Evaluating Text with Natural Language Processing

By: Jordan Gates

Problem Statement

As a manager of multiple similar subreddits, how can you identify users who would be interested in one of the other subreddits that you manage? The ability to identify posts that would be better suited for another subreddit that you manage would help grow your communities and help users find more relevant info.

For this presentation we will be using posts from the ‘MTB’ and ‘gravelcycling’ subreddits. These subreddits are intended for mountain biking and gravel cycling enthusiasts. Because these categories are very similar and have overlapping communities, it should be a good example of how we can leverage NLP to identify the differences between the two.

Gravel Bike

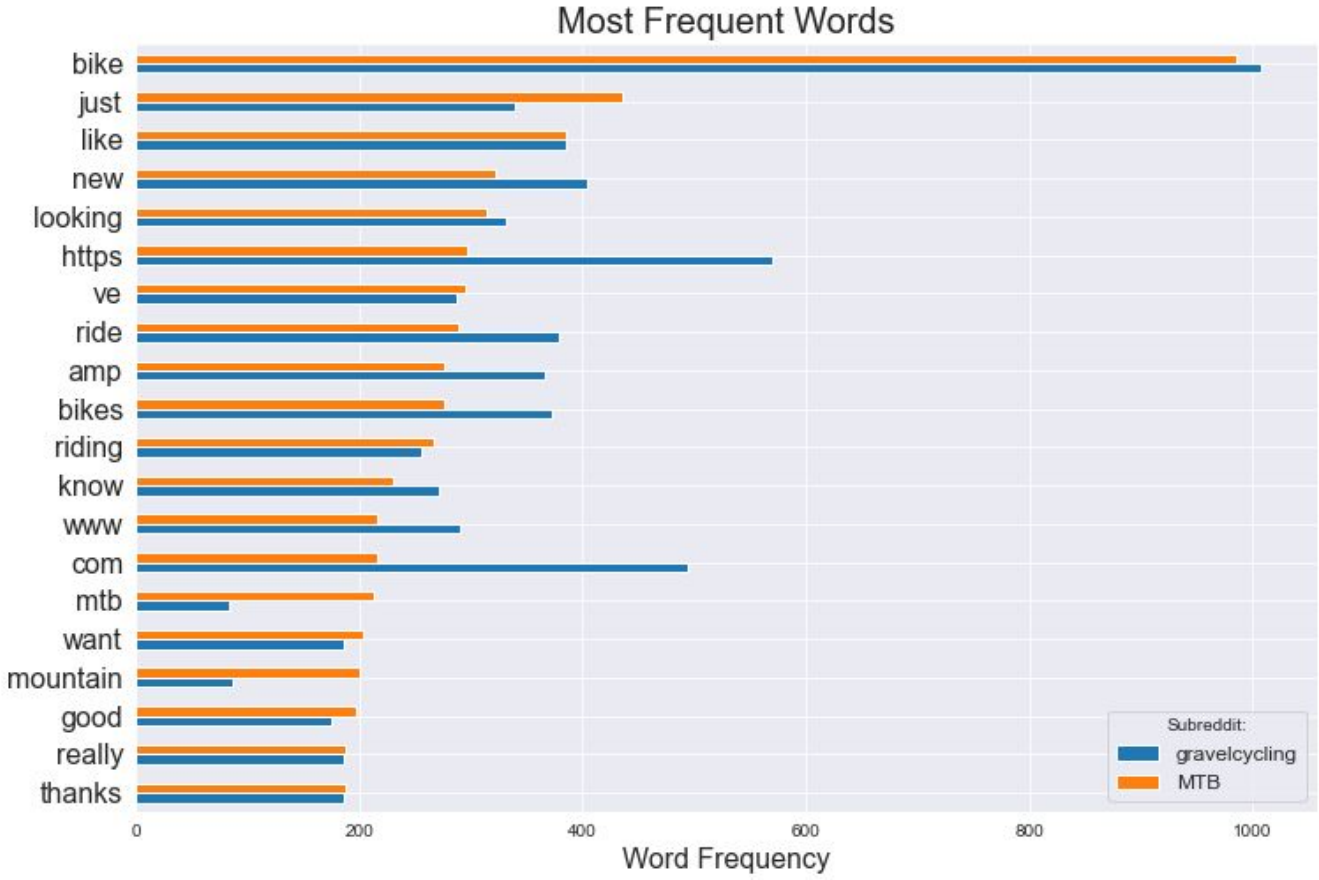
Mountain Bike

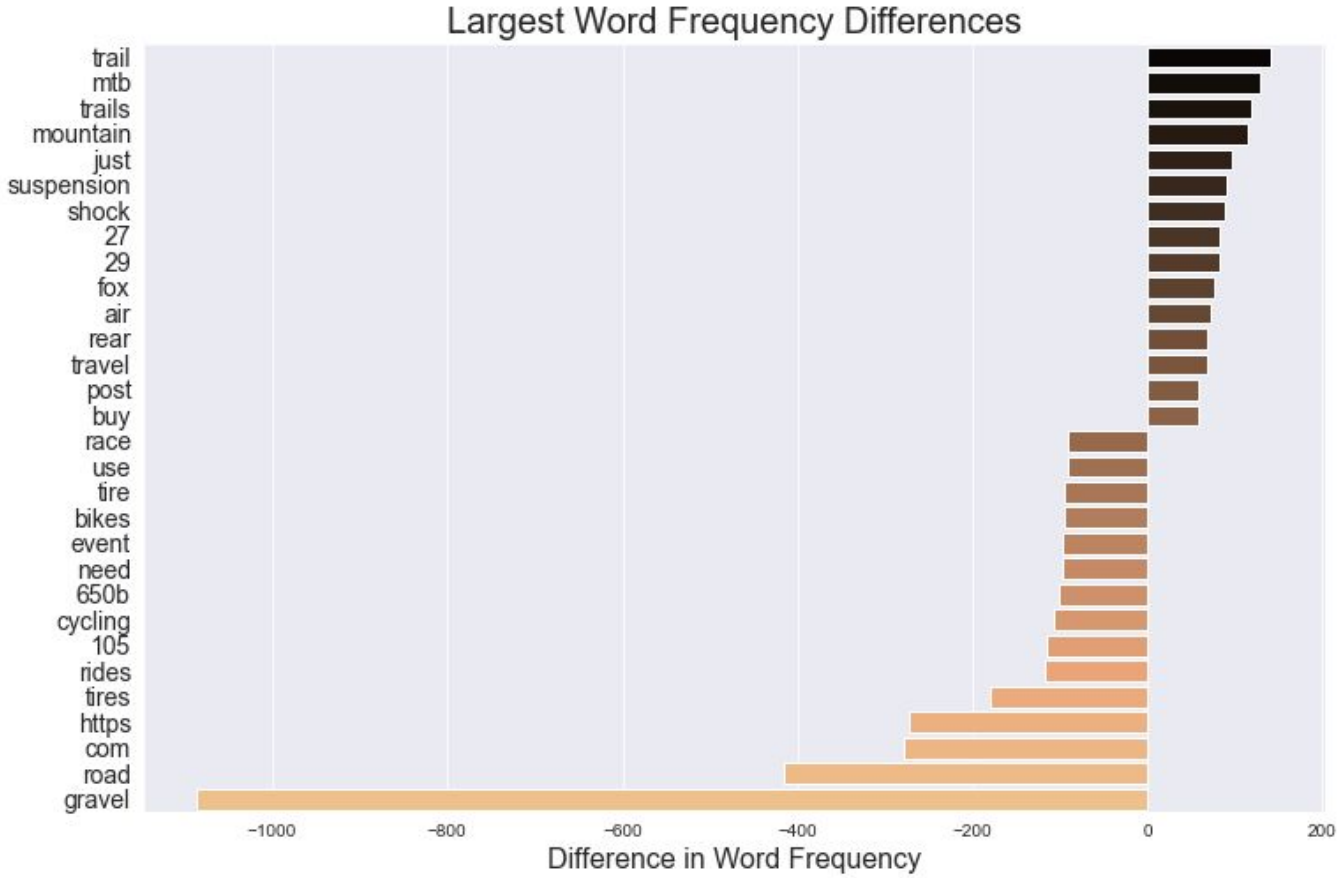
After scraping the data from each subreddit there were 1000 posts from each subreddit, giving us a **baseline accuracy of 50%** (Eg. If we predict that every post is from the ‘gravelcycling’ subreddit, we would be right 50% of the time.)

In order to outperform this baseline I started by using the TF-IDF Vectorizer to format the text from each post in a way that is meaningful to the Logistic Regression model that I started with.

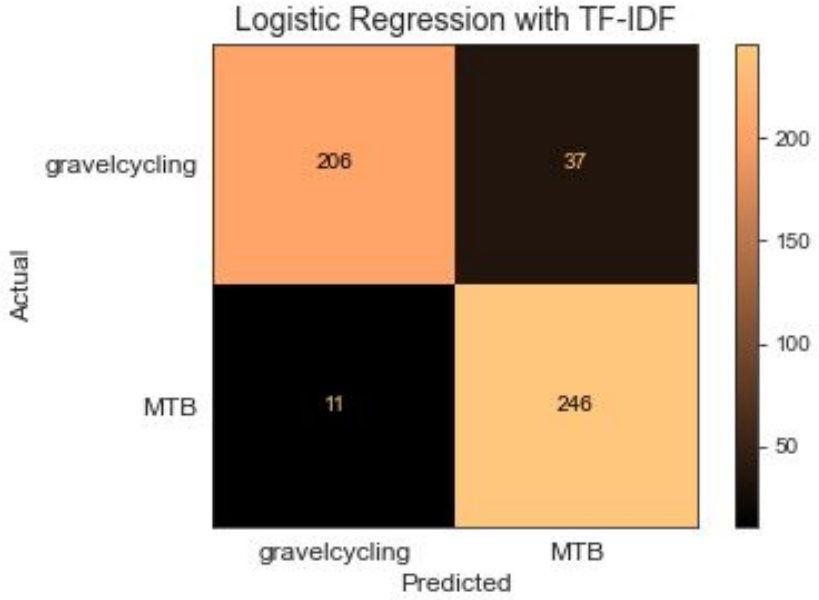
The result was an accuracy score of 89% on testing data - a noticeable improvement over the baseline accuracy of 50%.

But can this be improved?


Most Frequent Words
bike
just
like
new
looking
https
ve
ride
amp
bikes
riding
know
www
com
mtb
want
mountain
good
really
thanks
0
200
400
600
800
1000
Word Frequency
Subreddit:
gravelcycling
MTB

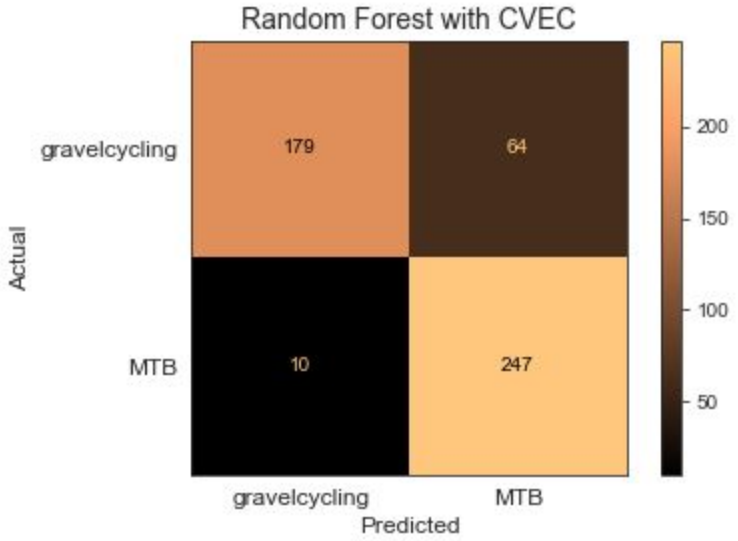

Largest Word Frequency Differences
trail
mtb
trails
mountain
just
suspension
shock
27
29
fox
air
rear
travel
post
buy
race
use
tire
bikes
event
need
650b
cycling
105
rides
tires
https
com
road
gravel
−1000
−800
−600
−400
−200
0
200
Difference in Word Frequency

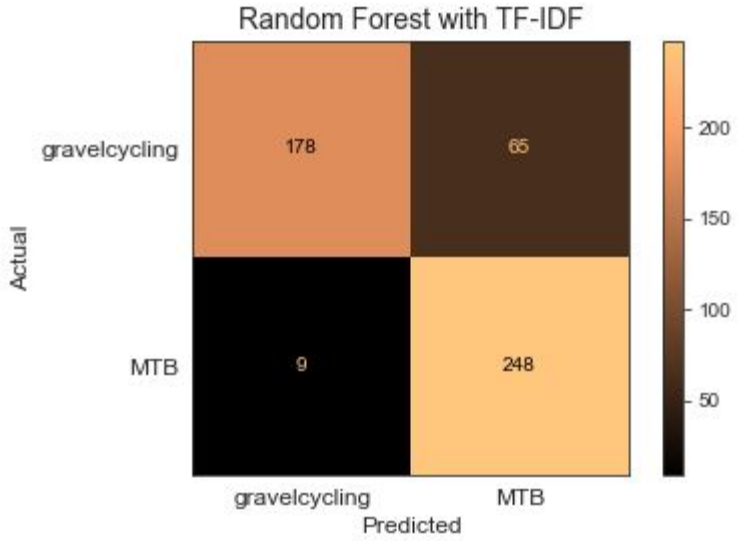
By using the additional stopwords that were selected from the previous visualizations, the first model's accuracy improved from 89% to 90.4%



We can see that there were 37 instances where the model incorrectly predicted that a post was from the MTB subreddit, and 11 instances where it incorrectly predicted that a post was from the gravelcycling subreddit.

How is this information useful?





The Random Forest models also did well with the CVEC and TF-IDF Vectorizers, with accuracy scores around 85%. Both of these models tended to predict that a post was from the ‘MTB’ subreddit more frequently than the alternative.

Summary

The most accurate of the models that were tested was the Logistic Regression model with TF-IDF Vectorizer, coming in at 90.4% accuracy - this would be the production model.

By examining the incorrect predictions of this model, you have a much shorter list of posts to review that are the most likely to be relevant to the other subreddit.

Moving forward, some potential improvements to this program could be:

- Using not only the post content, but the title text and comments as well
- Implementing OCR to analyze any photos that have text in them
- Creating a csv file that contains the incorrectly predicted posts, making it easy to look through them and decide which ones would be better suited for another subreddit.