**Predicting Suicide Tendency through NLP:**

**A Text Analytics Approach**

Xinwen Wu and Xiaoxuan Gu

Marshall School of Business, University of Southern California

DSO 560: Text Analytics and Natural Language Processing

Professor Lorena Martin

Dec 2, 2023

# 

# **Abstract**

This study employs Natural Language Processing (NLP) techniques to discern suicide-related posts within online textual data. Utilizing sentiment analysis, semantic analysis, and TF-IDF vectorization, it aims to identify nuanced patterns indicative of suicidal tendencies. Leveraging data from Reddit's "SuicideWatch" and "r/teenagers" subreddits, it undergoes preprocessing, sentiment and semantic analyses, and TF-IDF vectorization. It then trains models (Naive Bayes, Random Forest, XGBoost, Neural Network) to predict suicidal tendencies, achieving accuracy scores ranging from 79.3% to 88.8%. The feature importance analysis reveals critical terms in identifying suicide-related content. The study fills a gap in literature by employing comprehensive NLP approaches and diverse models for enhanced predictive performance. While facing challenges like imbalanced data and computational constraints, the findings advocate for the use of NLP in mental health monitoring and suggest proactive intervention strategies to aid individuals exhibiting signs of distress.

*Keywords*: suicidality, natural language processing, detection

# 

# **Introduction**

## Background

Mental health concerns, especially suicidal tendencies among adolescents in low-and middle-income countries, are substantial (Uddin et al., 2019). The internet has revolutionized monitoring methods, with online networking posts offering a valuable resource for emotional status detection. Analyzing these posts allows for the identification of potential suicidality, utilizing technology and artificial intelligence. This intersection of mental health and online platforms provides an opportunity for proactive intervention and support, contributing to a safer online community.

## Objective

The primary objective of this project is to leverage Natural Language Processing (NLP) and text analytics techniques to identify potential suicide-related posts within online text data. This involves employing advanced NLP techniques and machine learning methods to locate specific keywords, patterns, and linguistic markers indicative of suicidal tendencies within textual content available on online platforms.

This endeavor includes utilizing sentiment polarity and subjectivity scores to discern nuanced differences between suicide-related and non-suicidal posts, alongside the application of semantic analysis techniques to uncover underlying topics within the textual content. Additionally, this study aims to employ TF-IDF vectorization to preprocess and convert textual data into numerical features. Subsequently, by leveraging these transformed features, the objective is to train predictive models that accurately classify and predict suicide-related tendencies present within the text-based dataset. In general, this project aligns with the overarching goals of NLP and text analytics, aiming to contribute to the field by addressing a critical societal issue through computational analysis of text data.

## Literature Review

Studies have found that online social behavior is related with depression and suicidality. Youth with mental health issues tend to involve themselves more on social networking websites and seek help from the internet (Memon et al., 2018). Gunn and Lester also noticed twitter posts were related to suicide in 24 hours through LIWC methods (Gunn & Lester, 2015). O’Dea also discovered that twitter was used by individuals to express suicidality (O'Dea et al., 2015).

Certain studies and methods succeeded to utilize public networking data to distinguish suicidal posts as well. One study successfully applied statistical learning methods to distinguish the tendency of suicidality with twitter data (O'Dea et al., 2015). Jashinsky examined 1.7 million tweets data and proposed that Twitter may be a viable tool for real-time monitoring of suicide risk factors, and individuals who are at risk for suicide may be detected through social media (Jashinsky et al., 2014). Computerized sentiment analysis and data mining was useful to identify users at risk of suicide (Christensen et al., 2014). Nadeem utilized logistic regression to predict depression with twitter posts data and reached a f1 score of 84% (Nadeem, 2016).

Previous studies utilizing surveys and psychological experiments for suicidality prediction seem to have relatively low predictive power. Machine learning models with psychological indicators reached an F1 score of 0.60 (Wen et al., 2023).

## Study Objectives: Addressing Literature Gaps

This study expects to discover patterns related to suicidality, including keywords, pattern, polarity and subjectivity. By implementing multiple natural language processing methodologies, the goal is to reach higher predictive power than previous studies.

## Hypotheses

Null Hypothesis (H0): Suicide-related texts do not exhibit significantly more negative sentiments or lower subjectivity compared to non-suicide texts. Additionally, advanced NLP models cannot accurately classify or predict suicidal tendencies based on textual data.

Alternative Hypothesis (H1): Suicide-related texts demonstrate a significantly higher prevalence of negative sentiments and lower subjectivity in comparison to non-suicide texts. Moreover, utilizing advanced NLP models enables accurate classification and prediction of suicidal tendencies from text data.

# **Methods**

## Source of Text Data

For this project, the text data was sourced from Kaggle, specifically from a dataset named "Suicide Detection" available in CSV format. The dataset originated from posts within the "SuicideWatch" subreddits on Reddit. These posts were collected using the Pushshift API. The dataset contains posts made to "SuicideWatch" from December 16, 2008 until January 2, 2021. All posts gathered from "SuicideWatch" were labeled as suicide related. Non-suicide posts were sourced from the r/teenagers subreddit. This subreddit is one of the largest communities on Reddit, created by and for teenagers.

## Data Collection Methods

The original dataset consists of three columns: index, text (representing the posts), and class (indicating the label, either suicide or non-suicide), with a total of 232,074 rows. For the purposes of this study, a sample of 5,000 cases was randomly selected from the dataset. Initial exploratory data analysis involved creating graphical representations to count labels and depict the distribution of word counts in both suicide and non-suicide posts.

**Text Preprocessing**

Following the initial data exploration, numerical labels were assigned to the dataset, with non-suicide posts marked as ‘0’ and suicide-related posts denoted as ‘1’. This differentiation was crucial for subsequent analysis, allowing for the classification and prediction of suicidal tendencies based on the text data.

Following labeling, the text underwent several essential preprocessing steps. This included tokenization, where the text was segmented into smaller units or tokens, often words, for further analysis. Additionally, common words known as stopwords were removed to focus on more meaningful content. Lemmatization, another step, aimed to normalize words to their base or root form, ensuring consistency in the data for accurate analysis. The text was also cleaned by removing Unicode characters, special symbols, and non-alphanumeric characters, ensuring data uniformity and suitability for NLP tasks. These preprocessing steps collectively refined the text data, preparing it for sentiment analysis, topic modeling, and predictive modeling to identify potential suicidal tendencies within the dataset.

**Sentiment Analysis and Semantic Analysis**

Following the preprocessing phase, sentiment analysis was conducted to derive sentiment polarity and subjectivity scores for each post. These scores were further analyzed and compared between suicide and non-suicide labeled posts to understand the overall attitude and sentiments expressed. Statistical tests, such as z-tests, were employed to evaluate the significance of differences observed, including the subsequent assessment of corresponding p-values.

Moreover, semantic analysis using LDA was applied to uncover latent topics within the text content. The LDA model generated two distinct topics along with their associated weights, providing insights into underlying themes within the textual data.

**Utilizing TF-IDF/PCA/LSA and Model Evaluation**

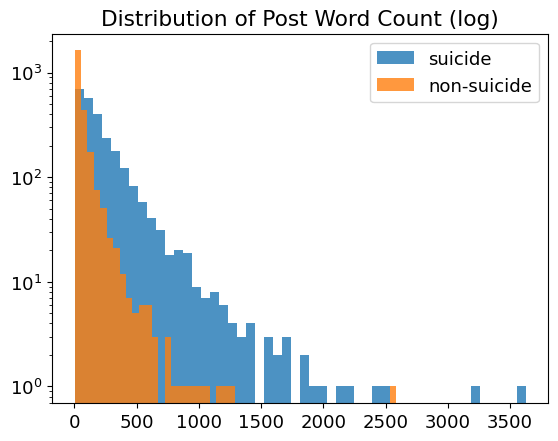
TF-IDF vectorization was employed to convert the preprocessed textual data into numerical features, which were then augmented with sentiment scores. Following this, we proceeded with the Latent Semantic Analysis (LSA) on the transformed data to unveil latent semantic structures. Additionally, Principal Component Analysis (PCA) was utilized for dimensionality reduction, focusing on the weights of the initial 3 Principal Components (PCs). Subsequently, the transformed data, along with polarity and subjectivity scores, underwent model training and testing using various classifiers such as Multinomial Naive Bayes, Random Forest, XGBoost, and Neural Network. Each model's performance was evaluated using metrics like accuracy, classification report (providing precision, recall, F1-score, and support statistics), and confusion matrices to understand model behavior and predictive capabilities.

Additionally, the feature importance metrics from the XGBoost model was utilized to identify significant features contributing to the model's predictions. The features with importance scores greater than zero were examined to determine their relevance in the classification task.

**Results**

## Exploratory Data Analysis

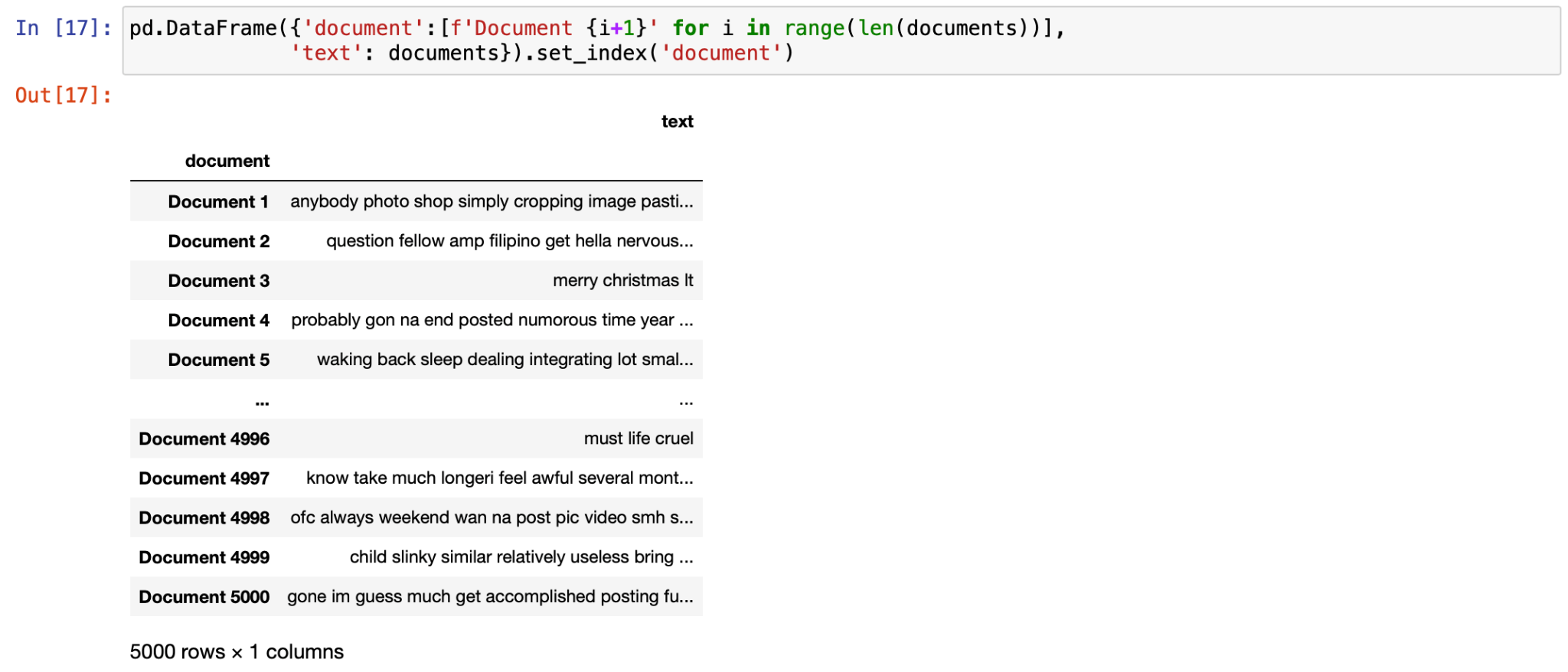
During the data exploration phase, the occurrences of suicide and non-suicide labels were counted, revealing an equitable distribution between these categories. Moreover, the analysis delved into the word count within each post for both labels. The subsequent visualization vividly presented the distribution of word counts, notably highlighting that posts associated with suicide tended to encompass a higher word count compared to non-suicidal posts.

A graph of suicide and non-suicide count

Description automatically generated

## Text Preprocessing Result

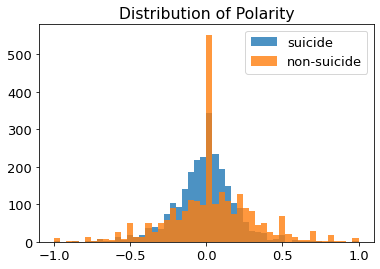
Following the initial exploration of the data, a series of text preprocessing steps were conducted. The screenshot below displays the transformed text post preprocessing, encompassing tokenization, removal of stopwords, and lemmatization. The words were processed, rejoined, and converted into lists—maintaining the original post structure.



## Sentiment Analysis Result

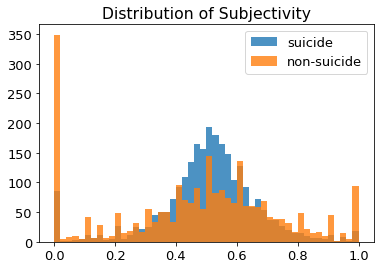
The preprocessed text was then analyzed for sentiment. The sentiment analysis unveiled notable differences between suicide and non-suicide posts. As shown below, the sentiment polarity revealed a substantial contrast, indicating that suicide-related posts possess a significantly lower average polarity (-0.015) compared to non-suicide posts (0.030) at a 1% significance level. This disparity suggests a more negative and passive tone within suicide-related posts, emphasizing a prevailing sense of pessimism. The chart below displays the polarity distribution for both groups. Non-suicide posts lean more towards the positive side than suicidal posts, implying a higher occurrence of positive sentiments. They also show a broader spread, indicating a wider array of emotions compared to suicide-related posts. Moreover, non-suicide posts have a prominent peak at a polarity value of 0, suggesting less emotional content than suicidal posts.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean of Polarity | Z-Statistics | P-value |
| Suicidal Posts | -0.0146 | -6.417 | 0.000 |
| Non-suicide Posts | 0.0300 |  |  |



Additionally, the sentiment subjectivity analysis highlighted a significant variance, showcasing that suicide-related posts exhibit a markedly higher average subjectivity (0.50) compared to non-suicide posts (0.47) at a 1% significance level. This discrepancy suggests that suicide-related posts tend to delve more into personal emotions and subjective viewpoints, portraying a greater focus on individual feelings and attitudes. Regarding the subjectivity distribution chart, it's evident that non-suicide posts exhibit greater dispersion compared to suicidal posts. They notably peak at a subjectivity value of 0, suggesting a more objective expression in these posts.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean of Subjectivity | Z-Statistics | P-value |
| Suicidal Posts | 0.503 | 5.835 | 0.000 |
| Non-suicide Posts | 0.466 |  |  |



## Semantic Analysis Result

Below results from the semantic analysis revealed two distinct topics within the text content. Topic 0 seems to revolve around general life experiences and emotions. Words like "like," "get," "know," "would," "want," "one," "life," and "year" suggest discussions related to personal thoughts, experiences, and reflections on life events over time.

Topic 1 appears to focus on emotions and personal experiences as well. Keywords such as "want," "feel," "like," "know," "life," "time," "even," and "friend" indicate discussions concerning emotional states, desires, and interpersonal connections. These words suggest a discourse about emotions, desires, and relationships in life. Both topics delve into emotions, personal experiences, and life circumstances, indicating a significant focus on these themes within the textual content.

Topics:

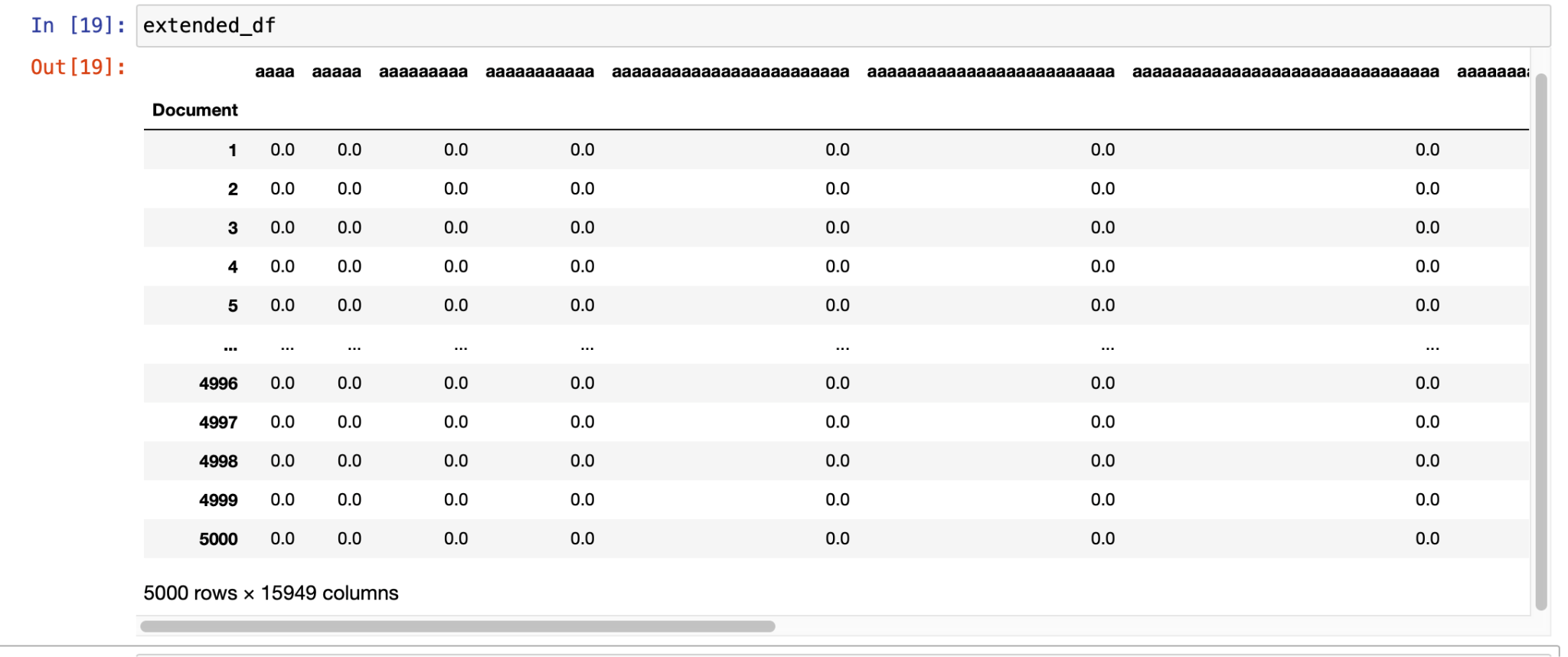
Topic 0: 0.015\*"like" + 0.009\*"get" + 0.009\*"know" + 0.008\*"would" + 0.008\*"want" + 0.007\*"one" + 0.007\*"life" + 0.007\*"year"

Topic 1: 0.014\*"want" + 0.014\*"feel" + 0.013\*"like" + 0.012\*"know" + 0.011\*"life" + 0.010\*"time" + 0.008\*"even" + 0.007\*"friend"

## TF-IDF Vectorizer

After that, TF-IDF vectorization was implemented to convert the preprocessed text into numerical features. The TF-IDF result displays a numerical representation of the preprocessed text. It's structured as a matrix with 5000 rows (documents) and 15949 columns (unique words or features). Each cell in the matrix corresponds to a specific word's TF-IDF score within the respective document.

This matrix enables the quantification of how relevant certain words are within individual documents compared to the entire dataset. Words with higher TF-IDF scores in a particular document are considered more important or unique to that document. Conversely, lower scores indicate less significance or commonality across the corpus.

****

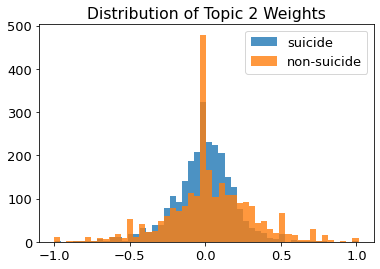
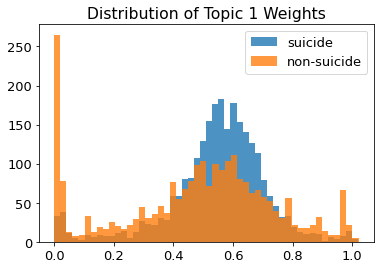
## Latent Semantic Analysis

After employing word vectorization, the analysis was extended by integrating the Latent Semantic Analysis (LSA) model to uncover latent semantic structures within the data. In the pursuit of discerning suicidal posts, two topics were defined to investigate whether the LSA model could effectively capture the distinguishing features of such posts. The outcomes revealed a noteworthy pattern: suicidal posts exhibited markedly higher weights in Topic 1, while displaying lower weights in Topic 2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean of Weights (Topic 1) | Z-Statistics | P-value |
| Suicidal Posts | 0.5470 | 11.561 | 0.000 |
| Non-suicide Posts | 0.4744 |  |  |

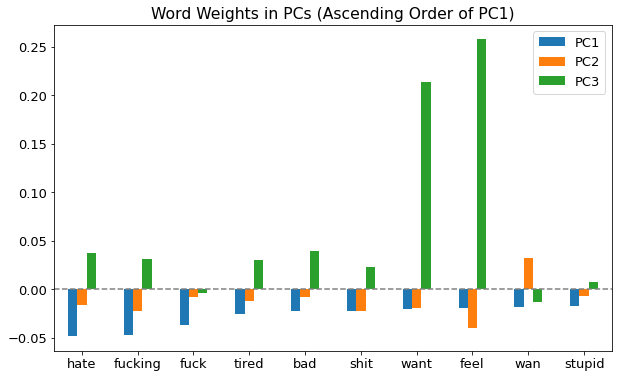
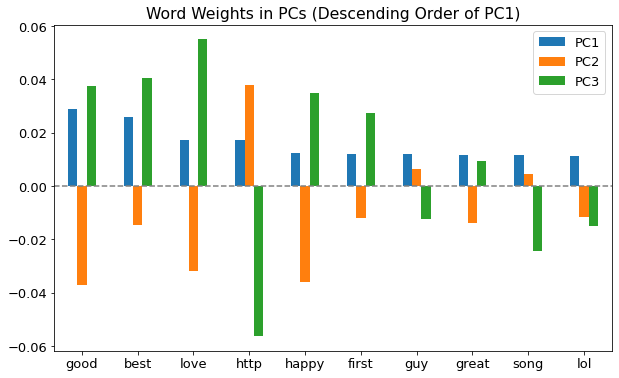
|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean of Weights (Topic 2) | Z-Statistics | P-value |
| Suicidal Posts | -0.009 | -6.252 | 0.000 |
| Non-suicide Posts | 0.035 |  |  |

However, while this differentiation in topic weights indicates a discernible pattern in suicidal posts, it's crucial to acknowledge that relying solely on the LSA model might not be sufficient for robust suicidality prediction within the corpus.



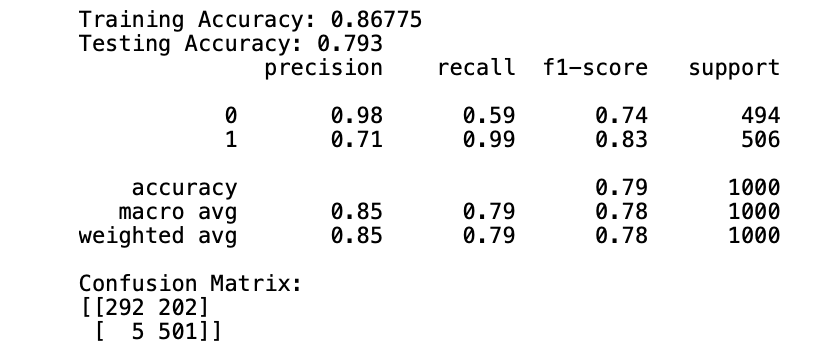
## Principal Components Analysis

The study further employed PCA for dimension reduction, and examined the weights of the first 3 PCs. For PC1, positive attitude was addressed, in which words like good, best, love, happy and great possessed higher positive weights. On the other hand, negative words such as hate, tired, bad and stupid had higher negative weights in PC1. For other PCs, a similar pattern of positive and negative words was also observed.



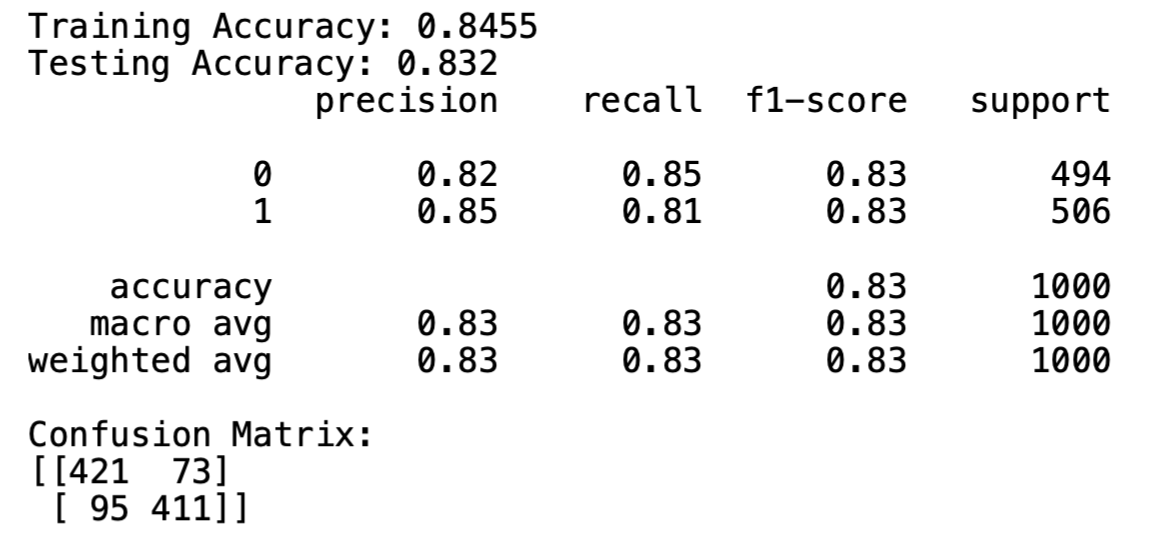
**Naive Bayes**

The transformed data from TF-IDF vectorization, combined with sentiment scores, was then utilized to train and evaluate several classification models. The Naive Bayes model achieved an 86.78% training accuracy and 79.3% testing accuracy in distinguishing between suicide-related and non-suicide-related posts. It displayed high precision (98%) for non-suicide posts but lower recall (59%). Conversely, for suicide-related posts, it showed lower precision (71%) but higher recall (99%). The model correctly identified 292 non-suicide posts and 501 suicide-related posts but misclassified 202 non-suicide posts as suicide-related and 5 suicide-related posts as non-suicide. Overall, the model showed potential but needed improvement to reduce misclassifications.

****

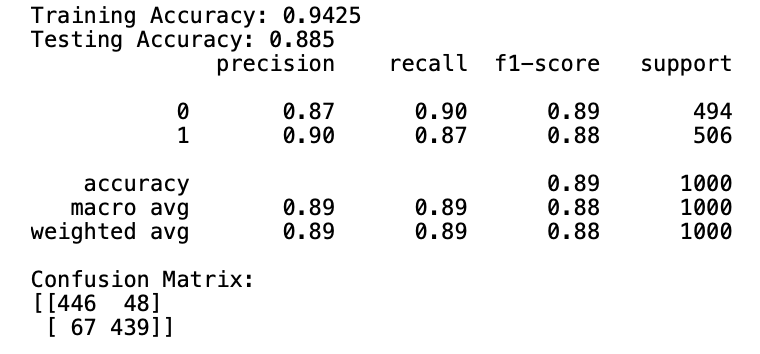
**Random Forest**

The Random Forest model displayed promising performance with 84.55% training accuracy and 83.2% testing accuracy in classifying suicide-related and non-suicide-related posts. It demonstrated balanced precision and recall for both classes, achieving 82% precision and 85% recall for non-suicide posts, and 85% precision and 81% recall for suicide-related posts. The model correctly identified 421 non-suicide posts and 411 suicide-related posts, while misclassifying 73 non-suicide posts as suicide-related and 95 suicide-related posts as non-suicide. Overall, it exhibited a well-rounded performance, showing comparable accuracy for both classes but with some misclassifications that might need further refinement.

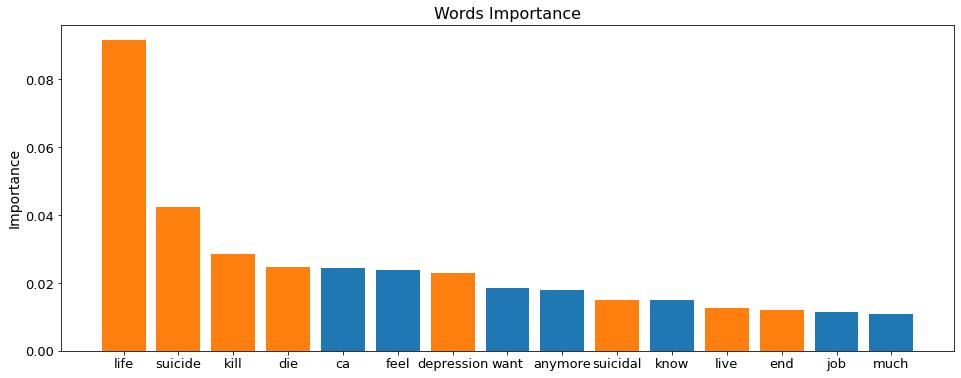
****

**XGBoost**

The XGBoost model showcased strong performance in classifying suicide and non-suicide posts, achieving a notable accuracy of 94.25% in training and 88.5% in testing. This model demonstrated balanced precision and recall for both classes, obtaining 87% precision and 90% recall for non-suicide posts, and 90% precision and 87% recall for suicide-related posts. It correctly identified 446 non-suicide posts and 439 suicide-related posts. However, it misclassified 48 non-suicide posts as suicide-related and 67 suicide-related posts as non-suicide. Overall, the model showcased robust performance, showing higher accuracy compared to previous models, but still exhibiting some misclassifications that could be further refined.

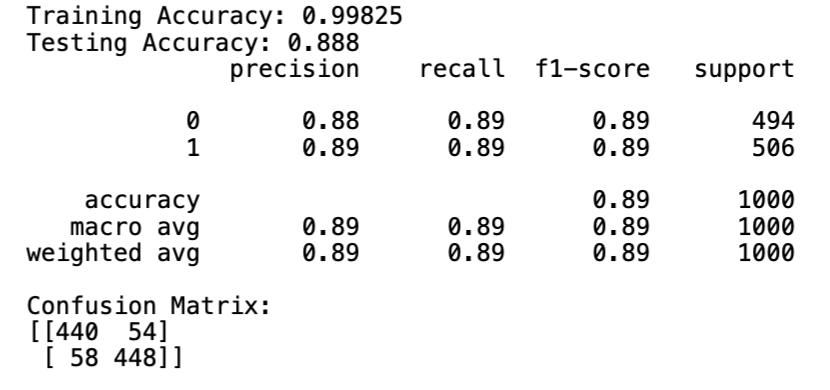


The XGBoost feature importance metrics, showcasing importance scores larger than 0, underline critical terms in identifying suicide-related content. Notably, "Life" holds the highest importance, followed by "suicide," "kill," "die," and "ca," suggesting their significant role in classifying such posts. Conversely, words like "two," "make," "people," "got," and "actually" have minimal impact on the classification process, indicating their lower relevance in discerning between suicide and non-suicide content. These important terms play a pivotal role in distinguishing and categorizing textual content for classification purposes.



**Neural Network**

The Neural Network model achieved high accuracy on both training (99.83%) and testing (88.8%) datasets, showcasing its ability to effectively classify suicide and non-suicide posts. The precision, recall, and F1-score metrics for both classes are approximately 0.89, indicating a balanced performance in predicting both categories. The model accurately classified 440 non-suicide posts and 448 suicide-related posts. However, it misclassified 54 non-suicide posts as suicide-related and 58 suicide-related posts as non-suicide. This model performed well in identifying suicide and non-suicide posts, displaying high accuracy but with some misclassifications that need improvement.



# **Discussion & Insights**

## Compared with Previous Studies

Similarities: This project shares the same methodology with literature utilizing natural language processing to predict the suicidality. In comparison with the literature, both methodologies utilized TF-IDF for feature extraction and suicidality prediction, resulting in high f1-scores and accuracy in both cases. Nadeem reached a f1 score of 84% for depression identification on twitter data (Nadeem, 2016). O'Dea reached a f1 score of 75% for distinguishing suicidal posts (O'Dea et al., 2015). In this study, a notable f1 score of 89% was attained.

Difference: In prior literature, natural language processing techniques have predominantly emphasized singular methodologies, often centering on approaches such as TF-IDF. However, in this study, a holistic approach was adopted by incorporating diverse variables such as TF-IDF, polarity, subjectivity, and word count into the learning model. This comprehensive integration has endowed the model with enhanced predictive capabilities.

Furthermore, the dataset used in this study is sourced from Redshift, presenting a notable departure from previous studies primarily conducted on Twitter data, which imposes a stringent 140-character limit on each post. In contrast, the Redshift dataset allows for a significantly higher word count per post. Given the comparatively expansive nature of the documents in this dataset, there is a higher likelihood that the model in this study can demonstrate superior performance compared to previous studies, which were constrained by the character limitations inherent in Twitter data.

Moreover, exploration was extended by implementing various models, such as XGBoost, Random Forest, and Neural Network. This strategic diversification enabled the utilization of the inherent predictive power offered by complex models. As a result, the comprehensive approach yielded a substantial improvement in both overall accuracy and F1 score compared to previous models. This underscores the effectiveness of the multi-faceted methodology in advancing the predictive performance of natural language processing models.

## Noteworthy Findings

During the feature importance analysis conducted in the study, several words related to suicidality were notably observed. Examples of these words include "life," "suicide," "kill," and so forth. The pronounced polarity and subjectivity differences conclusively demonstrate the efficacy of sentiment analysis in distinguishing between suicidal and non-suicidal posts. In particular, posts related to suicidality exhibit a greater degree of pessimism and subjectivity compared to non-suicidal posts. This observation is consistent with the expected expression of depression and personal attitudes toward life in suicidal content.

## Implications for Practice

This study has achieved successful predictions of suicidality within the Redshift dataset, allowing the identification of specific keywords that indicate suicidal tendencies. This investigation illuminates a discernible relationship between social networking posts and self-harm. Consequently, institutions can leverage these natural language processing techniques to proactively monitor and identify potential suicidality, thereby implementing preventive measures to avert such occurrences. The insights gleaned from this study offer a valuable foundation for the application of advanced linguistic analysis in the realm of mental health monitoring and intervention.

## Actionable Recommendation

To monitor websites for potential signs of suicidality or depression, the integration of natural language processing tools with machine learning methodologies proves invaluable. This combined approach enables the identification of linguistic cues indicative of potential suicidality. Upon detection, targeted mental health assistance can be offered, whether deliberately or unconsciously. This assistance may manifest in the form of personalized recommendations or the provision of mental health resources, including therapeutic interventions, aimed at providing timely support to individuals exhibiting signs of distress.

## Study Limitations

Access to comprehensive data in this study is somewhat constrained, as certain users may choose not to share their posts for further analysis, and others may abstain from using social networking websites or apps altogether. Additionally, a limiting factor in the research is the computational power available, which hinders the full exploration of the dataset's potential. The model's effectiveness and accuracy could have been further enhanced with a more extensive dataset if computational resources were not a limiting factor.

Furthermore, it is crucial to note that the current dataset utilized in this study is balanced. In reality, however, highly imbalanced data is more prevalent in this domain, as a substantial proportion of posts are unrelated to suicidality. Hence, for future implementations, addressing the challenges posed by imbalanced data should be a focal consideration to ensure the model's robustness in real-world scenarios.

# 

# **References**

Uddin, R., Burton, N. W., Maple, M., Khan, S. R., & Khan, A. (2019). Suicidal ideation, suicide

planning, and suicide attempts among adolescents in 59 low-income and middle-income countries: A population-based study. *The Lancet Child &amp; Adolescent Health*, *3*(4), 223–233. https://doi.org/10.1016/s2352-4642(18)30403-6

Wen, X., Kang, Y., Li, Y., Yuan, M., Li, Y., Xu, B., Su, P., & Wang, G. (2023). Machine

learning based suicidality risk prediction in early adolescence. *Asian Journal of Psychiatry*, *88*, 103716. https://doi.org/10.1016/j.ajp.2023.103716

Memon, A., Sharma, S., Mohite, S., & Jain, S. (2018). The role of online social networking on

deliberate self-harm and suicidality in adolescents: A systematized review of literature. *Indian Journal of Psychiatry*, *60*(4), 384. https://doi.org/10.4103/psychiatry.indianjpsychiatry\_414\_17

O’Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015).

Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183–188. https://doi.org/10.1016/j.invent.2015.03.005

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., &

Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, *35*(1), 51–59. https://doi.org/10.1027/0227-5910/a000234

Christensen, H., Batterham, P. J., & O’Dea, B. (2014, August 12). *E-health interventions for*

*suicide prevention*. MDPI. https://www.mdpi.com/1660-4601/11/8/8193

Nadeem, M. (2016, July 25). *Identifying depression on Twitter*. arXiv.org. https://arxiv.org/abs/1607.07384

Gunn, J. F., & Lester, D. (2015). Twitter postings and suicide: An analysis of the postings of a

fatal suicide in the 24 hours prior to death. *Suicidologi*, *17*(3). https://doi.org/10.5617/suicidologi.2173