How Not To Be An Asshole 101: A Scientific Analysis

By Chandler Kinch

The Problem

- Writing can be difficult, and it can be difficult to convey a message properly
- Wouldn’t it be nice to have a way to check if a author is coming off negatively?
- What words and tones should be avoided or used?
- Could aid with emails, texts, speeches and more

The Data

The Data

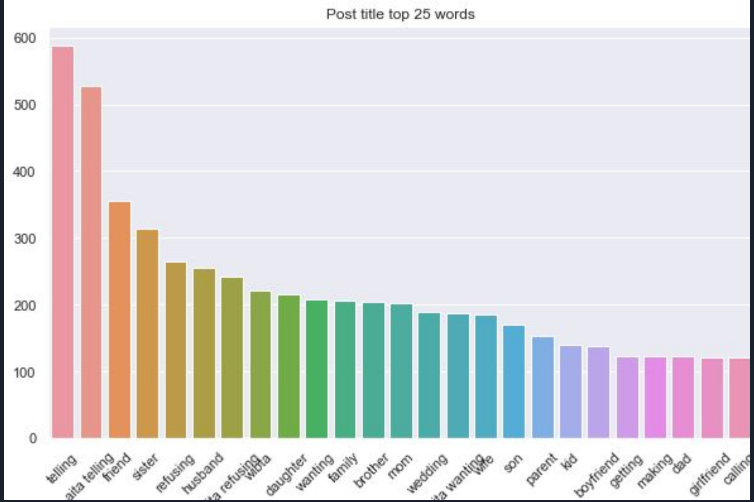
- Using data from r/AmITheAsshole
- Pulled ~5700 posts using PRAW
- Editing for flair left ~4500 posts, 539 asshole posts and 4023 non-asshole posts
- Posts are split into “Title” text and “Body” text
- Added word count, ease of readability and grade level scores

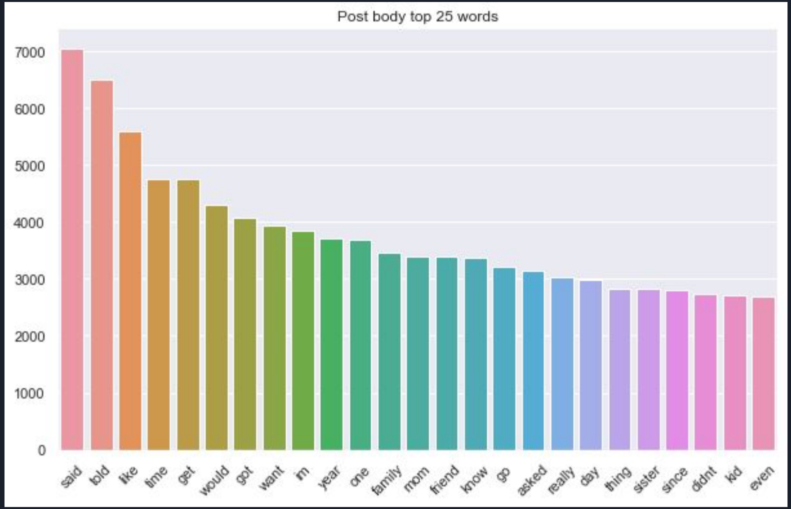
Vectorizing

- Bag of words method
- Removed stop words, numbers and punctuation
- Only kept words and phrases that occurred at least six times. Phrases were kept up to three words long.
- Body data has vocab of 17,868
- Title data has vocab of 892
- Removed ‘AITA’ for title data as it had 3319 occurrences. Next most had 588.

Most Frequently used words

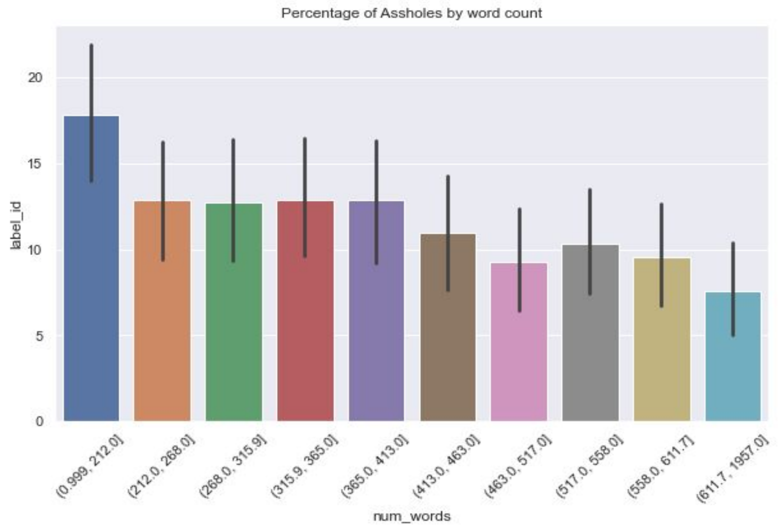
- Words revolve around conversation
- Friends and family also hot topics





Interesting Trend in Word Count

- Chart shows percentage of assholes by word counts within equal sized buckets
- P-value <0.01
- Makes intuitive sense, assholes explain less



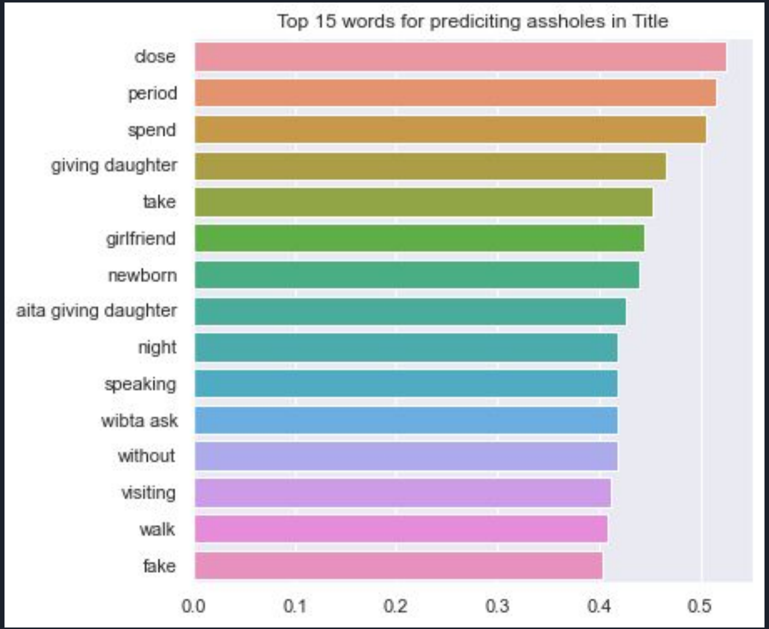
Most Predictive Words

Finding Most Predictive Words: Methodology

1. Create identity matrix the size of title vocab and body vocab.
    - This effectively makes a list of documents each containing one word that no other document contained.
2. Predict on this matrix with a simple Logistic Regression model.
3. Order words by the predicted probability for a given class
4. Repeat for both Title and Body text documents

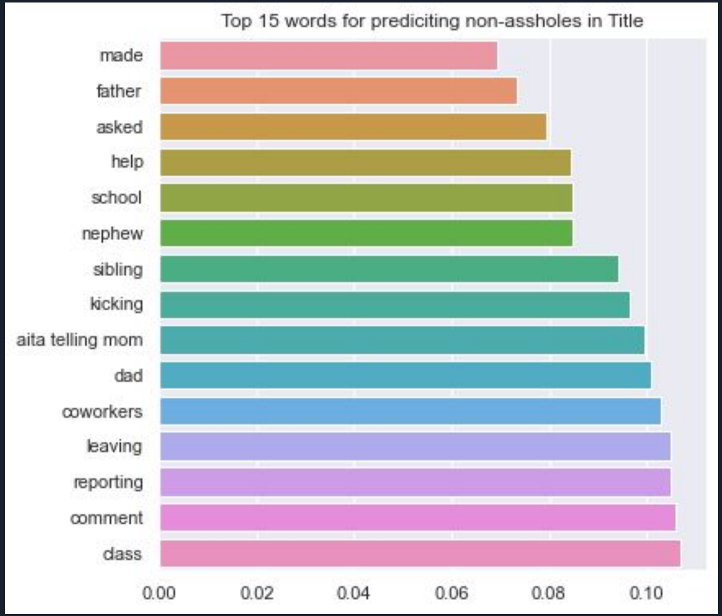
Top Title Asshole Words

- If you are asking if you’re too close, you probably are.
- Spend tends to refer to time, not money.



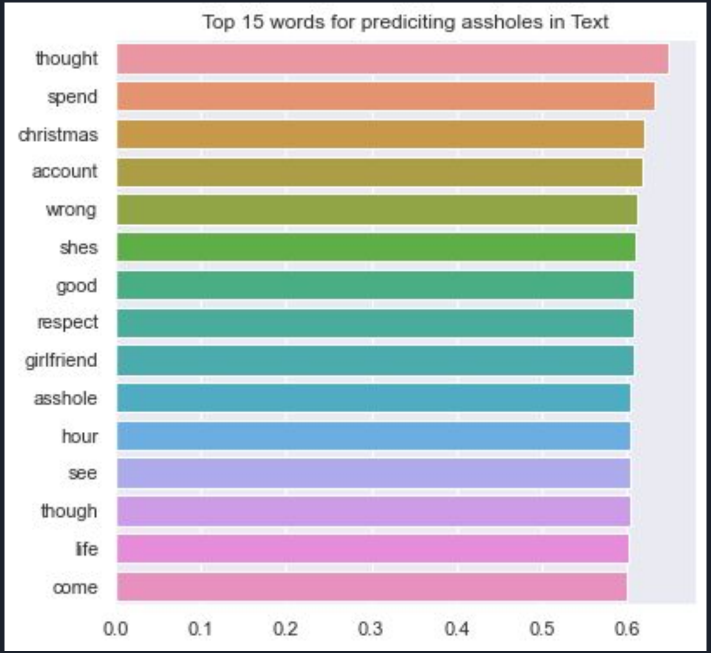
Top Title Non-Asshole Words

- Ask only appears once in asshole titles.
- Reddit seems to think just asking isn't offensive in and of itself.
- Requesting help also isn't seen as an asshole activity.



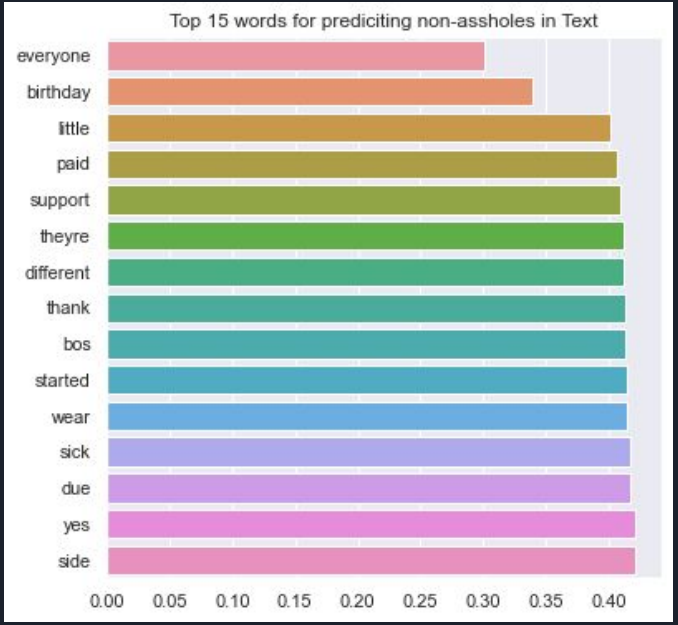
Top Body Asshole Words

- Thought seems to convey a lack of understanding.
- Wrong used to denote moral correctness. Assuming this made lead to its inclusion on the asshole list.
- Account used to denote reddit account, maybe for throwaways.



Top Body Non-Asshole Words

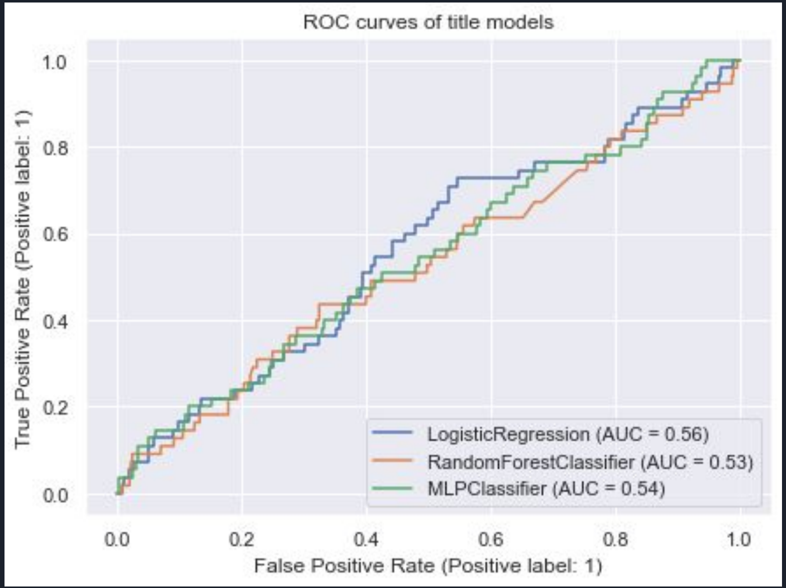
- Everyone used in edits to comments.
- Non-asshole may have more interaction with comments
- Little used to downplay issue at hand.



Modeling

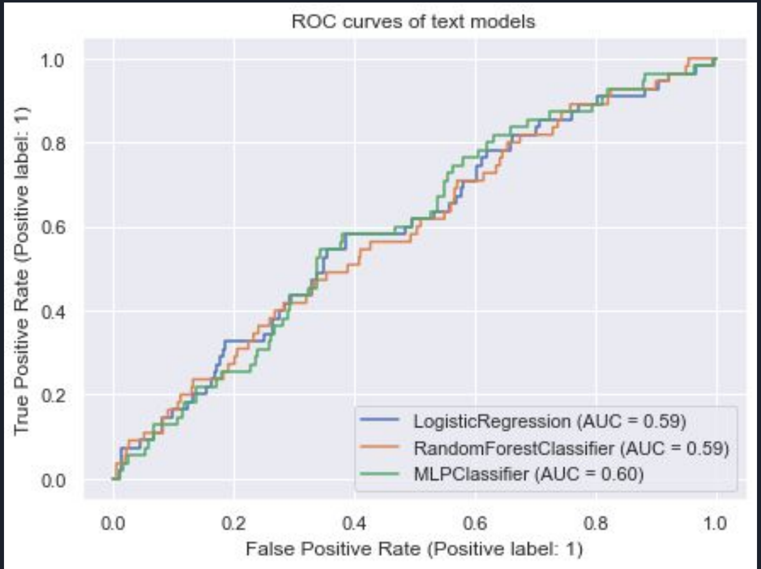
Title Data

- Grid searched three models.
- Chose ROC AUC as metric
- Best model was Logistic Regression
- Finding the asshole may be difficult to predict



Body Data

- Best model was MLP
- Slightly better scores than title data. Larger vocabs may make it easier to predict



How not to be an Asshole

Things to Avoid:

- Making assumptions
- Spend time with friends and family.
- Not supplying details

Things to do:

- Ask questions
- Request Help
- Be humble enough to realize your problems may be small
- Use plenty of details

Conclusion

- Predictive words and model provide powerful insights on what topics and words convey negative feelings
- How to improve analysis?
    - More Data!
    - Have models work together.
    - Set up rules based on patterns found during EDA.