SUICIDE SEVERITY RISK PREDICTION USING BERT

M.M.E. VAN DER LEE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2056209

COMMITTEE

Federico Zamberlan
dr. Juan Sebastian Olier Jauregui

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2022

SUICIDE SEVERITY RISK PREDICTION USING BERT

M.M.E. VAN DER LEE

Abstract

Due to the COVID-19 pandemic, the number of suicides among Dutch people under the age of 30 has increased by 15% compared to previous years. This while suicide was already the second most common mortality cause among adolescents. Since during this time almost all social interaction had to take place via the internet, the popularity of online platforms increased. People began sharing not only their positive but also their negative feelings on these platforms, including feelings of suffering and suicidal ideation. Early detection of these is considered most effective in preventing potential suicide attempts. Therefore, given the rising rate of suicide deaths and the increasing importance of social media, mining these platforms could play a major role in early suicidal ideation detection (SID). Since most studies regarding SID approach this as a binary classification problem, this research examines whether suicidal ideation can also be detected using a multi-class classification scheme. This scheme is based on the Columbia Suicide Severity Rating Scale (C-SSRS) and consists of five different suicide severity classes, with each class indicating a higher degree of suicide risk. In addition, this research will implement the fairly new Bidirectional Encoder Representations from Transformers (BERT) method. This in order to answer the following research question: *Can a BERT-based multi-class classification model provide an accurate prediction of one's suicide severity risk?* The dataset used in this research is the Gold Standard dataset created by Gaur et al. (2019). Previous studies show that BERT-based approaches can achieve state-of-the-art performance in many different NLP tasks, such as: question answering, natural language understanding, and generating text. Therefore, it was assumed that this method would also perform well on the multi-class SID task. However, this turned out not to be the case. The BERT-based model performed worse than the TF-IDF- and Word2Vec-based baseline models.

1 INTRODUCTION

1.1 *Introduction*

The COVID-19 pandemic, which haunts the world to this day, has brought many changes. These changes include rising unemployment rates, increasing reports of domestic violence, and an increase in adolescent social and material deprivation (Eurostats, 2022; Falk, 2020; Harvey, 2021). What all these changes have in common is that they can cause individuals a lot of psychological distress. This was confirmed by Qiu et al. (2020), who found in their study that 35% of their participants experienced psychological distress as a direct or indirect result of the COVID-19 pandemic. In turn, this can lead to a variety of mental health problems, such as post-traumatic stress disorder, anxiety, and depression (Ni et al., 2020). If left untreated, these serious illnesses can result in suicidal ideation and even suicide attempts. According to the Commissie Actuele Nederlandse Suïcideregistratie (2021), the number of suicides among Dutch people under the age of 30 has already increased by 15% compared to previous years. This while suicide was already the second most common mortality cause among adolescents (Weber, Michail, Thompson, & Fiedorowicz, 2017). Increases in mental illness and suicide deaths have also been observed during the Ebola and SARS epidemics (Hawryluck et al., 2004; Jalloh et al., 2018; Yip, Cheung, Chau, & Law, 2010). Sher (2020) attributes this increase to the general anxiety, social isolation and psychological distress associated with a pandemic. Therefore, even though it is clear that the current pandemic poses a threat to our physical health, its negative social and economic consequences may also threaten our mental health.

Research indicates that an important resilience factor against these mental diseases, and therefore also suicide, is social support from relatives and peers (Sippel, Pietrzak, Charney, Mayes, & Southwick, 2015). Unfortunately, during the imposed contact restrictions it was more difficult to obtain this. It also made it more difficult to seek professional psychological help. Since during this time almost all social interaction had to take place via the internet, the popularity of online platforms increased. More and more people began using online platforms as an outlet for their feelings. Well-known platforms that people would use for this purpose are, for example, Reddit and Twitter. Due to the aforementioned changes and the anonymity that can be maintained on the internet, many people would not only share their positive but also their negative feelings. These include feelings of suffering and suicidal ideation. Early detection of these is considered most effective in preventing potential suicide attempts (Ji et al., 2020). Given the rising rate of suicide deaths and the increasing importance of social media,

mining these platforms could play a major role in early suicidal ideation detection (SID).

In order for systems to understand the mined data and detect suicidal thoughts, Natural Language Processing (NLP) methods have to be applied. One of the most frequently used NLP methods is the Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF converts the words into a vector space by converting each word to a weight ratio. This weight ratio is based on the number of occurrences of that word in the dataset used. Words that occur more often in the dataset have a larger weight ratio and are therefore more important, than words that occur less. TF-IDF is a form of word embedding, where semantic information is preserved and embedded into a vector space. Another word embedding technique is Word2Vec. This method does not only take into account the frequency and importance of the words in a document, but also their similarity. Words that are more similar, are represented closer together in vector space. However, both of these methods are independent of the context, the fairly new Bidirectional Encoder Representations from Transformers (BERT) method is not (Devlin, Chang, Lee, & Toutanova, 2018). BERT is an attention based transformer that learns contextual relations between words by bidirectional training. This means that the word embeddings are created by looking at the entire sequence at once, rather than a single word at a time. BERT has been pre-trained on a large corpus containing more than 1 billion words. As a result, BERT-based approaches achieve state-of-the-art performance in many different NLP tasks, such as: question answering, natural language understanding, and generating text (Wang et al., 2019; Yang et al., 2019).

Most studies regarding SID approach this as a binary classification problem. This research examines whether suicidal ideation can also be detected using a multi-class classification scheme. This scheme is based on the Columbia Suicide Severity Rating Scale (C-SSRS; Posner et al., 2008) and is designed by Gaur et al. (2019). It consists of five different suicide severity classes, with each class indicating a higher degree of suicide risk. This C-SSRS based multi-class classification scheme could make a big difference in the timely recognition of suicidal ideation. Considering not only a distinction is made between non-suicidal ideation and suicidal ideation, but also between suicidal indicator, suicidal behaviour, and suicide attempt. Nevertheless, this scheme does perform somewhat worse compared to the binary approach due to its multiple classes. The F1-score Sawhney, Manchanda, Mathur, Shah, and Singh (2018) obtained with their binary model was 0.83, while Gaur et al. (2019) only obtained a score of 0.64. Since the above-mentioned BERT method has not been used for multi-class SID before and has already shown to work well on other NLP tasks,

it could be interesting to see whether this method can also provide an accurate assessment of one's suicide severity risk. Therefore, the aim of this research is to provide more information about the performance of a BERT-based model on the C-SSRS based multi-class scheme and to potentially improve overall performance. To summarize, the following question will be answered:

> *Can a BERT-based multi-class classification model provide an accurate prediction of one's suicide severity risk?*

To be able to determine this, the research question will be divided into two sub-questions.The first sub-question is:

> *Do the posts belonging to the five different classes of suicide severity, differ in content?*

This can be answered by examining whether the Reddit posts belonging to the different categories, differ in their textual features. This can be explored using, for example, Latent Semantic Analysis and calculating the cosine similarity between the categories. When the similarity is high, it can be argued that the classes may not differ in their textual features, and thus in their content. This is of importance because the BERT-based model must be able to recognize textual differences in the classes to be able to differentiate between them. The second sub-question is:

> *Does a BERT-based multi-class classification model outperform multi-class classification models based on TF-IDF or Word2Vec?*

To be able to answer this, a BERT-based model will be compared to a few baseline models. These baseline models will implement the TF-IDF and Word2Vec methods. The performance of each model will be evaluated and compared using the following customized evaluation metrics: Recall, Precision, Ordinal Error, and F1-score. These metrics take into account the ordinal ordering of the suicide severity levels.

1.2 *Related Works*

Traditional SID approaches are based on psychological interviews and questionnaires (Weber et al., 2017). There have already been some studies using this theoretical approach in machine learning techniques to automate the SID process. For example, Ji, Yu, Fung, Pan, and Long (2018) used multiple machine learning techniques to analyse online user content and to detect suicidal thoughts. Huang et al. (2014) also explored the linguistic features of social media posts from suicidal individuals. They did this using a self-created psychological lexicon to detect and count the number

of positive, neutral and negative words per post. When they combined this with a Support Vector Classifier, they achieved a F1-score of 68.3%. Wood, Shiffman, Leary, and Coppersmith (2016) examined the tweet content of 125 twitter users who had attempted suicide. When they compared these with posts from non-suicidal users, they were able to distinguish with a 70% certainty those who had attempted suicide and those who had not, using a simple linear classifier. So, based on user-generated content, such as social media posts, attempts have already been made to differentiate suicidal individuals from non-suicidal ones. These approaches are mainly focused on the identification of textual patterns in the social media posts. This can be done by applying machine learning techniques to various Natural Language Processing (NLP) methods.

As mentioned, most studies regarding SID approach this as a binary classification problem. One could either be suicidal or not. However, this is not always so straightforward. In addition, timely recognition of SID appears to be of great importance for preventing suicide attempts (Ji et al., 2020). Gaur et al. (2019) recognized this problem and were the first ones to approach this not as a binary problem, but as a multi-classification problem. They made a distinction of ones suicide severity risk using the Columbia Suicide Severity Rating Scale (C-SSRS). This scale is primarily used in clinical settings and based on this, one can be divided into three severity classes; Suicidal Ideation (ID), Suicidal Behaviour (BR), or Suicide Attempt (AT). Gaur et al. (2019) found that since the C-SSRS is designed for clinical settings it is not fully comprehensive for social media content. For example, in a clinical setting, it is usually suicidal individuals who talk about suicide. However, on social media non-suicidal individuals may also participate in these conversations to provide support for others who are suicidal. To include this, Gaur et al. (2019) added two additional classes to the existing C-SSRS to ensure proper suicidality risk assessment. The social media posts could therefore fall into one of the following classes: Supportive (SU), Suicide Indicator (IN), Suicidal Ideation (ID), Suicidal Behavior (BR), or Actual Attempt (AT). With an F1-score of 64%, their research showed that this way of distinguishing can provide a reasonably accurate prediction of someone's suicide severity risk.

2 METHOD

2.1 *Data*

The dataset used in this research is the Gold Standard dataset created by Gaur et al. (2019). This dataset consists of 500 Reddit posts, each belonging to a different user and taken from 15 different mental health related subreddits. Only the users who actively participated in the SuicideWatch subreddit were selected. However, if the user also posted in one of the other mental health subreddits and their content showed a cosine similarity of greater than 0.6, then these posts were concatenated to their SuicideWatch post. This enriched the dataset with more details about the users and their mental state. For the creation of this dataset Gaur et al. (2019) also used negation detection to filter out non-suicidal users. This is crucial to avoid confusing the model with sentences like: *"I will not kill myself just for embarrassing myself in front of the whole school"*. This sentence does not indicate that someone is suicidal, but it may make the model more prone to producing false positives. After this, the 500 posts were annotated to one of the 5 different suicide severity levels. This annotation was done based on the C-SSRS and performed by four practicing psychiatrists. The average pairwise agreement between them was 0.69 and the group-wise agreement was 0.73.

The different classes of increasing suicide risk severity to which a post could be annotated were: Supportive (SU), Suicide Indicator (IN), Suicidal Ideation (ID), Suicidal Behavior (BR), or Actual Attempt (AT). The Supportive (SU) category is defined as the individuals who participate in the SuicideWatch subreddit, but show no suicidal characteristics or other risk factors themselves. These people often offer support or advice to the people posting on the SuicideWatch subreddit. The Suicide Indicator (IN) category is defined as the individuals who participate in the SuicideWatch subreddit in the same way the SU individuals do, but the IN individuals might also show particular risk factor due to sharing personal experiences. For example, they share their experience with divorce, chronic illness or death, which are risk indicators on the C-SSRS, but only to show empathy to users who express suicidal ideation rather than expressing suicidal thoughts oneself. The Suicide Ideation (ID) class includes all individuals who have thoughts of suicide and show risk factors themselves. If a person confesses in their Reddit post to having active or historical self-harm behavior, or active suicide plans, or a history of psychiatric hospitalization, then that person falls under the Suicidal Behavior (BR) category. The last category is Actual Attempt (AT), this category includes all individuals who attempted suicide, whether this was successful or not. The annotated data

consists of 22% SU posts, 20% IN posts, 34% ID posts, 15% BR posts, and 9% AT posts. It can therefore be concluded that there is a large difference between the size of the categories in the dataset. All posts that were gathered and used to create this dataset, were published between 2005 and 2016.

2.2 *Pre-Processing*

The first pre-processing step was the removal of HTMLs, URLs, punctuations, symbols, numbers, stopwords and whitespaces from the Reddit posts. Hereafter, contractions were replaced, text converted to lowercase and tokenized. To ensure that the words that have the same root are grouped together and to avoid the creation of different word embeddings for words with the same definition, words are lemmatized. This was done using the Natural Language Toolkit (NLTK) WordNet Lemmatizer in combination with the corresponding Part Of Speech tag of each individual word. The last step of the pre-processing was the removal of all the words that occurred just once across all posts, these words were very rare and most of the time a result of typos.

2.3 *Feature Extraction*

After pre-processing some statistical and textual features were extracted from the posts. First, the Suicide Severity classes were factorized from lowest to highest suicide risk, with SU being 0 and AT being 4. Also a textual sentiment score per post was calculated using the AFINN method. This method maps words using their psycholinguistic features. AFINN consists of a large dictionary of words along with their corresponding affective score ranging from -5 to 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. The sentiment score of each post was calculated by adding all the scores for the entire post together. The number of words, the number of tokens and the average word length per post was also calculated. All these scores were standardized using the Scikit-learn StandardScaler method.

In addition, some Natural Language Processing (NLP) methods were applied. NLP methods enable machine learning models to extract meaningful information from textual data. After applying these methods, the textual data will be represented as a vector space. The NLP methods that were used in this research are: The Text Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and Bidirectional Encoder Representation from Transformers (BERT). The TF-IDF method shows how important a word is in relation to the entire dataset. This method ensures that words

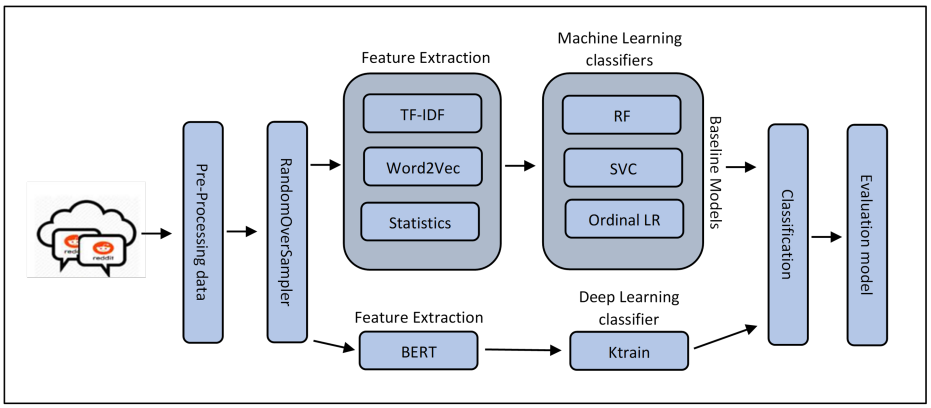


Figure 1. Overview of the architecture of the different classification models.

that matter little are excluded and important words are selected for further analysis. Word2Vec is applied with the Continuous Bag of Word (CBOW) architecture, this method not only looks at term frequency but also at similarity. Thus, words that are more similar are likely to be closer together in vector space. However, where the Word2Vec method creates word embeddings that are independent from context, BERT creates word embeddings that are context aware. BERT is a pre-trained transformer that generates contextual embeddings by looking at entire sequences at once, rather than a single word at a time. Figure 1. shows the overview of the architecture used in this research, this gives a good idea of how the different NLP methods have been applied.

2.4 *Model Architecture*

2.4.1 *Imbalanced Data*

As mentioned before, the dataset used in this research is very imbalanced. This means that there is an unequal class distribution. In particular, the BR and AT categories occur less frequent in the dataset. Since these are the groups that are of great importance, because these are the groups that the models should properly recognize, a number of methods will be applied to solve this inequality in class distribution. Firstly, the dataset will be split in a train and test set in a stratified way. Which means that the train and test set contain the same proportions of class labels as the original dataset. This ensures that all categories are included in the training of the model and that no class is accidentally excluded while sampling. After this the training set will be sampled using the RandomOverSampler method from Imbalanced-learn. This method oversamples the minority classes by picking samples at random with replacement. As a result, each class label appears 137 times in the training set and the total training set consists of

685 Reddit posts. This way of balancing the training set was done prior to every model implementation.

2.4.2 *BERT model*

For the implementation of BERT, a wrapper from a Keras-Tensorflow library was used. This wrapper is known as Ktrain (Maiya, 2020). Ktrain is a lightweight wrapper that helps build and train neural networks. Furthermore, the distilled version of the BERT base model, also know as distilBERT, was used. This version is 40% smaller and therefore less computational extensive than BERT. DistilBERT is trained on the same corpus as BERT and despite being smaller, this version retains over 95% of BERT's performance and runs 60% faster (Sanh, Debut, Chaumond, & Wolf, 2019). The first step in the implementation of BERT, was the use of the 'DistilBERT' pre-processing mode. This mode is built-in in Ktrain and can directly convert the textual input data into features. Since the BERT model has its own pre-processing mode, the model is given the raw unprocessed textual data. Only the label classes were factorized beforehand. After that the pre-trained BERT model was loaded with a randomly initialized final dense layer. Due to the fact that none of the layers are frozen, the weights of all the layers will be updated during training. To be able to get the best results, the optimal learning rate and batch size had to be determined. The BERT authors recommend fine-tuning the model with a batch size ranging from 8 to 128 and a learning rate of 3e-4, 1e-4, 5e-5, or 3e-5 (Turc, Chang, Lee, & Toutanova, 2019). They also recommended the usage of 4 epochs. Hence, after trying different combinations the best results were obtained with the model that was fine-tuned with a batch size of 32 and a learning rate of 5e-5. The number of epochs was set to 4.

2.4.3 *Baseline models*

The BERT model will be compared with the following three baseline models: Random Forest (RF), Support Vector Classifier (SVC) and Ordinal Logistic Regression (OLR). As can be seen in Figure 1. there are also three different kind of feature extraction methods that will be used as input for these baseline models. These extraction methods are used alone or in combination with each other. The three possible input options are: TF-IDF, TF-IDF + Word2Vec, or TF+IDF + Word2Vec + Statistics. Statistics include the standardized sentiment score, number of words, number of tokens, and average word length features. Each model will be given these different input options, to see whether this might influence the performance in some kind of way.

The fact that RF's are quite accurate with few parameters to tune and can be applied in a wide range of classification problems, makes them very popular. A RF model can also work with small sample sizes and high-dimensional data. They are easily parallelizable, which makes them suitable for dealing with large problems (Biau & Scornet, 2016). A RF classifier is actually a combination of a lot of single Decision Trees (Oshiro, Perez, & Baranauskas, 2012). This causes the model to be more robust to noise and more accurate than just one single Decision Tree. Thus, each post will be classified using the majority vote of these trees.

The SVC was originally developed for binary classification problems (Cortes & Vapnik, 1995). The SVC tries to separate two classes through a linear decision boundary. The optimal decision boundary is defined as the boundary that minimizes the generalization error. Thus, the boundary that has the greatest margin between the two classes. To be able to use this method for the multiclass classification problem of this paper, the one vs. one method will be applied. This method applies SVC to all possible pairs of classes. So, for n classes, $n(n-1)/2$ different models will be generated. In this case, this means that 10 models will be generated and each post will be classified according to the majority vote of these models.

Both RF and SVC were used in the research by Gaur et al. (2019). These models will therefore also be used in this study. However, since both the RF and the SVC do not take into account the ordinal nature of the classes, an OLR model will also be used as a baseline. An ordinal variable is a variable that has a clear ordering of the category levels. The suicide severity classes have this ordering, with SU being the least severe and AT the most severe class.

2.5 *Evaluation Metrics*

Gaur et al. (2019) proposed some customized evaluation metrics for the suicide severity classification problem. They defined False Positives (FP) as the ratio of the number of times the predicted suicide risk severity level (S^p) is greater than the actual level (S^a) in the test data (N_T). The False Negatives (FN) was defined as the ratio of the number of times the predicted suicide risk severity level (S^p) is less than the actual level (S^a) in the test data (N_T). The Ordinal Error (OE) was defined as the ratio of number of times where the difference between the actual severity level (S^a) and the predicted severity level (S^p) is greater than 1. This measures the tendency of a model to predict a low level severity for individuals who actually have a high level severity, such as BR or AT. In order to make a good evaluation of the performance and to compare it with the study

$$FN = \frac{\sum_{i=1}^{N_T} I(S_i^a > S_i^p)}{N_T} \qquad FP = \frac{\sum_{i=1}^{N_T} I(S_i^p > S_i^a)}{N_T} \qquad OE = \frac{\sum_{i=1}^{N_T} I(\Delta(S_i^a > S_i^p) > 1)}{N_T}$$

of Gaur et al. (2019), this research will use the same evaluation metrics. Therefore, the definitions of FP, FN and OE will be:
All models were performed with a 5-fold hold-out cross-validation and their performance will be evaluated with their corresponding F1-score, Recall, Precision and Ordinal Error score on the hold-out test set.

3 RESULTS

3.1 *Data Analysis Results*

In order to answer the first sub-research question, the different classes of suicide severity were analyzed and compared. In Figure 2. the two hundred most used n-grams per class are presented in word clouds. At first glance, it seems that all classes are quite different in their word usage. Furthermore, as can be seen in Figure 3. the average AFINN score per class also differs quite a bit. This score was calculated by adding all scores of each post together. Thus, a difference in average sentiment score could indicate different word usage per group. To further analyze this, the chi-square test was used to find the most predictive n-grams and bigrams for each category. For the SU class, the n-grams *'redditors'*, *'cheesy'* and bigrams *'truly happy'* and *'therapist help'* proved to be the most predictive. The n-grams *'shake'*, *'drama'* and bigrams *'thing sort'*, *'suppose make'* are most predictive for the IN class. While for the ID class the n-grams *'depersonalization'*, *'hobby'* and bigrams *'feel completely'*, *'couple week'* are most predictive. The most predictive n-grams and bigrams for BR are *'squeeze'*, *'oxy'*, *'eat disorder'*, and *'look help'*. Lastly, the most predictive n-grams and bigrams for the AT class are *'fruit'*, *'prolong'*, *'fail time'*, and *'suicide attempt'*.

For the SU and IN class, it can be seen that especially n-grams and bigrams with a relatively positive and supportive meaning are predictive. The n-grams and bigrams for the suicidal classes are of a more negative nature. Lastly, the cosine similarity score between the groups was calculated. This to see whether the suicide severity classes are close to each other in terms of their content. All the posts that had the same suicide severity label were concatenated, resulting in one document per label in which all posts are merged. Hereafter, the vector space of the documents was calculated using the TF-IDF method with singular value decomposition. This reduced the vector space, making it easier to calculate their similarity. The results of this cosine similarity calculation can be found in Figure 4. These results

Figure 2. Most frequently used words per suicide severity class.

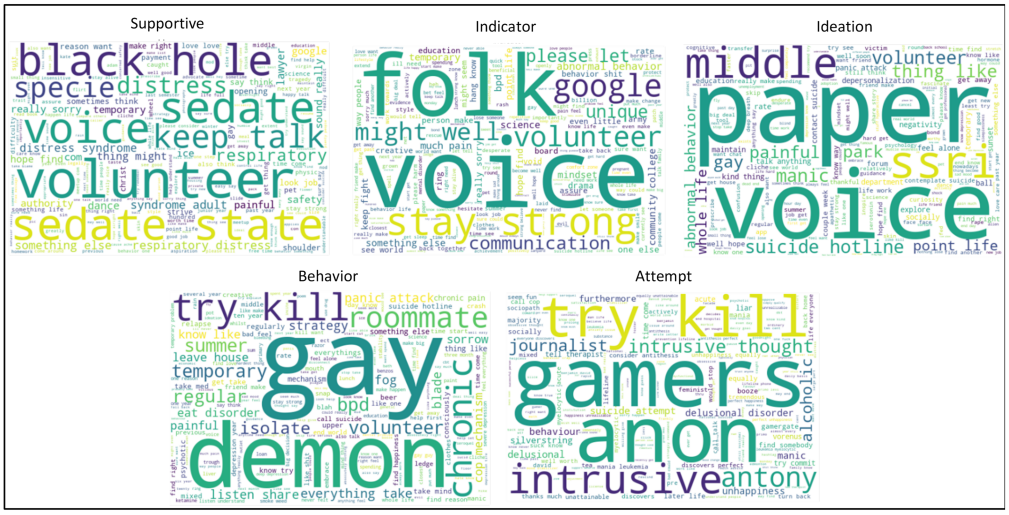


Figure 3. Mean AFINN score per group.

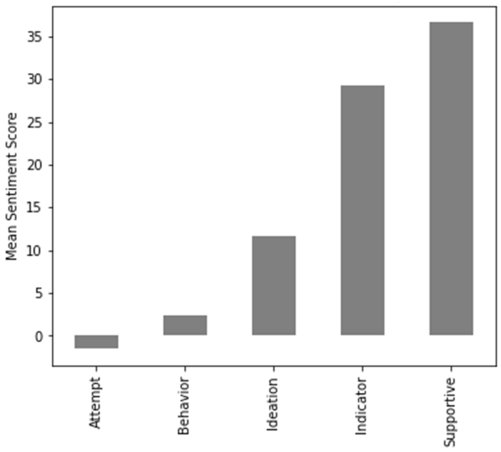
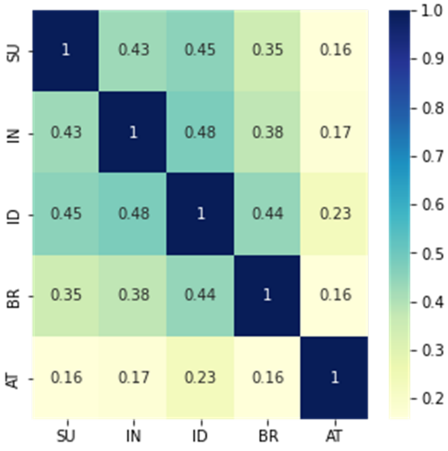


Figure 4. The cosine similarity between groups.



show that the similarity scores range from 16% to 48%, indicating that each class differs from the other classes by at least 52% or more.

3.2 *Classification Results*

For evaluation, the Precision, Recall, F1-score and Ordinal Error score of each model is considered. Table 1. shows the results of these measures for each model and each input option. It is observed that the performance of the Support Vector Classifier, based on the F1-score, is better compared to the other two baseline models. The Support Vector Classifier showed a higher precision when given the TF-IDF and Word2vec as input. Nevertheless, this model does show a high Ordinal Error, which indicates that 15%

Table 1. Performance results of each model.

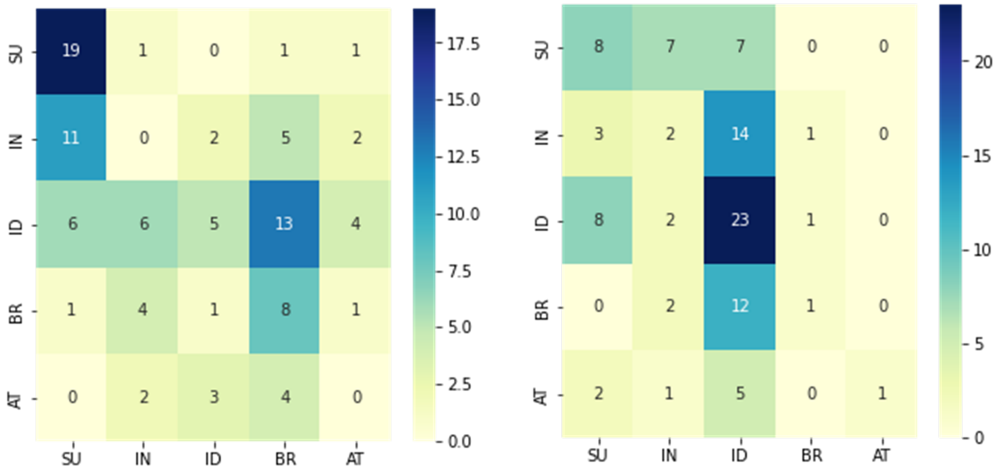
Methods	Feature Type	Precision	Recall	F1-Score	OE
RF	TF-IDF	0.56	0.48	0.52	0.18
	TF-IDF + Word2Vec	0.49	0.46	0.48	0.17
	Statistics + TF-IDF + Word2Vec	0.48	0.44	0.46	0.17
SVC	TF-IDF	0.57	0.51	**0.54**	0.15
	TF-IDF + Word2Vec	**0.58**	0.51	**0.54**	0.15
	Statistics + TF-IDF + Word2Vec	0.57	0.51	**0.54**	0.15
LR	TF-IDF	0.48	0.54	0.51	**0.09**
	TF-IDF + Word2Vec	0.49	**0.55**	0.52	**0.09**
	Statistics + TF-IDF + Word2Vec	0.49	**0.55**	0.52	**0.09**
NN	BERT	0.39	0.52	0.44	0.12

RF = Random Forest, SVC = Support Vector Classifier, LR = Linear Regression, NN = Neural Network.

of the individuals have been given a suicide severity level that deviated one or more severity levels from their actual suicide severity level. Suggesting that this model predicts the severity score often lower than it actually is, and thus underestimating the severity of suicide risk. The Linear Regression model shows the lowest Ordinal Error, only 9% of the individuals are given a deviant severity score of greater than 1. This corresponds to the somewhat higher Recall and lower Precision score, indicating that this model predicts the suicide severity scores often higher than they actually are. The Random Forest model has with 18% the highest Ordinal Error. Whereas, the Recall of this model is the lowest of all. Showing that this model has the tendency to predict low severity scores for individuals that actually have high severity scores. For the baseline models, each input option yielded nearly the same results. Indicating no significant difference in performance. The BERT-based simple Neural Network model achieved a F1-score of 0.44, which is the lowest of all models. Interesting is that the BERT-based model has a low Precision, but a relatively high Recall and low Ordinal Error. Indicating that this model is more likely to overestimate than underestimate the suicide severity score.

In Figure 5. The confusion matrix of the BERT-based model is compared to the best performing baseline, the SVC with TF-IDF and Word2Vec as input. The vertical line from the top left corner to the bottom right corner shows the number of correct classifications. As can be seen, the BERT-based model correctly classifies 32 individuals and the SVC model correctly classifies 35 individuals. The BERT-based model mainly correctly classifies SU individuals. These correct classification account for 59% of the total number of correct classifications. Whereas, the SVC model mostly correctly classifies the ID individuals. These classifications account for 66% of the total correct classifications. The SVC model almost never correctly predicts

Figure 5. Confusion matrix of left the BERT-based NN and right the best performing baseline (SVC). The Y-axis: actual level, X-axis: predicted level of suicide severity.



the BR and AT individuals. In fact, this model almost never predicts a posts as belonging to BR or AT. The confusion matrix shows that especially the SU, IN and ID labels are given as predictions. The SVC also seems to have difficulty distinguishing IN and BR individuals from ID individuals. Since IN and BR individuals are mostly predicted as ID. The BERT-based model appears to have trouble distinguishing IN individuals from SU individuals, as well as ID individuals from BR individuals.

3.3 *Error analysis*

In the case of predicting a person's suicide risk, producing a false negative can have greater consequences than a false positive. Given that a false negative can lead to someone actually committing suicide. As mentioned above, the best baseline, SVC with TF-IDF and Word2Vec as input, actually underestimates the suicide severity score. Whereas, the BERT-based model seems to overestimate the suicide severity score. This indicates that the SVC model produces more false negatives than the BERT-based model. So despite the fact that the BERT-based model does perform worse, the errors that this model makes have less serious consequences than the errors that the SVC model makes.

4 DISCUSSION

This study attempted to answer the following research question: *Can a BERT-based multi-class classification model provide an accurate prediction of one's suicide severity risk?* To be able to answer this, this question was divided into two sub-questions. The first sub-question is: *Do the posts belonging to the five different classes of suicide severity, differ in content?* The results of the dataset analysis show that the cosine similarity score between each category is at most 48% or less. This indicates that each suicide severity class consists for at least 52% or more of different content. The two hundred most used n-grams per class also appear to be quite different. In addition, the average sentiment score per class does also differ significantly. As well as the most predictive n-grams and bigrams per class. These n-grams and bigrams seem to represent the core of each category well. The Supportive and Suicide Indicator classes are mostly predicted by positive and uplifting words. While, the Suicide Ideation, Suicide Behavior, and Suicide Attempt classes are mainly predicted by negative words. The Supportive category has on average also a much higher sentiment score than the Suicide Attempt category. It is therefore clear that there seems to be a difference in textual features, and thus content, across the five different classes of suicide severity. Indicating that the models used should be able to differentiate between the different classes of suicide severity.

The second sub-research question is: *Does a BERT-based multi-class classification model outperform multi-class classification models based on TF-IDF or Word2Vec?* This question has been answered by comparing the performance of the BERT-based model to the performance of a number of baseline models. This comparison shows that the BERT-based model performs worse than the other models. The BERT-based model seems to overestimate suicide severity in its predictions. This makes individuals more likely to receive a higher severity risk label than they actually should have. The BERT-based model is therefore more likely to produce a false positive than a false negative. Whereas, the best performing baseline, SVC with TF-IDF and Word2Vec as input, is more likely to produce false negatives. In the case of predicting one's suicide risk, producing a false negative can have greater consequences than a false positive. So although, the SVC model performs better in terms of accurate predictions, the BERT-based model makes less severe errors than the SVC model. The BERT-based model does appear to have difficulty distinguishing Supportive individuals from Suicide Indicator ones, as well as individuals that show Suicide Ideation from Suicidal Behavior.

Previous research has shown that BERT-based approaches can achieve state-of-the-art results on various NLP tasks (Wang et al., 2019; Yang et

al., 2019). Therefore, it was assumed that this method would also perform well on the multi-class SID task. However, this turned out not to be the case. This could be due to the complex linguistic expressions used in the Reddit posts. Complex linguistic expressions include rhetorical questions and sarcasm. Research has shown that BERT-based models have difficulty understanding these types of expressions (Shih et al., 2021), since the actual meaning of these sentences is the opposite of their literal meaning. Examples of sentences from the dataset that represent this kind of language well are: *"So should I wait on the slightest change it does get better?"* and *"I am a teen living with parents that control every aspect of my life, what is the point of me even living it."*. The questions asked are not real questions, but rhetorical ones. The user already knows the answer. Nevertheless, the BERT-model will interpret these as literal questions and this can influence the model in its predictions.

Another reason for the poor performance of the BERT-based model may be that although oversampling was used, the model was still unable to learn enough features from the oversampled categories. Resulting in ill-defined classification groups. This could be the case for the Suicidal Behavior and the Suicide Attempt categories. Since both these categories together only made up 21% of the total dataset. This could also explain why the BERT-based model often predicts a post as belonging to the Suicidal Ideation class. Given that this label made up 34% of the dataset and therefore the model had more training material for this class. However, this is not a complete explanation as the Suicide Attempt class has the lowest cosine similarity compared to the other classes. Meaning that this class should be clearly distinguishable, because it shares almost no content with any other class. Yet the opposite turned out to be true. The Suicide Attempt posts were almost never correctly predicted. This was not only the case in the BERT-based model, but also in the other baseline models.

Even though the BERT-based model did not perform as expected, the performance of the baseline models is comparable to the results obtained by Gaur et al. (2019). Their models also seem to never properly classify the Suicide Behavior and Suicide Attempt posts. As mentioned above, this is also the case for the baseline models used in this research. It is therefore questionable whether this way of distinguishing between different levels of suicide severity actually works well. In addition, this also raises the question of whether this dataset contains enough posts to actually make a good distinction between the levels of suicide severity. Since mainly the Supportive and Suicide Ideation classes are correctly predicted and the dataset consists for a large part of these classes. Indicating that this could be due to the model being able to learn enough textual features to differentiate them. In summary, the models may have difficulty predicting

the other classes due to ill-defined classification groups. However, another possibility is that there is no clear distinction to be made. In the latter case, this may mean that models will never achieve the desired SID accuracy with this multi-class scheme. Moreover, since the baseline models can only manage to correctly predict two of the five classes, this still seems to amount to a binary classification scheme. Where a distinction is made between non-suicidal and suicidal, as done in previous studies on SID (Huang et al., 2014; Ji et al., 2018; Wood et al., 2016).

For this reason, one of the main limitations in this study is the size of the dataset used. This dataset was created by Gaur et al. (2019) and consists of 500 Reddit posts. To provide more clarity on whether this multi-class SID scheme works, it should be tested on more data. However, the labels of the data set were assigned by four practicing psychiatrists using the C-SSRS based scheme. So the problem is not finding new data, but assigning labels to this new data. As this should be done in the same way as Gaur et al. (2019) did and is very time consuming. Therefore, another limitation of this study is time, since this paper had to be conducted in a short time period. With more time, the dataset could be expanded and the performance of the BERT-based could also be compared to, for example, other Neural Networks. It might therefore be interesting for follow-up studies to extent the dataset, despite the disappointing results of the multi-class SID scheme in this study. It may also be interesting to compare the BERT-based model with other neural networks that use TF-IDF or Word2Vec, to gain more insight into the difference in performance on the multi-class SID task.

4.1 *Conclusion*

Previous research has shown that BERT-based approaches can achieve state-of-the-art results on various NLP tasks. Therefore, it was assumed that this method would also perform well on the multi-class SID task. However, this turned out not to be the case. There are a number of possible reasons for this. The first reason may be the complex linguistic expressions used in the Reddit posts. BERT-based models seem to have difficulty understanding these. Another reason may be that although oversampling was used, the model was still unable to learn enough features from the oversampled categories. Resulting in ill-defined classification groups. Lastly, it can be the case that there is no clear distinction to be made in the Reddit posts. To provide more clarity on this, and thus on whether this multi-class SID scheme works, it actually should be tested on a larger dataset. However, this was not possible due to time constraints. For future research, it might therefore be interesting to extend the dataset and see if this affects the performance of the BERT-based model on the multi-class SID task.

REFERENCES

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

C-SSRS; Posner, K., Brent, D., Lucas, C., Gould, M., Stanley, B., Brown, G., ... others (2008). Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, *10*.

Commissie Actuele Nederlandse Suïcideregistratie. (2021). *Geen toename in totaal aantal suïcides, extra aandacht voor jongeren noodzakelijk.* https://www.113.nl/actueel/geen-toename-totaal-aantal-suicides-extra-aandacht-voor-jongeren-noodzakelijk.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eurostats. (2022). *Youths: 7% severely materially and socially deprived.* https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220210-1.

Falk, G. (2020). *Unemployment rates during the covid-19 pandemic:*. Congressional Research Service.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., ... Pathak, J. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference* (pp. 514–525).

Harvey, R. (2021). The ignored pandemic: The dual crises of gender-based violence and covid-19.

Hawryluck, L., Gold, W. L., Robinson, S., Pogorski, S., Galea, S., & Styra, R. (2004). Sars control and psychological effects of quarantine, toronto, canada. *Emerging infectious diseases*, *10*(7), 1206.

Huang, X., Zhang, L., Chiu, D., Liu, T., Li, X., & Zhu, T. (2014). Detecting suicidal ideation in chinese microblogs with psychological lexicons. In *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* (pp. 844–849).

Jalloh, M. F., Li, W., Bunnell, R. E., Ethier, K. A., O'Leary, A., Hageman, K. M., ... others (2018). Impact of ebola experiences and risk perceptions on mental health in sierra leone, july 2015. *BMJ global health*, *3*(2), e000471.

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.

Ji, S., Yu, C. P., Fung, S.-f., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, *2018*.

Maiya, A. S. (2020). ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Ni, M. Y., Yao, X. I., Leung, K. S., Yau, C., Leung, C. M., Lun, P., ... Leung, G. M. (2020). Depression and post-traumatic stress during major social unrest in hong kong: a 10-year prospective cohort study. *The Lancet*, *395*(10220), 273–284.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition* (pp. 154–168).

Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B., & Xu, Y. (2020). A nationwide survey of psychological distress among chinese people in the covid-19 epidemic: implications and policy recommendations. *General psychiatry*, *33*(2).

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*.

Sawhney, R., Manchanda, P., Mathur, P., Shah, R., & Singh, R. (2018). Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 167–175).

Sher, L. (2020). The impact of the covid-19 pandemic on suicide rates. *QJM: An International Journal of Medicine*, *113*(10), 707–712.

Shih, C.-F., Tseng, Y.-H., Yang, C.-W., Chen, P.-E., Chou, H.-Y., Tan, L.-H., ... Hsieh, S.-K. (2021, October). What confuses BERT? linguistic evaluation of sentiment analysis on telecom customer opinion. In *Proceedings of the 33rd conference on computational linguistics and speech processing (rocling 2021)* (pp. 271–279). Taoyuan, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). Retrieved from https://aclanthology.org/2021.rocling-1.35

Sippel, L. M., Pietrzak, R. H., Charney, D. S., Mayes, L. C., & Southwick, S. M. (2015). How does social support enhance resilience in the trauma-exposed individual? *Ecology and society*, *20*(4).

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Wang, Q., Liu, P., Zhu, Z., Yin, H., Zhang, Q., & Zhang, L. (2019). A

text abstraction summary model based on bert word embedding and reinforcement learning. *Applied Sciences*, *9*(21), 4701.

Weber, A. N., Michail, M., Thompson, A., & Fiedorowicz, J. G. (2017). Psychiatric emergencies: assessing and managing suicidal ideation. *Medical Clinics*, *101*(3), 553–571.

Wood, A., Shiffman, J., Leary, R., & Coppersmith, G. (2016). Language signals preceding suicide attempts. In *Chi 2016 computing and mental health workshop, san jose, ca.*

Yang, W., Xie, Y., Tan, L., Xiong, K., Li, M., & Lin, J. (2019). Data augmentation for bert fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*.

Yip, P. S., Cheung, Y., Chau, P. H., & Law, Y. (2010). The impact of epidemic outbreak. *Crisis*.