Predicting Subreddit Posts Based on Titles

Erik Lindberg

General Assembly Project 3
8/28/2020

Problem Statement

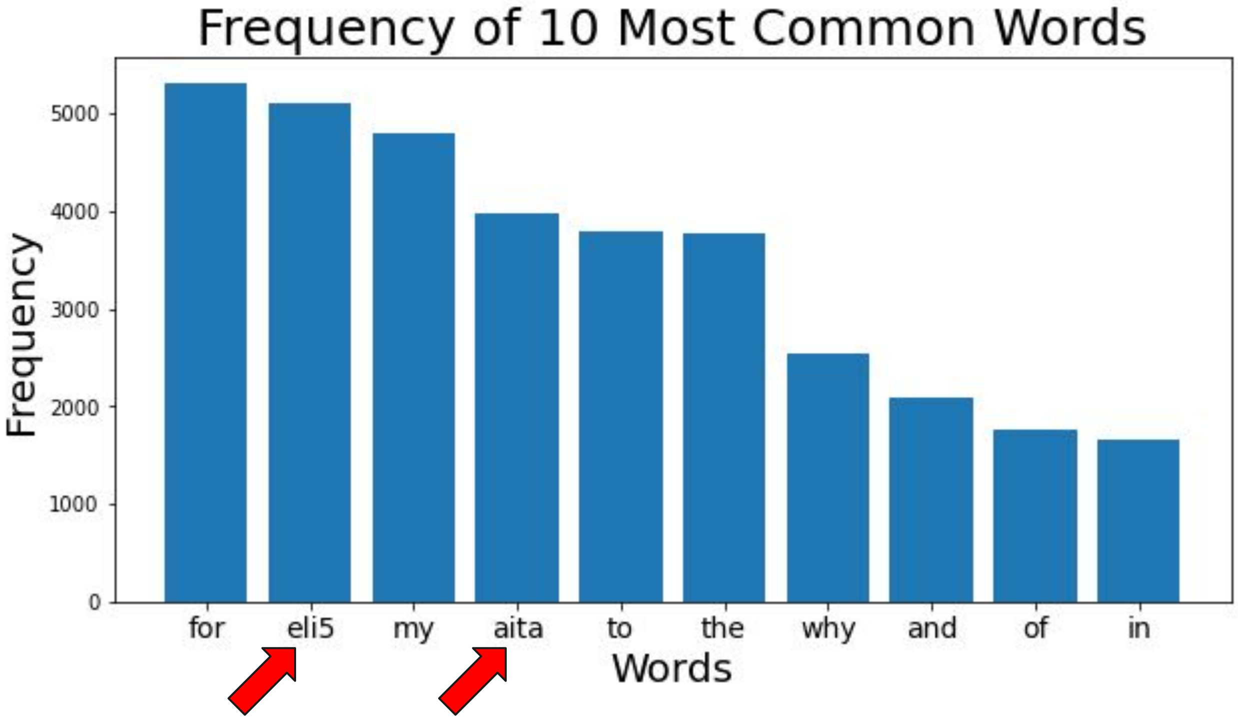
The Gardner Newspaper Company has recently opened an advice column in their weekly newsletter. The public response has been overwhelming and they have been inundated with questions. To accommodate this increased volume, Gardner Newspaper Company created two separate groups: one to answer interpersonal questions, and another to explain technical concepts simply.

Our consulting group was tasked with creating an algorithm to transfer the questions to the correct team.

Data Visualization

Looking at the most common words from posts on AITA and ELI5

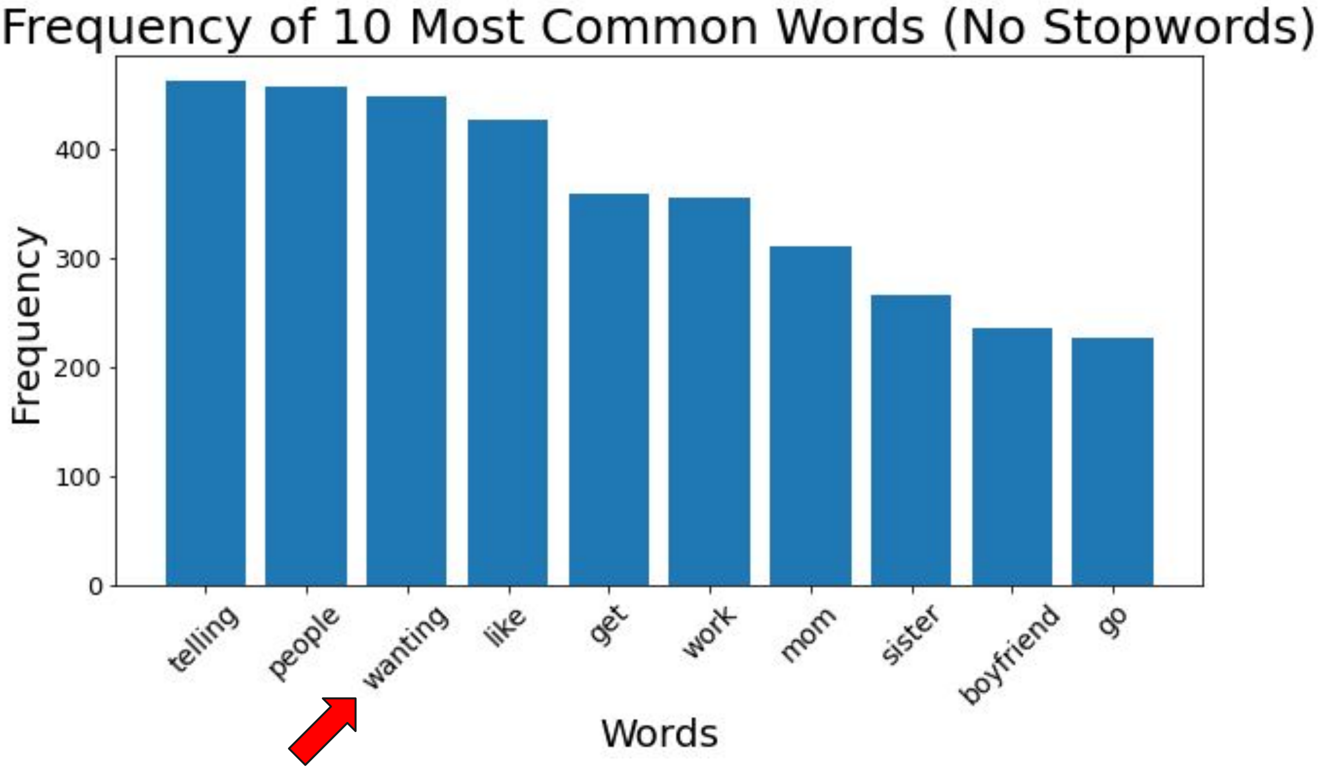
All of the most common words are typical stopwords except eli5 and aita. These two words are found in every post so they were added to the stopword list



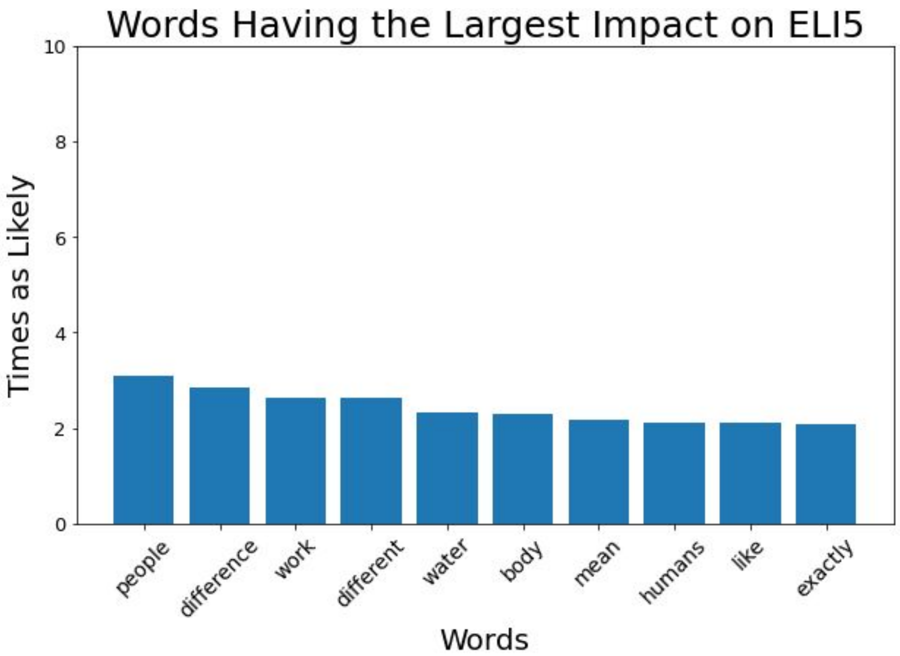
Data Visualization

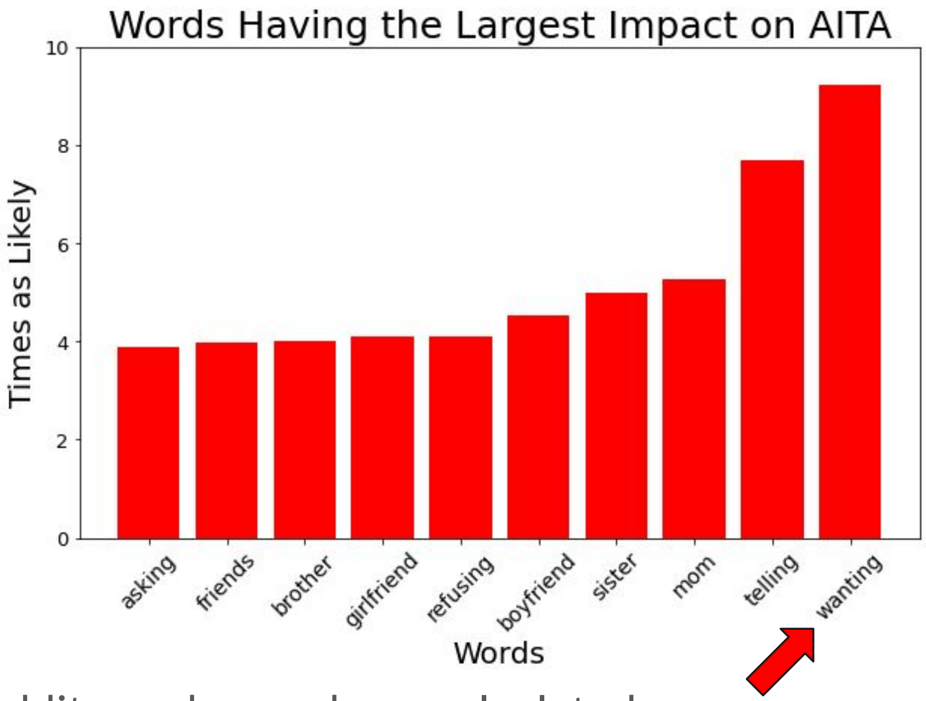
Without stopwords the most common words occur much less often

Additionally the common words in AITA are more often than the common words in ELI5



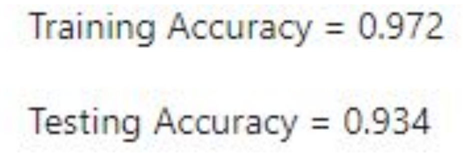
Data Visualization


Words Having the Largest Impact on ELI5


Words Having the Largest Impact on AITA

The highest impact words for each subreddit are shown above calculated from the logistic regression coefficients

Production Model

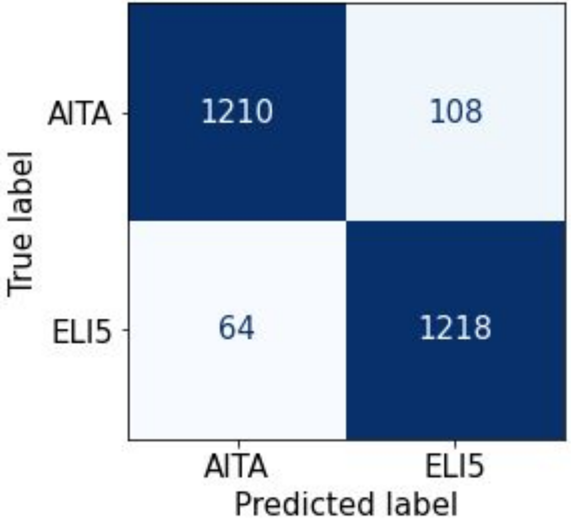


Training Accuracy = 0.972

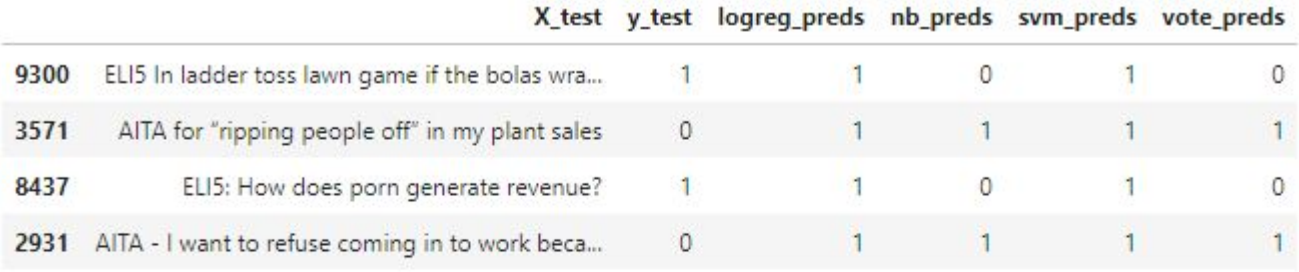
Testing Accuracy = 0.934

Logistic Regression
Naive Bayes
Support Vector Machine

} **Vote Classifier**



Overall the model predicts the testing data with 93% accuracy



	X_test	y_test	logreg_preds	nb_preds	svm_preds	vote_preds
9300	ELI5 In ladder toss lawn game if the bolas wra...	1	1	0	1	0
3571	AITA for "ripping people off" in my plant sales	0	1	1	1	1
8437	ELI5: How does porn generate revenue?	1	1	0	1	0
2931	AITA - I want to refuse coming in to work beca...	0	1	1	1	1

Each of the 3 component models make a prediction. The VoteClassifier then takes those probabilities and makes its own prediction

Conclusions

- AITA posts are easier to determine than ELI5 posts
- The Multinomial Naive Bayes model performs almost as well as the ensemble model with less computational resources required
- Personal and technical questions could be successfully identified and separated with 93% accuracy
- The model struggles to correctly identify technical questions based on personal concepts (ie sex, kids, marriage)

Future Plans

- Create a subsequent model which uses additional question text to provide more predictive power especially for technical questions