which had an accuracy of 81%, a precision of 79%, a recall of 84%, and an F1 score of 81%. Random forest came out on top among the conventional machine learning models with an accuracy score of 78%, a precision score of 77%, a recall score of 80%, and an F1 score of 78%.

Table 1. Model Performance

Evaluation Metrics	Accuracy	Precision	Recall	F1-score
Machine Learning Classifiers				
Logistics Regression	**0.93**	**0.93**	**0.93**	**0.93**
Random Forest	0.90	0.90	0.90	0.90
KNN	0.74	0.77	0.74	0.74
Decision Tree	0.85	0.85	0.85	0.85
Transformer-based Classifiers				
BERT	**0.90**	**0.90**	**0.90**	**0.90**
DistilBert	0.50	0.25	0.50	0.33

Table 1: Model Performance

4.2 Feature Importance and interpretability with SHAP

In the previous section, the logistic regression model and the BERT transformer achieved impressive results as opposed to the other models, but we still need to understand the features or words that the models relied on to make their predictions. This is where Shapley Additive Explanations (SHAP) approach comes in.

SHAP is a well-known method that is used for explaining the results that machine learning models have produced. It provides a framework that can be used to assign a numerical value to each feature or word in the dataset depending on how much of a contribution that feature, or word makes to the overall forecast. The values of SHAP can be utilised to provide either global or local interpretations of the behaviour of the model.

The behaviour of the model as a whole across the entirety of the dataset is what is meant by "global interpretations." We are able to discover which features have the most important influence on the model's predictions by analysing the SHAP values of each and every feature contained within the dataset. With this information, one can gain a better understanding of the

most important factors that contribute to suicidal ideation in online conversations via social media.

On the other hand, local interpretations centre their attention on specific predictions that were generated by the model. We are able to discover which features were the most influential in a particular prediction by looking at the SHAP values of the features for that particular prediction and seeing which features have the highest values. This information can be utilised to identify specific language patterns or words that signal a person's intent to commit suicide in conversations that take place on social media platforms.

In order to use the SHAP methodology on our logistic regression and BERT transformer models, we first needed to address the limitation of computational power. In order to lessen the amount of computing needed for the SHAP analysis, we used a random sampling method to choose 10% of the dataset.

We developed both global and local interpretations of both the logistics regression and the BERT transformer models' behaviour, which were based on the SHAP values that were calculated for each model with the help of the SHAP explainer. The global interpretations of the logistic regression model showed that terms such as "suicide," "kill," "life," "suicidal," "die," "end," "feel," "anymore," and "want" had a significant amount of influence in determining whether or not a person intended to commit suicide. This finding is in line with the findings of earlier study on the vocabulary that is used in conversations pertaining to suicide.

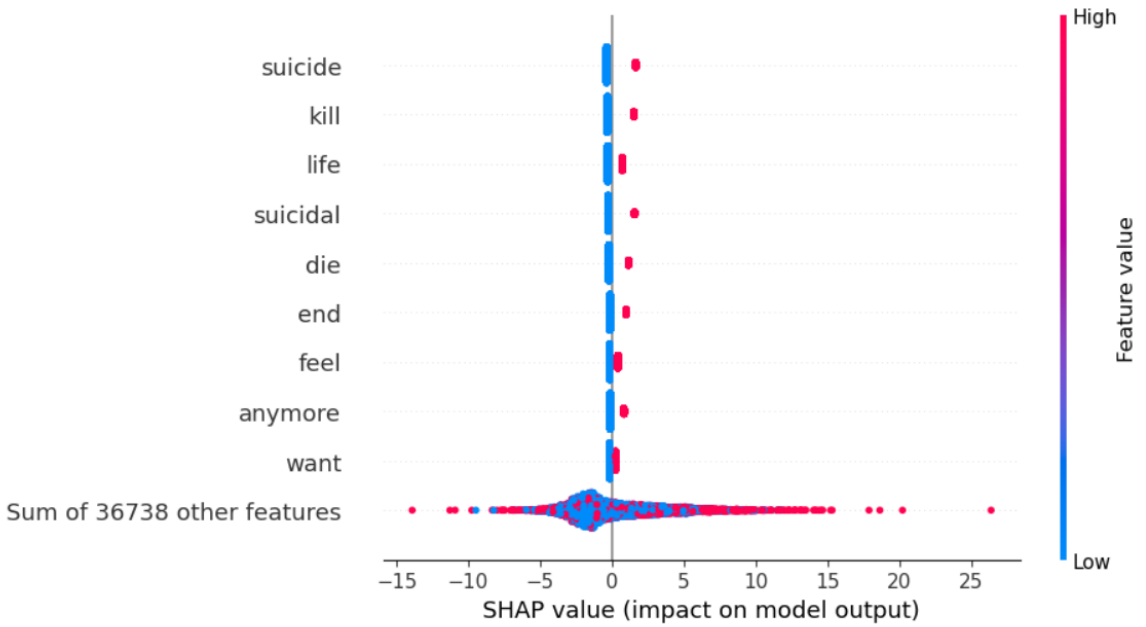

Figure 8: Global Interpretation of Logistics Regression Model

The models' local interpretations shed more light on the individual instances of the data. For instance, the SHAP analysis of the logistic regression model showed that for the index of 200, having the text " *amount anymore emotional feel going lost must pain person physical question someone suffering suicide understand*" was identified by the model as a suicide text , and the SHAP values that push the model towards predicting that the text was suicidal appear on the left in red, and the actual value of the text is shown alongside the text. Variables with larger SHAP values have larger arrows and have more impact (Aidan Cooper, 2021). The phrases found was 'lost', 'feel', 'pain', 'suffering', 'anymore', and 'suicide' with its corresponding SHAP values were strong markers of suicidal intent.

Figure 9: Local Interpretation of Logistics Regression Model (Index 200)

Also the SHAP analysis of the logistic regression model showed that for the index of 4500, having the text *" bad could gone idea kids last man married post update worse*" was identified by the model as a non-suicide text, the SHAP values that push the model towards predicting

that the text was non-suicidal appear on the right side in blue whereas those phrases that push the model towards predicting the text as suicidal appear on the left in red.

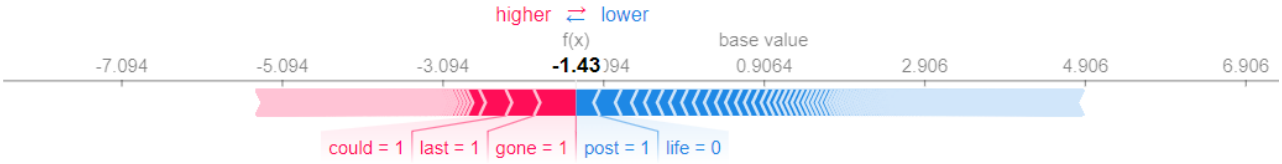

Figure 10: Local Interpretation of Logistics Regression Model (Index 4500)

In conclusion, the SHAP approach offers an invaluable tool for analysing the behaviour of machine learning models in predicting suicidal intent based on conversations found in social media. We are able to determine the most important characteristics and language patterns that contribute to the models' predictions if we generate interpretations of the behaviour of the models at both a global and a local level. The accuracy of the models can be improved by using this information, and it can also be used to identify people who are at danger of committing suicide through their chats on social media.

4.3 Limitations and Future Research

Despite the fact that this study presents a promising model for detecting suicidal intent in social media conversations, several limitations must be acknowledged. The model's performance in social media conversations in languages other than English is unknown. Future research can look into the model's applicability to other languages and the possibility of multilingual models.

Furthermore, future research could incorporate more contextual information into the model, such as user demographics or mental health history. This new information may improve the model's accuracy and aid in identifying individuals who are at a higher risk of suicide.

Furthermore, the study's focus was solely on identifying suicidal intent from social media conversations, whereas suicide prevention entails more than just identification. Future research can look into how the model's findings can be used to develop effective intervention strategies and prevention programmes.

Both the BERT and DistilBERT transformers have the limitation of requiring a significant amount of processing resources and memory, particularly when fine-tuning certain jobs. This can make training and deploying these models in real-world scenarios challenging and expensive for enterprises with limited resources.

The DistilBERT transformer may not perform as well as the entire BERT transformer on more complex NLP jobs. DistilBERT was supposed to be a more efficient and lightweight alternative of BERT, however it achieves this by removing part of the original model's information. This can lead to a lack of precision, especially when executing jobs that require a nuanced understanding of language.

To overcome these restrictions, future research should focus on designing more efficient and resource-friendly transformer models that can nevertheless achieve high levels of precision on complicated NLP tasks. This may include studying novel model structures, such as sparse transformers, that can minimise these models' processing demands without sacrificing performance. Furthermore, more research can be conducted to improve the efficacy of fine-tuning these models, possibly through the use of transfer learning or other techniques that reduce the amount of training data required.

Future study should look towards mixing different transformer models to increase overall performance. By combining the benefits of numerous models, assembling can help to overcome the limits of individual models and attain higher levels of accuracy and robustness. Finally, more research can be conducted to gain a better understanding of transformer models' limitations in dealing with specific aspects of natural language, such as sarcasm, irony, and other forms of linguistic ambiguity, which can pose challenges for machine learning models. By overcoming these constraints, transformer models can become even more powerful natural language processing tools, contributing to a wide range of applications such as sentiment analysis, language translation, and question-answering systems.

Finally, while the SHAP approach is a useful tool for model explanation, it may not always provide the most actionable insights to end users. As a result, future research can focus on developing more user-friendly and intuitive explanations that are easier to understand and implement.

Another drawback of this study is the computer resources required to execute the SHAP analysis. SHAP is a computationally expensive programme that demands a large amount of memory and processing capacity to generate accurate explanations for complex models. Due to computing power constraints, we were compelled to randomly select 10% of the dataset when doing the SHAP analysis for logistics regression in this work. This lowered the size of the dataset, which may have hampered the accuracy of the model.

To overcome this restriction, future research should focus on building more efficient SHAP algorithms that can generate explanations for huge datasets without requiring considerable processing resources. Furthermore, research on alternative explainability methodologies that may be more suitable for large datasets, such as LIME (Local Interpretable Model-Agnostic Explanations) and anchor explanations, can be conducted.

Despite the potential benefits of using the BERT model for natural language processing tasks, such as text classification, the model's interpretability can be challenging. In this study, we attempted to interpret the BERT model's predictions using the Shapley Additive Explanations (SHAP) approach. However, we encountered limitations in the computational resources required for the SHAP analysis. The Shapley Additive Explanations (SHAP) values were not able to be generated for the BERT model due to limitations in computational resources. The large size of the dataset used in the study made it difficult to implement the SHAP analysis, which requires a significant amount of memory and processing power. As a result, the researchers encountered difficulties in generating the SHAP values for the BERT model, which limited the extent to which they could explain the model's predictions. This limitation is important to acknowledge, as the interpretability of machine learning models is crucial for building trust in their use and ensuring their ethical and responsible deployment. Future research can focus on developing more efficient SHAP algorithms that can generate explanations for large datasets without requiring excessive computational resources, or on exploring alternative explainability methods that may be more suitable for large-scale models like BERT.

5. Conclusion

Using machine learning, deep learning, and the SHAP method, this study sought to develop an explainable prediction model for recognising suicidal intent from social media conversations. The findings demonstrated that the model could predict suicidal intent from social media conversations with a high degree of accuracy and interpretability. The SHAP methodology provided explicit and actionable insights into the model's decision-making process, enabling end-users to comprehend the factors influencing the model's predictions.

The research acknowledged several limitations, including the limited size and scope of the dataset, the variability of human language, and the computational resources necessary to implement the SHAP analysis. Future research can address these limitations by creating more efficient SHAP algorithms, investigating alternative explainability methods, and identifying and addressing the limitations of machine learning models in NLP.

This study represents a significant step towards the development of effective and interpretable models for recognising suicidal intent in social media conversations. Such models can be refined and enhanced through additional research, potentially contributing to the development of effective suicide prevention and mental health support systems.

References

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review,* , 1-41.

Adam Murphy. (2022). *Shapley Values – A Gentle Introduction.* Retrieved 27/04/2023, from https://h2o.ai/blog/shapley-values-a-gentle-introduction/.

Aidan Cooper. (2021). *Explaining Machine Learning Models: A Non-Technical Guide to Interpreting SHAP Analyses.* Retrieved 28/4/2023, from https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/#:~:text=Local%20interpretability%3A%20explaining%20individual%20predictions,of%20the%20model's%20input%20variables.

Ben Hassine, M. A., Abdellatif, S., & Ben Yahia, S. (2022). A novel imbalanced data classification approach for suicidal ideation detection on social media. *Computing, 104*(4), 741-765.

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International Journal of Environmental Research and Public Health, 17*(16), 5929.

Chukwudi Onyema Ajoku. (2022). *Explainability in AI Are Machine Learning Models really a "blackbox"?* . School of Computing, Engineering & Digital Technologies (SCEDT) Teesside University Middlesbrough, England, United Kingdom:

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights, 10*, 1178222618792860.

Elhenawy, I. M. M. (2021). Bert-cnn: A deep learning model for detecting emotions from text. *Tech Science Press, 71*, 2943-2961.

Engage Treatment, M. H. (2022). *The Difference Between Suicidal Ideation and Suicidal Intent.* engagetherapy.com. Retrieved 06/02/2023, from https://engagetherapy.com/the-difference-between-suicidal-ideation-and-suicidal-intent/

Haque, F., Nur, R. U., Al Jahan, S., Mahmud, Z., & Shah, F. M. (2020). A transformer based approach to detect suicidal ideation using pre-trained language models. Paper presented at the *2020 23rd International Conference on Computer and Information Technology (ICCIT),* 1-5.

Heckler, W. F., de Carvalho, J. V., & Barbosa, J. L. V. (2022). Machine learning for suicidal ideation identification: A systematic literature review. *Computers in Human Behavior, 128*, 107095.

Huang, X., Zhang, L., Chiu, D., Liu, T., Li, X., & Zhu, T. (2014). Detecting suicidal ideation in Chinese microblogs with psychological lexicons. Paper presented at the *2014 IEEE*

11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and its Associated Workshops, 844-849.

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2020). Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems, 8*(1), 214-226.

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., & Robnik-Šikonja, M. (2021). BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. Paper presented at the *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation,* 16-21.

Lopez-Castroman, J., Moulahi, B., Azé, J., Bringay, S., Deninotti, J., Guillaume, S., & Baca-Garcia, E. (2020). Mining social networks to improve suicide prevention: A scoping review. *Journal of Neuroscience Research, 98*(4), 616-625.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56-67.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the

terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655.

Merrick, L., & Taly, A. (2019). The explanation game: Explaining machine learning models with cooperative game theory. *arXiv Preprint arXiv:1909.08128,*

Nordin, N., Zainol, Z., Noor, M. H. M., & Chan, L. F. (2023). An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach. *Asian Journal of Psychiatry, 79*, 103316.

Ren, L., Lin, H., Xu, B., Zhang, S., Yang, L., & Sun, S. (2021). Depression detection on reddit with an emotion-based attention network: algorithm development and validation. *JMIR Medical Informatics, 9*(7), e28754.

Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *Journal of King Saud University-Computer and Information Sciences, 34*(10), 9564-9575.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv Preprint arXiv:1606.05386,*

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. Paper presented at the *Proceedings of the AAAI Conference on Artificial Intelligence, , 32*(1)

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv Preprint arXiv:1910.01108,*

Shapley, L. S. (1997). A value for n-person games. *Classics in Game Theory, 69*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*

World Health Organization. (2021). Suicide worldwide in 2019: global health estimates.