UCONN | SCHOOL OF BUSINESS

Saving Reddit

Team 5 : | Kush Patel | Raj Thakkar | Amisha Aggarwal | Gaurav Borad | Sai Kiran Reddy

OPIM 5671

10-21-2024

Executive Summary

This project uses text mining techniques to identify the risk of suicidality by analyzing Reddit posts from individuals. The dataset comprises 500 anonymized posts, each categorized into labels such as “Supportive,” “Behaviour,” “Attempt,” “Ideation,” and “Behavior.” Our objective is to apply data analysis to identify individuals at the highest risk for self-harm or suicide, as early detection through proper analysis could potentially save lives.

The dataset contains unique identifiers for each post, text content, and labels that categorize posts based on suicidality indicators. Given the smaller size of the dataset, we encountered issues with low representation in some categories. To address this, we combined similar categories, merging “Behavior” with “Attempt” and “Indicator” with “Ideation” to ensure more balanced data. The data was then split into 60% training, 20% validation, and 20% testing sets. We applied text filtering, followed by text parsing, and conducted text clustering using a low singular value decomposition (SVD) resolution with exactly three clusters.

To analyze the data, we experimented with three models: Decision Tree, Regression, and Memory-Based Reasoning (MBR). We employed various term weighting techniques such as entropy, mutual information, and inverse document frequency (IDF). The best-performing model was the Decision Tree using entropy-based term weighting, which resulted in an accuracy of 64% and a misclassification rate of 36%. Despite the small dataset size, the Decision Tree model outperformed other methods and classified significant suicidality indicators.

One of the main challenges encountered was the limited size of the dataset, which caused some key terms to be omitted from the analysis. To account for this, we adjusted the term frequency threshold to ensure that terms appeared in at least two documents. Additionally, an imbalance in cluster weight distribution emerged, with one cluster containing less important terms being overemphasized. To overcome this, we also implemented multiple stoplists to filter out irrelevant words and to ensure that more meaningful clusters received proper attention.

This project demonstrates the potential of text mining in the field of mental health by enabling early detection of suicidal ideation. The insights gained from analyzing user posts can support interventions and provide timely assistance to individuals at risk. With further improvements such as larger datasets, refined model accuracy, and adding a more comprehensive mulit-term list: this approach could become a valuable tool in suicide prevention efforts.

Introduction

Background

Suicide is a critical public health issue, with millions of people worldwide struggling with mental health challenges that may lead to self-harm or suicide. Early identification of individuals at risk can significantly improve intervention efforts and save lives. In the digital age, user-generated content on social media platforms like Reddit offers a rich source of information that can be analyzed to detect early signs of suicidality. Analyzing this data through text mining techniques enables mental health professionals to better understand at-risk individuals and offer timely support.

This project focuses on text mining suicidality-related posts on Reddit, a platform where users often discuss a number of topics including personal issues. By analyzing these posts, we aim to identify patterns and predict the likelihood of self-harm or suicide attempts. This study not only advances the understanding of suicidality but also demonstrates the value of data analytics in preventing tragic outcomes through timely intervention.

Dataset

The dataset used in this project consists of 500 anonymized posts from Reddit users, each labeled into one of five categories: “Supportive,” “Attempt,” “Ideation,” “Indicator,” and “Behavior.” These labels reflect varying degrees of risk, from individuals seeking help to those expressing explicit suicidal intentions. Each post is linked to a unique identifier for the user and contains the text content of their post, along with its assigned label.

One of the key challenges with the dataset is its small size, which limits the depth of analysis. Additionally, some categories, such as “Attempt,” had fewer occurrences, which required merging with other related categories to increase statistical significance. Despite these limitations, the dataset provides valuable insights into the language patterns associated with different stages of suicidality.

Objective

The primary objective of this project is to identify the likelihood of suicidality by analyzing text posts. Through this, we aim to develop predictive models that can classify posts into relevant categories, ultimately identifying those at high risk of self-harm or suicide. By applying advanced text mining techniques, we seek to contribute to the growing field of mental health analytics, helping professionals intervene earlier and more effectively.

Our analysis employs methods like text clustering, term frequency weighting, and classification models to extract meaningful patterns from the data. The ultimate goal is to refine the models to achieve higher accuracy, despite the constraints of a small dataset, and offer insights that could be used in real-world applications for suicide prevention.

Data Info

Variable	Description
User	A unique identifier for each user or post. This ID has no numerical or ordinal significance; it is nominal data used solely for identifying individual posts.
Post	The text content written by Reddit users in their posts. This free-text data captures the thoughts, emotions, and experiences shared by users regarding suicidality.
Label	The assigned category or target label for each post, classifying it into one of five categories: "Supportive," "Attempt," "Ideation," "Indicator," or "Behavior." These labels are used to predict the level of suicidality risk.

Kaggle Dataset: Suicidality on Reddit

User	Post	Label
user-3	['I tried to kill my self once and failed badly cau	Attempt
user-18	['No need for thanks it just makes me happy th	Attempt
user-24	['Thank you so much for the advice. The only re	Attempt
user-25	['To update you guys friend called police in me	Attempt
user-30	['Came back home about 2 hours ago...', 'It is t	Attempt
user-43	['seems fun for someone who would be into it b	Attempt
user-46	['There is nothing else to share. Nothing can ch	Attempt
user-48	['Definitely not easy. I live in the Southeast US.	Attempt
user-61	['Hey man, You cant be convinced and I cann	Attempt
user-82	['I just took10 more. Okay I threw up a little bit I	Attempt
user-97	['And how does that make you feel? Depressed	Attempt
user-124	['Same, pm me and we can talk.', 'Hi. Im in the :	Attempt
user-142	['How long have you been depressed?', 'No one	Attempt
user-147	['Well the biggest factor preventing me from try	Attempt
user-166	['Sounds like you need closure.Ive been in som	Attempt
user-192	['Ive been hospitalized 3 times. Each time it ha	Attempt
user-218	['If youre still here, so am I. Ive been seriously D	Attempt
user-237	['I started anti-DP treatment this morning.Feel	Attempt
user-238	['Most people in life are not equipped with the s	Attempt

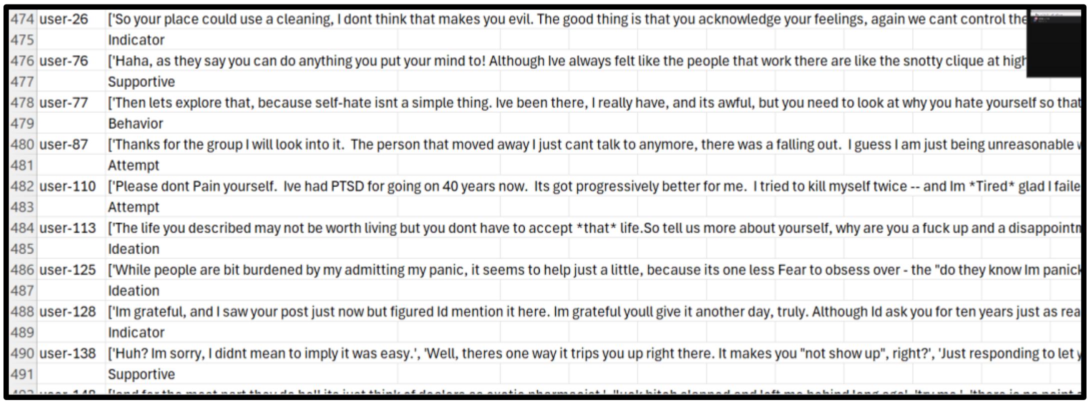
Sample Data

Data Exploration:

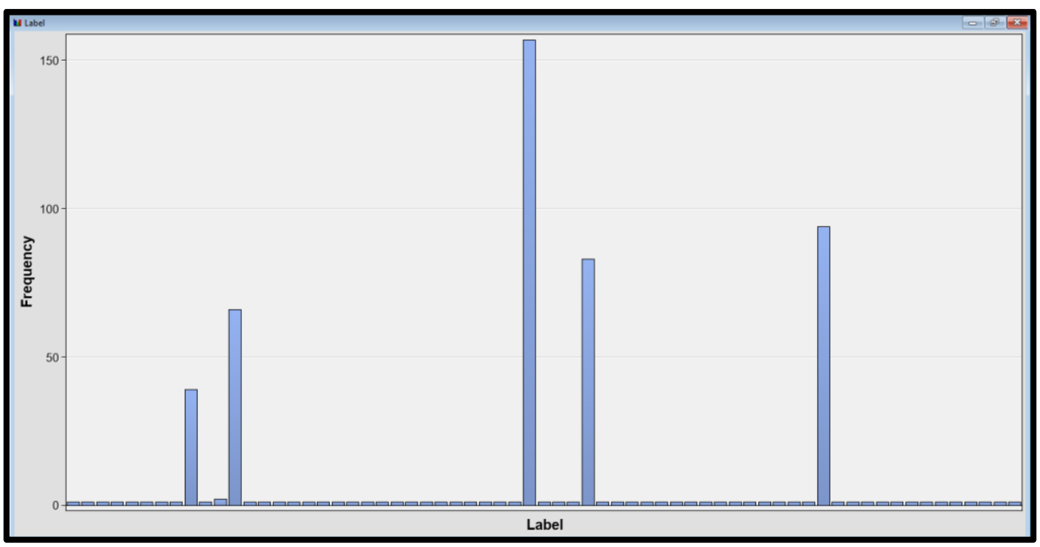
1. <u>Data Exploration and Initial Insights</u>

The project began with the importation of the dataset into SAS Miner using the File Import feature. After successfully importing, we utilized the Data Distribution feature within the Text Filter node to visualize the dataset. Our primary focus was on the target variable, “Label,” which is a categorical variable and contains five categories: "Supportive," "Attempt," "Ideation," "Indicator," and "Behavior."

Here, in raw data - a major inconsistency was discovered where some rows had merged columns for both the “Post” (text) and “Label” (target) variables. This inconsistency was addressed by separating and cleaning the affected rows, ensuring each post matched correctly with its label.

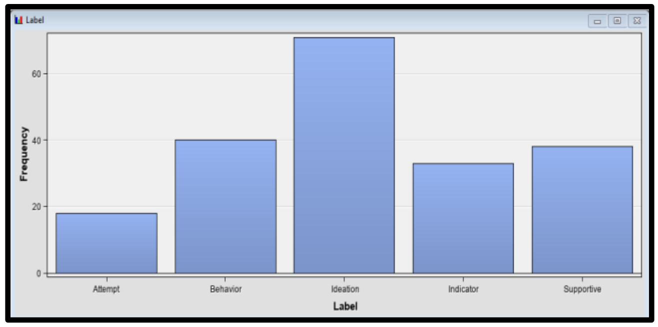


474	user-26	['So your place could use a cleaning, I dont think that makes you evil. The good thing is that you acknowledge your feelings, again we cant control the
475		Indicator
476	user-76	['Haha, as they say you can do anything you put your mind to! Although Ive always felt like the people that work there are like the snotty clique at high
477		Supportive
478	user-77	['Then lets explore that, because self-hate isnt a simple thing. Ive been there, I really have, and its awful, but you need to look at why you hate yourself so that
479		Behavior
480	user-87	['Thanks for the group I will look into it. The person that moved away I just cant talk to anymore, there was a falling out. I guess I am just being unreasonable
481		Attempt
482	user-110	['Please dont Pain yourself. Ive had PTSD for going on 40 years now. Its got progressively better for me. I tried to kill myself twice -- and Im *Tired* glad I faile
483		Attempt
484	user-113	['The life you described may not be worth living but you dont have to accept *that* life.So tell us more about yourself, why are you a fuck up and a disappointm
485		Ideation
486	user-125	['While people are bit burdened by my admitting my panic, it seems to help just a little, because its one less Fear to obsess over - the "do they know Im panick
487		Ideation
488	user-128	['Im grateful, and I saw your post just now but figured Id mention it here. Im grateful youll give it another day, truly. Although Id ask you for ten years just as rea
489		Indicator
490	user-138	['Huh? Im sorry, I didnt mean to imply it was easy.', 'Well, theres one way it trips you up right there. It makes you "not show up", right?', 'Just responding to let y
491		Supportive

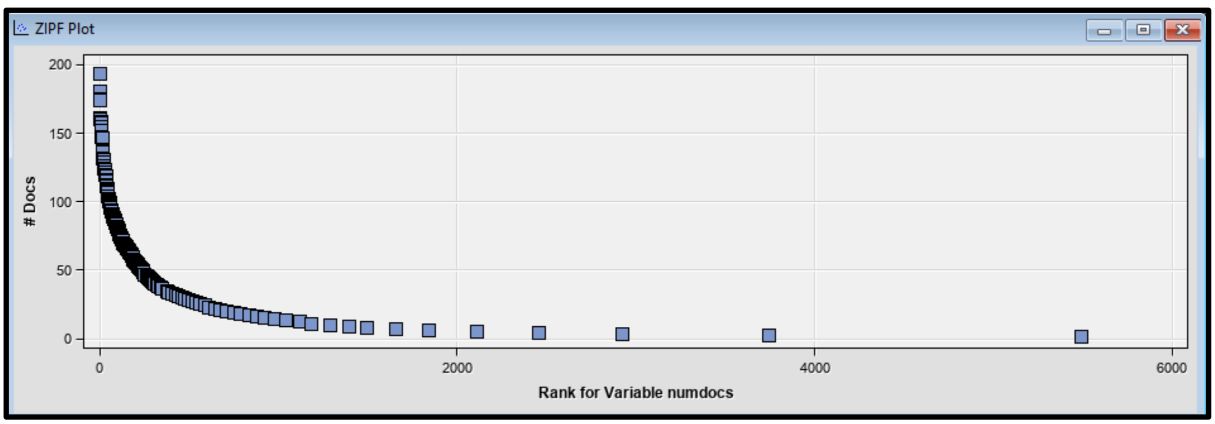


<u>Discrepancies in some of the users where Posts and Labels got merged</u>

However, early data exploration revealed an imbalance in the distribution of these categories. Critical labels such as “Attempt” and “Behavior” were underrepresented, while categories like “Ideation,” “Indicator,” and “Supportive” dominated the dataset.

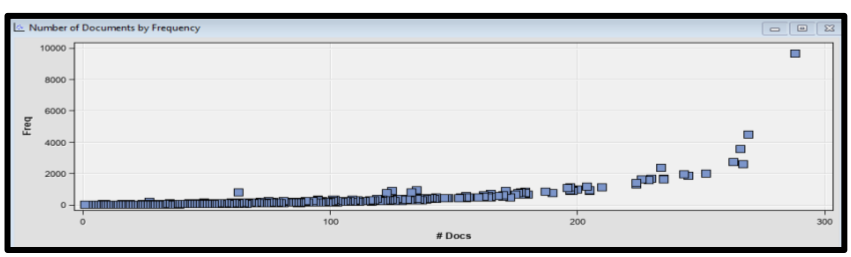


To assess the dataset's readiness for text mining, we applied Zipf's Law to evaluate the distribution of terms and reviewed document frequency plots. Zipf's plot confirmed that the terms followed a power-law distribution, indicating the presence of a few frequently used terms and many infrequently used ones—a common characteristic in text data.



Zipf Plot

The document frequency plot shows the distribution of document frequencies in the text dataset. The x-axis represents the number of documents containing specific terms, while the y-axis indicates how often those terms appear across the dataset. The skewed distribution suggests that many terms appear in few documents, while a few terms are present in many documents, a common pattern in text mining.

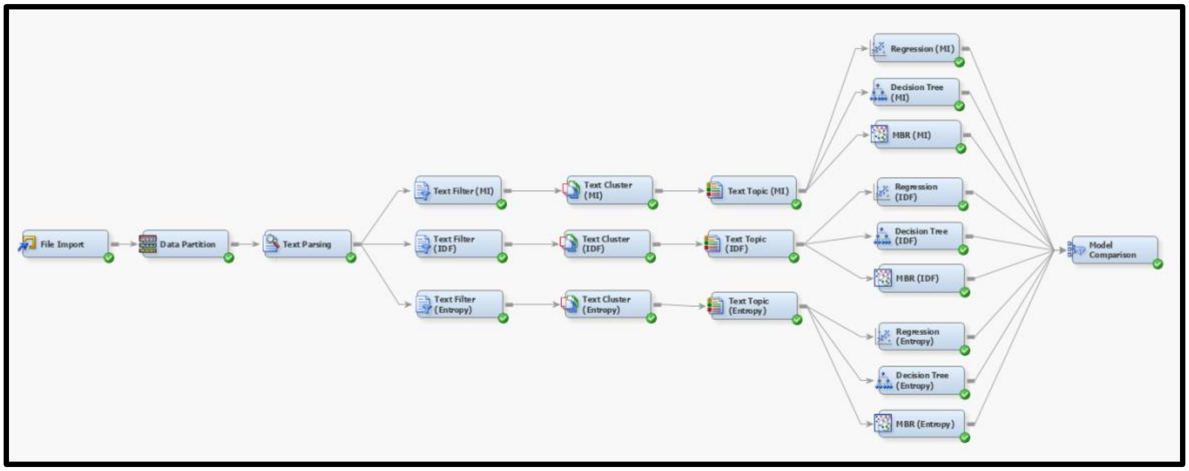


We tried to run our initial model with below specifications:

1. Data Partition : 40% Training, 30% validation, 30% test
2. Term Filters : Min. no. of documents appearance in document - 100

3. Text Cluster Specifications: High SVD Resolution with Exact 5 clusters (as 5 categories)

Initial model runs on this raw dataset yielded poor performance, as demonstrated by the results from the Decision Tree model using Inverse Document Frequency (IDF). The model showed a high misclassification rate of 63%, resulting in an accuracy of only 37%. Given the low quantity of data in critical categories like "Attempt" and "Behavior," it became evident that the imbalanced target variable was impacting the model's predictive power.



Initial Diagram

Fit Statistics

Selected Model

Selected Model	Predecessor Node	Model Node	Model Description	Selection Criterion: Valid: Misclassification Rate
Y	Tree2	Tree2	Decision Tree (IDF)	0.610738
	Tree3	Tree3	Decision Tree (Entropy)	0.624161
	Tree	Tree	Decision Tree (MI)	0.630872
	Reg3	Reg3	Regression (Entropy)	0.657718
	Reg2	Reg2	Regression (IDF)	0.677852
	Reg	Reg	Regression (MI)	0.691275
	MBR	MBR	MBR (MI)	0.691275
	MBR2	MBR2	MBR (IDF)	0.711409
	MBR3	MBR3	MBR (Entropy)	0.718121

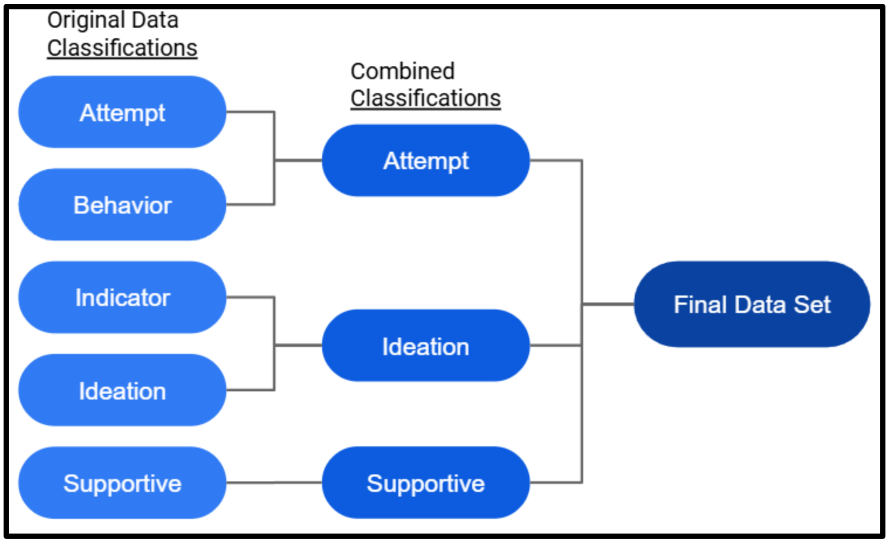
Initial Results (Best Model - Decision Tree (IDF))

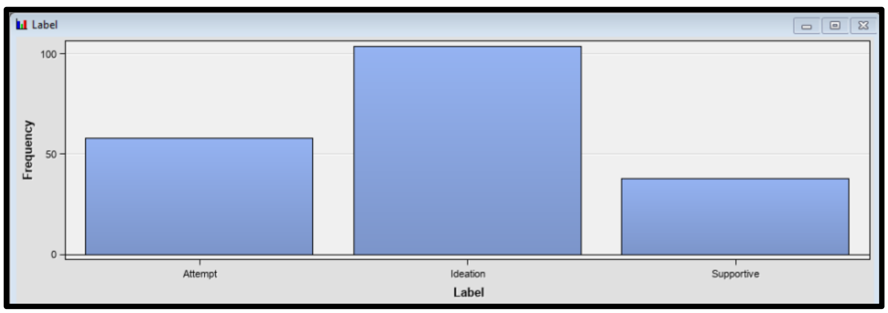
Cluster ID	Descriptive Terms	Frequency
1	+friend +find +good +feel im +life +live +talk +know +help +want +thing dont +time ive ...	25
2	better always +find +talk +day +friend ive cant +live +year +time +good +help +people +life ...	83
3	+know +thing dont +want +life +day im +feel always +time +year +good +help +live +people ...	70
4	cant im ive +know +time +feel dont +people always +live +want better +talk +year +day ...	8
5	+people +feel im +help +year dont +find +thing +want +life +know cant better ive +talk ...	10

Clusters (Exact 5)

2. Category Merging and Data Refinement:

To improve prediction accuracy for key categories, we opted to merge several underrepresented labels. Specifically, we combined “Attempt” and “Behavior” into a single “Attempt” category, and “Ideation” and “Indicator” into a unified “Ideation” category, leaving “Supportive” as a separate category. This restructuring ensured a more balanced dataset with meaningful representation across the target labels, thus improving the potential for model performance.





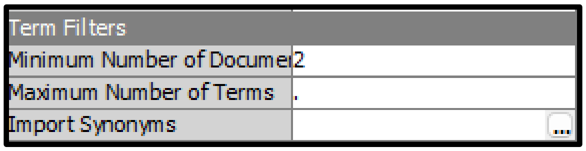
Modified Target Categories

We partitioned the data into 60% training, 20% validation, and 20% testing sets. Given the small size of the dataset, this split allowed us to maximize the training data for better model learning while maintaining adequate test and validation sets.

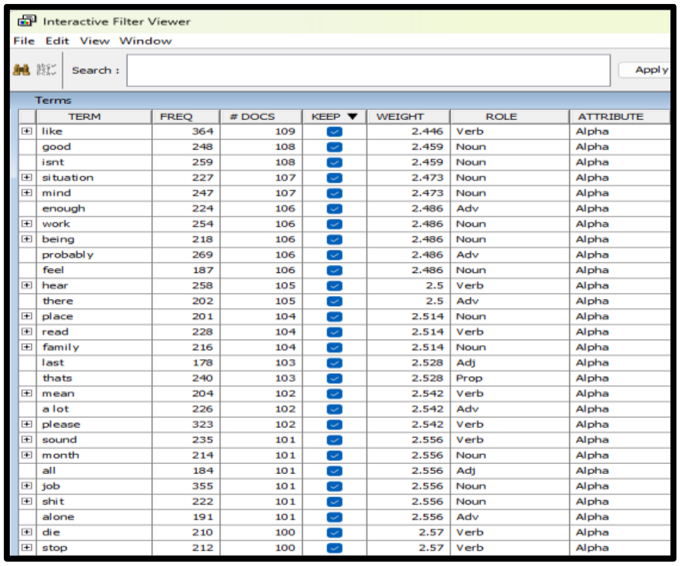
Data Set Allocations	
Training	60.0
Validation	20.0
Test	20.0

3. Data Cleaning and Text Filtering

The next step involved refining the dataset through text filtering to eliminate irrelevant terms and enhance modeling quality. Initially, we set a minimum document frequency of 100, but this threshold excluded many important, low-frequency terms due to the limited size of the dataset. To capture more meaningful terms, we adjusted the threshold to a minimum of two document appearances, allowing us to retain infrequent but potentially significant terms related to suicidality.



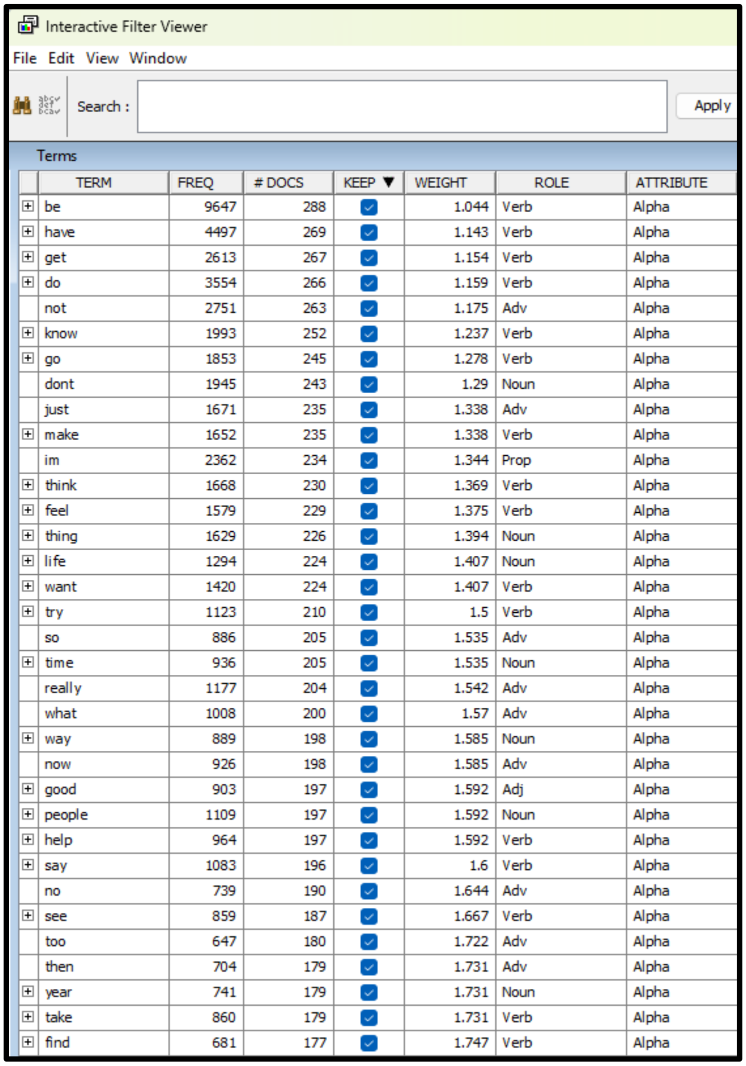
Term Filters	
Minimum Number of Documer	2
Maximum Number of Terms	.
Import Synonyms	...



TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
like	364	109	✓	2.446	Verb	Alpha
good	248	108	✓	2.459	Noun	Alpha
isnt	259	108	✓	2.459	Noun	Alpha
situation	227	107	✓	2.473	Noun	Alpha
mind	247	107	✓	2.473	Noun	Alpha
enough	224	106	✓	2.486	Adv	Alpha
work	254	106	✓	2.486	Noun	Alpha
being	218	106	✓	2.486	Noun	Alpha
probably	269	106	✓	2.486	Adv	Alpha
feel	187	106	✓	2.486	Noun	Alpha
hear	258	105	✓	2.5	Verb	Alpha
there	202	105	✓	2.5	Adv	Alpha
place	201	104	✓	2.514	Noun	Alpha
read	228	104	✓	2.514	Verb	Alpha
family	216	104	✓	2.514	Noun	Alpha
last	178	103	✓	2.528	Adj	Alpha
thats	240	103	✓	2.528	Prop	Alpha
mean	204	102	✓	2.542	Verb	Alpha
a lot	226	102	✓	2.542	Adv	Alpha
please	323	102	✓	2.542	Verb	Alpha
sound	235	101	✓	2.556	Verb	Alpha
month	214	101	✓	2.556	Noun	Alpha
all	184	101	✓	2.556	Adj	Alpha
job	355	101	✓	2.556	Noun	Alpha
shit	222	101	✓	2.556	Noun	Alpha
alone	191	101	✓	2.556	Adv	Alpha
die	210	100	✓	2.57	Verb	Alpha
stop	212	100	✓	2.57	Verb	Alpha

We initially used the Filter Viewer to manually review term frequencies and make decisions about which terms to keep or exclude. This involved a careful, case-by-case examination to filter out non-relevant terms. However, we soon realized this manual approach was too time-consuming and inconsistent, especially with a dataset that required a systematic approach for effective filtering. To address this, we experimented with various customized stop lists, iterating through different versions to find one that would remove common, non-predictive terms without omitting meaningful ones. After testing multiple lists, we identified the most effective stop list, which successfully filtered out irrelevant terms—like overly common words—while retaining terms essential to the prediction task. This process of trial and adjustment ultimately enabled us to refine the dataset more efficiently and ensure that it included only meaningful, predictive terms.

<u>Added Multiple Stop-Lists</u>



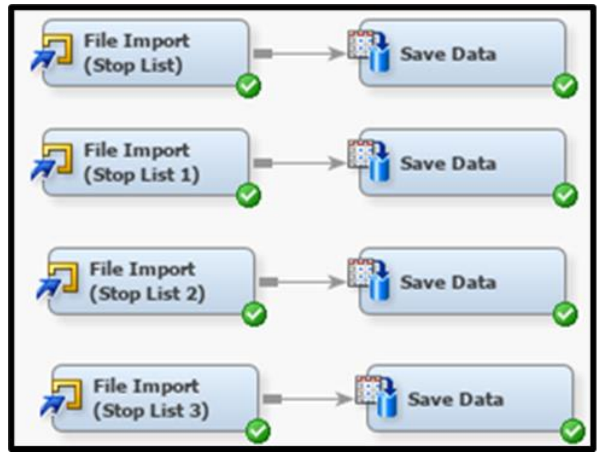
Interactive Filter Viewer

File Edit View Window

Search : Apply

Terms

TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
be	9647	288	✓	1.044	Verb	Alpha
have	4497	269	✓	1.143	Verb	Alpha
get	2613	267	✓	1.154	Verb	Alpha
do	3554	266	✓	1.159	Verb	Alpha
not	2751	263	✓	1.175	Adv	Alpha
know	1993	252	✓	1.237	Verb	Alpha
go	1853	245	✓	1.278	Verb	Alpha
dont	1945	243	✓	1.29	Noun	Alpha
just	1671	235	✓	1.338	Adv	Alpha
make	1652	235	✓	1.338	Verb	Alpha
im	2362	234	✓	1.344	Prop	Alpha
think	1668	230	✓	1.369	Verb	Alpha
feel	1579	229	✓	1.375	Verb	Alpha
thing	1629	226	✓	1.394	Noun	Alpha
life	1294	224	✓	1.407	Noun	Alpha
want	1420	224	✓	1.407	Verb	Alpha
try	1123	210	✓	1.5	Verb	Alpha
so	886	205	✓	1.535	Adv	Alpha
time	936	205	✓	1.535	Noun	Alpha
really	1177	204	✓	1.542	Adv	Alpha
what	1008	200	✓	1.57	Adv	Alpha
way	889	198	✓	1.585	Noun	Alpha
now	926	198	✓	1.585	Adv	Alpha
good	903	197	✓	1.592	Adj	Alpha
people	1109	197	✓	1.592	Noun	Alpha
help	964	197	✓	1.592	Verb	Alpha
say	1083	196	✓	1.6	Verb	Alpha
no	739	190	✓	1.644	Adv	Alpha
see	859	187	✓	1.667	Verb	Alpha
too	647	180	✓	1.722	Adv	Alpha
then	704	179	✓	1.731	Adv	Alpha
year	741	179	✓	1.731	Noun	Alpha
take	860	179	✓	1.731	Verb	Alpha
find	681	177	✓	1.747	Verb	Alpha



Text Clustering

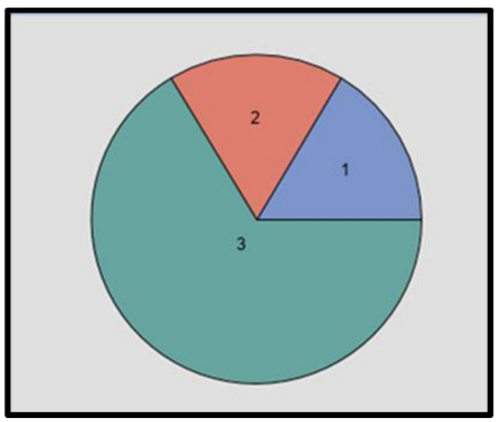
	A	B	C	D	E	F	G	H
1		TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBU
2961		manageable	5	5	N	0.717	Adj	Alpha
2962	+	drift	5	5	N	0.717	Verb	Alpha
2963		hs	5	5	N	0.717	Prop	Alpha
2964	+	scream	5	5	N	0.717	Noun	Alpha
2965		beat	5	5	N	0.717	Noun	Alpha
2966		someones life	5	5	N	0.717	Noun Group	Alpha

<u>Filter Viewer to keep or delete terms</u>

We applied text clustering to group the data based on similarity, which helped structure the dataset for better predictive modeling. To manage the high dimensionality of the text data, we experimented with different Singular Value Decomposition (SVD) resolutions, ultimately selecting a dimension of 100, which provided the best balance between performance and noise reduction. This dimensionality reduction technique allowed us to focus on the dataset's most relevant features, improving the model's performance by filtering out less important information.

Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Exact
Number of Clusters	3
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15

We also tested different numbers of clusters to see how they affected the model's accuracy, initially exploring both maximum clusters and an exact match to our target variable categories. After merging our target variable into three categories, we found that specifying three clusters in the clustering algorithm aligned well with our goals, particularly in distinguishing the "Attempt" category, which is essential for predicting suicidality. By refining the dataset with a customized stop list to remove irrelevant terms, we minimized noise from non-predictive words, further enhancing the clustering process and ultimately improving the model's accuracy.

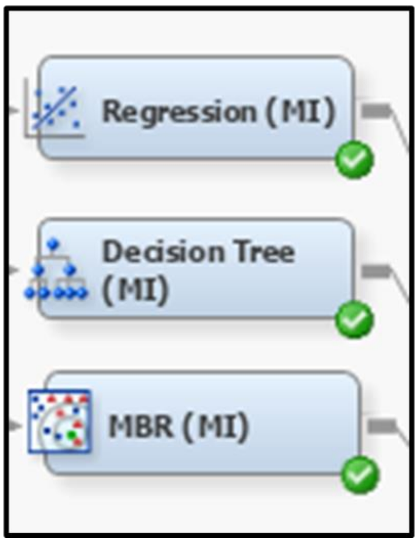


Model Development:

In the modeling phase of the suicidality data project, we applied three models—Regression, Decision Tree, and Memory-Based Reasoning (MBR)—each under three different term-weighting strategies: Mutual Information, Inverse Document Frequency (IDF), and Entropy. The objective was to predict suicidal behavior based on text posts.

1. Frequency Weighting: Log with Mutual Information

- **Term Weighting Strategy**: Mutual Information was chosen as it captures the amount of information a word contributes toward predicting the target variable (suicide behavior or ideation).



Models Used:

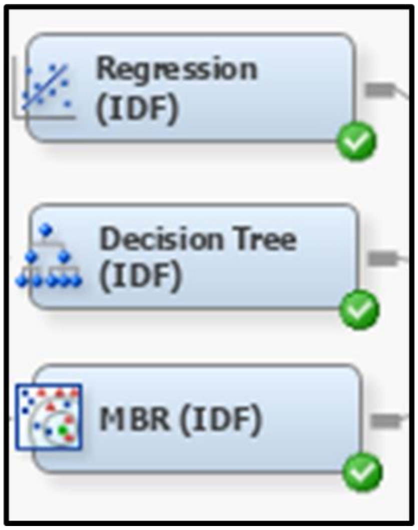
Regression: The text data was transformed into a structured format, allowing a logistic regression model to classify posts into risk categories.

Decision Tree: We built a decision tree with Mutual Information to understand which words contribute the most to predicting suicidality. The tree structure provided insights into key predictors like 'depression' or 'hopelessness.'

MBR (Memory-Based Reasoning): This model worked by comparing the input post with past similar posts. Mutual Information helped in identifying significant words to match against similar posts.

2. Frequency Weighting: Log with Inverse Document Frequency (IDF)

- **Term Weighting Strategy**: IDF was used to penalize common terms across the dataset and give weight to terms that are rare but important for predicting suicidality.



Models Used:

Regression: Here, IDF improved the ability of the regression model to emphasize rare but critical terms such as 'self-harm' or 'suicide attempt.'

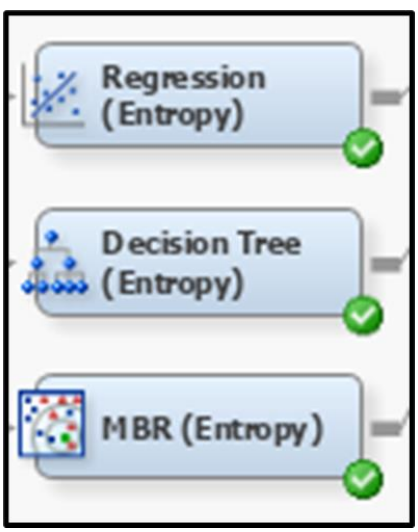
Decision Tree: We constructed a decision tree with IDF, limiting it to 25 leaves to prevent overfitting due to small sample size. The focus was on identifying rare but crucial terms that might be overlooked with standard frequency measures.

MBR: In the case of MBR, IDF helped match the test post with past instances where rare but significant terms were present, thus improving accuracy.

○

3. Frequency Weighting: Log with Entropy

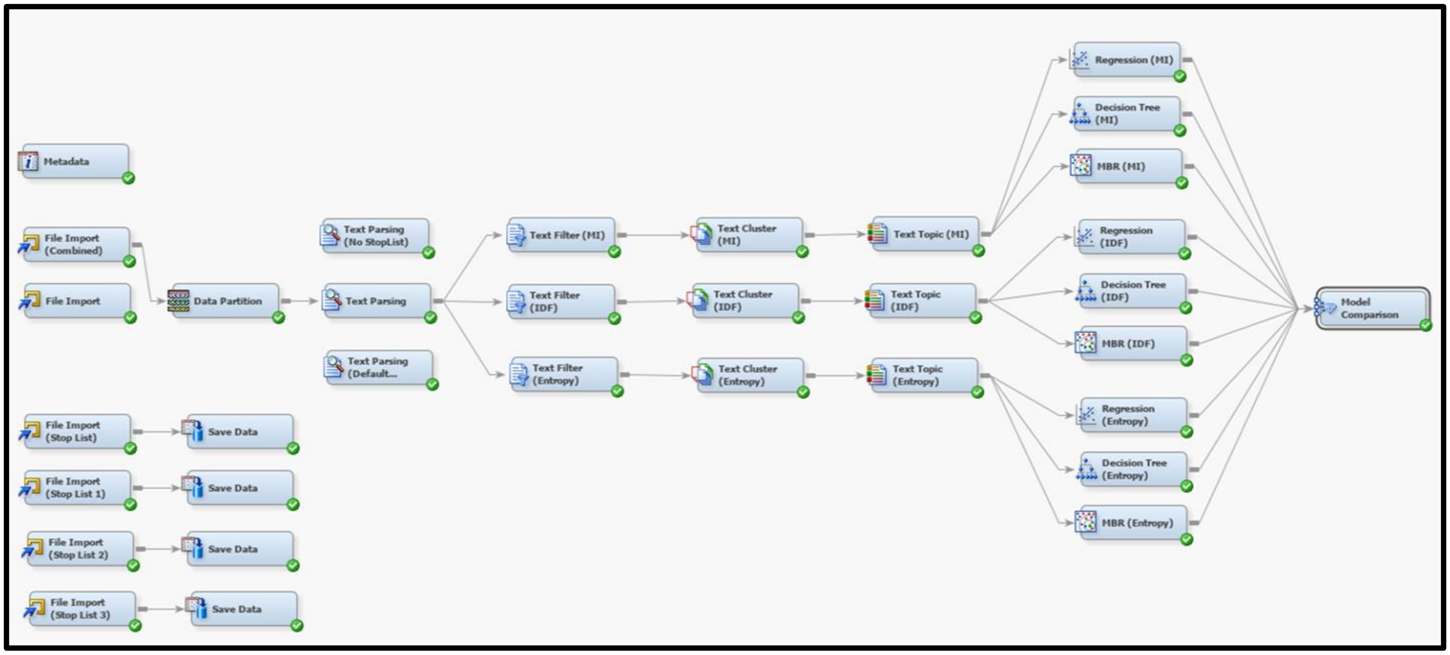
- **Term Weighting Strategy**: Entropy was used to measure the unpredictability or uncertainty of word occurrences across the posts. Words contributing more uncertainty (i.e., less predictable but relevant) were given higher weights.



- **Models Used**:
  - **Regression**: The entropy-weighted regression model tried to capture unpredictability in word use related to suicide ideation, focusing on terms that appeared sporadically but indicated high-risk behavior.
  - **Decision Tree**: The Decision Tree using entropy allowed for the classification of high-risk categories by balancing the spread of key terms like 'pain' or 'loss' that signaled suicidal ideation or behavior.
  - **MBR**: Entropy also improved the MBR model by considering the unpredictability of term appearances in previous cases. This helped to identify patterns in posts with high uncertainty about suicidal intent.

Modeling Diagram

Each model and term-weighting combination contributed a unique perspective on how textual patterns related to suicidality could be analyzed and predicted. These models laid the groundwork for identifying which approach could yield the highest accuracy, leading to more focused and efficient interventions.



Model Comparison:

The best-performing model in our analysis was the **Decision Tree using Entropy**, with a **Misclassification Rate** of 0.36 (36%) and an accuracy of 64%. This model excelled because **Entropy** captures the unpredictability of term distributions, allowing it to better differentiate between posts related to different levels of suicidality risk. By focusing on terms that provided the most uncertainty reduction, this model was able to capture subtle linguistic patterns within the posts. The accuracy of 64% signifies that it could effectively classify the posts with a reasonable degree of precision, making it the most reliable model for predicting suicide attempts or related behavior.

In contrast, the other models, including the **Decision Tree with Mutual Information** (41% Misclassification Rate) and various **Regression** and **Memory-Based Reasoning (MBR)** models, performed relatively weaker. The **Regression models** had misclassification rates ranging from 42% to 45%, and the **MBR models** were the weakest, with rates from 47% to 55%. These models struggled with capturing the complex relationships between terms and the target variable, likely due to their limitations in handling the intricacies of the text data. Despite moderate performance, they could not match the effectiveness of the Decision Tree with Entropy for this suicidality dataset.

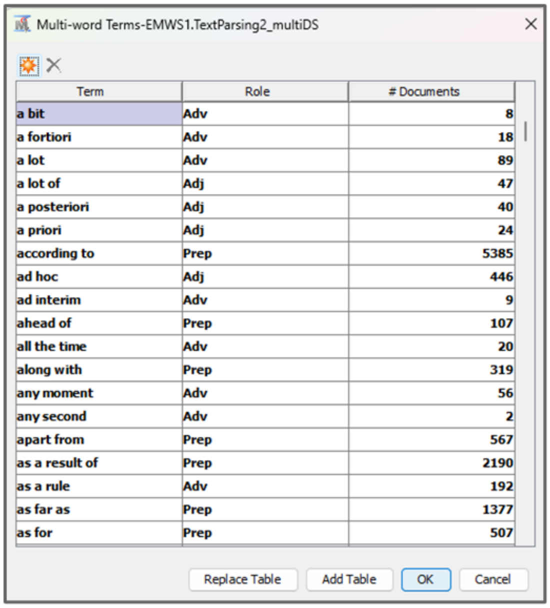
Model Description	Selection Criterion: Valid: Misclassification Rate
Decision Tree (Entropy)	0.36
Decision Tree (MI)	0.41
Regression (IDF)	0.42
Regression (Entropy)	0.43
Regression (MI)	0.45
Decision Tree (IDF)	0.46
MBR (MI)	0.47
MBR (IDF)	0.52
MBR (Entropy)	0.55

Why Is the Accuracy Still Low?

While the Decision Tree model with Entropy achieved the highest accuracy of 64%, the overall performance highlights the challenge of correctly classifying sensitive text around suicidality. A primary reason for lower accuracy is the nuanced language used in posts. For instance, a post such as, "**I can support you through these <u>suicidal</u> thoughts and help you overcome them**" would fall under the **supportive** category, while a post like, "**I want to end my life with <u>suicide</u>**" clearly belongs to the **attempt** category. However, due to the complexity of natural language, a model relying on individual terms or frequency may misclassify posts. Even with or without the word "suicide," phrases such as "end my life" may be misclassified because the model might not always understand the context or differentiate between types of intent.

This leads to higher misclassification, especially when terms like "thoughts" or "support" overlap between categories. The model's dependence on frequency and weightings like Entropy helps to some extent, but it still struggles with interpreting nuanced meanings, which impacts its ability to classify the posts with higher precision.

Future Improvements



Term	Role	# Documents
a bit	Adv	8
a fortiori	Adv	18
a lot	Adv	89
a lot of	Adj	47
a posteriori	Adj	40
a priori	Adj	24
according to	Prep	5385
ad hoc	Adj	446
ad interim	Adv	9
ahead of	Prep	107
all the time	Adv	20
along with	Prep	319
any moment	Adv	56
any second	Adv	2
apart from	Prep	567
as a result of	Prep	2190
as a rule	Adv	192
as far as	Prep	1377
as for	Prep	507

1. **Custom Multi-word Terms**:

Incorporating multi-word expressions like "end my life" or "reach out for help" can provide richer context and reduce misclassification. These phrases capture meaning beyond individual words and can help the model differentiate between categories like **attempt** and **supportive** more effectively.

2. **More Defined Stop List**:

A more refined stop list tailored to the suicidality context can remove terms that are less indicative of the target categories. This would help the model focus on words that carry higher semantic importance for accurate classification.

3. **More Data**:

Expanding the dataset would greatly enhance the model's learning ability. A larger, more diverse dataset could provide better training examples for distinguishing between subtle language variations in posts, ultimately improving model accuracy and robustness over time.

Conclusion:

This project demonstrated the potential of using text mining techniques to predict suicidality based on social media posts from Reddit. By analyzing free-text data, we were able to classify posts into categories representing various stages of suicidality, ranging from supportive behavior to direct attempts. Despite the limitations of a small dataset and category imbalances, the models, particularly the Decision Tree with entropy weighting, showed promise in identifying key risk factors for suicide.

One of the critical challenges encountered was the dataset's imbalanced category distribution, which necessitated merging similar categories to enhance model performance. Additionally, the need to carefully filter out irrelevant terms using customized stop lists and term frequency thresholds proved essential in refining the data for analysis. The application of low SVD resolution for dimensionality reduction was effective in clustering similar posts, contributing to a clearer understanding of the patterns within the dataset.

While the Decision Tree model emerged as the best-performing algorithm with an accuracy of 64%, it is important to note that the overall performance could be improved with larger datasets, more diverse training data, and enhanced feature engineering. By integrating additional factors such as user metadata or external signals (e.g., engagement metrics), future studies could offer deeper insights and greater predictive power.

In conclusion, this project highlights the value of text mining in mental health analytics, particularly in identifying individuals at risk of self-harm or suicide. With further improvements, such techniques can play a critical role in early detection and intervention, ultimately contributing to life-saving mental health support. The findings underscore the importance of integrating data-driven solutions in suicide prevention efforts, while also suggesting areas for further research and methodological refinement.