CPS803 Group 26: Project Report

Adriano Mariani
Ryerson University
Toronto, Canada
adriano.mariani@ryerson.ca

Arsalan Khuwaja
Ryerson University
Toronto, Canada
akhuwaja@ryerson.ca

Alize De Matas
Ryerson University
Toronto, Canada
adematas@ryerson.ca

Jasmine Joy
Ryerson University
Toronto, Canada
jasmine.joy@ryerson.ca

Our project uses machine learning techniques in detecting a probable suicide message based on social media posts. The posts were obtained from *r/SuicideWatch* and *r/Depression* Reddit communities and was used to train classical machine learning models such as Naïve Bayes, Support Vector Model and Logistic Regression to identify posts that indicated *suicide* vs *non-suicide*. We also experimented with deep learning using Multilayer Perceptron to train our data. The word associations derived from each model was used to identify posts with suicidal tendencies. In tokenization of text, BERT and CountVectorize methods were implemented in parallel. The tokenization method and machine learning model with best performance on three datasets was used for analysis of suicide. A word cloud was generated from the outputs of the test datasets. Key themes of language use emerge from the word cloud.

Index Terms – Machine Learning, Supervised Learning, Natural Language Processing, Naïve Bayes, Binary Classification, Logistic Regression, Suicide, Support Vector Model, Multinomial, Count Vectorize, BERT, Bernoulli.

I. INTRODUCTION

Social media platforms like Reddit provide spaces for people to build communities and share their thoughts, feelings, and emotions. Sharing ideas, memes, opinions, advice and being a part of communities are helpful coping mechanisms. The anonymity of the platforms and the presence of like-minded communities often provide people a sanctuary to vent about and share their vulnerabilities. Likewise, those who suffer from suicidal ideation also choose to exist and participate on such platforms, and they do reach out for help. However, very often such avenues go unnoticed by mental health professionals and the rest of the society. There is an opportunity to tap into such online communities and offer help and support for those seeking that connection.

Emile Durkheim was a French sociologist who presented a theory of suicide which was focused on a societal level. The key variables he identified consisted of social integration and social regulation, which affected four types of suicide. Egoistic suicide is when individuals who lack social integration and are detached from a social group, bonds, or society itself. These people are often isolated and lack a sense of belonging. Altruistic suicides are when individuals feel their death would benefit society because of being fully integrated into society. The third type of suicides includes Anomic suicides most often happen in societies that have minimal social regulation, which fails to instill a sense of meaning. The final type of suicide is Fatalistic where social regulations are extreme, and authority is oppressive and controlling so the individual rather die than continue living in their current situations. [2]

The psychological model of suicide is also known as the "escape theory"[3]. This model has a sequential process involving the following six steps: Falling short of standards (person fails to meet unrealistic life expectations), internalization of self-blame, aversive sense of self, negative affect and/or negative consequences (which manifests as depression, anxiety or anger), cognitive constriction (tunnel vision), reckless behaviours and lastly, absence of emotion and irrational thought (substance abuse, self-harm, risky behaviours and/or societal withdrawal). Some of the posts within our dataset and our findings will be correlated to these highlighted processes in the psychological model of suicide and the social theory of suicide.

According to United States National Institute of Health, the leading causes for suicide include depression, mental disorders, substance abuse, chronic pain, and exposure to violence. [3]

Furthermore, those suffering from clinical depression also suffer from irritability, restlessness, have frequent angry outbursts, withdraw from hobbies, consider and self-harm. There is also a noticeable difference in their use of language. Those with depression contained one or more of the following features in their use of language: [4]

1. Use of negative language was also more commonly associated with those suffering from depression.
2. Rumination, or repetitions of the same, negative information.
3. Excessive use of mental state verbs such as *think*, or words denoting causal relations.
4. Use of generalizing terms such as *everything*, *always*, *nothing*, *never* as also associated with their ambivalent emotional state.
5. Those suffering from depression also opt to use more descriptive, narrative and less often reasoning.

In simplest terms, those suffering from depression choose to engage with content that concerns their issues, and often come across as introspective, self-focused and overly critical about themselves.

Mental health diagnosis is a multi-variate problem and inferring suicidal ideation from social media posts are prone

to error when a lot of information often gets lost in social media communications. The errors could arise from lack of available information, context, intention of user, etc. However, monitoring such subreddit communities could provide more insight into the challenges and concerns that users regularly face.

Mental illnesses are a diverse and complex topic such that several aspects of illnesses such as depression are still being studied. The lack of physically visible and tangible symptoms compounded by the presence of cognitive biases add more layers of ambiguity to the field thereby increasing the risk of incorrect diagnoses. Hence, due to the ambiguous and complex nature of the topic, this project will not identify if the suicidal ideation is passive or active and will only focus on identifying a post as *suicide* or *non-suicide*. Additionally, the datasets obtained were already labelled as *suicide* or *non-suicide* based on criteria not known to us.

II. METHOD

In the project, a social media post is analyzed for suicidal ideation. Since we only evaluate for if a post is inferring suicide, this is a classification problem. And considering there are only two categories, this becomes a binary classification problem.

A. Datasets

Three datasets were used for this project. All three datasets were obtained from Kaggle.

1. *"Suicide and Depression Detection"* contains 232,074 unique values and classifies a user as suicidal or non-suicidal based on the text of their reddit post. This dataset contains accurate labels associated with each post. The dataset was used as a training dataset, and as a test dataset in the train-test-split experiment, where 25% of the data was used for testing and 75% for training. The dataset contains 116011 (50%) non-suicide records and 116031 (50%) suicide records and is a balanced dataset.

2. *"Depression_suicide"* contains 20,364 unique values and contains posts from *r/depression* and *r/SuicideWatch*. For the purposes of our project, we considered *r/SuicideWatch* posts to be "*suicide*", and *r/depression* posts to be "*non-suicide*". The dataset was used as a test dataset. The blind labelling of posts based on subreddits are not accurate, as there are *suicide* posts in *r/depression* and *non-suicide* posts in *r/SuicideWatch*. Hence, the dataset is not expected to yield highly accurate results. However, analysis of results for this dataset may yield insights into relationship between depression and suicide. The dataset contains 10371 (50.93%) *r/depression* records and 9992 (49.07%) *r/SuicideWatch* records and is a balanced dataset.

3. "*Suicide_notes*" contains 464 unique values. The notes were written by users who were confirmed with suicidal tendencies. The dataset is unbalanced, and all posts are labelled as *suicide*. The extract contains posts as of August 2021.

TABLE 1: DATASETS

Datasets	Intended Use	Rows	Description
suicide_detection	Training	232,074	Data from Kaggle[5]. Reddit posts labelled as *suicide*, *non-suicide*.
depression_suicide	Training & Test	20,364	Data from Kaggle[6]. Not labelled. Reddit posts from *r/depression* and *r/suicidewatch*.
suicide_notes	Test	464	Data from Kaggle[7]. Notes written by users confirmed with suicidal tendencies.

III. CLASSIFICATION ALGORITHMS

A. Pre-processing

1. Using BERT
   - BERT (Bi-directional Encoder Representation Transformer) learns a language model by encoding input text into a sequence of tokens based on the contextual relationships between words.
   - Preprocessing was done on input data by removing whitespace, non-alphanumeric characters, and Unicode characters.
   - BERT was used to obtain word embeddings of the input sentences in vector form. The vectors represent the input features used in the classification and training algorithms.
   - Due to constraints of the BERT model, vectors of size greater than 512 were removed from training and testing.
   - The tokenized output of BERT was used as input for classifiers.
2. Using Count Vectorizer
   - CountVectorizer is a tokenizer that transforms text into a vector based on the frequency of each word that occurs in the entire text. Since frequencies of words were being considered, preprocessing of the data was not necessary.
   - The tokenized output from CountVectorizer was used as input for classifiers.

B. Classification

The classification and training algorithms considered for this project were: Naive Bayes, Support Vector Model, and Logistic Regression. The algorithms are trained to classify a text input as suicidal or non-suicidal. The results from each algorithm will then be compared.

1. Naïve Bayes is a supervised classification algorithm where the variables are independent of each other. It is simple, fast, scalable, and ideal for small datasets. The datasets used in the experiments are not large, therefore a supervised classifier with high bias needed to be used. Since Naïve Bayes does relatively well with smaller datasets, Naïve Bayes was picked as a model for the project. We used *sklearn* Python library to access the *BernoulliNB* and *MultinomialNB* models. Although

GaussianNB was a good algorithm for text classification, it was not considered because the distribution of the datasets on the project are not normally distributed, and therefore is not a normal classification.

2. Support Vector Machine is a supervised classification algorithm that finds the hyperplane that optimally separates a dataset into two distinct classes. It can easily convert a linear problem into a non-linear problem using the kernel trick. Although SVM is not necessarily good for natural language processing, it is a good model for binary classification. Additionally, LinearSVC is the best linear classifier and finds the decision boundary easily. Hence, this model was considered for the project, and the *svm.LinearSVC* model is accessed through the *sklearn* Python library.

3. Logistic Regression is a classification algorithm used to predict the probability that the outcome is equal to 1 (*suicide*). It models predictions through use of a sigmoid or *relu* function, and classifies the posts are *suicidal* or *non-suicidal*. Logistic Regression uses historical data to predict the likelihood of events. Since the output is binary and is commonly used for solving binary classification problems, this model was considered for this project. We used the *sklearn* Python library to access the *LogisticRegression* model. [8]

4. Multilayer Perceptrons are the classical type of neural network composed of layers of neurons. This deep learning model has 5 layers altogether. The first layer has 100 neurons, second layer is a dropout layer, third layer has 10 neurons, fourth is a dropout layer and fifth is the output layer with 1 neuron. The accuracy of the metrics and binary cross entropy are used as a loss function. Additionally, as a result of working with a small network, as data is proportionally small, and dropouts are there to prevent overfitting. Logistic Regression and Multinomial Naïve Bayes from the classical machine learning models performed well. MLP was chosen as a model for this project to observe how well a simple neural network model would perform in comparison to the classical models. For outputs from BERT, the *MLPClassifier* model from *sklearn* Python library was used. For outputs from CountVectorizer, the *Sequential* model from *keras* Python library was used.

C. Post Processing

Datasets with predictions were divided into three categories: true positive, false positive and false negative for further analysis. The word cluster obtained in each category was analyzed further. The *nltk* python library was used to tokenize words and build a word dictionary. The *nltk* library was considered for this task due to its' ability to gather and classify unstructured texts and because it is widely used in pre-processing of natural text.

Accuracy was the primary performance metric that was considered. However, a confusion matrix was used to identify clusters of words to describe behaviour of suicide ideation. Additionally, the precision and recall rates determined the quality of the model used.

IV. Results & Discussion

A. Experiment 1: Splitting Training dataset into training and test

Expectation:

The testing set in this experiment would have the highest accuracy. Models that perform with high accuracies would be considered for Experiment 2 and 3.

Result:

1. Using BERT for tokenization provided consistent results for all the classical ML models and performed slightly worse with MLP.
2. Bernoulli Naïve Bayes was the only model that saw accuracy decline with the use of CountVectorizer for tokenization. This is expected because unlike Multinomial Naïve Bayes, Bernoulli Naïve Bayes tests for the presence or absence of a feature. For our purposes this would be a poor model as the feature set is built from the gathered posts and are not explicitly defined enough to test for the presence and absence of a feature. In contrast, Multinomial Naïve Bayes classifies a document based on the counts it finds of multiple features.
3. Although Logistic Regression performed the best with the use of CountVectorizer for tokenization, with the exception of Bernoulli Naïve Bayes, all other classical models performed well.
4. For SVM, the test accuracy was improved by using regularization. After experimenting with various values for the regularization term, we found that a regularization term equal to 0.1 yielded the highest accuracy on the test set.
5. For MLP, the train accuracy increases with each epoch and reaches very high levels close to 100%, meanwhile test accuracy increases and then decreases after some epochs due to overfitting.
6. The highest improvement was noted for the Multilayer Perceptron.
7. All models show high recall rates and high precision rates. This implies that there were fewer false positives and fewer false negatives. The model is accurately predicting the desired outcome.

Table 3: Performance of ML models on suicide_detection.csv (used for training & test)

	BERT Accuracy	**Count Vectorize Accuracy**	**Δ Accuracy**	**Precision**	**Recall**
Logistic Regression	81.08	93.19	+12.11	97.02	96.59
Bernoulli Naïve Bayes	82.33	77.75	-4.58	84.95	99.86
Multinomial Naïve Bayes	82.33	90.25	+7.92	95.98	96.36

Support Vector Machine	82.33	92.55	+10.22	97.06	94.85
Multilayer Perceptron	82.07	94.52	+12.45	97.16	95.20

B. Experiment 2 – Testing on skewed dataset, Suicide_notes

Expectation:

The model that labels the largest number of posts as suicide is considered the most accurate model of this experiment.

Result:

1. The suicide_notes.csv is a heavily skewed dataset. Models were not expected to perform well with this dataset. However, models that performed well in Experiment 1 and in Experiment 2 are learning to identify suicide posts well.
2. There was an overall improvement in performance when CountVectorizer was used for tokenization and MLP model saw the highest improvement at 59% and Bernoulli Naïve Bayes saw the least improvement at 6%.
3. Among the classical ML models, Multinomial Naïve Bayes had the most improvement at 59%. The best classical model for this experiment is Multinomial Naïve Bayes with CountVectorize used for tokenization.
4. The MLP model outperformed all the other models with both BERT and CountVectorizer, and an incredibly high accuracy rate of 97.84% for the skewed dataset.
5. Except for Bernoulli Naïve Bayes, all models show high precision rates and high recall rates. This implies that there would be no false positives, and fewer false negatives.

Table 2: Performance of ML models on suicide_notes.csv

	BERT Accuracy	**Count Vectorize Accuracy**	**Δ Accuracy**	**Precision**	**Recall**
Logistic Regression	59.50	83.41	+23.91	100.0	82.33
Bernoulli Naïve Bayes	52.86	59.05	+6.19	100.0	59.05
Multinomial Naïve Bayes	36.16	94.83	+58.67	100.0	94.82
Support Vector Machine	59.04	78.66	+19.62	100.0	78.45
Multilayer Perceptron	60.41	97.84	+37.43	100.0	80.82

C. Experiment 3 - Testing on unlabelled dataset with posts from r/depression and r/SuicideWatch

Expectation:

Assumption that *r/depression* and *r/SuicideWatch* are independent. The model that performed well in Experiment 1 and 2 would perform well in Experiment 3 if the assumption is correct. If the assumption is incorrect, the range of performance of the models would be similar. If a model performs with high accuracy, it is likely able to distinguish between *r/depression* and *r/SuicideWatch* posts.

Result:

1. BERT and CountVectorizer provide consistent results for all the classical ML models. The range of 45%-55% on accuracies was expected due to the assumptions made in the design of the dataset.
2. Among classical models, the largest improvement for accuracy was seen in Logistic Regression and SVM models. However, MLP saw a considerable improvement at 52% with the use of CountVectorizer tokenizer.
3. Among classical models, Multinomial Naïve Bayes performed the best with BERT, however linear regression and SVM performed the best with CountVectorizer.
4. MLP faced the same issues that the other classifiers did when using BERT. However, the MLP model learned the most with the CountVectorize model. This is in part due to the use of keras for MLP modelling instead of the MLPClassifier from *sklearn.neuralnetworks*. The accuracy of the MLP model is similar to that of Experiment 2 with the skewed dataset. This implies that the MLP model is learning how to distinguish between *r/SuicideWatch* and *r/depression* posts.
5. All models show high recall rates and low precision rates. This implies that there would be higher false positives, and fewer false negatives.

Table 4: Performance of ML models on r/depression and r/SuicideWatch posts

	BERT Accuracy	**Count Vectorizer Accuracy**	**Δ Accuracy**	**Precision**	**Recall**
Logistic Regression	44.98	54.61	+9.63	52.26	88.01
Bernoulli Naïve Bayes	45.02	45.30	+0.28	45.38	56.35
Multinomial Naïve Bayes	45.45	49.43	+3.98	49.21	95.84
Support Vector Machine	44.86	54.58	+9.72	52.40	85.47
Multilayer Perceptron	45.89	97.98	+52.09	51.98	89.44

D. Project Analysis

For the scope of our project, the gathered posts are only targeted for indication of suicide ideation. The outputs generated are hence, labelled as *'suicide'* or *'non-suicide'*.

In the pre-processing stage, BERT and Count Vectorizer were used in parallel for tokenization. Results from Count Vectorizer performed better than BERT in all three experiments, and the most improvement was noted with the MLP models. For this project, Count Vectorizer is the better transformer as it transforms text into a vector based on the frequency of the words that occur in texts. This aligns with the features of depression identified earlier, where repetition,

use of negation, and generalizing terms were common. Models that consider repetitive use of generalizing terms and negation have a higher advantage for accurately identifying suicide ideation. Commonsense, pragmatic inferences are often a challenge for BERT and BERT is not very good at interpreting negation of a sentence[9]. Hence, extracting the sentiment of a post is not enough as the contextual relationship between words within the use of negation must also be considered.

Experiment 1 results were consistent for all models with BERT. This could be due to the limitations of BERT as mentioned above. Among Naïve Bayes models, Multinomial Naïve Bayes should be registering better accuracy rates in comparison to Bernoulli Naïve Bayes. However, accuracy was consistent. Therefore, the BERT method of tokenizing isn't providing reasonable results.

Among classifiers, Bernoulli Naïve Bayes performed consistently between BERT and Count Vectorizer while all other models saw a large improvement in performance with the use of Count Vectorizer. Bernoulli Naïve Bayes also performed consistently poorly with Count Vectorizer in all three experiments. Hence, Bernoulli Naïve Bayes was the worst performing model for this project.

Although the Multinomial Naïve Bayes model did well in Experiment 1 and 2, it faced the same challenge that the other classical models came across in Experiment 3. The dataset assumed that *r/SuicideWatch* only contains posts with suicide ideation and was hence labelled '*suicide'*, and *r/depression* only contains posts that are '*non-suicide'*. Since all the models have roughly 50% accuracy, this implies that *r/SuicideWatch* and *r/depression* are not independent. In fact, this result confirms that people who are depressed or suffer from suicidal ideation visit both the subreddit communities. This further indicates that depression and suicide are symbiotic.

V. Emerging Themes & Future Work

In our project proposal, we aimed to conduct a comparison study of posts that appeared in social media before the pandemic and during the pandemic. However, it was not possible to obtain posts that dated back to 2020, and therefore that segment of the project could not be completed. The dataset with posts from r*/SuicideWatch* and *r/depression* are reflective of the pandemic and were obtained as of August 2021.

A dictionary of words that were frequently used in each dataset was developed and a further breakdown of the cluster of words was performed. Words that appeared in false positives were similar to those that appeared in true positives, however the context of the use of words were different.

Within true positives, words such as: *please, life, much, someone, get, time, die, hours, ago, told, end, done, long, help, like, person, drank, took, joke, pain*, etc. appeared most often. Such words were used in the context of suffering and hopelessness: *life is a joke", "I want to die", "someone please help", "end life", "don't have much time", "done with life", "end the pain"*, etc. There is also a reference to time or the past, such as *"long time ago", "hours long"*, etc.

Within false positives, words such as: *anyone, right, private, please, message, wish, person, take, really?, cares, hurting, pain,* etc. appeared most often. Such words were used in the context of pain and social anxiety: *"anyone right now", "really?", "does anyone care", "wish wasn't hurting",* etc. There is some overlap with the use of among true positives and false positives. However, the context of usage in the clustered words differ. The model falsely predicted the posts that express pain and social anxiety as that of suicide ideation.

Within false negatives, words such as: *feel, nothing, hideousness, politicians, calmly, anything, see, culture, neighbours, everywhere, venting, name, look, observing, around, society, talk, life, like, redeeming*, etc. appeared most often. Such words were used in the context of loneliness and social dissatisfaction: *"observing everyone around me", "I observe hideousness", "as I look everywhere", "venting frustration", "I feel nothing", "I don't feel anything"*, etc. Words like culture, neighbours, politicians, society, everywhere indicate that the subject is frustrated and feels disconnected from society. This is in contrast with the word cluster of true positives, where the focus was regarding the existential nature of life, time, and hopelessness. The model incorrectly predicted that such posts were *"not-suicide"*, but frequent agitation and feeling disconnected from society are considered to be indications of depression.

Results of Experiment 3 along with the false negative word cluster suggest that depression has the potential to lead a person to suicide ideation. Therefore, depression must be taken seriously. Those who suffer from depression and suicide ideation may not reach out to others in real life, they are however reaching out to such communities online.

Although such language analysis has utility in the real-world, mental health diagnosis is a multivariate problem. And suicide ideation detection is only one aspect of it. To expand the project further the following modifications could be included:

1. The problem of suicide ideation could be expanded to a multi-class problem with more features.
2. Additionally detecting probable cause from inference of language could also be done.
3. Likewise, more characteristics of depression and suicide ideation could be added as features to perform a more accurate prediction of suicidal tendencies and each post could be further classified as active or passive suicide ideation.

REFERENCES

[1] Marci Littlefield, Borough of Manhattan Community College. *"Introduction to Sociology."* What Are the Types of Suicide given by Durkheim? | Introduction to Sociology, https://courses.lumenlearning.com/atd-bmcc-sociology/chapter/what-are-the-types-of-suicide-given-by-durkheim/.

[2] Dean PJ, Range LM. The escape theory of suicide and perfectionism in college students. Death Stud. 1996 Jul-Aug;20(4):415-24. doi: 10.1080/07481189608252790. PMID: 10160573.

[3] U.S. Department of Health and Human Services. (n.d.). *Frequently asked questions about suicide*. National Institute of Mental Health. Retrieved December 13, 2021, from https://www.nimh.nih.gov/health/publications/suicide-faq.

[4] Smirnova, D., Cumming, P., Sloeva, E., Kuvshinova, N., Romanov, D., & Nosachev, G. (2018, April 10). *Language patterns discriminate mild depression from normal sadness and euthymic state*. Frontiers in psychiatry. Retrieved December 13, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5902561/.

[5] Komati, N. (2021, 05 19). Suicide and Depression Detection. Retrieved from Kaggle: https://www.kaggle.com/nikhileswarkomati/suicide-watch

[6] Mashaly, M. (2020, 07 04). Suicide Notes. Retrieved from Kaggle: https://www.kaggle.com/mohanedmashaly/suicide-notes

[7] Rigoulet, X. (2021, 08 21). Reddit dataset: r/depression and r/SuicideWatch. Retrieved from Kaggle: https://www.kaggle.com/xavrig/reddit-dataset-rdepression-and-rsuicidewatch

[8] M. P. LaValley, "Logistic Regression," Circulation, vol. 117, no. 18, pp. 2395–2399, 2008.

[9] Devlin, Jacob, et al. 2019, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[10] A. Mariani, A. Khuwaja, A. De Matas, J. Joy (2021, 10 25). Suicide Detection Code Repo. Retrieved from Github: https://github.com/AdriMariani/CPS803-Suicide-Detection