From Depression To Suicide, How The Way We Speak Predicts The Way We Feel

By Yoni Levine

Support Community

Take this life and improve it.

OVERVIEW

1. PROBLEM STATEMENT
2. DATA ACQUISITION
3. METHODS
4. FINDINGS
5. MOVING FORWARD
6. QUESTIONS

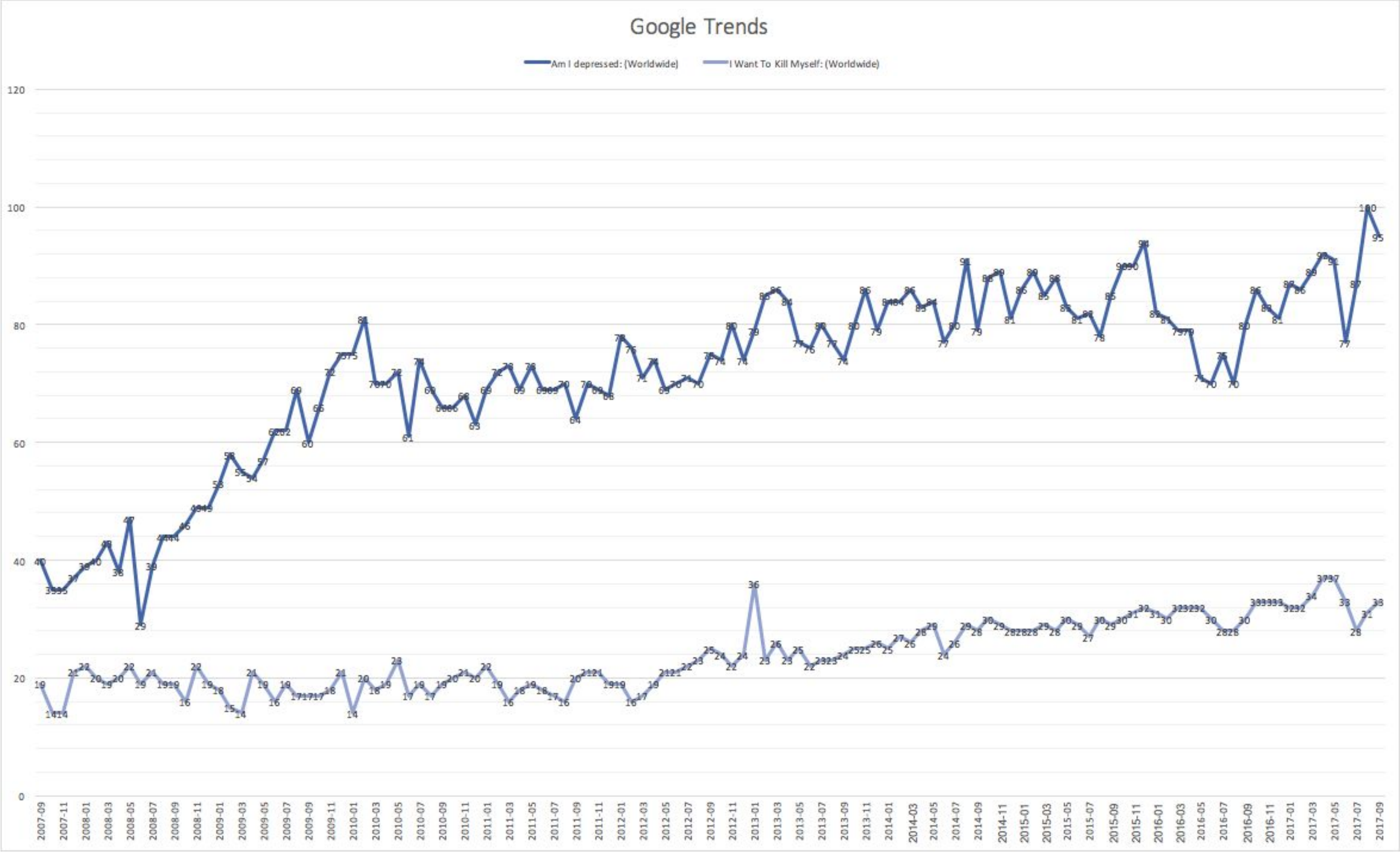
Problem Statement

The goal of my project was to determine if I could classify which users of a depression forum, would eventually become suicidal, based on the words that they used when describing their depression

Depression is a very broad ranging diagnoses and allowing healthcare professionals to determine which cases need the most attention, without waiting until things become very serious would be super useful to the field.

How common is depression?

- An NIMH study states that **nearly** 7 percent of US adults, and over ten percent of those between the ages of 18-25, have had a major depressive episode in the last year.
- According to the CDC less than one-third of Americans taking one antidepressant medication and less than one-half of those taking multiple antidepressants have seen a mental health professional in the past year.
- There were more than twice as many suicides (44,193) in the United States as there were homicides (17,793).
- Up to 9 % of people that have been diagnosed with depression in their lifetime will go on to complete suicide



And unfortunately it's still on the rise.

Getting My Data

- I scraped takethislife.com a forum for people dealing with depression and suicidal tendencies.
- There were a total of nearly 25,000 posts across the two forums.
- There were 5,719 users unique to the depression forum, 1093 that overlapped, and 2,210 that were unique to suicide.
- I also scraped the number of posts that a user posted, as well as the number of swedish fish.
- I worked exclusively with the text from the depression and appended a dummy-variable if that user had also posted in suicide.
- In total there were 17,400 posts and half of them were written by users who had also written in the suicide forum.

Methods
NLP

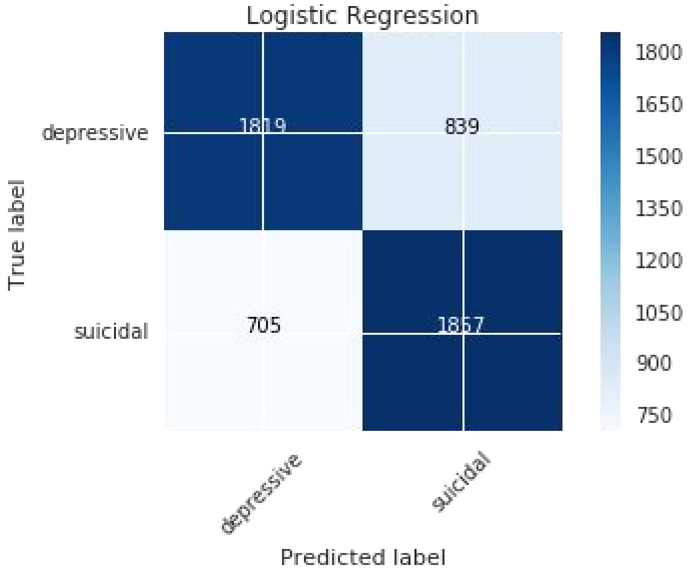
- Used NLTK for preprocessing, I lemmatized my data but left in the stop words in for analysis.
- I used TF-IDF to see if we could compare what words are more unique to each group as opposed to how many times a group says a word.
- Used n-grams ranges from 1-4 and took all numbers out of my text

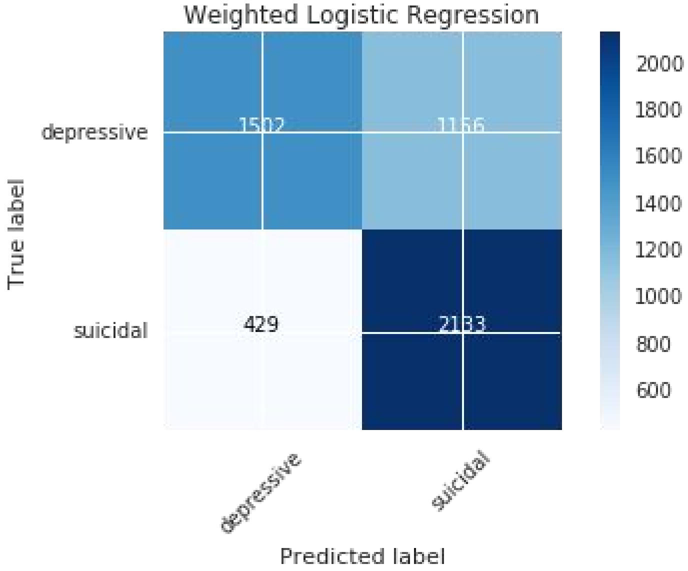
Modeling

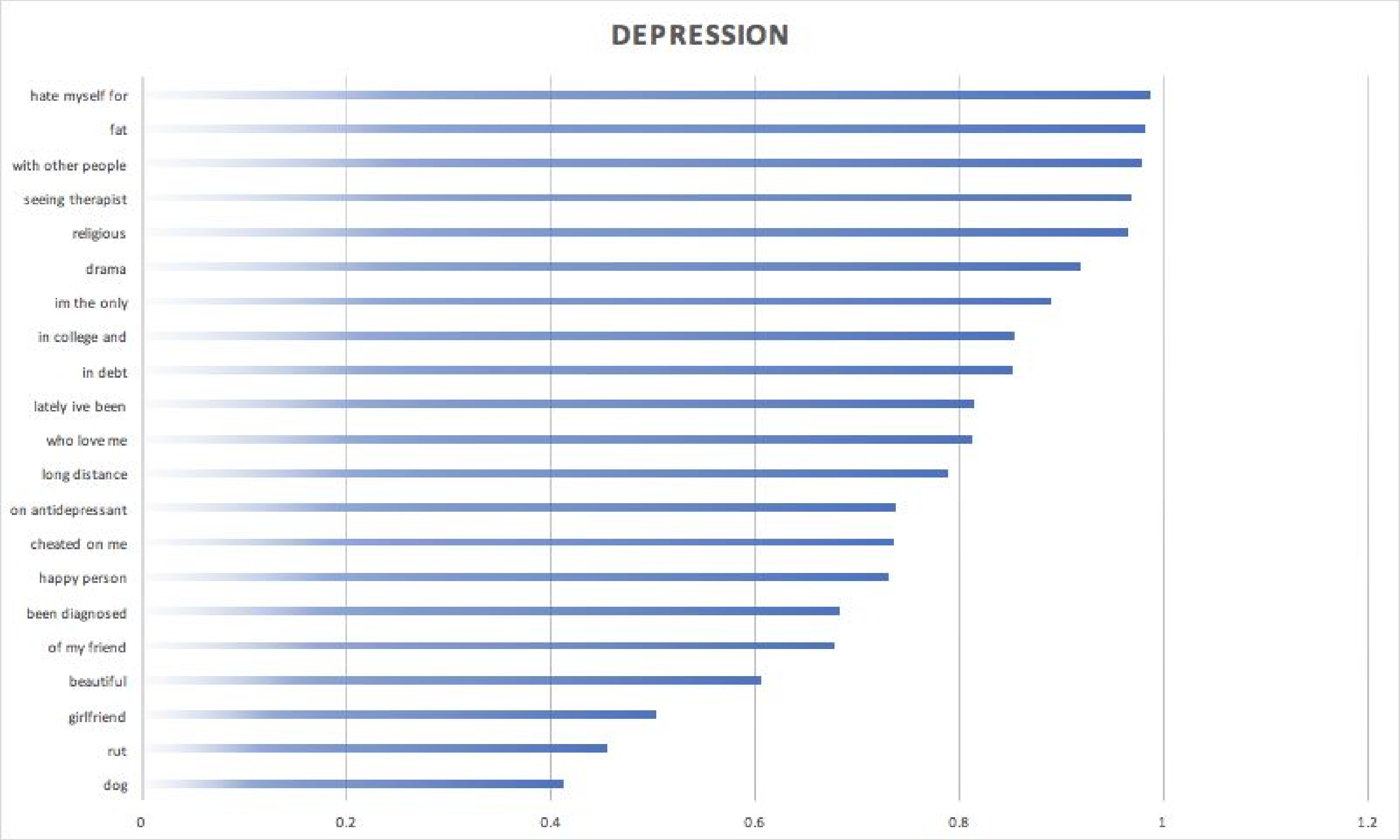
- I tried out several different models that generally work well with text and ultimately went with logistic regression, my model got an ROC-AUC score of 78 with a recall of 72.
- I had to decide if I wanted to weigh my classes differently in order to achieve maximum recall in my suicide class, even at the risk of misclassifying non-suicidal people.
- I made an additional model where I weighted my models probability of predicting suicide at 1.5 as probable compared to depression, this model had the same ROC-AUC score as my previous one but the recall was 83.

Logistic Regression

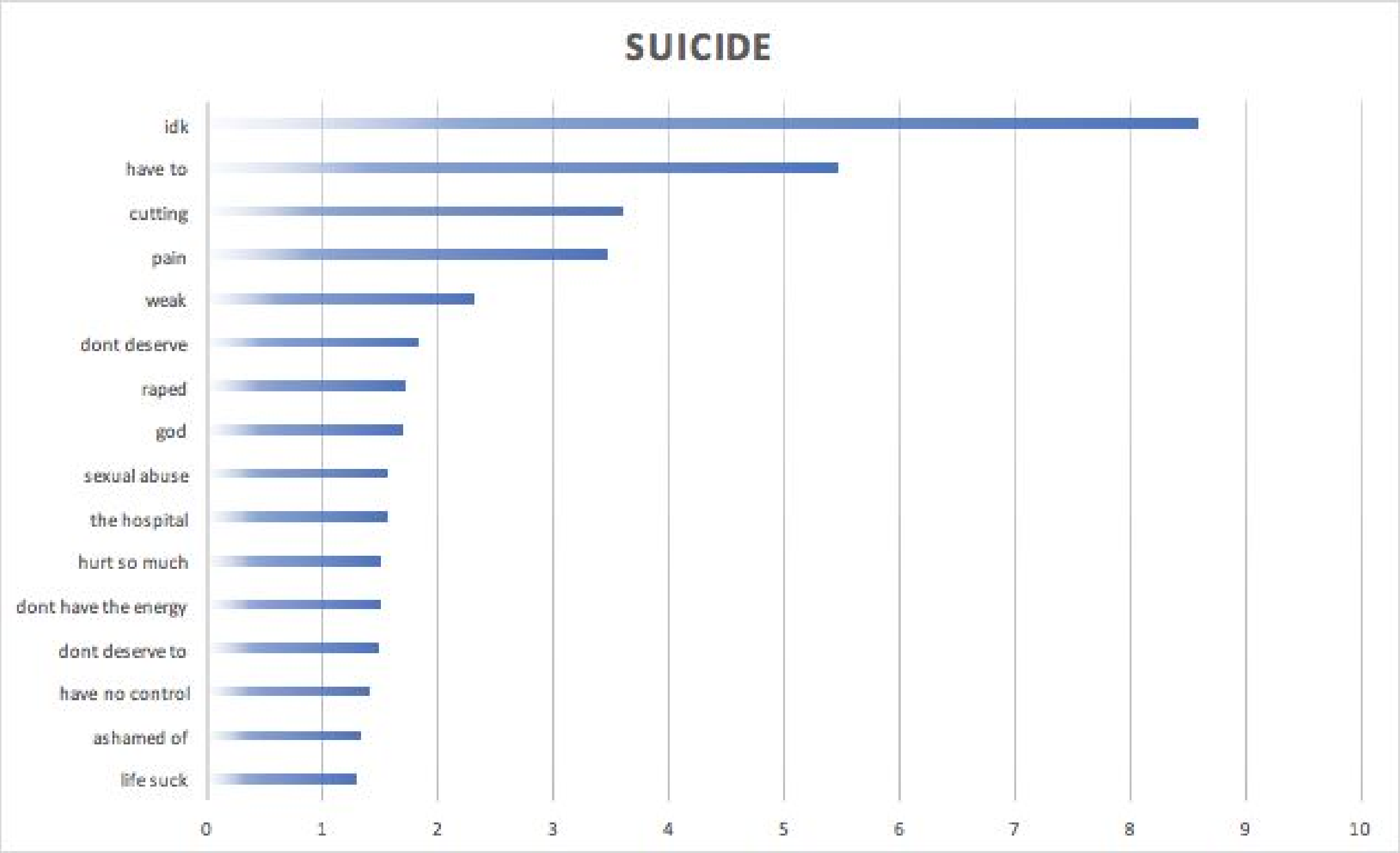
Weighted Classes







Depressive Speech Feature Importance



Suicidal Speech Feature Importance

Moving Forward

- I would like to get a hold of both industry specific stop-words as well as look into an informal texting stop-words list
- Try tagging my data to different parts of speech and see if I could infer something from that
- Continue to train my model on other corpora from sites such as Redit
- Further investigate which posts my model misidentified and try to discern why
- Try to get my model into a deployable form

Thank you for joining me.

Any Questions ?