Classifying Question Types

Jonathan Andrews

Building a better web browser

Browsers use web scraping to search through html for websites with things similar to words and sentiment in your search.

There might be some kind of benefit for a web browser to be able to analyze the question or words being passed into it to determine the type of search result that you get.

These question types shall be split into general inquiry, and emotional inquiry.

A general inquiry will be a question that is just for interest as opposed to searching for personal responses.

An emotional inquiry will be a question that searches for an answer to the persons question related to themself

Problem Statement

Is it possible to classify inquiry type in order to begin to consider making a web browser based on inquiry type instead of relationship to key words.

Data Collection

Data is from

https://www.reddit.com/r/askreddit, general Inquiry

https://www.reddit.com/r/Explainlikeimfive, general inquiry

https://www.reddit.com/r/AmItheAsshole/, emotional inquiry

Types of classification

Naive Bayesian Classification

Logistic Regression Classification

Random Forest Classification

Naive Bayesian Classification

Very good classifier but assumes that the data is normally distributed and that all information is unrelated to other pieces of information.

It also has a high bias which is useful for this type of information as text information can become high variance very quickly

For this dataset the data is generally normally distributed in length and character count but the information due to it being speech based is inherently related to itself.

Nonetheless Bayesian Classification is able to give good results on this data classifying correctly on the testing dataset 99.896% of the time

Logarithmic Regression Classification

Logarithmic Regression would be my next choice for this type of classification problem for the same reasons that we used naive bayesian classification in that it is high bias which can correct for the inherent variance of natural language processing.

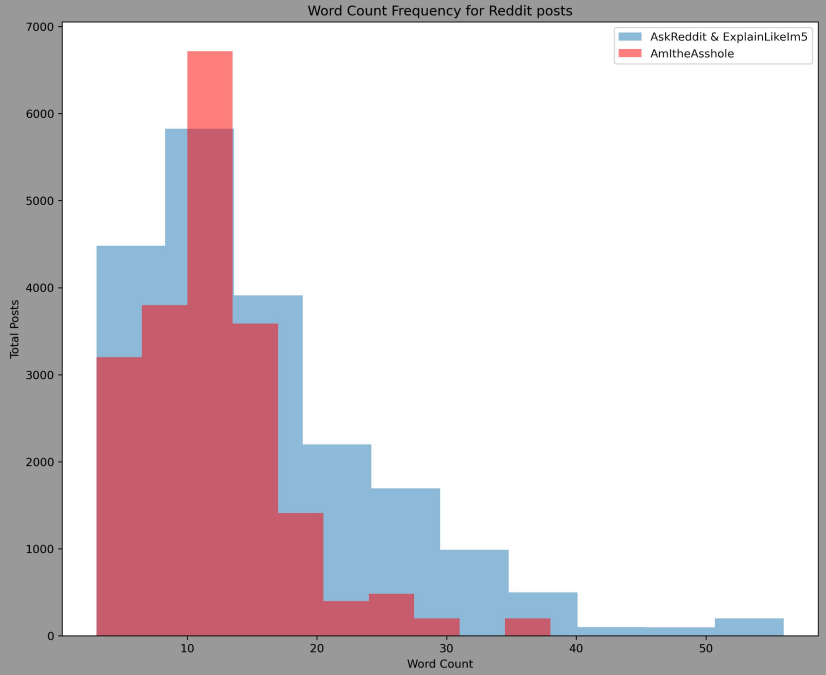
I also used Logarithmic Regression because you can get the coefficients out of it easily which makes it more manageable when drawing conclusions from the data.

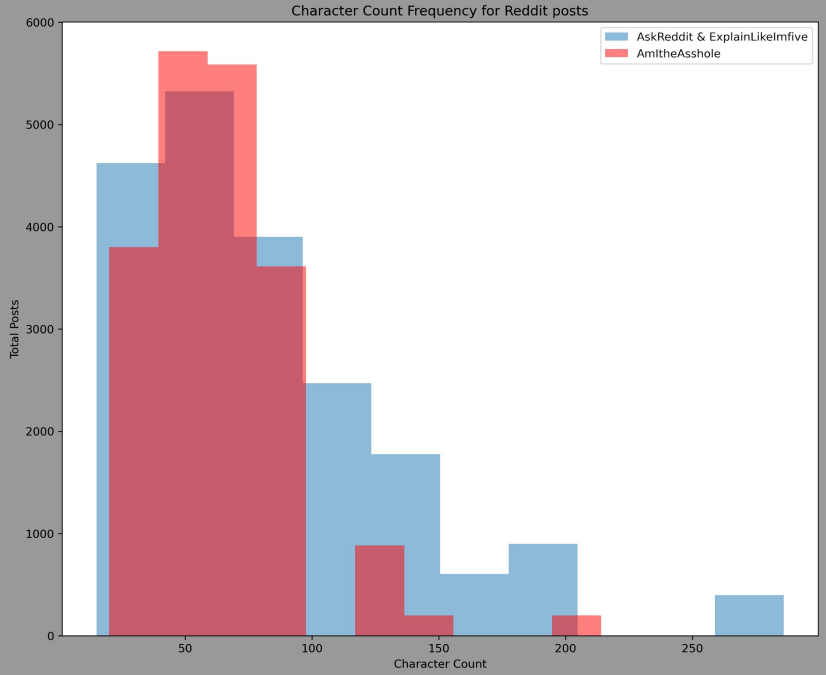
Random Forest Classifier

Random forest is being used as a contrast to the other classifiers as it is moderate variance and low bias.

This is being used to see if there is anything in the data that is making it to simple for the algorithm to correctly predict which of the classes to place our questions into

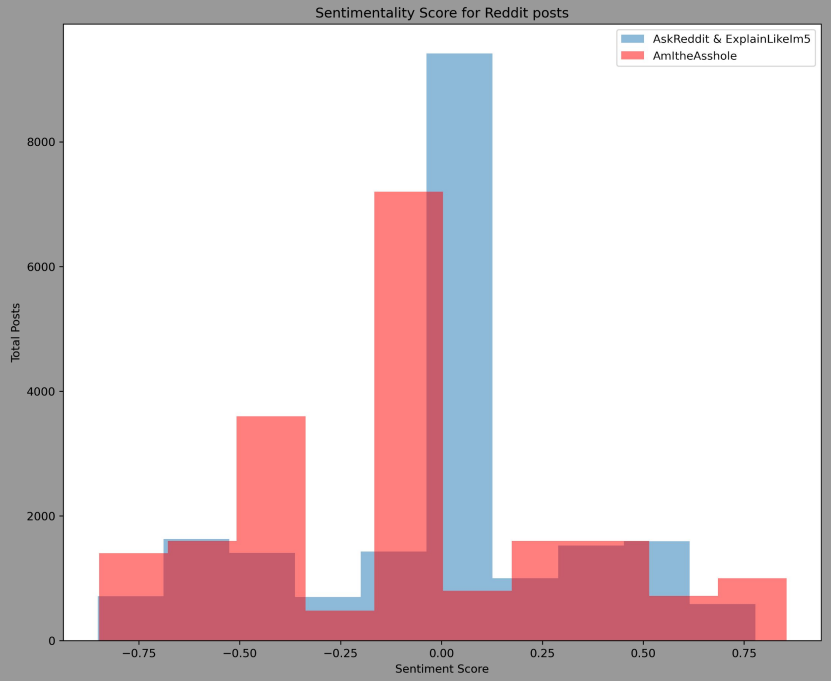
Gaussian Distribution





Non-Gaussian Distribution

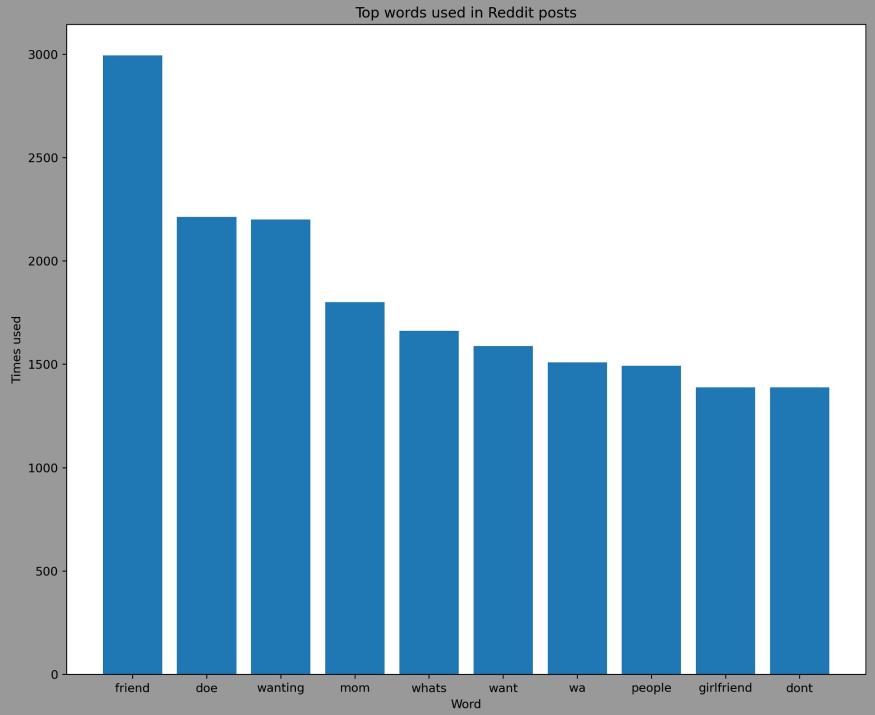
Due to the nature of emotional inquiries they tend to be varied in their sentimentality score to have more of an aggressive or positive sentiment than inquiry based questions



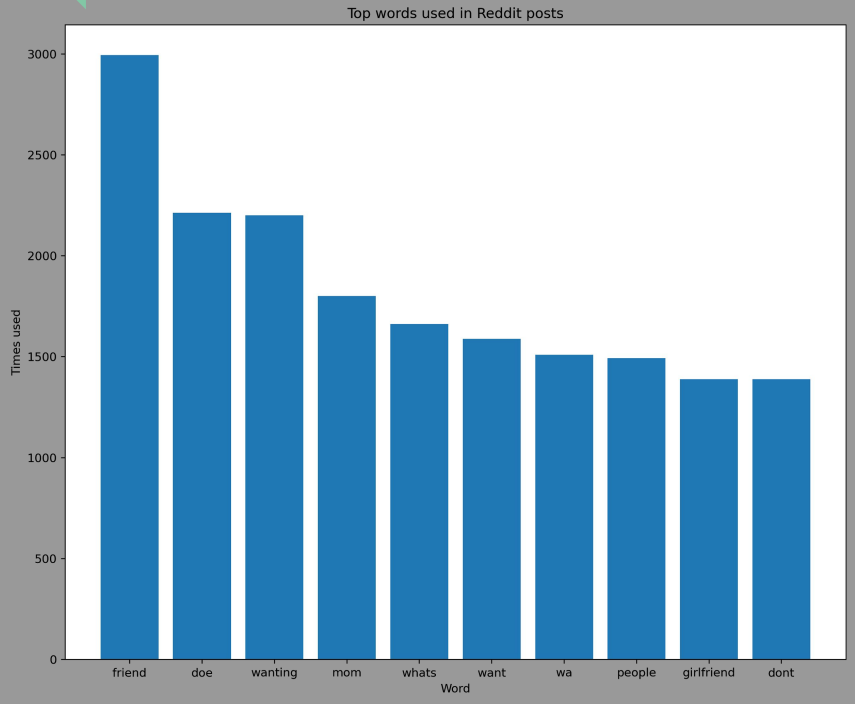
Word Density

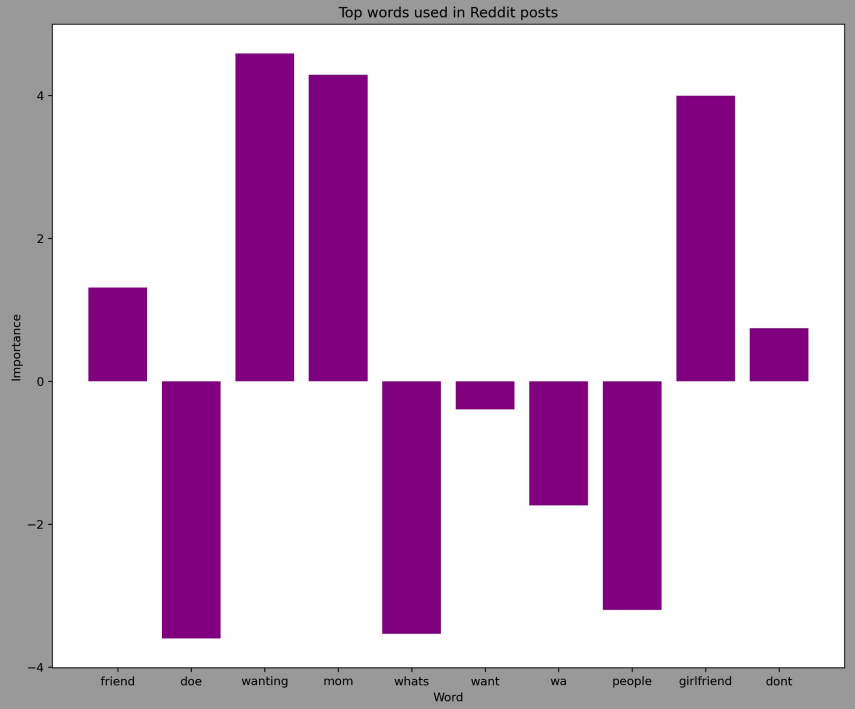
In general the most dense words were words that had little implication on the final answer though strangely friend, mom, girlfriend, and wanting are all very high importance words that ended up being used quite often.

It probably has to do with the subreddit chosen and the exact type of demographic they serve or allow on their subreddit

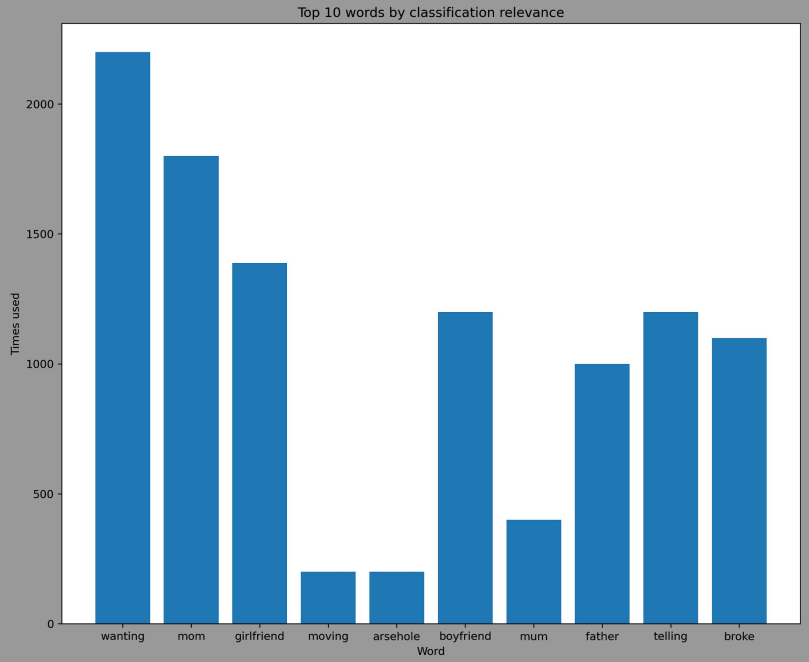


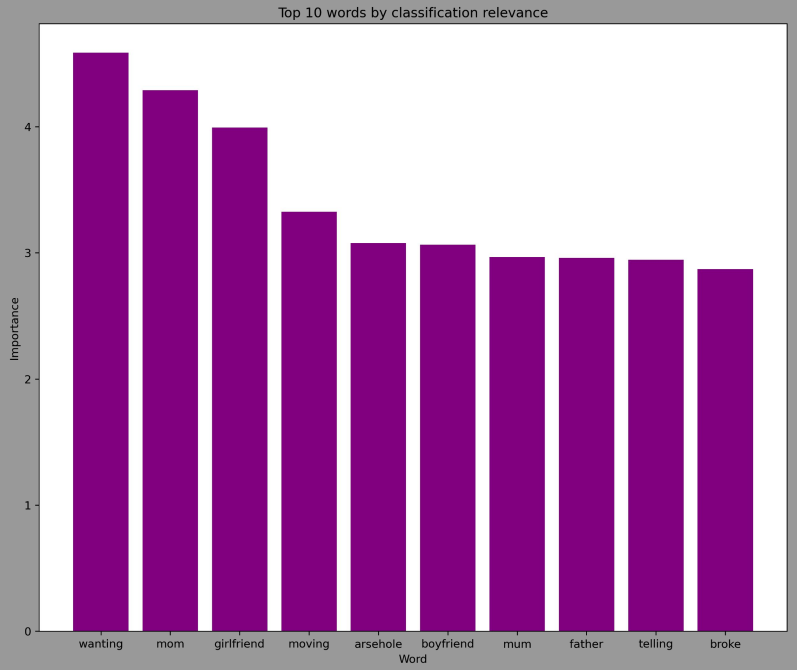
Density vs Importance





Most relevant words





Ways to improve the data

I would create more classes such as informational inquiry, based on the idea of just gaining general information about a subject or pure answer inquiry for questions that only want a specific answer instead of a general explanation.

The data is very difficult to judge because of the moderation and rules placed on any question asked on reddit so there is quite a bit of restriction to real world questions that people would ask a web browser.

In order to improve the data more subreddits should be surveyed and probably expanded beyond pure question asking subreddits to questions in every subreddit so you can get a more varied sample