

**IS450 Text Mining and Language Processing**

**Academic Year 2023/24 Term 2**

**Suicide Prevention: Analysis of Social Media Posts**

**Section: G2**

**Submitted by:**

**Group 1**

|  |  |
| --- | --- |
| **Full Name** | **Campus ID** |
| Law Wen Xin  (wenxin.law.2021@scis.smu.edu.sg) | 01415713 |
| Jin Zi Long  (zilong.jin.2023@exchange.smu.edu.sg) | 01396692 |
| Ong Swee Long  (slong.2022@scis.smu.edu.sg) | 01414567 |
| Wong Dehou  (dehou.wong.2021@scis.smu.edu.sg) | 01387549 |
| Hsu Thitsar Lwin (tlhsu.2021@scis.smu.edu.sg) | 01432424 |

[1. Introduction: Background and Motivation 4](#_Toc163678923)

[2. Dataset 4](#_Toc163678924)

[3. Related Work 4](#_Toc163678925)

[3.1 Detection of Depression-Related Posts in Reddit Social Media Forum 4](#_Toc163678926)

[3.2 Proactive Suicide Prevention Online (PSPO): Machine Identification and Crisis Management for Chinese Social Media Users With Suicidal Thoughts and Behaviors 4](#_Toc163678927)

[4. Solution Overview 5](#_Toc163678928)

[5. Solution Details 5](#_Toc163678929)

[5.1 Data Pre-processing 5](#_Toc163678930)

[5.2 Classification 6](#_Toc163678931)

[5.2.1. Random forest 6](#_Toc163678932)

[5.2.2 Naive Bayes 6](#_Toc163678933)

[5.2.3 Support Vector Machine (SVM) 6](#_Toc163678934)

[5.2.4 Challenges and Solutions 6](#_Toc163678935)

[5.3 Topic Modelling 6](#_Toc163678936)

[5.3.1 Non-negative Matrix Factorization (NMF) 7](#_Toc163678937)

[5.3.2 BERTopic 7](#_Toc163678938)

[5.3.3 Latent Dirichlet Allocation (LDA) 7](#_Toc163678939)

[5.3.4 Challenges and Solutions 7](#_Toc163678940)

[6. Results and Analysis 8](#_Toc163678941)

[6.1 Comparison of Classification Models 8](#_Toc163678942)

[6.1.1 Evaluation of the Models 8](#_Toc163678943)

[6.2 Comparison of Topic Modelling Models 8](#_Toc163678944)

[6.2.1 Evaluation of the Models 8](#_Toc163678945)

[6.2.2 NMF Results and Analysis 8](#_Toc163678946)

[6.2.3 BERTopic Results and Analysis 9](#_Toc163678947)

[6.2.4 LDA Results and Analysis 9](#_Toc163678948)

[6.2.5 Overall Comparison of coherence score, intertopic distance map, and the topics 9](#_Toc163678949)

[6.2.6 BERTopic Insights on Cluster 9 “Self-Image and Physical Appearance” 9](#_Toc163678950)

[7. Demonstration 10](#_Toc163678951)

[7.1 Web Application for Classification of Suicidal Language 10](#_Toc163678952)

[7.2 Topic Modelling Visualization for BERTopic 10](#_Toc163678953)

[8. Business analysis 10](#_Toc163678954)

[9. Discussions and Gap Analysis 11](#_Toc163678955)

[9.1 Discussions 11](#_Toc163678956)

[9.1.1 What worked well 11](#_Toc163678957)

[9.1.2 What did not work well (Error Analysis) 11](#_Toc163678958)

[9.2 Gap Analysis 12](#_Toc163678959)

[10. Use of generative AI 12](#_Toc163678960)

[11. Future Work and Conclusion 12](#_Toc163678961)

[11.1 Future Work 12](#_Toc163678962)

[11.2 Conclusion 13](#_Toc163678963)

[12. Project Experiences/ Reflections 13](#_Toc163678964)

[12.1 Law Wen Xin 13](#_Toc163678965)

[12.2 Jin Zi Long 13](#_Toc163678966)

[12.3 Ong Swee Long 13](#_Toc163678967)

[12.4 Wong Dehou 13](#_Toc163678968)

[12.5 Hsu Thitsar Lwin 14](#_Toc163678969)

[13. References 15](#_Toc163678970)

[14. Appendix 17](#_Toc163678971)

# Introduction: Background and Motivation

Suicide is a tragic and deeply personal issue impacting people of all ages and communities. Over 700,000 lives are lost to suicide every year, making it a leading cause of death globally among 15 to 29-year-olds (WHO, 2023). In our home of Singapore, we have witnessed a concerning surge in suicide rates, with a record high of 476 suicides in 2022 (CNA, 2023), highlighting the urgent need for more effective intervention strategies.

While appreciating the measures from the Ministry of Education and the Health Promotion Board (Ministry of Health, Singapore, 2023), we believe there is still more work to be done. In today's digital era, social media has become a platform for individuals to share their personal struggles, including those related to mental health. Our project aims to leverage text mining techniques such as Topic Modelling and Classification to parse through the depths of Reddit.

Through careful classification of language patterns and an exploration of the underlying topics in suicide-related conversations, we aim to provide mental health professionals and support networks with the insights they need for timely and effective outreach. These insights are the keys to deploying mental health resources more thoughtfully and crafting the most resonating interventions. Our goal is to make suicide prevention a proactive strategy, not just a reactive effort, ultimately safeguarding our community's well-being.

# Dataset

For our project, we used a dataset sourced from Kaggle. It includes 232,074 Reddit posts and contains two main classes: "suicide" and "non-suicide." The posts labeled as "suicide" are from the r/SuicideWatch subreddit, while the "non-suicide" class includes posts from r/teenagers and other subreddits. The “non-suicide” class serves as a control group for comparison. The dataset is balanced and contains the same amount of data for each class.

# Related Work

## 3.1 Detection of Depression-Related Posts in Reddit Social Media Forum

The study by Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang from Dalian University of Technology, China, aimed to identify the indicators of depression in posts shared on the Reddit discussion forum. The study utilized various machine learning classification models such as Logistic Regression, Support Vector Machine, Random Forest, Adaptive Boosting, and Multilayer Perceptron (MLP). The study found that analyzing posts by combining LIWC, LDA, and bigram features resulted in significant improvements in the performances of the models. The MLP classifier with those combined features performed the best, with an accuracy of 91% and an F1 score of 0.93.

## 3.2 Proactive Suicide Prevention Online (PSPO): Machine Identification and Crisis Management for Chinese Social Media Users with Suicidal Thoughts and Behaviours

The study done by Xingyun Liu, Xiaoqian Liu, Jiumo Sun, Nancy Xiaonan Yu, Bingli Sun, Qing Li, and Tingshao Zhu employed several machine learning models, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR) to identify people at risk of suicide on a Chinese social media platform. It was found that the SVM model with a combination of n-gram features, domain knowledge, and theory-motivated features (related to aspects like personality and depression) performed the best. People identified by the model received messages with emotional support, resources, and counselling options. After the intervention, their language showed a potential decrease in suicidal thoughts, indicating a potential reduction in suicidal ideation.

We will be applying the ideas and techniques learned from these studies, as well as other articles on text mining techniques, to our project to identify suicidal language on social media and generate insights that are beneficial to society.

# Solution Overview

Our team decided to work on Classification and Topic Modelling for this dataset. To do so, we followed the tasks in the solution overview diagram below. First, we started with data pre-processing to prepare the data set for model training. We chose three models for each task and trained the models. We hyper-tuned the models to get the optimal results for each model before we compared and chose the best model for each task. From the results of the selected model for each task, we extracted insights that our stakeholders can use, and we developed visualizations and demonstrations to capture these insights that will benefit our stakeholders.

A diagram of a model

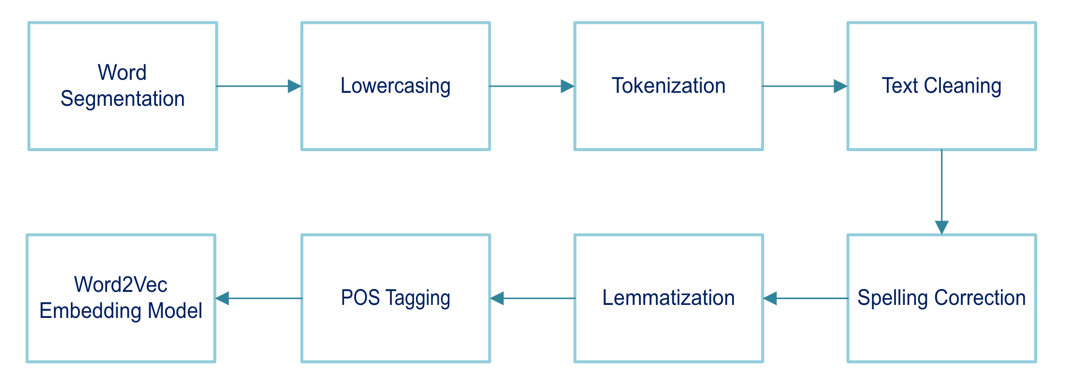
Description automatically generated

# Solution Details

## 5.1 Data Pre-processing

For our data pre-processing, we began by segmenting the text into individual words using Wordninja, which made it easier to apply further processing. We then converted all the characters into lowercase to ensure consistency. Next, we split the text into individual words or tokens and embarked on a meticulous text-cleaning process. This involved replacing abbreviations, removing elements such as stop words, punctuation, numbers, special characters, URLs, emails, emojis, and symbols, as well as attending to any white spaces within the text.

To improve data accuracy, we addressed typos and misspellings through spelling correction. We also used lemmatization to reduce words to their base forms, promoting consistency and reducing the dimensionality of our features. Finally, we implemented part-of-speech (POS) tagging to identify each word's grammatical function, such as nouns and verbs. To capture the semantic meaning and relationships between words, we utilized a Word2Vec word embedding model. This comprehensive pre-processing process ensured that our data was clean and standardized for further analysis.



## 5.2 Classification

Our goal here is to determine whether posts are suicidal. We decided to use Random Forest, Naïve Bayes, and Support Vector Machine (SVM) models for classification. To ensure consistency, we set the random\_state property for all models to 42.

### 5.2.1. Random Forest

In this model, we first train it with TF-IDF features from the text data. After that, we included sentiment scores as additional feature engineering. Using a grid search, we got the best fit for the hyperparameters n\_estimators, max\_depth, min\_samples\_split, and min\_samples\_leaf. We also considered using a randomized search but found that the parameters from the grid search were better. The hyperparameter values produced by grid search are 100, 20, 5, and 1, respectively. Moving on, we then trained the random forest model using these parameters.

### 5.2.2 Naive Bayes

We use multinomial Naive Bayes and trained it using TF-IDF and sentiment score features. To fine-tune the model, we declared several alpha values, from 0.5 to 2.0. Using grid search, the accuracy results for the best alpha value was 2.0. We then trained the model using an alpha of 2.0.

### 5.2.3 Support Vector Machine (SVM)

With SVM, we used a linear kernel. We considered a non-linear kernel but realized that since this classification produces a binary output of either suicidal or non-suicidal, we felt that the line is clearly distinguishable, and a linear kernel is sufficient. Like the previous two models, we trained it with TF-IDF and sentiment score features. In addition, we improved the model using fine-tuned hyperparameters of regularization parameter (c), kernel choice, and gamma. These hyperparameter values were produced by randomized search. Grid search would have also worked, but it took too long to run, which is why we decided to drop it.

### 5.2.4 Challenges and Solutions

Implementing grid search for parameter fine-tuning for SVM took too long (16 hours run time and it is still running), so we decided not to pursue the model. In addition, we were also unsure about using sentiment scores as a feature. Still, we ultimately decided to do so as sentiments tend to be negative and, therefore, lead to a suicidal classification for our model. (Teixeira et al., 2021).

## 5.3 Topic Modelling

Our goal here is to extract topics from the dataset so that we can identify the trends for suicidal and non-suicidal data. These extracted topics and their accompanying frequent words can give us insights into the suicidal data and are beneficial for our stakeholders. The three models that we have decided on are Non-negative Matrix Factorization (NMF), BERTopic, and Latent Dirichlet Allocation (LDA).

### 5.3.1 Non-negative Matrix Factorization (NMF)

**Model chosen:** For NMF, we tried to run the code on both raw and cleaned data. Both go through a series of our own cleaning, such as using English stopwords and contraction dictionaries. Then, we run it using n-grams between 1 and 2 and iterate through a number of topics (up to 75) to find the optimal number of topics with the highest coherence score. Using that, we reran it to get the list of top words and the intertopic distance map.

**Parameters chosen:** For NMF, we eventually chose cleaned data, which was limited to 100K rows due to time constraints. After running through the series, we found that the optimal number of topics is 10. There are some overlapping topics, but the top words indicate significant semantic differences regardless.

### 5.3.2 BERTopic

**Model chosen:** BERTopic is relatively straightforward. We trained it on the entire uncleaned data to keep the meaning of the words, and before training the BERTopic, the vectorizer model implemented the removal of stop words from the data.

**Parameters chosen:** For BERTopic, we used a minimum topic size of 300, which resulted in 64 topics, many of which were overlapping.

### 5.3.3 Latent Dirichlet Allocation (LDA)

**Model chosen:** We decided to train 9 models, all possible variations between (unigram, bigram, trigram) and (suicide class, non-suicide class, combination of both classes). Cleaned data that has undergone preprocessing is chosen to train the LDA model. We tried using raw data to train the LDA with the intention of keeping the original meaning of the phrases, but there were too many spam words, which led to the frequent words in the topics making little sense.

**Parameters chosen:** The only tuning that we did for LDA was changing the number of topics to get the optimal coherence score and ensure that the topics are distinct (small amount of overlap in the “intertopic distance map”).

### 5.3.4 Challenges and Solutions

**Lack of computing power:** Online workspaces like Deepnote take forever to run the model, and the Kernal will usually fail halfway through the training. We resolved this by transitioning to using our local drives and Jupyter Notebook to train the models. With bigrams and trigrams, the code takes up a lot of computational power to form them. Hence, we decided to use the first 100,000 rows of data instead of the complete data set to train the LDA models.

Hyper-tuning the models to find the optimal number of topics for the best coherence score required a long loop through. To overcome the limited time and computing power, we first chose the best model, which is LDA Bigram trained on 100k data. We then ran the model at intervals of 5 topics, from 5 to 10, 15 to 20 topics, and chose the number that resulted in the highest coherence score.

**Too many topics on the intertopic distance map:** With 64 topics for BERTopic, there were too many topics that were overlapping. Hence, we grouped the overlapping topics to form a single topic.

**Forming Bigrams and Trigrams:** We had some coding challenges when forming bigrams and trigrams with functions available from Gensim. Hence, we resorted to searching online forums, and, in the end, we used an enumerate function to form the bigrams and trigrams.

**Poor output:** Our initial output from bigrams and trigrams LDA needed to be more readable because we used lemmatized cleaned data. We then implemented the Prof’s advice on LDA best practices, which resulted in readable outputs. The steps we took were to form the bigrams and trigrams first from unprocessed data before removing any meaningless bigrams or trigram phrases. We then trained the LDA model.

# Results and Analysis

## 6.1 Comparison of Classification Models

We compared the performance of three classification models, Random Forest, Naïve Bayes, and SVM, in classifying whether texts were suicidal or non-suicidal. We evaluated the models using accuracy, precision, recall, and f1-score.

### 6.1.1 Evaluation of the Models

The evaluation results of the classification models can be found in Appendix 6.1.1. Overall, the SVM model has the best scores, as it tops the score for every metric compared to the other models. However, since the difference in scores was minimal, it was inadequate to rely on the metrics to determine the best model. Thus, we decided to create new texts and have the models evaluate them.

To decide on the best model, we generated some suicidal texts, which can be found in Appendix 6.1.2. From the results, we can see that the Naive Bayes model classifies most of them accurately despite not having a high score. Thus, through human evaluation, we determined that the Naive Bayes model is the best and used it for the demonstration.

## 6.2 Comparison of Topic Modelling Models

### 6.2.1 Evaluation of the Models

As BERTopic does not have a coherence score, human evaluation is the deciding factor. The model that produces the most significant number of distinct topics that are coherent and meaningful will be the best model for topic modelling. In this case, BERTopic performed the best in terms of distinct and coherent topics and words.

|  |  |  |  |
| --- | --- | --- | --- |
|  | NMF | BERTopic | LDA Bigram |
| Coherence Score | 0.351 | Na | 0.7668 |
| Examples of topics | 1) Generic feelings e.g. bored, sad, hate  3) Suicidal thoughts (but not very clear) | 2) Media and Entertainment Engagement  3) Social Interaction and Companionship Seeking  6)Mental Health Challenges and Coping Mechanisms  9)Self-Image and Physical Appearance | 1) Suicidal thoughts  2) People and feelings |
| Human Evaluation | Generally better topics than LDA Bigrams | Generates the best topics among all 3 models | Topics did not fare well as compared to the other 2 models |

### 6.2.2 NMF Results and Analysis

Using NMF, we first experimented with the model on the raw data, with a bit of basic cleaning on our own. By iterating through the different topic numbers and finding their corresponding coherence score, we got the following results (Appendix 6.2.2A). Surprisingly, 10K rows provided a trend of increasing coherence score against the number of topics, while 100K rows have a decreasing trend. However, the optimal number of topics is similar, at around 50.

Subsequently, we experimented on the cleaned data, and we saw a smaller optimal number of topics in general, with the best coherence score being slightly better than the previous results (Appendix 6.2.2B). Hence, we used these to generate an Intertopic Distance Map (IDM) for each of these datasets. We ultimately determined that the 100K cleaned data is the best as it has a decent number of topics with generally coherent top words, and there is less overlap between topics on the IDM. However, we ought to compare this with the results from other models to determine its effectiveness.

### 6.2.3 BERTopic Results and Analysis

BERTopic generated 64 topics, many of which overlap (Appendix 6.2.3A). Hence, we combined the overlapping topics to form a single topic (Appendix 6.2.3B). An alternative way to combine the topics for BERTopic was to use Hierarchical Clustering (Appendix 6.2.3C).

Among the 12 topics, there was a mix of suicidal and non-suicidal topics. Suicidal topics such as cluster 6: “Mental Health Challenges and Coping Mechanisms” have frequent words such as “suicidal thoughts” and “want kill”. For non-suicidal topics like cluster 2 “Media and Entertainment Engagement”, it has frequent words like “watch anime” and “play minecraft”. Overall, the 12 topics are meaningful, and frequent words are coherent.

### 6.2.4 LDA Results and Analysis

We tried to optimize the LDA coherence score by varying the number of topics. Unigram LDA had a better coherence score with 5 topics or below or with 19 topics (Appendix 6.2.4A). However, we chose 10 topics because that is one of the higher and more stable points. For unigrams, even though the number of topics has been optimized, the coherence score is still low at 0.3778.

For bigrams, trigrams, and a combination of both, we used a fixed number of 10 topics. They have a higher coherence score than unigrams, with Bigrams having the highest coherence score of 0.7668 (Appendix 6.2.4B). Some were based on the intertopic distance map. Hence, we combined the overlapping topics to form a single topic.

**Human Evaluation:** There are 6 unigram LDA topics (Appendix 6.2.4C), which mainly have little meaning on their own. There are 4 bigram LDA topics (Appendix 6.2.4D), 2 of which make sense and 2 of which do not. There are 4 trigram LDA topics extracted (Appendix 6.2.4E), which performed poorly as the frequent words are repeated across the topics. There are 4 topics for bigrams and trigrams LDA (Appendix 6.2.4F), and they do have some meaningful insights.

Given that the Bigram gave a better coherence score and more relevant topics (by human evaluation), we will use the Bigram model. We trained the LDA model on suicidal and non-suicidal documents together as LDA models that were trained in a single class gave us very similar topics.

### 6.2.5 Overall Comparison of Coherence Score, Intertopic Distance Map, and Topics

Comparing the 3 models, BERTopic emerged as the winner. We are looking for a model that produces the greatest number of distinct topics that are meaningful and coherent. Based on the intertopic distance map, BERTopic produced 12 distinct clusters (which became topics). After conducting human evaluation, BERTopic’s topics were the most coherent, and each topic’s frequent words were relevant and made sense.

### 6.2.6 BERTopic Insights on Cluster 9 “Self-Image and Physical Appearance”

|  |  |  |
| --- | --- | --- |
| Cluster 9  Self-Image and Physical Appearance | Topics: 11,34,45,47,59 | ['trans people', 'gay people', 'super straight', 'im trans', 'feel like', 'gender dysphoria', 'im transgender', 'im ugly', 'look like', 'youre ugly', 'good looking', 'self esteem', 'long hair', 'cut hair', 'hair like’, 'look like', 'feel like', 'im just', 'look mirror', 'good looking', 'dont want', 'plastic surgery'] |

From the lists of top frequent words related to cluster 9, we extracted a list of unique phrases that are closely related to the topic such as ['trans people', 'gay people' ...... 'look mirror', 'plastic surgery'].

Looping through each Reddit post, we try to find the number of times the words in the list appear in suicidal data and non-suicidal data. We then count (with the help of Python) the number of documents that are suicidal and non-suicidal, which have the above words in their document.

The words that appeared in suicidal documents were twice as many as the words appearing in non-suicidal documents. The words appeared 3080 times in suicidal documents and 1525 times in non-suicidal documents. However, the number of suicidal documents is about half of the number of non-suicidal documents that contain these words. There are 56 suicidal documents and 99 non-suicidal documents. This would average about 55 suicidal words per suicidal document and 15.4 suicidal words per non-suicidal document.

This would mean that each suicidal document in this topic has around 3.6 times more words related to self-image and physical appearance as compared to non-suicidal documents. If any document has a relatively high number of words from the above list of relevant frequent words (anything above 30 words), it could mean that the person is suicidal, and help needs to be given to the person.

# Demonstration

## 7.1 Web Application for Classification of Suicidal Language

The demonstration for the classification text mining task involved developing a web application designed to identify potential suicidal ideation within the written text. (Appendix 7.1A) Users can submit a sentence, which is then analyzed using the best model, Naive Bayes, after evaluation. The analysis prioritizes the display of relevant resources on the results page if the sentence is being predicted as suicide. (Appendix 7.1C) These resources include links to crisis hotlines, mental health websites, and other support services.

Using this web application to predict suicide sentences will aid crisis hotlines and mental health organizations in detecting at-risk individuals with suicidal thoughts early and providing necessary aid to them.

## 7.2 Topic Modelling Visualization for BERTopic

**a. Interactive Charts to show the distribution of the topics extracted (Appendix 7.2A)**

From the distribution of the topics, NGOs can identify suicidal topics from non-suicidal topics. Topics that have a large distribution indicate that it is an important topic that affects many people. Hence, NGOs can then run campaigns to address suicidal topics that have a large distribution in the pie chart.

**b. Word Cloud of each Topic (Appendix 7.2B)**

As the word cloud shows the frequent words of the topic, NGOs can concentrate on topics that are more suicidal in nature. NGOs can train their employees and volunteers to look out for people who use frequent words in suicidal topics, as these people could be highly suicidal and require help.

# Business analysis

With our project, the Singapore government and non-profit mental health organizations such as Samaritans of Singapore can be aware of the issues closely related to suicides. As shown in our analytics earlier, the topic “Self-Image and Physical Appearance” is a factor that affects potential suicides. From this, mental health welfare classes, which teach the importance of self-image and physical appearance, can be conducted. Our classification model can be used as an API for our aforementioned stakeholders to closely monitor social media texts to aid in the early identification of suicides.

# Discussions and Gap Analysis

## 9.1 Discussions

If we had access to more advanced techniques, we could probably train a more accurate model to determine suicides. For instance, context is hard to capture and could impact classification. For example, the text “I want to live” may seem like a non-suicide case at first, but it may also be interpreted as suicide as it may reflect a person’s negative mental state while wanting to live. The cost of false negatives (suicidal texts being predicted as non-suicidal) outweighs the cost of false positives (non-suicidal texts predicted as suicidal), and misclassification of a text can cost a life. Therefore, it would be great if we could train a super-accurate model with advanced techniques.

As suicide is a crucial pressing public health issue, we wanted to dive deeper and analyze the various topics related to suicide. However, we were limited by time and could only dive into the topic of Self-Image and Physical Appearance. If more time was given, we would definitely study the other topics in depth as well.

### 9.1.1 What worked well

Since we are all computing students, we all have backgrounds in Python, and it was easy for us to integrate it into our project. As mentioned above, we also used platforms such as DeepNote where possible, making it easier to share our work.

Visualization libraries are easy to use and provide the information we need to improve our models. For example, if an “intertopic distance map” shows the overlap between many topics, it tells us that the number of topics is too high.

The team worked well together with good communication and frequent Zoom calls to check in on the project’s progress.

### 9.1.2 What did not work well (Error Analysis)

As previously stated, we had limited computational power, which meant that we were unable to explore certain models.

The NMF model for topic modelling did not give coherent groups of words; hence, it was difficult to find suitable topic labels for most of the clusters. It was also very challenging to find ways to visualize using Intertopic Distance Map, as there were many incompatibilities between different versions of the pyLDAvis library and the sklearn library. We had to change the code from the library module itself to make them work.

For the classification demonstration, the web application for detecting suicide sentences, we encountered situations where the prediction of the sentences changes due to the change in grammar we use. For example: “I want to jump down the building” compared to “I am going to jump down the building.” Both sentences carry suicidal intent but are predicted differently through the model.

Another error during the classification process was, for example: “I want to die” compared to “I wan to die". Both sentences are the same. However, they are predicted differently with the lack of “t” in the “want.”

To overcome these errors in the classification process, more training data would be required to aid the accuracy of the prediction.

## 9.2 Gap Analysis

**Language Limitation:** The dataset we used only includes the English language, so our trained models are optimized for English-speaking populations only. This makes it difficult to use our models in multilingual contexts where people use other languages.

**Literacy Requirement:** Our models depend on the users' ability to read and write, which makes it challenging for illiterate or low-literacy groups to benefit from them. This limitation excludes a significant part of the world's population and limits our models' widespread applicability and impact.

**Unexplored Repeated Words:** Our current analysis could not delve deeper into the significance of sequentially repeated words within the dataset, a potential avenue for uncovering hidden patterns or meanings in the data. This oversight underscores the need for future research to explore these potential insights, which could be crucial for understanding nuances in communication or emphasizing certain aspects of the text.

**Limited Age Demographic:** The dataset does not offer a comprehensive representation of all age groups. It mainly focuses on Reddit users, who are usually between the ages of 18 and 49. Therefore, the models may need to capture the linguistic patterns and nuances present in the language use of individuals outside this age group. This could result in a reduction in our models' effectiveness and accuracy when applied to a broader population.

# Use of generative AI

Firstly, we used generative AI for ideas on different text-mining tasks that can be done for our dataset. Next, we used generative AI to help us explore and fine-tune the models. In addition, we also used it to generate prompts to test out the evaluation of models. Lastly, we also used it to generate ideas for the implementation of our demonstrations.

# Future Work and Conclusion

## 11.1 Future Work

**Broader Data Collection:** To improve the quality and representativeness of our models, we intend to incorporate data from a broader range of social media platforms and online forums. This will help capture a broader spectrum of linguistic expressions and ensure that our models are trained on a dataset that reflects diverse communication styles.

**Model Refinement:** We aim to refine our models for continuous improvement through ongoing research and incorporating feedback from users and mental health experts.

**Multilingual and Multi-Model Approach:** Recognizing the multicultural nature of online platforms, we can expand our project to support multiple languages. This will ensure that our tools are inclusive and capable of assisting a wider audience.

**Experimenting with Alternative Word Embedding Models:** Given the suboptimal performance of the Word2Vec model, we plan to experiment with other word embedding techniques, such as the Glove model. Exploring various embedding models will allow us to compare and identify the most effective approach for capturing semantic meaning and relationships within text. This comparative analysis will be crucial in selecting an embedding technique that best suits the nuances of our dataset and improves the overall performance of our predictive models.

## 11.2 Conclusion

Our group aimed to leverage text mining to identify signs of suicidal ideation on social media, with a focus on the Reddit application. Despite the challenges and limitations inherent in our models and dataset, we have established a foundational framework for understanding and addressing suicide through our analysis. Our classification and topic modelling models have shown promising results, and we have provided a platform that can be expanded upon by mental health professionals and support networks for more nuanced intervention strategies.

While we recognize that our work is only a step towards the larger goal of proactive suicide prevention, it is a step in the right direction. Through this project, we hope to contribute to the ongoing dialogue and efforts in mental health care. We are optimistic about the potential of machine learning and natural language processing to save lives and provide support to those in need. Moving Forward, we will work on advancing our models and methodologies to make a meaningful impact on suicide prevention in Singapore and beyond.

# Project Experiences/ Reflections

## 12.1 Law Wen Xin

Through this project, I’ve learnt to use different models for classification, some learnt in class, while some was outside of class. I’ve also learnt to implement hyperparameter tuning to the different models, but due to lack of computational power, I had to reduce the number of parameters to tune. Doing the web application was something new as I’d never used Flask before. The whole process of learning how to use Flask and implementing it (with the help of teammates) was a ride that brought satisfaction as the web application came to life. Overall, it was an enjoyable and fruitful experience to be able to work with my teammates on this project.

## 12.2 Jin Zi Long

After this project, I understood the value of good communication and teamwork. Instead of doing our own models, working together and exchanging ideas was an effective way to discover new methodologies quickly. I’ve learnt a lot from my peers about how to use machine learning models and how to convert it into web. I’ve also learnt how to solve compatibility issues with different versions of libraries, which required me to read and change the code of the Python module itself, building many invaluable soft skills like collaboration and tenacity.

## 12.3 Ong Swee Long

I learnt that what we learn in the labs and actual implementation differs based on our project needs. To do the project well, there are many resources and past work online which we can refer to and adapt them to meet our own project needs. I learnt that anyone could do text mining but to do it well, one must be open-minded to trying new text mining methods. I also need to keep trying without giving up, as models can take hours to train and hyper-tune, and the results could still be disappointing even after hours spent training the model. This project was a good experience, and I will continue using text mining for my future project.

## 12.4 Wong Dehou

Overall, the project aligned with the course objectives in teaching us the various concepts of text mining, from the start of gathering and cleaning data to performing the different types of analysis. I was able to apply the theories learned during class and put them into practice in this project. I was also very fortunate to work with my teammates, who are very knowledgeable and capable. I have learnt many things while working with them, and it was a fantastic experience.

## 12.5 Hsu Thitsar Lwin

During this project, I gained hands-on experience in applying machine learning models to real-world data, and I saw how tough it can be to clean up text data from social media posts. This project was challenging, but it taught me valuable lessons in NLP and text mining that I know will be useful later on. The rewarding moments came during our demonstrations when we saw how the outcomes came together after all the hard work we put in over the semester. I really enjoyed working with my team and regularly meeting together to talk about our progress and any issues we faced. It really showed me how good communication and teamwork can lead to great results.

# References

Channel News Asia. (2023, July 1). 476 suicides reported in Singapore in 2022; highest in more than 20 years. <https://www.channelnewsasia.com/singapore/suicides-reported-singapore-476-2022-highest-more-20-years-3597791>

Liu, X., Liu, X., Sun, J., Yu, N. X., Sun, B., Li, Q., & Zhu, T. (2019). Proactive suicide prevention online (PSPO): Machine identification and crisis management for Chinese social media users with suicidal thoughts and behaviors. <https://www.jmir.org/2019/5/e11705/>

Ministry of Health, Singapore. (2023, October 5). Launch of National Mental Health and Well-being Strategy. <https://www.moh.gov.sg/news-highlights/details/launch-of-national-mental-health-and-well-being-strategy>

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of Depression-Related Posts in Reddit Social Media Forum. <https://www.researchgate.net/publication/332215254_Detection_of_Depression-Related_Posts_in_Reddit_Social_Media_Forum>

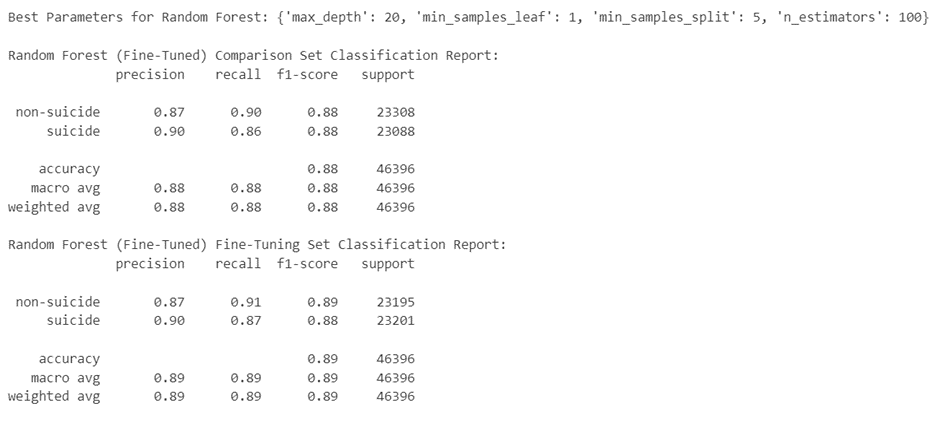
World Health Organization. (2023, Aug 28). *Suicide.* <https://www.who.int/news-room/fact-sheets/detail/suicide>

Teixeira, A.S., Talaga, S., Swanson, T.J. *et al.* Revealing semantic and emotional structure of suicide notes with cognitive network science. *Sci Rep* **11**, 19423 (2021). <https://doi.org/10.1038/s41598-021-98147-w>

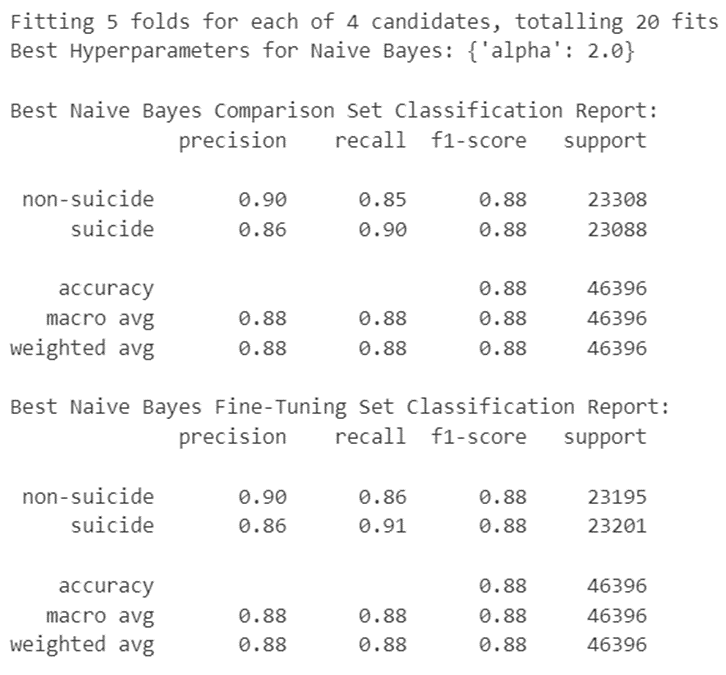
# Appendix

6.1.1 Comparison of Fine-tuned Classification Models

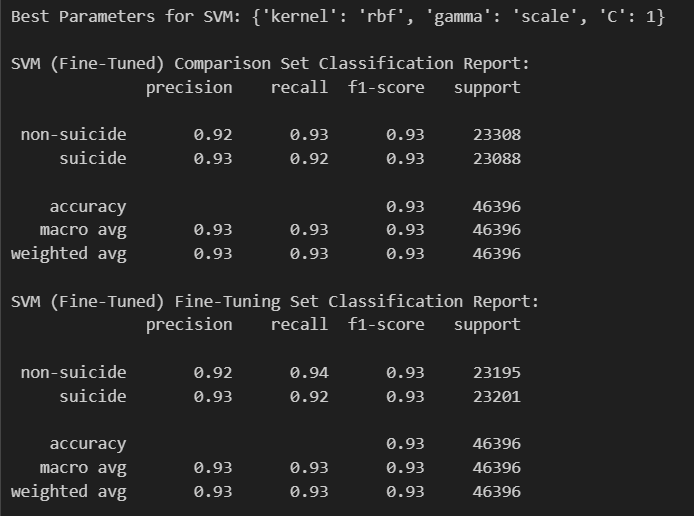
Random Forest



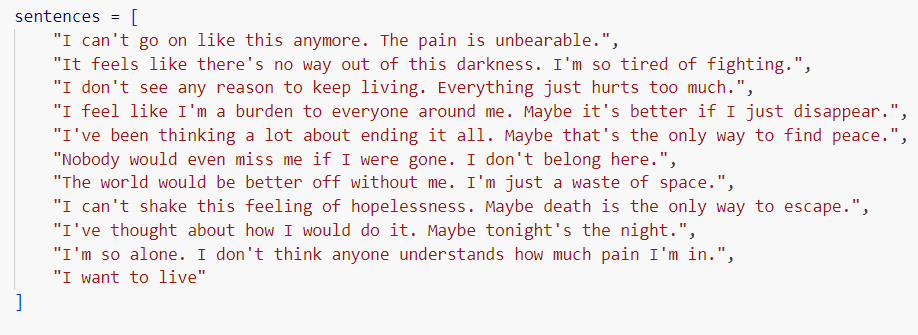
Naive Bayes

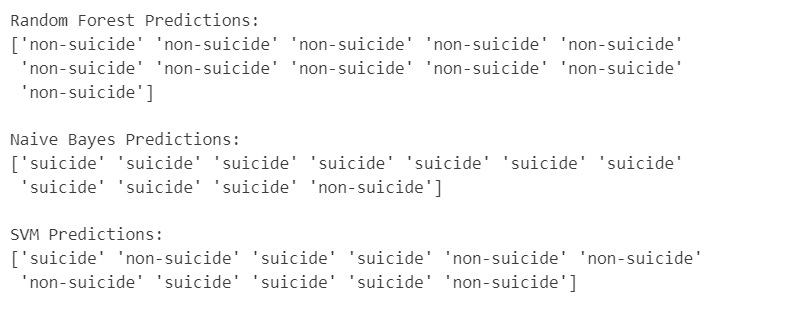


SVM



Appendix 6.1.2 Prompts & Results





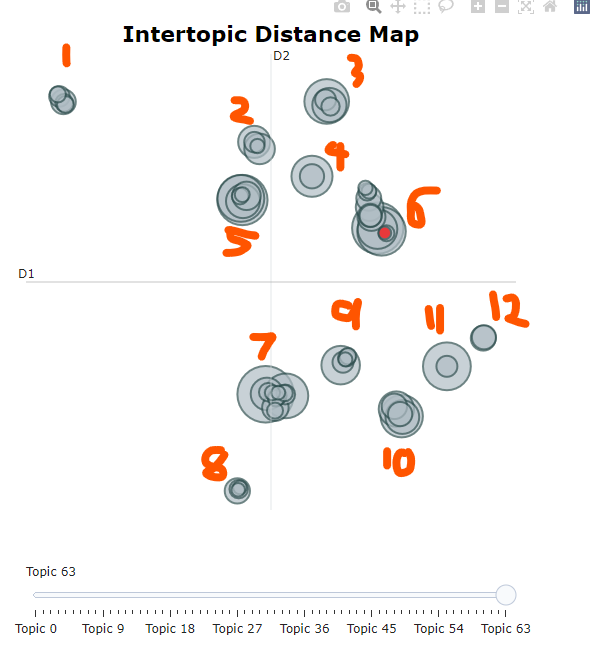
Appendix 6.2.2A - NMF Coherence Scores, Frequent Words and Intertopic Distance Map for raw data

|  |  |
| --- | --- |
| 10K rows | 100K rows |
|  |  |
|  |  |

Appendix 6.2.2B - NMF Coherence Scores, Frequent Words and Intertopic Distance Map for cleaned data

|  |  |
| --- | --- |
| 10K rows | 100K rows |
|  |  |
|  |  |
|  |  |

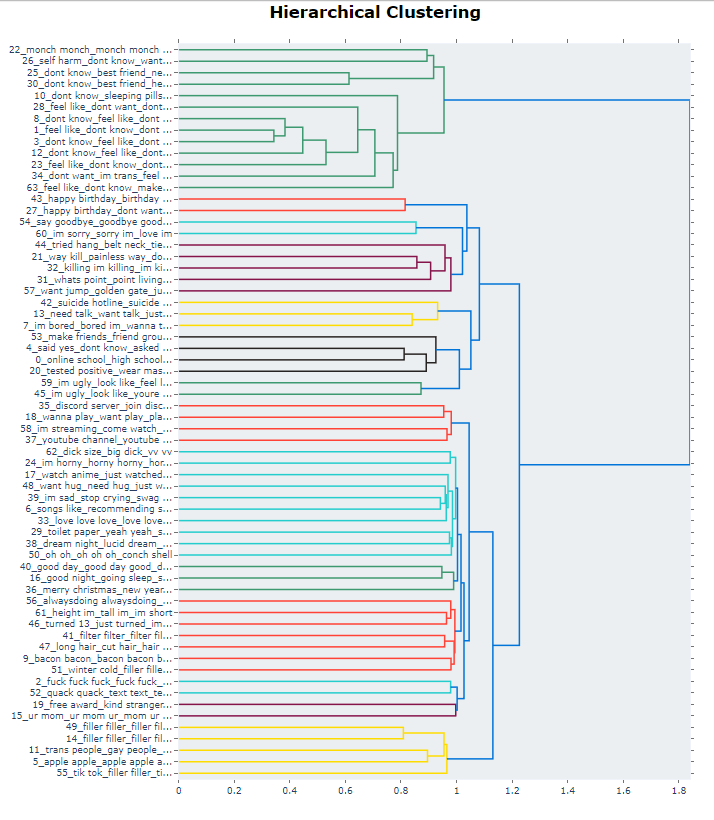
Appendix 6.2.3A - BERTopic Intertopic Distance Map



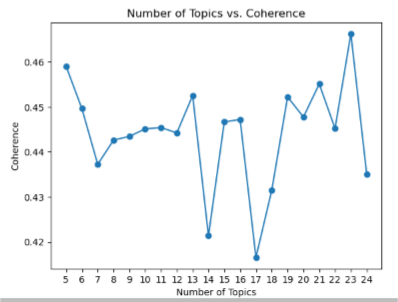
Appendix 6.2.3B - BERTopic All the 12 topics with frequent words

|  |  |  |
| --- | --- | --- |
|  | Topics in Cluster | Frequent Words from the topics |
| Cluster 1  Celebratory Events and Personal Reflections | 27,36,43,46,51 | ['happy birthday', 'dont want', 'im going', 'feel like', 'dont know', 'birthday im', 'just want', 'birthday coming', 'im just', 'dont think']  ['merry christmas', 'new year', 'happy new', 'christmas merry', 'christmas merry christmas', 'merry christmas merry', 'happy new year', 'money fuck money', 'fuck money fuck', 'money fuck']  ['happy birthday', 'birthday today', 'today birthday', 'birthday birthday', 'happy birthday birthday', 'birthday happy', 'birthday happy birthday']  ['turned 13', 'just turned', 'turning 20', 'turn 20', 'teenage years', 'turned 20', 'turning 18', 'im turning 20', 'just turned 13']  ['winter cold', 'cold winter', 'winter cold winter', 'cold winter cold', 'raining outside', 'cold weather'] |
| Cluster 2  Media and Entertainment Engagement | 17,18,37,58 | ['watch anime', 'just watched', 'ive watched', 'iron man', 'captain america', 'just finished', 'watching anime', 'im watching', 'attack titan', 'tony stark']  ['wanna play', 'want play', 'play minecraft', 'does nintendontgenesis does', 'nintendontgenesis does', 'does nintendontgenesis', 'nintendontgenesis does nintendontgenesis', 'people play', 'im bored', 'game play']  ['youtube channel', 'youtube video', 'montage parodies', 'yt channel', 'started youtube', 'started youtube channel', 'youtube videos', 'check video', 'million subscribers', 'unus annus']  ['im streaming', 'come watch', 'im live', 'live twitch', 'streaming twitch', 'come join', 'come hang', 'twitch come', 'im live twitch', 'streaming minecraft'] |
| Cluster 3  Social Interaction and Companionship Seeking | 7,13,35,42 | ['im bored', 'bored im', 'wanna talk', 'im bored im', 'wanna chat', 'bored im bored', 'chat im', 'want talk', 'talk im', 'want chat']  ['need talk', 'want talk', 'just need', 'just need talk', 'just want talk', 'talk im', 'just want', 'really need', 'really need talk', 'just talk']  ['discord server', 'join discord', 'join discord server', 'wanna join', 'dm link', 'ping ping', 'ping ping ping', 'want join', 'play games', 'memes play']  ['suicide hotline', 'suicide prevention', 'suicide hotlines', 'called suicide', 'national suicide', 'dont know', 'dont want', 'called suicide hotline', 'need talk', 'calling suicide'] |
| Cluster 4  Miscellaneous Personal Topics | 9,29 | ['bacon bacon', 'nowdrink water', 'water nowdrink', 'water nowdrink water', 'nowdrink water nowdrink']  ['toilet paper', 'yeah yeah', 'smell like', 'piss piss', 'yeah yeah yeah', 'smells like', 'coordinate piss', 'need pee', 'taking shit', 'flex tape'] |
| Cluster 5  Social Media Engagement and Online Behavior | 2,5,14,19,49,55 | ['fuck fuck', 'cheese cheese', 'cecil cecil', 'jake paulfuck']  ['apple apple', 'text post', 'post weekend', 'post images', 'got banned', 'text post weekend', 'social media']  ['free award', 'kind stranger', 'kind stranger edit', 'stranger edit', 'stranger edit thanks', 'edit thanks', 'thanks gold kind', 'gold kind', 'thanks gold', 'gold kind stranger']  ['draw im', 'art class', 'want draw', 'tell draw', 'know draw', 'ill draw', 'draw ill', 'digital art']  ['tik tok', 'filler filler', 'tik tok bad', 'tok bad', 'hate tiktok', 'tiktok bad', 'trash trash', 'like tiktok', 'trash trash trash'] |
| Cluster 6  Mental Health Challenges and Coping Mechanisms | 1,3,10,20,21,26,31,32,44,53,56,57,63 | ['feel like', 'dont know', 'dont want', 'just want', 'im just', 'suicidal thoughts', 'want die', 'like im', 'im tired', 'im going']  ['dont know', 'feel like', 'dont want', 'high school', 'im going', 'im just', 'just want', 'just dont', 'like im', 'dont think']  ['dont know', 'sleeping pills', 'just want', 'just took', 'dont want', 'im going', 'want die', 'im just', 'im drunk', 'feel like']  ['tested positive', 'wear mask', 'got covid', 'social distancing', 'wear masks', 'dont want', 'positive covid', 'corona virus', 'covid 19', 'dont know']  ['way kill', 'painless way', 'dont want', 'buy gun', 'im going', 'pull trigger', 'best way', 'just want', 'dont know', 'way die']  ['self harm', 'dont know', 'want cut', 'just cut', 'feel like', 'dont want', 'cut deep', 'just want', 'started cutting', 'self harming']  ['whats point', 'point living', 'whats point living', 'reason live', 'point life', 'life worth', 'worth living', 'dont want', 'meaning life', 'dont point']  ['killing im killing', 'im killing im', 'killing im', 'im killing', 'want kill', 'want die', 'kill want kill', 'want kill want', 'kill want', 'dont want']  ['tried hang', 'belt neck', 'tie noose', 'dont know', 'dont want', 'rope im', 'rope neck', 'partial suspension', 'buy rope', 'im going']  ['make friends', 'friend group', 'talk people', 'friends like', 'new friends', 'feel like', 'new school', 'online friends', 'friends people', 'make new friends']  ['alwaysdoing alwaysdoing', 'alwaysdoing alwaysdoing alwaysdoing', 'lose weight', 'push ups', 'doing pushups', 'losing weight', 'gain weight', 'weight loss', 'pushups decide', 'doing pushups decide']  ['want jump', 'golden gate', 'jump bridge', 'gate bridge', 'golden gate bridge', 'going jump', 'tall building', 'just jump', 'jumping bridge', 'thinking jumping']  ['feel like', 'dont know', 'make friends', 'dont want', 'like im', 'high school', 'im just', 'social anxiety', 'dont think', 'just want'] |
| Cluster 7  Daily Life and Casual Conversations | 0,6,16,22,38,39,40,48,50,52,54,60 | ['online school', 'high school', 'online classes', 'online class', 'dont know', 'wish luck', 'im gonna', 'teacher just']  ['songs like', 'recommending songs', 'yeah yeah', 'wilbur villain', 'villain wilbur', 'villain wilbur villain', 'wilbur villain wilbur', 'favorite song', 'na na', 'music taste']  ['good night', 'going sleep', 'sleep schedule', 'going bed', 'fall asleep', 'good morning', 'hours sleep', 'im tired', 'im going', 'im going sleep']  ['monch monch', 'dont know', 'title title', 'title title title', 'dont want', 'meow meow', 'dog just', 'dog died', 'cat just']  ['dream night', 'lucid dream', 'sleep paralysis', 'weird dream', 'dream dream', 'wet dream', 'just dream', 'night dream', 'felt like', 'looked like']  ['im sad', 'stop crying', 'swag swag', 'feel sad', 'sad sad', 'like crying', 'im crying', 'feel like', 'really sad']  ['good day', 'good day good', 'day good day', 'day good', 'great day', 'doing today', 'day going', 'hows day', 'guys doing', 'hows doing']  ['want hug', 'need hug', 'just want', 'just want hug', 'free hugs', 'hug im', 'virtual hug', 'hug just', 'hug hug', 'wanna cuddle']  ['oh oh', 'oh oh oh', 'conch shell', 'magic conch shell', 'magic conch', 'wrote poem', 'poem wrote', 'lie lie', 'feel like monster', 'feel like']  ['quack quack', 'text text', 'fun fact', 'fuck mosquitoes', 'fishing rod', 'hawk moth']  ['say goodbye', 'goodbye goodbye', 'im sorry', 'want say', 'goodbye people', 'goodbye im', 'wanted say', 'wanted say goodbye', 'just wanted say', 'good bye']  ['im sorry', 'sorry im', 'love im', 'im sorry im', 'love im sorry', 'end life courage', 'life courage end', 'life courage', 'courage end life', 'im going'] |
| Cluster 8  Sexuality and Gender Discussions | 24,41,61,62 | ['im horny', 'horny horny', 'horny horny horny', 'horny jail', 'horny post', 'horny im', 'horny just', 'horny like', 'horny people', 'want horny']  ['boys skirts', 'want wear', 'im wearing']  ['height im', 'tall im', 'im short', 'filler filler', 'short im', 'year old', 'average height', 'tall people', '69 inches']  ['dick size', 'big dick', 'vv vv', 'vv vv vv', 'big boobs', 'small dick', 'boobs boobs', 'having big', 'thigh pics', 'penis size'] |
| Cluster 9  Self-Image and Physical Appearance | 11,34,45,47,59 | ['trans people', 'gay people', 'super straight'', 'mods gay', 'im gay', 'reddit mods gay', 'dont care', 'reddit mods']  ['dont want', 'im trans', 'feel like', 'dont know', 'just want', 'im going', 'im just', 'gender dysphoria', 'im transgender', 'dont think']  ['im ugly', 'look like', 'youre ugly', 'good looking', 'ugly people', 'ugly just', 'called ugly', 'im pretty', 'fucking ugly', 'self esteem']  ['long hair', 'cut hair', 'hair like', 'ass hair', 'facial hair', 'hair cut', 'hair im', 'ass haircut', 'yee yee', 'look like']  ['im ugly', 'look like', 'feel like', 'im just', 'look mirror', 'dont know', 'leave house', 'good looking', 'dont want', 'plastic surgery'] |
| Cluster 10  Life Uncertainties and Relationship Dynamics | 8,12,15,23,28 | ['dont know', 'feel like', 'dont want', 'just want', 'im just', 'years old', 'high school', 'im going', 'like im', 'just dont']  ['dont know', 'feel like', 'dont want', 'just want', 'best friend', 'im just', 'im going', 'like im', 'just dont', 'months ago']  ['ur mom', 'ur mom ur', 'mom ur mom', 'mom ur', 'dont know', 'year old', 'mom just', 'living room', 'dad just', 'mom mom']  ['feel like', 'dont know', 'dont want', 'just want', 'im just', 'dont think', 'im going', 'best friend', 'want die', 'just dont']  ['feel like', 'dont want', 'dont know', 'year old', 'just want', 'im just', 'just dont', 'im going', 'years ago', 'dont think'] |
| Cluster 11  Romantic Interests and Relationships | 4,33 | ['said yes', 'dont know', 'asked crush', 'girl like', 'crush said', 'theres girl', 'best friend', 'ask crush', 'really like', 'like girl']  ['love love love', 'love love', 'got girlfriend', 'want gf', 'got gf', 'girlfriend im', 'want girlfriend', 'guy knows guy'] |
| Cluster 12  Support for Friends in Crisis | 25,30 | ['dont know', 'best friend', 'need help', 'commit suicide', 'know help', 'dont want', 'help friend', 'shes going', 'wants kill', 'doesnt want']  ['dont know', 'best friend', 'help friend', 'need help', 'commit suicide', 'doesnt want', 'hes going', 'want help', 'dont want', 'know hes'] |

Appendix 6.2.3C - BERTopic Hierarchical Clustering



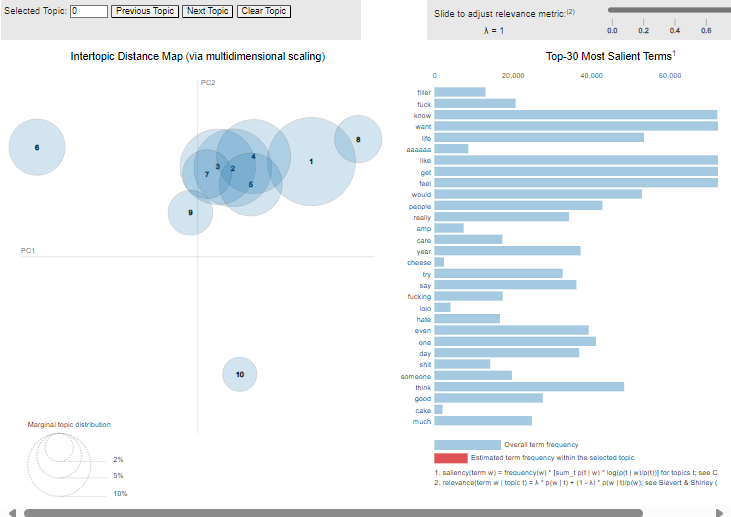
Appendix 6.2.4A - Hypertuning LDA Unigram to get higher coherence score



Appendix 6.2.4B - LDA coherence score table

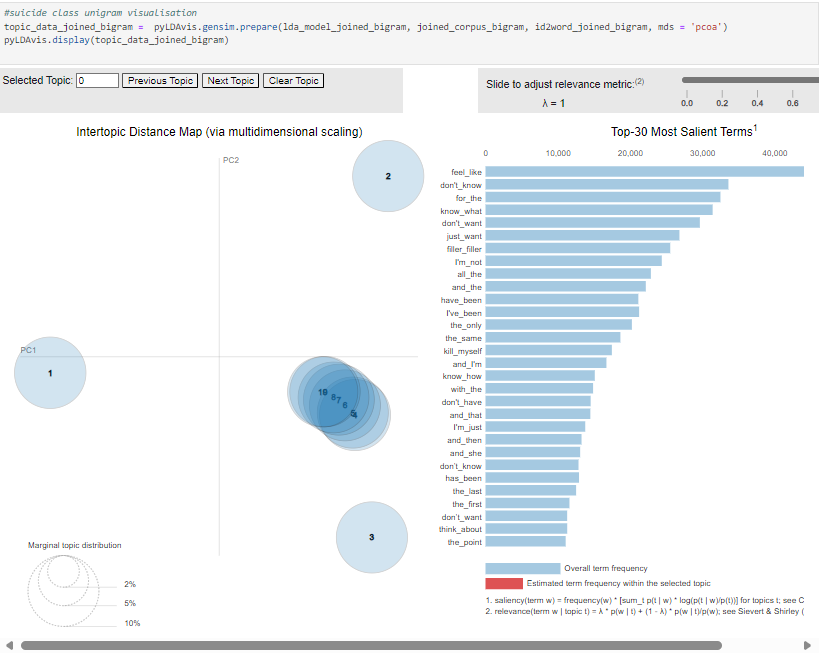
|  |  |
| --- | --- |
|  | Coherence Score |
| Unigram model | 0.3778 (10 topics) |
| Bigram model | 0.7668 (10 topics) |
| Trigram model | 0.6950 (10 topics) |
| Bigrams with Trigrams | 0.7263 (10 topics) |

Appendix 6.2.4C - Unigram intertopic distance map with topics extracted



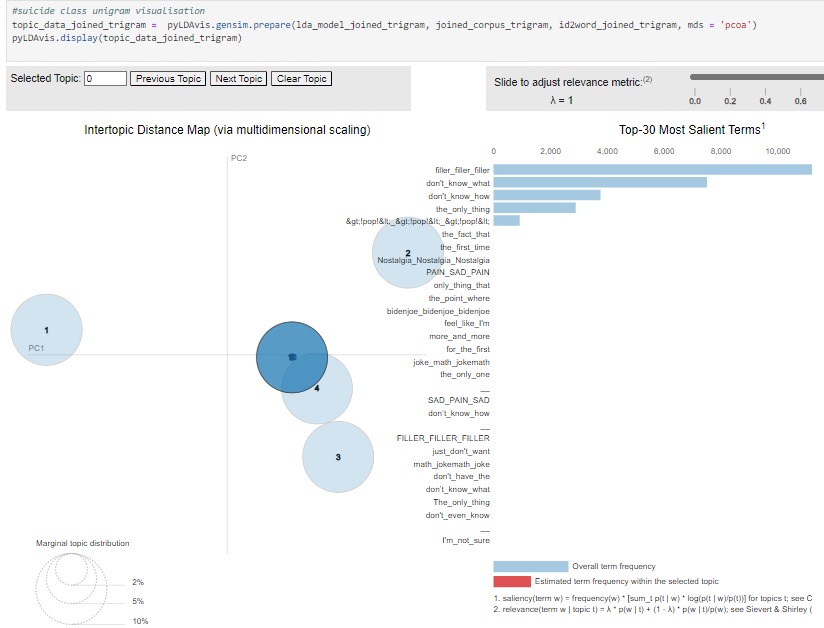
|  |  |  |
| --- | --- | --- |
|  | Topics in Cluster | Frequent Words from the topics |
| Cluster 1  Feel bad about life | 1 | get, like, make, feel, say, want, one, really, fuck, life |
| Cluster 2  Friend and people | 2,3,4,5,7 | want, get, feel, like, people, know, life, think, well, make, day, would, say, one, year, friend, good, fuck, amp, paul |
| Cluster 3  Actions | 6 | know, people, want, get, would, like, say, think, make, friend |
| Cluster 4  Others | 8 | like, know, want, life, would, get, feel, think, people, even |
| Cluster 5  Swearing life | 9 | filler, fuck, like, get, feel, life, want, think, year, know |
| Cluster 6  Feeling | 10 | feel, like, want, get, know, even, life, make, think, would |

Appendix 6.2.4D - Bigram intertopic distance map with topics extracted



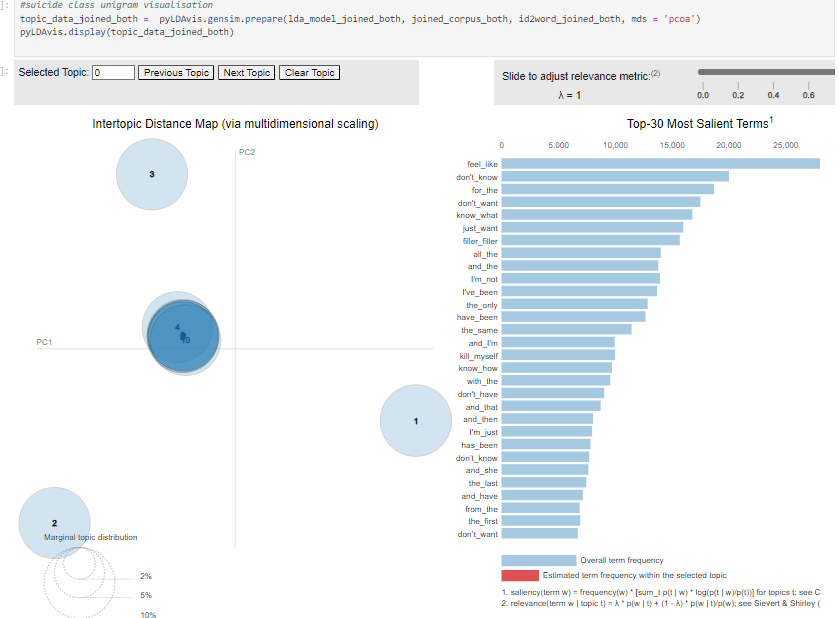
|  |  |  |
| --- | --- | --- |
|  | Topics in Cluster | Frequent Words from the topics |
| Cluster 1  Suicidal thoughts | 1 | I'm not, kill myself, the last, she was, that she, the world, but the, you can, you are, but they |
| Cluster 2  People and feelings | 2 | for the, I'm just, the point, and now, and they, life and, the people, away from, just feel, right now |
| Cluster 3  Others | 3 | just want, have been, and she, the time, know that, I'm going, the best, i'm not, The only, and I'm |
| Cluster 4  Others | 4,5,6,7,8,9,10 | feel like, I've been, the only, know how, with the, and have, and not, friends and, they are, the other, know what, the end, don't think, the way, kill myself, year old, school and, her and, I've tried, get the, don't know, think about, don't want, was the, but I'm, myself and, don't even, with her, don't have, care about, that I'm, and just, people who, for me, about how, the most, about the, felt like, more than, deal with, you have, out and, the next, and it's, that the, talk about, was going, me, and, and the, and that, and then, from the, have the, the past, how much, just don't, feels like, right now, all the, the same, and I'm, has been, the first, thinking about, only thing, too much, have any |

Appendix 6.2.4E - Trigram intertopic distance map with topics extracted



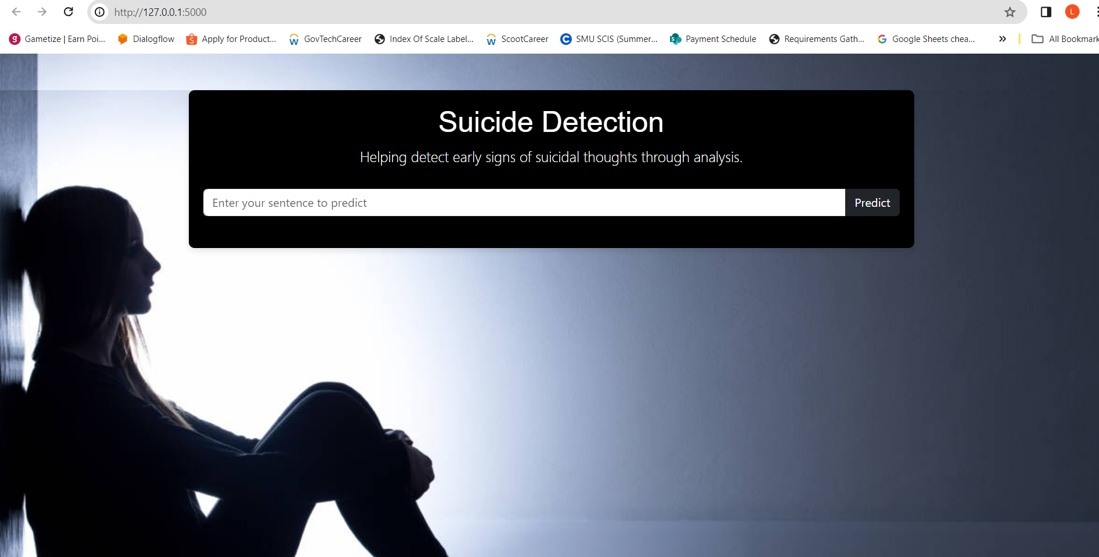
|  |  |  |
| --- | --- | --- |
|  | Topics in Cluster | Frequent Words from the topics |
| Cluster 1  Pain and Sadness | 1 | The only thing, Nostalgia Nostalgia Nostalgia, don't know what, PAIN SAD PAIN, SAD PAIN SAD, the first time, the fact that, don't know how |
| Cluster 2  Pain and Sadness | 2 | Nostalgia Nostalgia Nostalgia, PAIN SAD PAIN, don't know what, SAD PAIN SAD, the only thing, the point where, the first time, bidenjoe bidenjoe bidenjoe |
| Cluster 3 Pain and Sadness | 3 | Nostalgia Nostalgia Nostalgia, don't know what, SAD PAIN SAD, PAIN SAD PAIN, the only thing, the first time, the fact that, don't know how |
| Cluster 4 Pain and Sadness | 4 | Nostalgia Nostalgia Nostalgia, don't know what, SAD PAIN SAD, PAIN SAD PAIN, the only thing, don't know how, the first time, the point where |
| Cluster 5 Pain and Sadness | 5,6,7,8,9,10 | filler filler filler, >!pop!<*>!pop!<*>!pop!<, Nostalgia Nostalgia Nostalgia, the fact that, PAIN SAD PAIN, don't know what, SAD PAIN SAD, don't know how, the only thing, bidenjoe bidenjoe bidenjoe, the first time, SAD PAIN SAD, only thing that, the point where, Nostalgia Nostalgia Nostalgia, feel like I'm |

Appendix 6.2.4F - Bigtams + Trigram intertopic distance map with topics extracted

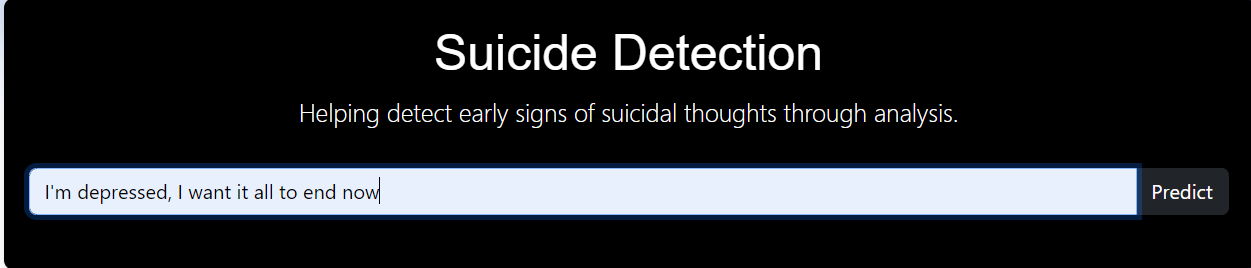


|  |  |  |
| --- | --- | --- |
|  | Topics in Cluster | Frequent Words from the topics |
| Cluster 1  Uncertainty and time | 1 | for the, don't want, know what, and that, I'm just, and she, think about, the time, and now, that she |
| Cluster 2  About life | 2 | filler filler, all the, don't have, and have, the world, deal with, just don't, don't even, friends and, you have |
| Cluster 3  Suicidal thoughts | 3 | feel like, I'm not, kill myself, and then, don't want, people who, too much, The only, about the, even though |
| Cluster 4  Reflecting | 4,5,6,7,8,9,10 | I've been, the same, know that, the past, and just, the end, the people, the next, now and, talk about, care about, don't think, and not, out and, you are, into the, not even, with this, me, and, the last, I'm going, life and, for me., only thing, the other, and all, with her, was going, will never, don't know, and I'm, the first, the point, and they, more than, the best, how much, right now, know how, with the, from the, that I'm, about how, have any, just feel, her and, not sure, just want, the only, have been, has been, have the, thinking about, was the, myself and, over the, she was, but I'm, feels like, fact that, year old, time and, don't have, with me., depression and, it's not |

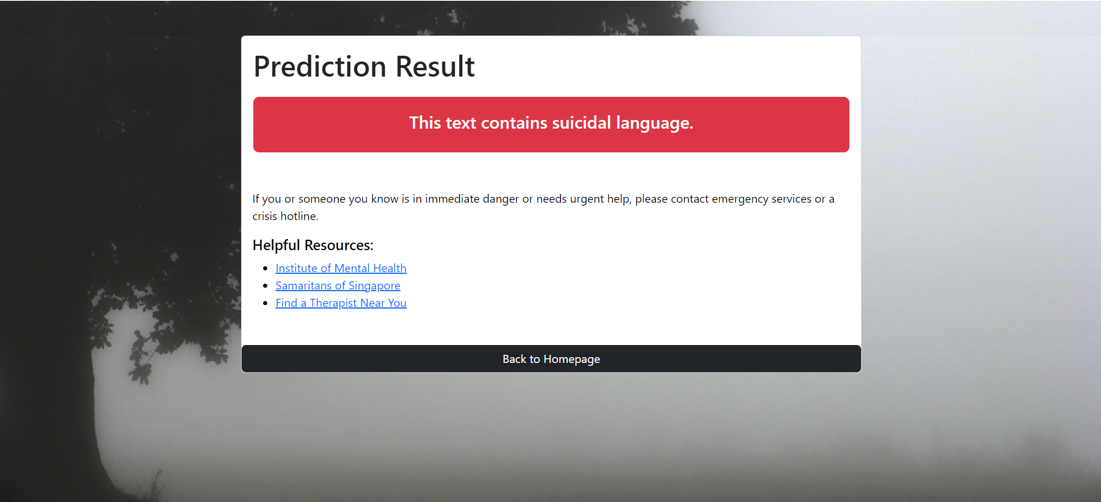
Appendix 7.1A - Web Application Homepage



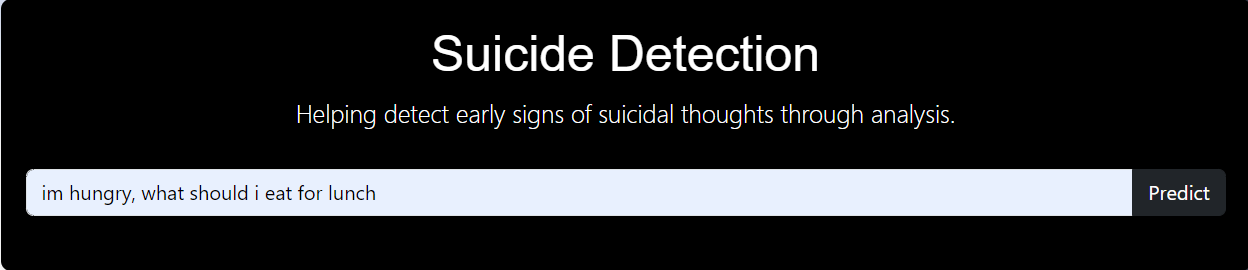
Appendix 7.1B - Inputting suicide sentence



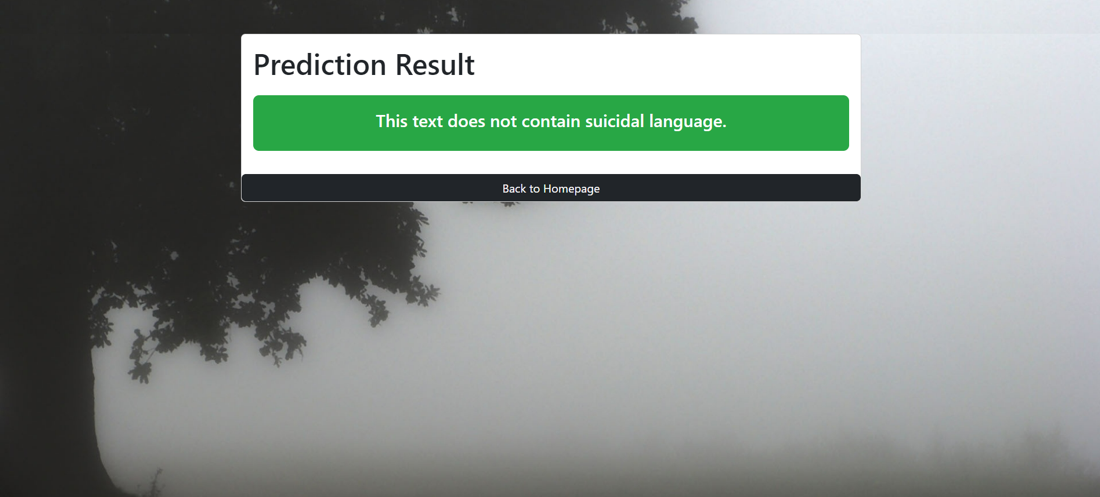
Appendix 7.1C - Result page for suicide prediction



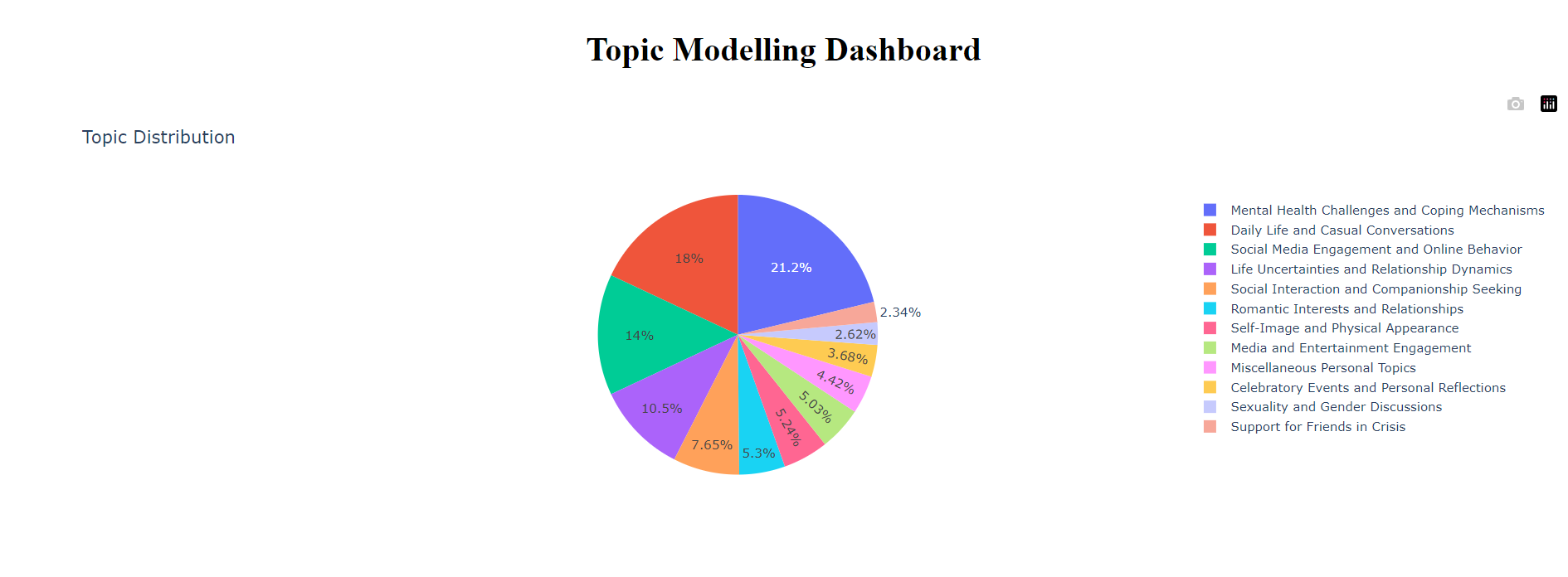
Appendix 7.1D - Inputting non-suicide sentence

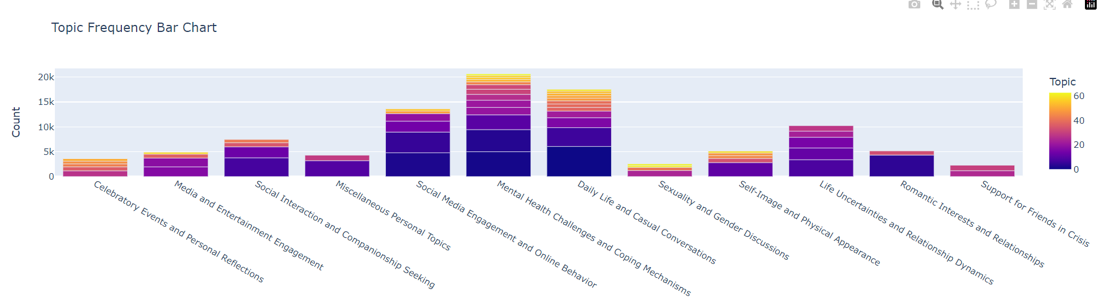


Appendix 7.1E - Result page for non-suicide prediction



Appendix 7.2A - BERTopic topic distribution Pie Chart





Appendix 7.2B - Word Cloud for BERTopics’ topics

