Group U: Shogo Toyonaga, Dongwon Lee, Abel Habte, Jonathan Ramos

EECS 4404/5327 Project Part 5 Report

Abstract

In this paper, our group proposes a text-based binary classifier that predicts if a social media post contains thoughts of suicidal ideation. The model input consists of text that is pre-processed using a TFIDVectorizer [1]. The associated output labels are suicide and non-suicide. Using fine tuned Random Forest [2], Multinomial Naive Bayes [3], and Logistic Regression models produced positive results on the Suicide and Depression Detection dataset [4] that was used for training and validation. The minimum obtained accuracy on the training data among the three models was 88% and the maximum was 95%. Similarly, the minimum accuracy on the testing dataset was 86% and the maximum was 94%. The results demonstrate that our computationally-light models can perform suicide-ideation classification reasonably well. These models have the implication of being able to help developing countries by providing computationally cheap, easy-to-use models that can identify suicidal-intention. A lack of resources (i.e., models) that can identify suicidal intention and warning signs has been a highlighted bottleneck for developing countries in their ability to successfully intervene in suicide attempts [5].

Introduction

1. What is your application?

WellnessOracle is a text-based binary classifier that aims to predict a social media user's mental wellbeing based on their previous posting and/or comment history. It aims to determine signs of underlying depression and/or apathy so that the machine learning model can quickly determine if a user is mentally at-risk to themselves or others.

2. What are the assumptions/scope of your project?

We assume that our dataset was composed of comments that were primarily written by adolescents and young adults (aged 15-44 years [6]) as they are the target audience of social networking websites [7]. Additionally, it was found that older users (aged 50+ years [7]) of these applications were more passive in their interactions, favouring viewing content over contributing media and discussion [7]. As a result, their writing style and choice of vernacular may not have been well-emphasized in the learning process of our application.

3. Justify why is your application important?

Suicide, especially for the younger generation, has become an increasingly pervasive problem [8]. It has become the second-most leading cause of fatality [5] among young adults, worldwide.

Our application is significant as it can be used to scrape social media networks and quickly identify suicidal sentiment in text-based posts and/or comments. This is significant as social media has become a useful tool to look into the insights of its users mental state [5]. A previous study [9] found that both suicide attempters and completers had left notes prior to taking further action. As a result, it then becomes critical to quickly identify these notes and intervene as soon as possible. Additionally, it was found that the majority of suicides occurring in developing countries were due to insufficient identification of warning signs and resources [5]. The aforementioned limitations which inhibit the classification of suicidal behaviours and intention is what our app proposes to address. We offer a variety of machine learning models and a holistic application that is (1)

computationally light, (2) easily accessible, (3) intuitive to use, and lastly, (4) scalable to new data.

4. Similar applications

Tadasse et al [5] developed an ensemble model composed of a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) [5]. It used a variety of word embeddings such as n-grams and TF-IDF to classify whether Reddit posts contained suicidal thoughts. It had the highest accuracy (93.8%) among other classical models such as XGBoost and SVM [5].

Additionally, there have been similar proposed applications which use prediction models such as Logistic Regression and Naive Bayes on other social media platforms (i.e.,Twitter) [6]. However, these approaches have only garnered accuracies of 74 - 80% among text-based suicide classification tasks.

Our work differs from the related work by leveraging a custom-tuned Random Forest Classifier that leverages TF-IDF Word embeddings and low computational costs with relatively high accuracy (86%). It is computationally less demanding to use, store, and train, in comparison to models that leverage neural networks [10]. Lastly, the application extends from previous work in (1) providing an intuitive front-end graphical user interface (GUI) to operate and (2) supplying readily-accessible testing data through integrated web scraping of Reddit and Quora. The outputs of our web scraping functionalities can then be used to retrain the model, thus, promoting continued, incremental learning [11].

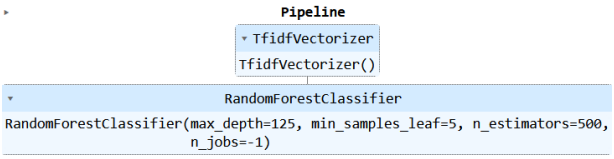
5. Adjustments to part 1

The first adjustment involved comparing our original model of choice (Random Forest [2]) against alternatives (Logistic Regression, Multinomial Naive Bayes [3]).
The second adjustment involved implementing web scraping utilities to test the model on easily-accessible, unseen data.

Methodology

1. Design/pipeline

Comparative Design Study: To analyze the model's results on unseen data, we will have an author evaluate the new data obtained from scraping an equal number of Reddit posts from York University, University of Toronto, Waterloo, and McGill (n=92). The authors classification will be compared against the predictions that the models make. The spreadsheet that outlines the results can be accessed here: https://docs.google.com/spreadsheets/d/13iuYASrK4oeUY-iVZnejvd6yndolZLQ2LtO_crGffkg/edit#gid=0
Design Pipeline: This design/pipeline can be best be illustrated by the following diagram:



1. The input text is pre-processed using the TF-IDF Vectorizer [1] from the sci-kit learn library. The default constructor is used meaning that (1) the expected content passed is expected to be of type string (2) all characters are converted to lowercase prior to tokenization, and (3) the default behaviour of stopwords is used.
2. The pre-processed text is passed to the model for fitting when training itself with train_X and train_y.

Group U: Shogo Toyonaga, Dongwon Lee, Abel Habte, Jonathan Ramos

2. Dataset

We obtained a Suicide and Depression Detection dataset that was used to train and validate our model from Kaggle [4]. It can accessed through the following link:
https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

The dataset properties can be summarized below by the following table.

Label	Count
suicide	116037
non-suicide	116037
Dataset Dimension	(232074. 2)

Additionally, our testing/validation dataset was enhanced by web scraping posts from social media platforms such as Reddit and Quora. This was accomplished by using Python libraries and dependencies such as PlayWright[12] and BeautifulSoup4[13].

We used TF-IDF[1] to pre-process both the raw and testing data features. The data was read and stored through a csv file using a Panda Dataframe and dropping any rows from consideration if there were missing values. The csv file was stored and accessed through a Google Drive directory. We were getting its dimensionality to construct the relationship between the obtained text and resulting classification. The data was pre-processed this way because we want to look for keywords in each post that may imply the proper classification and keep count of them.

3. Model training

The input is the pre-processed training data from a Kaggle dataset [4] and the testing data from a partition of the dataset in addition to the scraped social media posts. The techniques involved in initially training the model included trial-and-error until a set of good parameters were identified. After identifying the best parameters, we found their optimal values by using GridSearchCV[14] with the cross-validation generator set to 5, resulting in 160 fits.
Lastly, the output is a binary classification of either non-suicide and suicide.
As for the training process, no changes were made from what was planned in part 1.

4. Prediction

With a trained model, the model predicts the output by using TF-IDF[1] to obtain the frequency of certain keywords (the extracted features) to conclude an appropriate classification for any given new example. For example, a post with frequent use of the word "depressed" could result in the suicidal classification whereas a post with frequent use of the word, "happy" could result in the non-suicidal classification. A word cloud is procedurally generated within the Google Colaboratory environment to visualize the word relationships to the classification tasks.

Group U: Shogo Toyonaga, Dongwon Lee, Abel Habte, Jonathan Ramos

Results

1. Evaluation

To evaluate our model, we separated our dataset into a portion of training and testing sets to assess the model's performance. The training set used consisted of 80% of the data, and the remaining 20% allotted to the testing set. Additionally, to fine-tune the parameters we used the GridSearchCV[14] techniques on all models, ensuring our models are optimal.

By leveraging web scraping tools such as PlayWright [12] and BeautifulSoup4 [13], we saw an opportunity to implement a Reddit, Quora, and Twitter scraper. We were able to automate the process and feed new unseen data to the model for evaluation. This allowed us to run a comparative rather than a user study.
For our model comparisons, we used Random Forest [2] and compared it against Multinomial Naive Bayes[3] and Logistic Regression, popular choices for text classification and baseline comparisons.

Random Forest [2], our initial model of choice, stood out by its ability to capture more intricate, nonlinear relationships between features and class labels. In particular, we chose this model as it would perform better when expanding the scope of this work (i.e., the vocabulary encapsulated by the training set).

While Random Forest did not always outperform Logistic Regression and Multinomial Naive Bayes in simpler text classification, its ability to scale and handle more complex sentences makes it a compelling choice for future tasks where complexity and subtle interactions play an important role in achieving accurate predictions.

The metric used is accuracy, which measures the proportion of correctly classified instances or predictions over the total number of instances in the dataset [15]. Accuracy offers a general overview of our models "correctness", that is to say how well it makes correct predictions.

2. Results

The model scores are as follows

The graph on the left displays the training accuracies for the respective models. The graph on the right displays how accurate it was in classifying our scraped comments from University subreddits(n=92) against one of the author's text-based classifications.

The results are as follows:

Model Accuracy for Suicide and Depression Detection Dataset

Model	Training Data (%)	Testing Data (%)

Random Forest	**88**	**86**
Naive Bayes	**94**	**91**
Logistic Regression	**95**	**94**

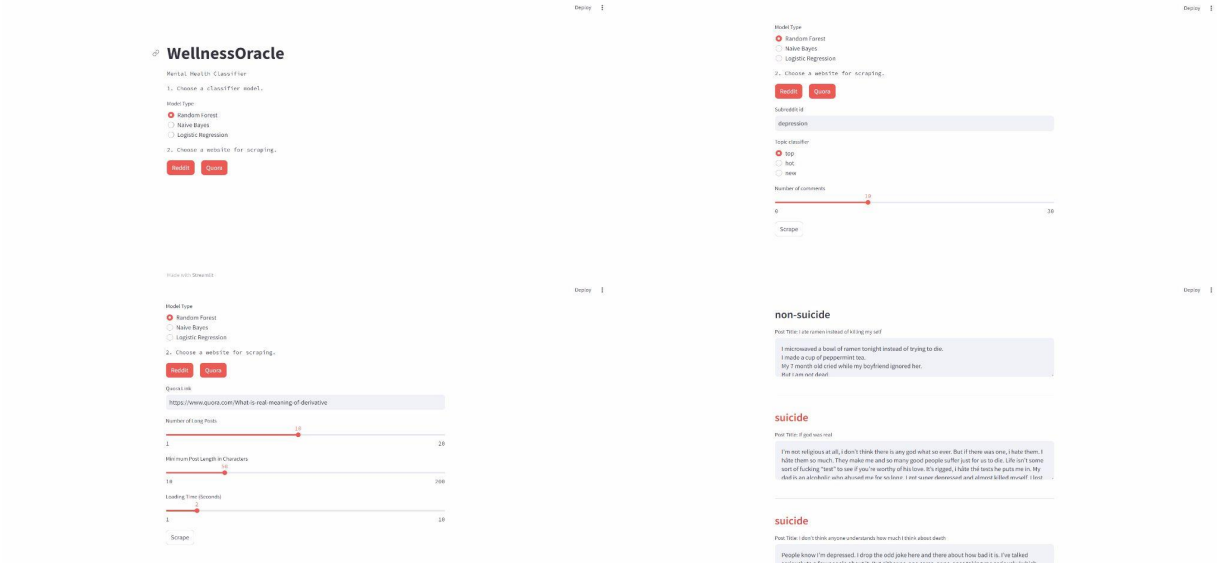
Model Accuracy for Scraped Website Posts:

Model	**Accuracy (%)**
Random Forest	**87**
Naive Bayes	**65**
Logistic Regression	**94**

Web UI

For intuitive use of the application, we built Web UI with python frontend library Streamlit.

1. A user selects one of Random Forest, Naive Bayes, and Logistic Regression as a machine learning model.
2. The user selects one of Reddit and Quora for web scraping.
3. After choosing a website, the user can set configuration, like subreddit id or number of posts to scrape.
4. The user clicks the 'Scrape' button to scrape posts and see the predicted results. Posts and prediction result 'suicide' or 'son-suicide' will show up for each post.



Discussions

1. Implications

The results suggest that our goal has been achieved. Our trained computationally cheap machine learning models could identify suicidal ideation in text classifications with high accuracy. There were several factors that contributed to that. Firstly, it was crucial that we used a comprehensive and diverse dataset for obtaining accurate results on the machine learning models. Secondly, it was crucial that the hyperparameters for each

model were fine tuned by using GridSearchCV [14] in order to prevent over and/or underfitting.

2. Strengths

- Computationally Lightweight
  - *WellnessOracle* uses one of Random Forest, Logistic Regression, Multinomial Naive Bayes models according to the user's choice. They are computationally cheaper compared to neural network models like LSTM and CNN which were leveraged in previous work [5]. It ensures that our application can be run even on low-level hardware.
- Scalable to New Data
  - With the application's web scraping features, a user can predict on the text data from popular social media websites like Reddit, Quor, and Twitter just by making simple settings like inputting a subreddit id.
- Easily Usable
  - The application has a web UI built on Streamlit as an interactive interface, where users can choose a classifier, scrape test data from websites with just simple settings, and see the prediction results intuitively[16]. It allows users with varying technical proficiency to use the model with ease.
  - Our application is also provided as a Google Colab file so that anyone can run the application with their Google Colab. The user does not need to install any third-party software and/or dependencies to use our software-based solution.
- High Accuracy
  - Our logistic regression model achieved the accuracy of 0.94 for the test data. The highest accuracy from the similar application was Tadasse et al [5]'s application's 0.938. So, we made a model that surpasses similar applications for suicidal ideation text-classification.

3. Limitations

- Limited Web Scraping
  - Currently, our application scrapes posts from only Reddit, Quora, and Twitter. It can be developed further to support other websites and use-cases.
- Difficulties in capturing semantic information
  - In our application, TF-IDF Vectorization [1] was used. It does not understand the semantic meaning of text-based data unlike a Large Language Model, but just calculates the vector value of the data. So it has limitations in achieving very high accuracy and understanding semantic relationships.
  - The model has only been tested based on English-based prompts. It is unclear if this solution generalizes well to other languages and dialects.
  - The model will have difficulty in identifying uncommon vernacular (i.e., English Slang) as it may not be well-represented in the training data's text corpus. Such words could have significant indications/ hints towards suicidal ideation that could otherwise be ignored by the strict use of TF-IDF [1].
- Difficulties in capturing non-linear relationships
  - Multinomial Naive Bayes [3] relies on the assumption of conditional independence between features given the class label which can limit its effectiveness in capturing non-linear relationships that are present in language. Logistic Regression also fails to capture non-linear relationships in data due to its linear decision boundary.

4. Future directions

- Training on a dataset with more robust features. While TF-IDF provided useful insights, there may have been other important aspects to consider such as the user's personality (i.e., Dark Humour, Satire, Maturity / Age), or time of post (hour / month) [17].
- Implementing a moderation tool that flags user accounts on social media if they start posting consecutively harmful posts. The threshold can be set by the application maintainer.
- Implementing a feature that can merge scraped social media posts and comments into the dataset for manual annotation by an expert. In this way, we would be supporting incremental learning [11].
- Implementing a custom-trained large language model (LLM) that can generate persuasive messages which attempt to get flagged users to seek help for their specific problems. This feature aims to overcome the one-size-fits-all approach used in suicide prevention that is generally used [18].
- Evaluating the performance of the Random Forest model on multi-class text-based classification. While its accuracy was lower when testing for binary text classification, it was still able to better understand and classify more intricate and complex sentences compared to Multinomial Naive Bayes and Logistic Regression.
- Utilizing an ensemble of computationally cheap models and making suicide-ideation classifications based on majority-voting.
- Finetuning pre-existing neural networks from HuggingFace to perform our suicide-ideation classification.

Additional Questions

1. What feedback did you find useful from the peer evaluation?

- Change #A: Misc (Other)
  - **Feedback / Recommendation:** Expand features to more generalized suggestive langauge about the topic of suicide rather than explicit comments.
  - **Author Comment:** This is a very beneficial piece of advice as adding more features could help with the models accuracy on more difficult-to-classify sentences. Doing so would enable semantic understanding of a sentence at a very low-level.
- Change #B: Limitations
  - **Feedback / Recommendation:** Explain how the model might not understand how to interpret English slang.
  - **Author Comment:** We made sure to incorporate this feedback in the Limitations section.
- Change #C: Discussion
  - **Feedback / Recommendation:** Clarify which machine learning model the paper is focusing on and a suggestion to choose a model based on an algorithm.
  - **Author Comment:** This is reasonable feedback as it showed that there was ambiguity in our writing process. It was fixed by appending elaborations in the Implications section.
- Change #D: Misc (Other)
  - **Feedback / Recommendation:** It could benefit from incorporating more modern methods like deep learning models (e.g., autoregressive models) for text analysis.
  - **Author Comment:** We agree that this is certainly a good step in future work as deep learning offers more powerful applications and robustness to suicide ideation classification and prevention. If we had more time, our group would pursue using a tailored, computationally-light LSTM model to achieve this task and pair it with classifications made by our other three models. The final output would leverage weighted majority-voting.
- Change #E: Methodology
  - **Feedback / Recommendation:** There are some fundamental flaws in the selected dataset in that it is essentially pre-labelled (r/suicide vs. r/teenager), with no kind of confirmation (ex. suicide reports and/or human labelling).

    - **Author Comment:** The feedback falls outside of the scope of our project. Although, it could be interesting to explore this path in future related work.
- Change #F: Misc (Other)
    - **Feedback / Recommendation:** The abstract is very verbose. Some other parts are verbose too.
    - **Author Comment:** We agreed that certain areas could be trimmed down and summarized better. This will make the proposal more clear and easy to understand.
- Change #G: Methodology
    - **Feedback / Recommendation:** Determining if a given text is indicative of suicidal thoughts based on the frequency of words such as "depressed" may not be entirely accurate.
    - **Author Comment:** We do agree that using more complicated, pretrained models which understand semantic meanings and relationships (i.e., BERT) can be useful. In future work, we do see the benefit in either finding and or creating small and robust neural networks which can be used alongside our current models that use TF-IDF to provide a more well-rounded classification.

2. What changes did you make based on the feedback from peer evaluation?

- Change #A: General Suggestion
    - Feasibility: Unreasonable (Major Revision; Limited Time)
    - **Reflected Changes:** None were made at this time. Due to the limited remaining time, this change could not be incorporated into the project as it would require substantial data augmentation on the dataset and re-training of the three models.
- Change #B: Limitations
    - Feasibility: Reasonable (Small Writing Revision)
    - **Reflected Changes:** Minor sentence revision in the Limitations / Future Work.
- Change #C: Discussion
    - Feasibility: Reasonable (Small Writing Revision)
    - **Reflected Changes:** Minor sentence revision to remove ambiguity in Implications.
- Change #D: General Suggestion
    - Feasibility: Unreasonable (Major Revision; Limited Time)
    - **Reflected Changes:** None were made at this time. We will not be able to implement, train, and test a neural network at this time. There is simply not enough time to act on this feedback and ensure quality results. If we were to act on this advice, we would most likely pursue the creation and usage of a small LSTM model.
- Change #E: Methodology
    - Feasibility: Unreasonable (Out of Project Scope)
    - **Reflected Changes:** None were made at this time. The suggestion falls outside of the scope of sentiment analysis (classification).
- Change #F: General Suggestion
    - Feasibility: Reasonable (Small Writing Revision)
    - **Reflected Changes:** Minor revisions and rewriting across the document.
- Change #G: Methodology
    - Feasibility: Unreasonable (Major Revision; Limited Time)
    - **Reflected Changes:** Unfortunately, due to limited time and resources (Google Colaboratory restricting RAM to insufficient amount), we cannot implement or test large language models of reasonable complexity. If we had more time, we would try to utilize sharded pre-existing models from HuggingFace and then finetune them using FP16 on our suicide ideation classification task.

References

[1] "sklearn.feature_extraction.text.TfidfVectorizer," scikit-learn. Accessed: Nov. 11, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[2] "sklearn.ensemble.RandomForestClassifier," scikit-learn. Accessed: Nov. 11, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[3] "sklearn.naive_bayes.MultinomialNB," scikit-learn. Accessed: Nov. 11, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[4] "Suicide and Depression Detection." Accessed: Nov. 11, 2023. [Online]. Available: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

[5] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Suicide Ideation in Social Media Forums Using Deep Learning," *Algorithms*, vol. 13, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/a13010007.

[6] V. Leiva and A. Freire, "Towards Suicide Prevention: Early Detection of Depression on Social Media," in *Internet Science*, I. Kompatsiaris, J. Cave, A. Satsiou, G. Carle, A. Passani, E. Kontopoulos, S. Diplaris, and D. McMillan, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 428–436. doi: 10.1007/978-3-319-70284-1_34.

[7] C. Bell, C. Fausset, S. Farmer, J. Nguyen, L. Harley, and W. B. Fain, "Examining social media use among older adults," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, in HT '13. New York, NY, USA: Association for Computing Machinery, May 2013, pp. 158–163. doi: 10.1145/2481492.2481509.

[8] D. D. Luxton, J. D. June, and J. M. Fairall, "Social Media and Suicide: A Public Health Perspective," *Am. J. Public Health*, vol. 102, no. S2, pp. S195–S200, May 2012, doi: 10.2105/AJPH.2011.300608.

[9] T. M. DeJong, J. C. Overholser, and C. A. Stockmeier, "Apples to oranges?: A direct comparison between suicide attempters and suicide completers," *J. Affect. Disord.*, vol. 124, no. 1, pp. 90–97, 2010, doi: 10.1016/j.jad.2009.10.020.

[10] A. Sanyal, "Neural Networks Training with Approximate Logarithmic Computations," Medium. Accessed: Oct. 14, 2023. [Online]. Available: https://towardsdatascience.com/neural-networks-training-with-approximate-logarithmic-computations-44516f32b15b

[11] "Incremental learning," *Wikipedia*. Sep. 07, 2023. Accessed: Nov. 11, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Incremental_learning&oldid=1174328493

[12] "Installation | Playwright Python." Accessed: Nov. 03, 2023. [Online]. Available: https://playwright.dev/python/docs/intro

[13] L. Richardson, "beautifulsoup4: Screen-scraping library." Accessed: Nov. 03, 2023. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/

[14] "sklearn.model_selection.GridSearchCV," scikit-learn. Accessed: Nov. 11, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[15] "sklearn.metrics.accuracy_score," scikit-learn. Accessed: Nov. 11, 2023. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.metrics.accuracy_score.html

[16] "Streamlit • A faster way to build and share data apps." Accessed: Nov. 11, 2023. [Online]. Available: https://streamlit.io/

[17] D. G. Coimbra *et al.*, "DO SUICIDE ATTEMPTS OCCUR MORE FREQUENTLY IN THE SPRING TOO? A SYSTEMATIC REVIEW AND RHYTHMIC ANALYSIS," *J. Affect. Disord.*, vol. 196, pp. 125–137, 2016, doi: 10.1016/j.jad.2016.02.036.

[18] J. I. Meza and E. Bath, "One Size Does Not Fit All: Making Suicide Prevention and Interventions Equitable for Our Increasingly Diverse Communities," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 60, no. 2, pp. 209–212, 2021, doi: 10.1016/j.jaac.2020.09.019.