IST 736 Text Mining

Course Final Project

Bomin Powers

**Would tech be able to help?**

**: Suicide detection**

A picture containing text, indoor

Description automatically generated

Table of Contents

[1. Introduction 2](#_Toc121436179)

[2. Analysis 2](#_Toc121436180)

[2-1. About the data 3](#_Toc121436181)

[2-2. Preprocessing of the Dataset 3](#_Toc121436182)

[2-3. Explore the data 5](#_Toc121436183)

[2-4. Methodology 6](#_Toc121436184)

[**i.** Multinomial Naïve Bayes 7](#_Toc121436185)

[**ii.** Support Vector Machines (SVMs) 7](#_Toc121436186)

[**iii.** Logistic Regression 7](#_Toc121436187)

[**iv.** Decision Tree 7](#_Toc121436188)

[**v.** Random Forest 7](#_Toc121436189)

[3. Results 7](#_Toc121436190)

[4. Conclusions 12](#_Toc121436191)

[Reference 12](#_Toc121436192)

# Introduction

***Figure 1****. Average daily deaths in the US (2021) – suicide ranked in 11th*

*A picture containing chart

Description automatically generated*

According to a study by the Suicide Awareness Voices of Education (SAVE), "Suicide" ranks 12th in the United States for deaths. Worse, suicide is the third leading cause of death for Americans aged between 15 and 24. In 2020, 45,979 Americans died of suicide. In addition, about 25 suicide attempts result in one suicide death. This means that suicide attempts are far more than actual suicides have occurred. We can easily say that suicide, specifically depression, is a big social issue.

Suicide is one of the leading causes of death globally (not just US) and is notably a significant cause of death among young people. A suicide outcome is a complex combination of personal, social, and health factors. Therefore, suicide prevention is a challenge, requiring a systems approach incorporating public health strategies, screening at-risk individuals, targeted interventions, and follow-up for suicide survivors and those bereaved by suicide.

Due to the anonymity of online media and social networks, people tend to express their feelings and sufferings in online communities. Maybe we can detect and prevent suicide through this. In order to prevent suicide, it is necessary to detect suicide-related posts and users' suicide ideation in cyberspace by natural language processing methods. We focus on the online community called Reddit and classify users' posts with potential suicide and without suicidal risk through text features processing and machine learning-based methods.

# Analysis

## 2-1. About the data

The dataset is originally scraped from online community platform called Reddit; collection of posts from "SuicideWatch" and "depression" subreddits. Then a binary value of 'suicide' or 'non-suicide' was assigned by a psychiatrist to each post. So, we have two columns: a 'text' column for post text and a 'class' column for the suicidal label. The total number of posts that were scraped was 232,074.

***Figure 2****. Initial scraped data frame (2 columns and 232,074 rows)*

*Graphical user interface, text

Description automatically generated with medium confidence*

## 2-2. Preprocessing of the Dataset

Since we have enough data to get meaningful insights from the model, first, reduce the data into 20,000 rows, 10,000 rows from each suicidal class to save time for code running. We randomly sampled data into 20,000. Then set the stopwords with nltk stopwords of ‘ English’ and added other words that we thought didn’t necessarily for the model. We created a function that preprocessed text to our needs following steps were included: lemmatize, remove the symbols, trans all characters into lower case, remove email using regular expression and remove punctuation. This function was applied to 20,000 text data and took 6 minutes and 22 seconds.

***Figure 3****. After the text preprocessing*

Graphical user interface, text, email

Description automatically generated

The next preprocessing for the model is tokenize and vectorization. Performed unigram/bigram count vectorizer, unigram/bigram Boolean vectorizer, and TFIDF for Multinomial Naïve Bayes model, SVM models, and Logistic Regression models. For the Decision Tree model and Random Forest model, performed tokenizing and vectorizing the data in a different form since this model needs integer data, not a string.

***Figure 4****. Vectorized text data for DT and RF*

Table

Description automatically generated with medium confidence

## 2-3. Explore the data

Let’s explore the data as we finished the preprocessing. From the word clouds for suicide text, we can observe these words: life, feel, live, want, and love. It seems like they are trying to explanate their feelings of hope towards life. However, we can see many negative words such as kill, end, shit, help, bad, fuck, lose, leave, pain, etc.

***Figure 5****. Wordclouds for suicide text*

Text

Description automatically generated

***Figure 6****. Wordclouds for non-suicide text*

Text

Description automatically generated

From converting suicide and non-suicide documents into a count vectorizer, we got the top 20 words for each. Hopeful words were shown at the top of the list from the suicidal document.

***Table 1****. top 20 words from suicide and non-suicide documents*

|  |  |
| --- | --- |
| # top 20 words from the suicidal document  # Apply the function  get\_top\_n\_words(corpus\_su, n=20) | # top 20 words from the non suicidal document  # Apply the function  get\_top\_n\_words(corpus\_un, n=20) |
| [('want', 14004),  ('feel', 13293),  ('like', 11839),  ('know', 10811),  ('life', 9839),  ('think', 8086),  ('time', 7311),  ('people', 6346),  ('year', 6288),  ('try', 6070),  ('friend', 5941),  ('day', 5048),  ('get', 4949),  ('live', 4876),  ('help', 4776),  ('tell', 4729),  ('kill', 4603),  ('good', 4448),  ('bad', 4183),  ('die', 4178)] | [('like', 4651),  ('know', 2530),  ('want', 2316),  ('day', 1987),  ('jake', 1944),  ('paulfuck', 1941),  ('think', 1934),  ('people', 1930),  ('friend', 1831),  ('feel', 1754),  ('time', 1656),  ('get', 1633),  ('good', 1611),  ('talk', 1377),  ('tell', 1317),  ('school', 1291),  ('year', 1263),  ('guy', 1252),  ('girl', 1249),  ('come', 1228)] |

## 2-4. Methodology

To detect suicide from the documents, we tried 5 different classification models.

### Multinomial Naïve Bayes

The multinomial Naive Bayes classifier is suitable for classification with discrete features. The multinomial distribution usually requires integer feature Counts. However, in practice, fractional counts such as TF-IDF may also work. For this, we used sklearn.naive\_bayes MultinomialNB class. The output results for suicide detection modeling were recorded in the new data frame for evaluation.

### Support Vector Machines (SVMs)

Support vector machines is an algorithm that determines the best decision boundary between vectors belonging to a given group (or category) and those not. For this, we used sklearn.svm SVC class. The output results for suicide detection modeling were recorded in the new data frame for evaluation. Three different option kernels were: Linear kernel, Polynomial (degree 1) kernel, and RBF kernel

### Logistic Regression

Logistic Regression Model is a classification algorithm used to solve binary classification problems. It uses the weighted combination of the input features and passes them through a sigmoid function. For this, we used sklearn.linear\_model LogisticRegression class. The output results for suicide detection modeling were recorded in the new data frame for evaluation.

### Decision Tree

Decision tree learning is a supervised learning approach in statistics, data mining, and machine learning. In this formalism, a classification or regression decision tree is used as a predictive model to conclude a set of observations. For this, we used sklearn.tree.DecisionTreeClassifier class. The output results for suicide detection modeling were recorded in the new data frame for evaluation.

### Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operates by constructing many decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees. For this, we used sklearn. ensemble.RandomForestClassifier class. The output results for suicide detection modeling were recorded in the new data frame for evaluation.

# Results

## 3-1. Multinomial Naïve Bayes

For MNB, tried 8 different types of vectorizations vary: 1) Unigram count vectorizer 2) bigram count vectorizer 3) Unigram count vectorizer that removed stopwords set 4) bigram count vectorizer that removed stopwords 5) Unigram boolean vectorizer 6) bigram boolean vectorizer 7) Unigram boolean vectorizer that removed stopwords set 8) bigram boolean vectorizer that removed stopwords. Among these models, we found that MNB with Unigram boolean vectorizer data set showed the highest accuracy score.

***Figure 7****. MNB with 8 different types of vectorized dataset*



***Table 2****. Evaluating on Multinomial* *Naïve Bayes with* Unigram boolean vectorizer

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MNB** *on text data Report* | **Precision** | **Recall** | **F1-score** | **support** |
| **non-suicide** | 0.965559 | 0.791873 | 0.870133 | 2018.00000 |
| **suicide** | 0.820896 | 0.971241 | 0.889762 | 1982.00000 |
| **accuracy** | 0.880750 | 0.880750 | 0.880750 | 0.88075 |
| **macro avg** | 0.893227 | 0.881557 | 0.879948 | 4000.00000 |
| **weighted avg** | 0.893878 | 0.880750 | 0.879859 | 4000.00000 |

Out of all documents in the dataset that were predicted as non-suicidal, 96.5% were non-suicidal (: Precision). And the model predicted this outcome correctly for 79.1% of those non-suicidal texts (: Recall). Since the F1-score of 0.87 is relatively close to 1, it tells us that the model does a great job of predicting whether the text is suicidal or non-suicidal. We got 88.07% of accuracy score from MNB.

## 3-2. Support Vector Machines (SVMs)

For SVM model we tried three different kernels: linear, polynomial and RBF. We found that for this classification problem, linear kernel showed the best performance among the SVMs model: 90.9% of accuracy score. However, the polynomial kernel (degree 1) showed an 87.25% accuracy score, and the RBF kernel was an 88.75% of accuracy score.

***Table 3****. Evaluating on SVMs with linear kernel*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SVM** *on text data Report* | **Precision** | **Recall** | **F1-score** | **support** |
| **non-suicide** | 0.891943 | 0.932607 | 0.911822 | 2018.000 |
| **suicide** | 0.928042 | 0.884965 | 0.905992 | 1982.000 |
| **accuracy** | 0.909000 | 0.909000 | 0.909000 | 0.909 |
| **macro avg** | 0.909993 | 0.908786 | 0.908907 | 4000.000 |
| **weighted avg** | 0.909830 | 0.909000 | 0.908933 | 4000.000 |

## 3-3. Logistic Regression

From the Logistic Regression classification model, set the solver as ‘liblinear’. We got the best performance with the training data set from this classifier among all the classifiers. In addition to accuracy score, precision, recall, and f1-score all scored high, so overfitting of the model is suspected. With this train model, we got test results as well below (table5)

***Table 4****. Evaluating on LR with liblinear solver (training dataset)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LR** *on text data Report* | **Precision** | **Recall** | **F1-score** | **support** |
| **non-suicide** | 0.969123 | 0.986971 | 0.977965 | 7982.000000 |
| **suicide** | 0.986787 | 0.968695 | 0.977657 | 8018.000000 |
| **accuracy** | 0.977812 | 0.977812 | 0.977812 | 0.977812 |
| **macro avg** | 0.977955 | 0.977833 | 0.977811 | 16000.000000 |
| **weighted avg** | 0.977975 | 0.977812 | 0.977811 | 16000.000000 |

***Table 5****. Evaluating on LR with liblinear solver (test dataset)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LR** *on text data Report* | **Precision** | **Recall** | **F1-score** | **support** |
| **non-suicide** | 0.898170 | 0.948464 | 0.922632 | 2018.00000 |
| **suicide** | 0.944355 | 0.890515 | 0.916645 | 1982.00000 |
| **accuracy** | 0.919750 | 0.919750 | 0.919750 | 0.91975 |
| **macro avg** | 0.921263 | 0.919489 | 0.919638 | 4000.00000 |
| **weighted avg** | 0.921055 | 0.919750 | 0.919665 | 4000.00000 |

Test data prediction scores were not as higher as the training results, but still, it was relatively high, with a 91.97% of accuracy score.

## 3-4. Decision Tree

Decision Tree classifier, we got lower performance. However, we could observe which feature works for the classification. The first and second features important for the classifier were ‘life’ and ‘kill.’ The feelings of the two words are opposite. Followed to the other important feature, we could observe such meaningful words: feel, die, want, anymore, end, etc.

***Table 6****. Evaluating on DT model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **DT** *on text data Report* | **Precision** | **Recall** | **F1-score** | **support** |
| **non-suicide** | 0.784094 | 0.843865 | 0.812882 | 2991.000000 |
| **suicide** | 0.832075 | 0.769026 | 0.799309 | 3009.000000 |
| **accuracy** | 0.806333 | 0.806333 | 0.806333 | 0.806333 |
| **macro avg** | 0.808085 | 0.806446 | 0.806096 | 6000.000000 |
| **weighted avg** | 0.808157 | 0.806333 | 0.806075 | 6000.000000 |

***Table 7****. The important features from #1 to #40*

|  |  |
| --- | --- |
| 1. feature 22 (0.185702)  feature name: life  2. feature 19 (0.094523)  feature name: kill  3. feature 40 (0.069045)  feature name: suicide  4. feature 8 (0.060222)  feature name: feel  5. feature 5 (0.034184)  feature name: die  6. feature 46 (0.028931)  feature name: want  7. feature 23 (0.026601)  feature name: like  8. feature 0 (0.025622)  feature name: anymore  9. feature 24 (0.022602)  feature name: live  10. feature 6 (0.020871)  feature name: end  11. feature 31 (0.019890)  feature name: people  12. feature 43 (0.019808)  feature name: think  13. feature 20 (0.018833)  feature name: know  14. feature 10 (0.016327)  feature name: friend  15. feature 4 (0.015311)  feature name: day  16. feature 14 (0.015168)  feature name: good  17. feature 21 (0.014175)  feature name: leave  18. feature 41 (0.014140)  feature name: talk  19. feature 44 (0.013371)  feature name: time  20. feature 13 (0.013181)  feature name: get | 21. feature 45 (0.013159)  feature name: try  22. feature 17 (0.013062)  feature name: help  23. feature 29 (0.012617)  feature name: need  24. feature 27 (0.011092)  feature name: love  25. feature 12 (0.010948)  feature name: fucking  26. feature 34 (0.010802)  feature name: post  27. feature 1 (0.010559)  feature name: bad  28. feature 35 (0.010516)  feature name: right  29. feature 49 (0.010238)  feature name: year  30. feature 3 (0.009609)  feature name: come  31. feature 38 (0.009573)  feature name: start  32. feature 16 (0.009563)  feature name: hate  33. feature 42 (0.009512)  feature name: tell  34. feature 47 (0.009394)  feature name: well  35. feature 26 (0.009301)  feature name: look  36. feature 9 (0.009156)  feature name: find  37. feature 7 (0.009138)  feature name: family  38. feature 37 (0.009067)  feature name: shit  39. feature 11 (0.008796)  feature name: fuck  40. feature 36 (0.008316)  feature name: school |

## 3-5. Random Forest

For the Random Forest model, we used two different tfidf data sets. One with default set of tfidf vectorizer and the other one was tfidf vectorizer with ngram range of 1 to 2. The results were very close, but ngram range of 1 to 2 was slightly better. However, compared to other classification models, Random Forest was not performed well in this task.

***Table 8****. Evaluating on RF model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RF** *on text data Report* | **Precision** | **Recall** | **F1-score** | **support** |
| **non-suicide** | 0.784094 | 0.843865 | 0.812882 | 2991.000000 |
| **suicide** | 0.832075 | 0.769026 | 0.799309 | 3009.000000 |
| **accuracy** | 0.806333 | 0.806333 | 0.806333 | 0.806333 |
| **macro avg** | 0.808085 | 0.806446 | 0.806096 | 6000.000000 |
| **weighted avg** | 0.808157 | 0.806333 | 0.806075 | 6000.000000 |

## 3-6. Overall

Overall, comparing each classification’s best performance, Logistic Regression showed the highest accuracy score.

***Figure 9****. Bar plot of Accuracy score by models*

# Conclusions

There is always a possibility that AI will misjudge a person at risk of suicide. "false positive" means that someone is identified as at risk, but it is not. In this case, it implies a mis-attention to someone at risk of suicide. If you use "false negation," people at risk are not flagged. The risk of harm from false negatives and false positives is too significant for researchers to use AI to identify suicide risks before they are convinced it works.

Among these risks, Facebook used artificial intelligence to identify users at imminent risk of suicide. According to the website, Facebook allows users to report on things related to posts, including live Facebook videos, which may indicate they are in a crisis related to suicide. AI also makes the option of scanning posts and allowing users to report posts when deemed appropriate more prominent. Whether a user reports a post, AI can scan and flag Facebook posts and live videos. Facebook employees decide how to handle posts and videos flagged by users or AI. They can seek advice from a friend who created the post or from crisis support calls such as the National Suicide Prevention Lifeline, which launched a three-digit 988 number this month. Users can contact the crisis line directly through Facebook Messenger. If the post represents an urgent risk, Facebook can contact a police station near a potential crisis Facebook user. A police officer is then sent to your home to check your health.

The use of technology for detecting suicide risk can be definitely one promising solution. We are able to address many of the limitations of traditional detection and treatment of suicide risk. However, we must understand that there are risks and still lots of problems that we have to solve such as privacy concerns and ethical issues.

# Reference

1 SUICIDE STATISTICS [Web Link](https://save.org/about-suicide/suicide-statistics/#:~:text=There%20is%20one%20suicide%20death,and%20Black%20males%20(12.6).);

2 How Facebook AI Helps Suicide Prevention [Web Link](https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/);

3 Information and Communication Technology Use in Suicide Prevention: Scoping Review [Web Link](https://www.jmir.org/2021/5/e25288/);

4 Technology and the future of suicide prevention. [Web Link](https://psycnet.apa.org/record/2020-64368-016);