Suicide Detection

Natural Language Processing for Suicide Detection

Andrew Seal

Capstone 3 - Springboard

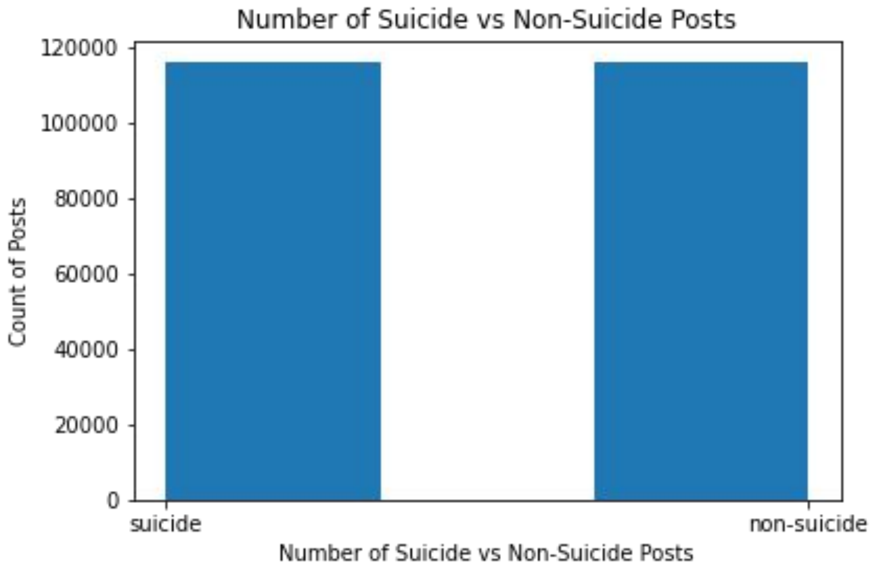
Background: Most people who die by suicide talk about it first. Many of these people can be helped.

Objective: Develop a model to interpret written text and flag when someone’s risk of suicide is high.

Stakeholders: Social media companies (schools, parents), online therapy providers such as Betterhelp, 7 cups, Talkspace

Dataset

- Dataset consisted of 232,074 social media posts from Reddit
  - Exactly half were from ‘Suicide Watch Subreddit
  - The other half were from ‘Teenager’ Forum
- Downloaded the dataset from kaggle.com
- Goal is to develop a model to predict the source of each social media post



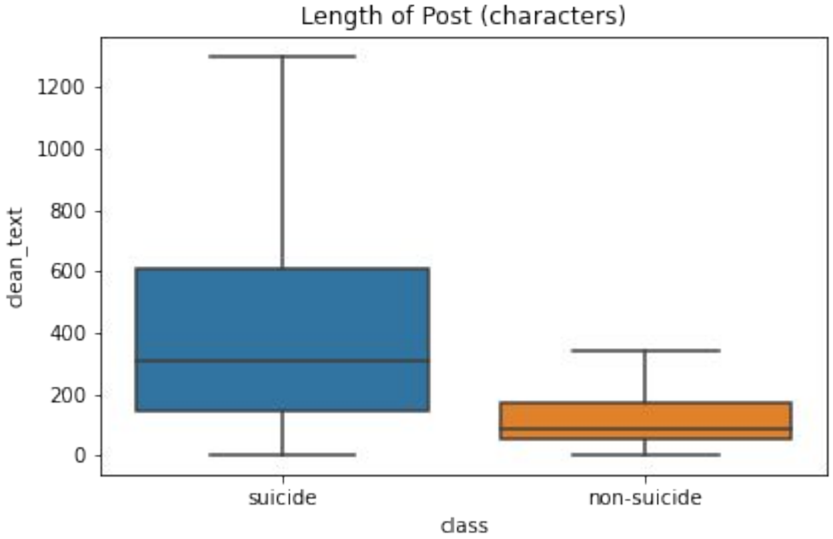
Data Wrangling and Data Cleaning

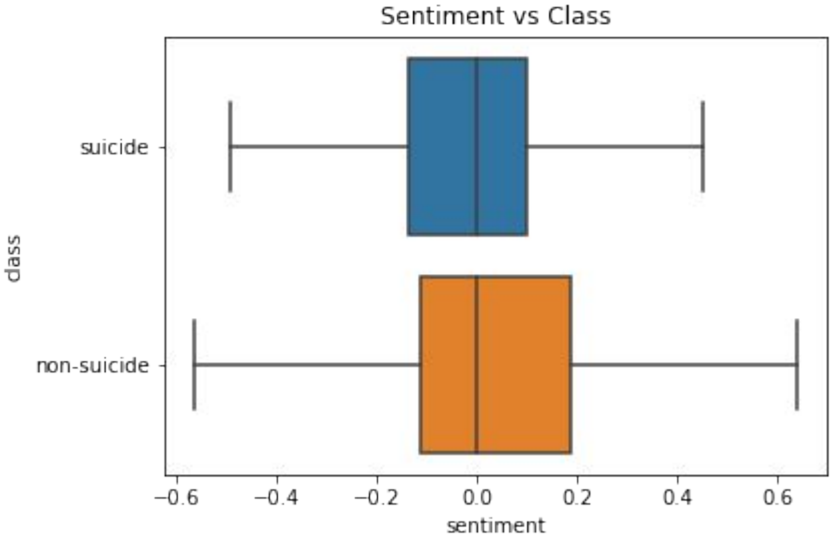
Confirmed there were no duplicate posts and no blank posts (no text)

Normalize Text:

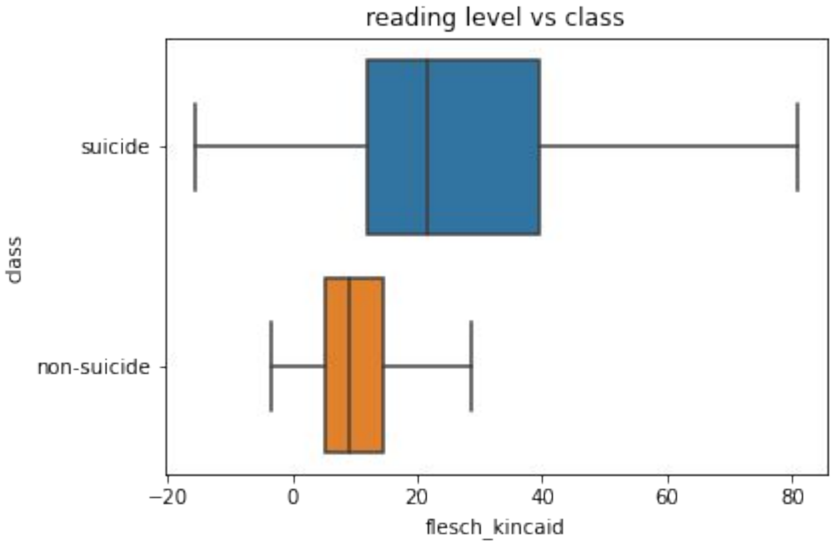
- Converting all text to lowercase
- Converting all hyperlinks and urls to standard text
- Converting all emojis and emoticons to text
- Removing punctuation and numerals
- Removing white spaces
- Ensuring all posts are written in the English language
- Expanding contractions
- Removing stopwords utilizing the spacy dictionary
- Lemmatizing verbs to convert them to their lemma or stem

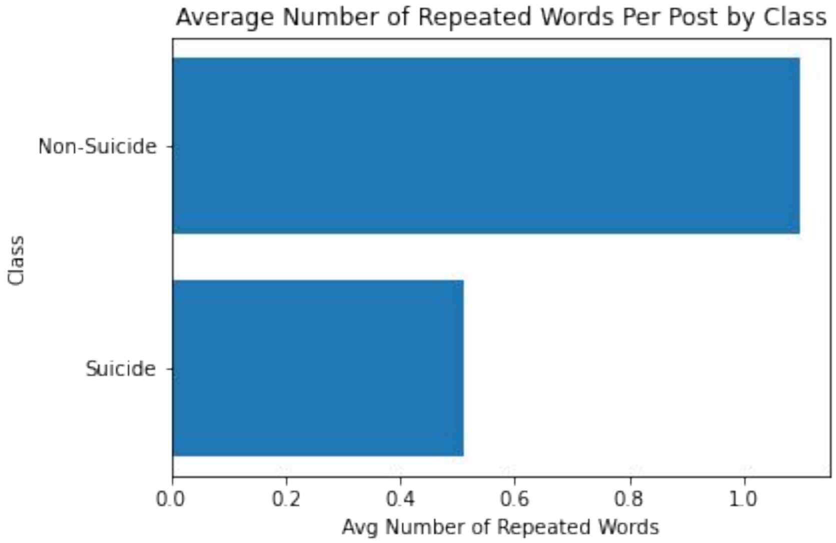
Exploratory Data Analysis


Length of Post (characters)

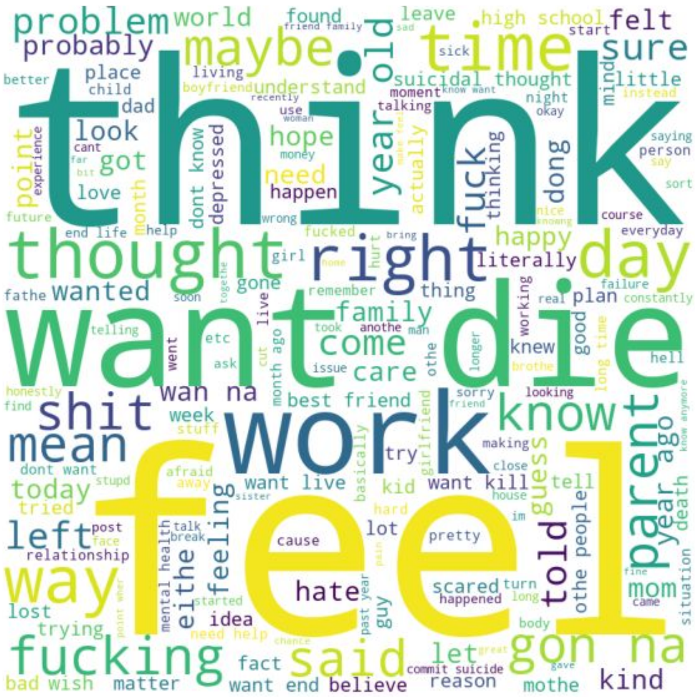

Sentiment vs Class

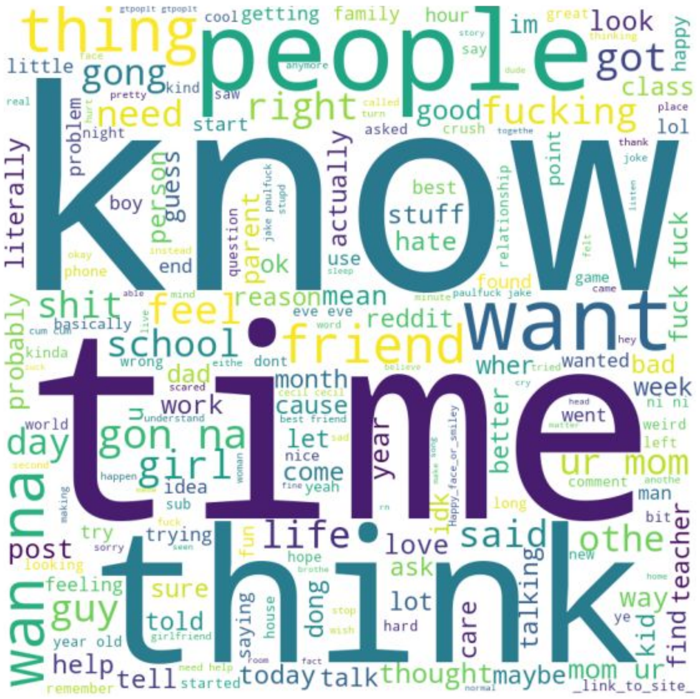
Exploratory Data Analysis - Continued



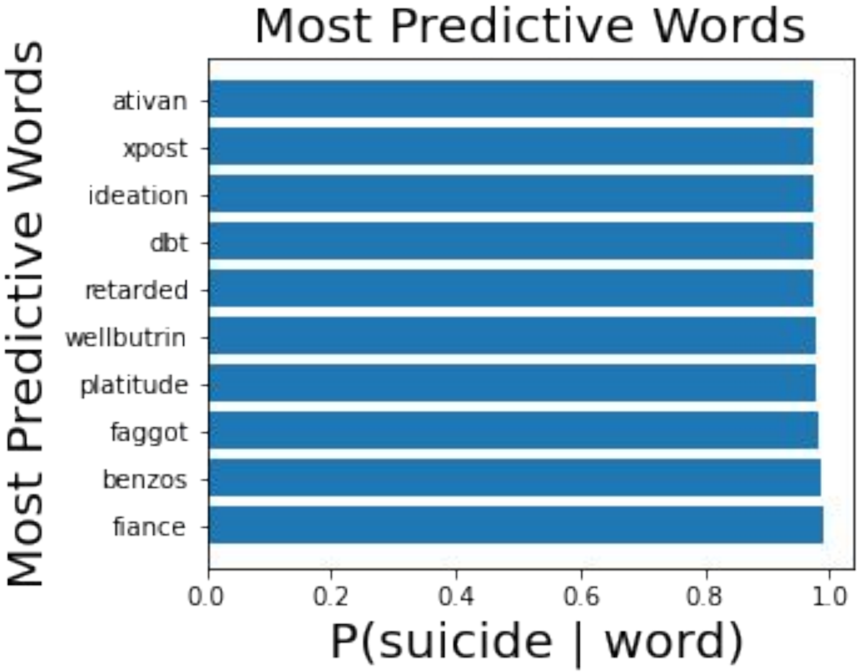


Vocabulary - Frequency





Vocabulary - Predictive



Overall the most predictive words seem to be used to talk about one or more of the following:

- Relationships gone bad (fiance)
- Talking about suicide (xpost, platitudes)
- Treatment for depression (dbt, benzos, ativan, wellbutrin)
- Self loathing (faggot, retarded)
- Suicide attempts (benzos, ativan, wellbutrin, ideation, fiance)

Modeling

Three different machine learning models:

1) Multinomial Naive Bayes
2) Random Forest
3) Logistic Regression

Count Vectorizer and TF-IDF Vectorizer

Text only and with All Features (text PLUS length, reading level, sentiment, repetition)

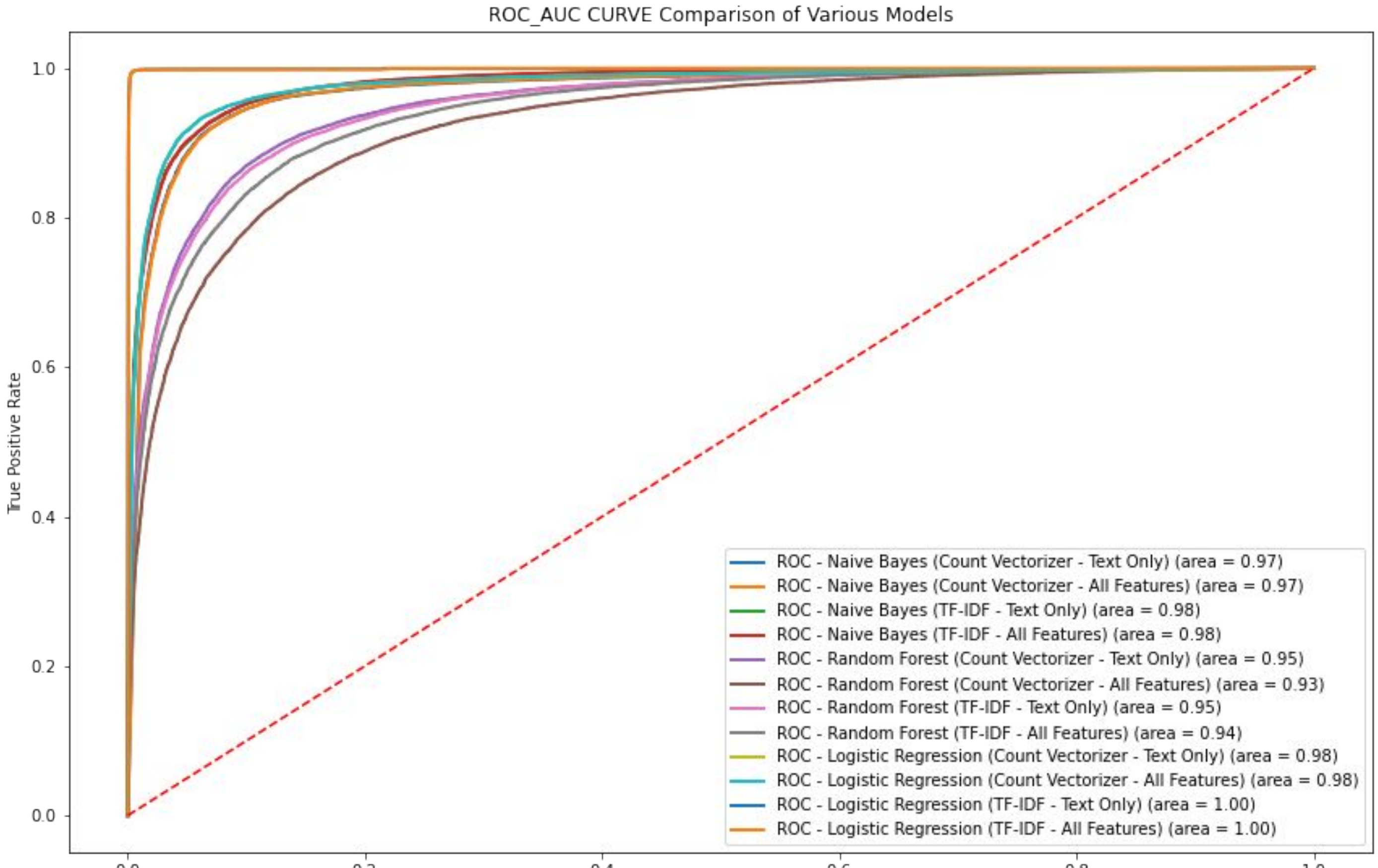
Total of 12 models analyzed

Modeling

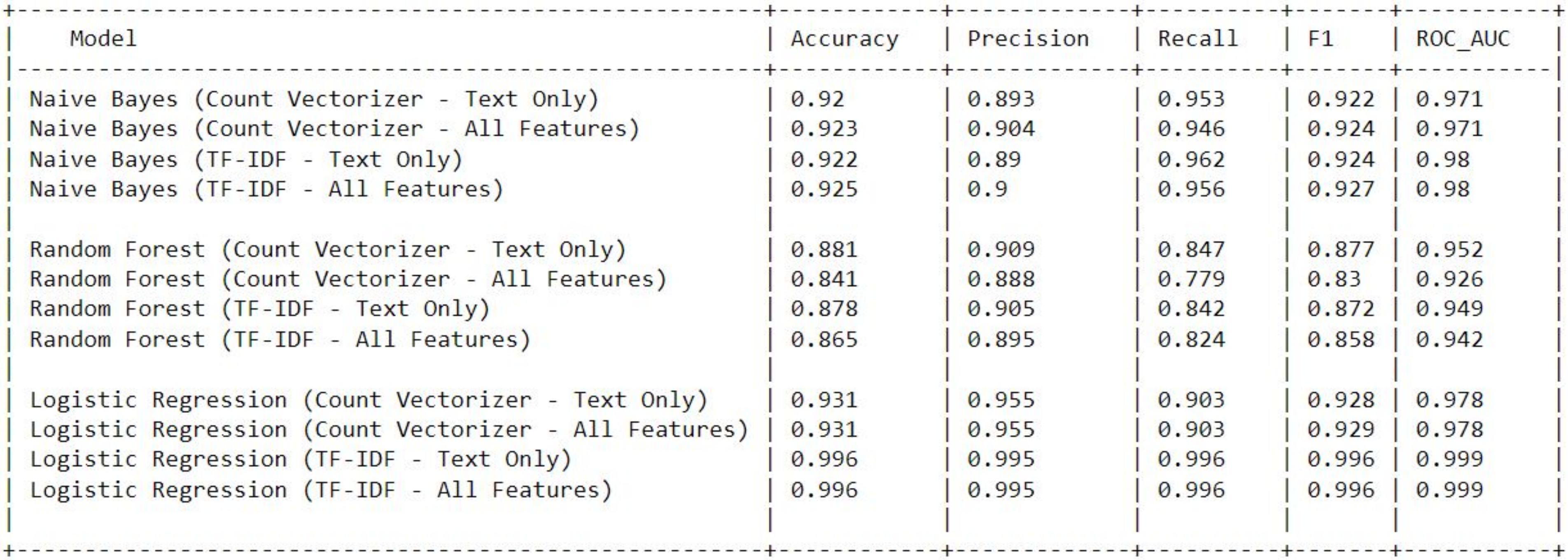
For all models:

- 75 / 25 Train Test Split
- Set min_df = 3
- Set n_grams = (1,2) - analyzed unigrams and bigrams
- Used Grid Search to Optimize Hyperparameters
    - For Multimomial Naive Bayes - ‘Alpha’
    - For Random Forest - ‘Max Depth’ and ‘N-Estimators’ (optimized on a subset)
    - For Logistic Regression - ‘C’

Model Metrics

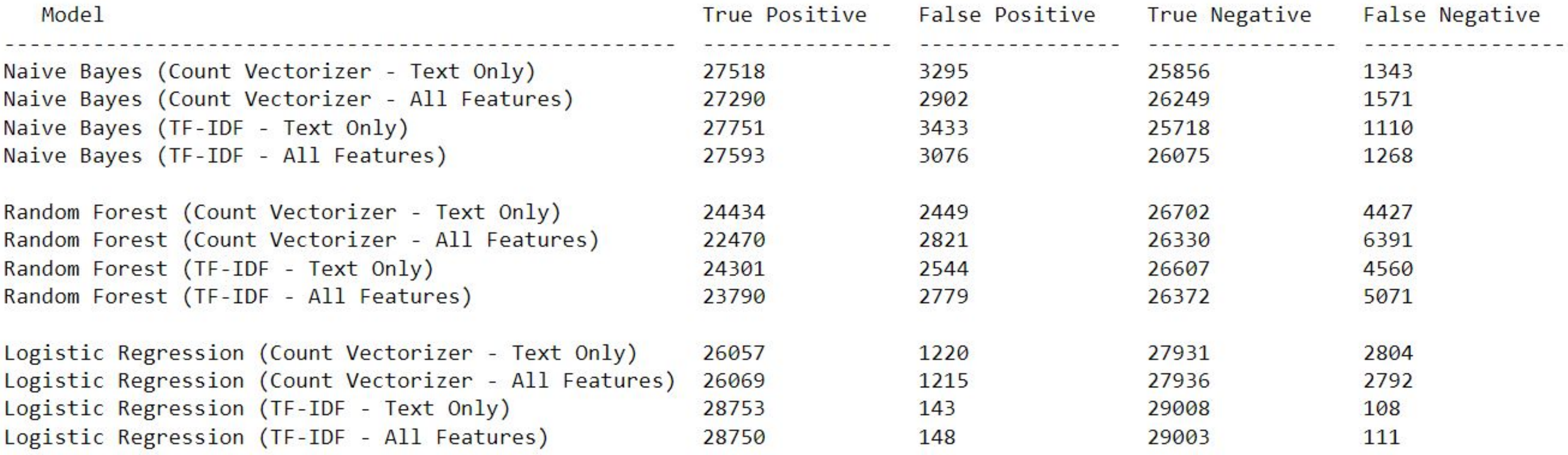


Model Metrics



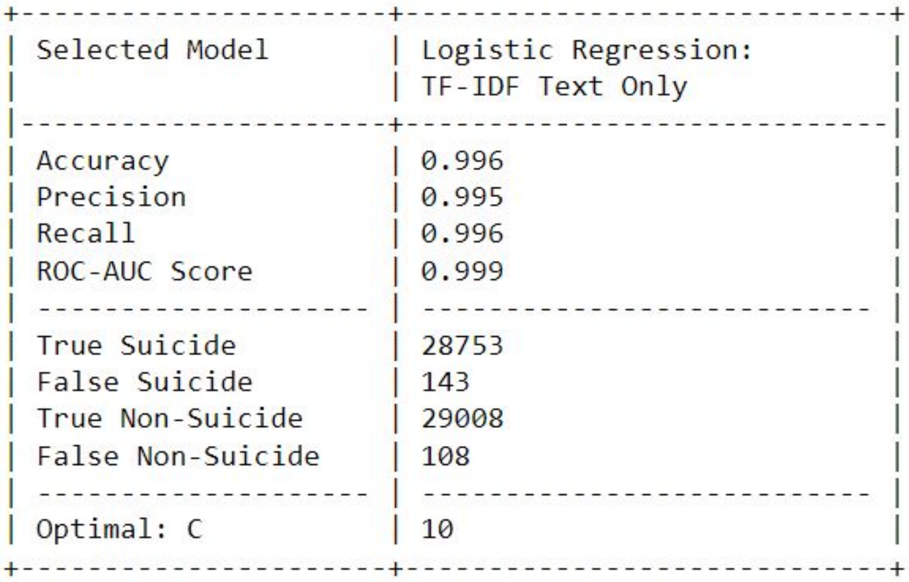
```
+-------------------------------------------------------+------------+-------------+----------+-------+-----------+
|   Model                                               | Accuracy   | Precision   | Recall   | F1    | ROC_AUC   |
|-------------------------------------------------------+------------+-------------+----------+-------+-----------
| Naive Bayes (Count Vectorizer - Text Only)            | 0.92       | 0.893       | 0.953    | 0.922 | 0.971     |
| Naive Bayes (Count Vectorizer - All Features)         | 0.923      | 0.904       | 0.946    | 0.924 | 0.971     |
| Naive Bayes (TF-IDF - Text Only)                      | 0.922      | 0.89        | 0.962    | 0.924 | 0.98      |
| Naive Bayes (TF-IDF - All Features)                   | 0.925      | 0.9         | 0.956    | 0.927 | 0.98      |
|                                                       |            |             |          |       |           |
| Random Forest (Count Vectorizer - Text Only)          | 0.881      | 0.909       | 0.847    | 0.877 | 0.952     |
| Random Forest (Count Vectorizer - All Features)       | 0.841      | 0.888       | 0.779    | 0.83  | 0.926     |
| Random Forest (TF-IDF - Text Only)                    | 0.878      | 0.905       | 0.842    | 0.872 | 0.949     |
| Random Forest (TF-IDF - All Features)                 | 0.865      | 0.895       | 0.824    | 0.858 | 0.942     |
|                                                       |            |             |          |       |           |
| Logistic Regression (Count Vectorizer - Text Only)    | 0.931      | 0.955       | 0.903    | 0.928 | 0.978     |
| Logistic Regression (Count Vectorizer - All Features) | 0.931      | 0.955       | 0.903    | 0.929 | 0.978     |
| Logistic Regression (TF-IDF - Text Only)              | 0.996      | 0.995       | 0.996    | 0.996 | 0.999     |
| Logistic Regression (TF-IDF - All Features)           | 0.996      | 0.995       | 0.996    | 0.996 | 0.999     |
|                                                       |            |             |          |       |           |
+-------------------------------------------------------+------------+-------------+----------+-------+-----------+
```

Model Metrics

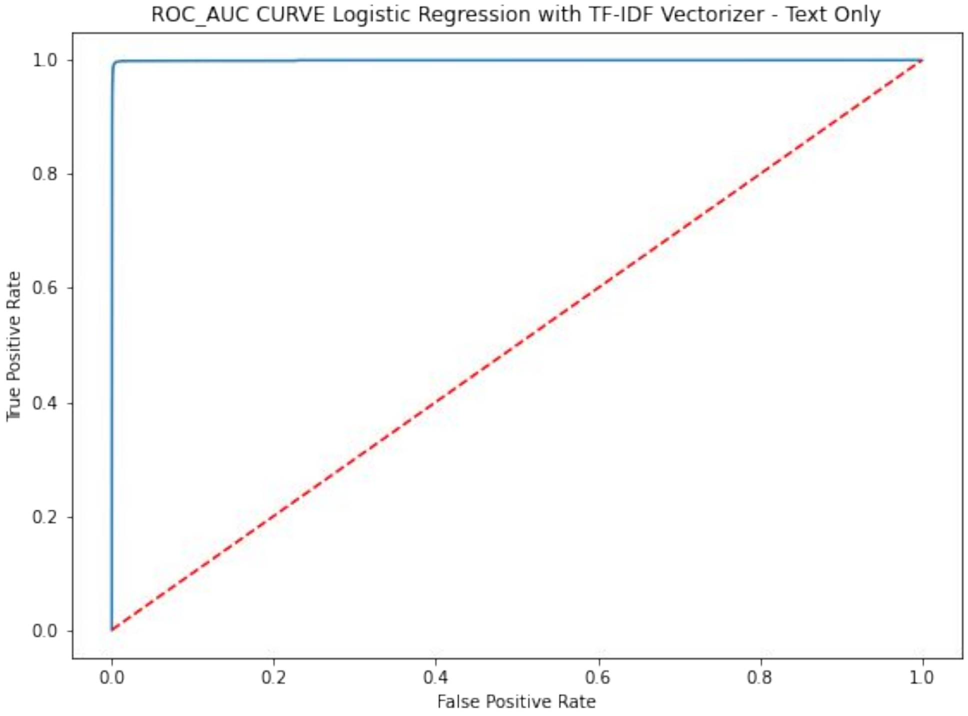


```
   Model                                                True Positive    False Positive    True Negative    False Negative
-----------------------------------------------------  ---------------  ----------------  ---------------  ----------------
Naive Bayes (Count Vectorizer - Text Only)             27518            3295              25856            1343
Naive Bayes (Count Vectorizer - All Features)          27290            2902              26249            1571
Naive Bayes (TF-IDF - Text Only)                       27751            3433              25718            1110
Naive Bayes (TF-IDF - All Features)                    27593            3076              26075            1268

Random Forest (Count Vectorizer - Text Only)           24434            2449              26702            4427
Random Forest (Count Vectorizer - All Features)        22470            2821              26330            6391
Random Forest (TF-IDF - Text Only)                     24301            2544              26607            4560
Random Forest (TF-IDF - All Features)                  23790            2779              26372            5071

Logistic Regression (Count Vectorizer - Text Only)     26057            1220              27931            2804
Logistic Regression (Count Vectorizer - All Features)  26069            1215              27936            2792
Logistic Regression (TF-IDF - Text Only)               28753            143               29008            108
Logistic Regression (TF-IDF - All Features)            28750            148               29003            111
```

Selected Model - Logistic Regression (TF-IDF: Text Only)



```
+----------------------+----------------------------+
| Selected Model       | Logistic Regression:       |
|                      | TF-IDF Text Only           |
|----------------------+----------------------------|
| Accuracy             | 0.996                      |
| Precision            | 0.995                      |
| Recall               | 0.996                      |
| ROC-AUC Score        | 0.999                      |
| -------------------- | -------------------------- |
| True Suicide         | 28753                      |
| False Suicide        | 143                        |
| True Non-Suicide     | 29008                      |
| False Non-Suicide    | 108                        |
| -------------------- | -------------------------- |
| Optimal: C           | 10                         |
+----------------------+----------------------------+
```



Mis-Identified Posts

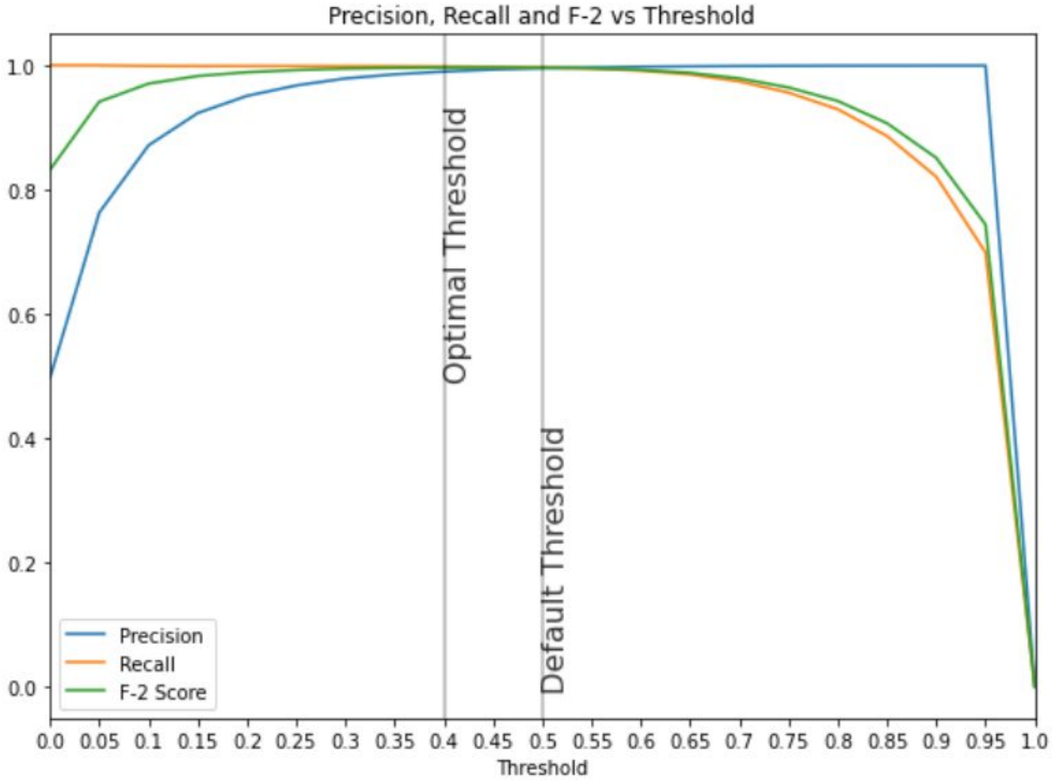
FALSE POSITIVE

- Primarily posts dealing with suicide that were posted in the teenager forum

FALSE NEGATIVE

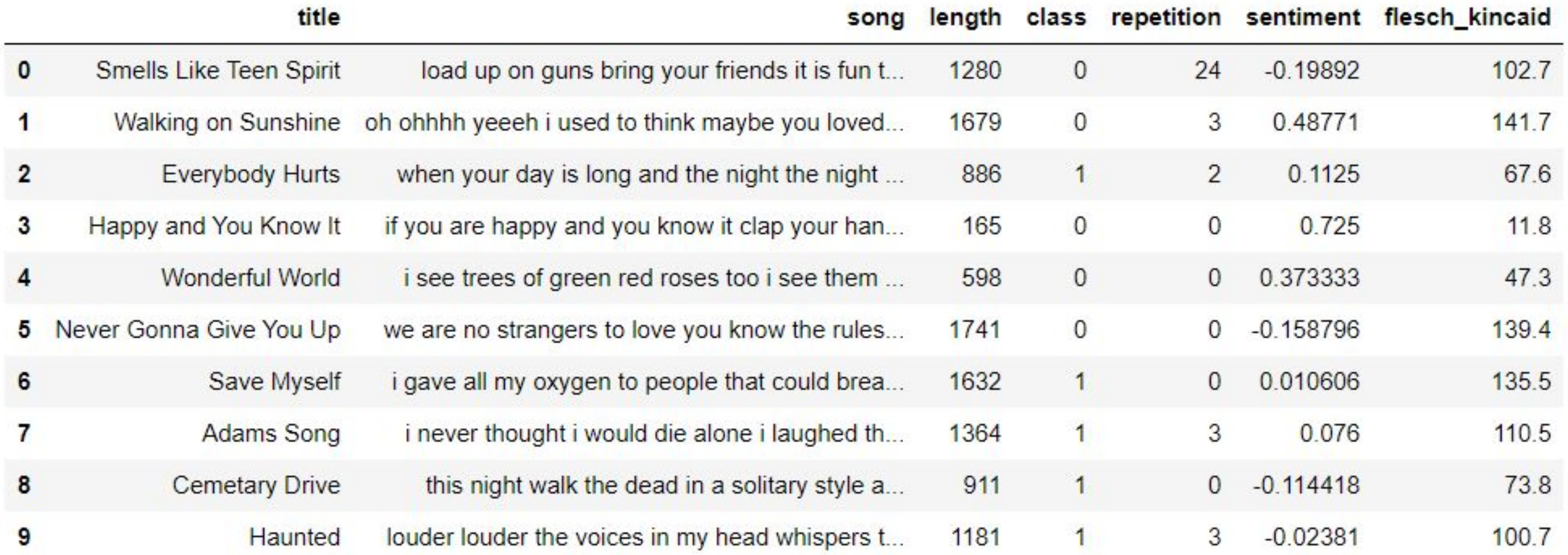
- Mostly very short texts
- Not explicit about suicide / intention
- Included misspelled words / lack of spacing between words

Thresholding

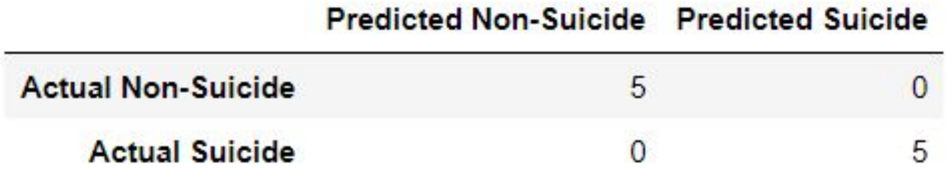


- Identify Threshold so that F-2 score is maximized
- Recall increases
- Precision decreases
- False Negatives decrease

Evaluate Text From a Different Source



	title	song	length	class	repetition	sentiment	flesch_kincaid
0	Smells Like Teen Spirit	load up on guns bring your friends it is fun t...	1280	0	24	-0.19892	102.7
1	Walking on Sunshine	oh ohhhh yeeeh i used to think maybe you loved...	1679	0	3	0.48771	141.7
2	Everybody Hurts	when your day is long and the night the night ...	886	1	2	0.1125	67.6
3	Happy and You Know It	if you are happy and you know it clap your han...	165	0	0	0.725	11.8
4	Wonderful World	i see trees of green red roses too i see them ...	598	0	0	0.373333	47.3
5	Never Gonna Give You Up	we are no strangers to love you know the rules...	1741	0	0	-0.158796	139.4
6	Save Myself	i gave all my oxygen to people that could brea...	1632	1	0	0.010606	135.5
7	Adams Song	i never thought i would die alone i laughed th...	1364	1	3	0.076	110.5
8	Cemetary Drive	this night walk the dead in a solitary style a...	911	1	0	-0.114418	73.8
9	Haunted	louder louder the voices in my head whispers t...	1181	1	3	-0.02381	100.7



	Predicted Non-Suicide	Predicted Suicide
Actual Non-Suicide	5	0
Actual Suicide	0	5

Summary

Evaluated 12 Models

Best model was Logistic Regression (TF-IDF Vectorizer - Text Only)

ROC-AUC Score = 0.999

Recall = 0.996 - Lowest Number of False Negative Predictions

Summary

Model seems to perform well enough to be useful.

Additional input on performance from potential stakeholders.

These include Social media companies (schools, parents), online therapy providers such as Betterhelp, 7 cups, Talkspace.

Next Steps

- Conduct additional analysis between more similar groups.
- Engineered features were of limited value - include other features?
- False negatives had misspellings / lack of spacing between words. Look into ways to correct these prior to prediction.
- More robust hyperparameter optimization.