Chandler Kinch

Final Report: Am I the Asshole Analysis

Problem Statement

r/AmITheAsshole is a popular forum on the website Reddit with a current subscriber count of 3.4 million. This subreddit is a place where people go to tell their stories, and to get an unbiased opinion on it. As the name of the subreddit so clearly states, they are looking to find out if they are the asshole of their story. Once a user posts their story to the subreddit, other users will respond with their input on the situation and whether or not the author is the story's asshole or not. Based on these responses and the upvotes these responses get, the author is labeled as the asshole, not the asshole or several other titles. My quest in this analysis is to create a classifier to skip the user input and get an instant response. In any situation in which one seeks to persuade, to evoke an emotional response, to win friends or influence people - it is essential not to come off as an asshole. Politicians, writers, lawyers, all could use this classifier to check if their written or spoken works could be perceived as assholery. Furthermore, it is often known that email and text communication often can be misinterpreted due to the lack of tone and body language information. Being able to automatically check if a piece of written communication could be interpreted in a negative light before sending it could at the very least help improve business dealings and, in more dire cases, save business relationships. Here, I seek to build a model to fill this need, or at the very least take the first steps towards doing so.

The Data

The data for this analysis were scraped from Reddit using the PRAW API. This API allows for many different ways to interact with subreddits, including pulling post information. For this analysis, I made several different pulls to get the necessary data as this API only allows pulls of ~1000 posts at a time. The most efficient way I could pull this data was pulling the data by their popularity over time. This was nice as it allowed easy access to posts with a lot of data attached to them, but it did limit posts to more recent ones.

After making the necessary pulls and deleting duplicates, there were 5777 different posts to work with. This was further cut down by the flair text associated with each post. This flair text describes the author of the post and is set by a bot on the subreddit. This bot counts upvotes and downvotes in the comments, and based on the labeling acronym used in the comments, assigns the flair text to the post. Because of the reliability and simplicity of these flairs, they will be used to label the data. There are multiple flairs that could be set. Among them are NTA(not the asshole), YTA(you're the asshole), ESH(everyone sucks here), NAH(no assholes here) and INFO(not enough info). For the purpose of this analysis, I cut out all posts that weren't labeled as 'not the asshole' or' asshole'. This left me with 4023 not the asshole posts and 539 asshole posts.

Now that the data were acquired and trimmed down to useful data points, I needed to make the data more friendly for analysis and for machine learning algorithms. For this analysis I decided to go with the bag of words method. This entails breaking down each post into a vector, where the columns represent a word in the vocabulary and the values represent how many times those words appear in the text. The first steps to getting this bag of words is removing numbers, punctuation and stop words. Stop words are words like ‘the’ ‘and’ or ‘I’. They appear frequently and don’t provide much to the analysis, so they are removed. The words are then lemmatized. This will break words down into their base form so that all forms of the word can be accounted for with one variable. The final step is to assemble the vocabulary and get the counts of each word for each post. Along with words, phrases up to three words long were counted as well. During this counting, words and phrases were excluded if they didn’t have at least 6 occurrences.

For this analysis, the title and body of posts will be analyzed separately. This is because the title and post body are different things and words could be used differently between the two or have different significances. This could mean that words have different weights depending on if they are used in the title or post body. After vectorizing the text, the title data had a vocabulary size of 892 words and the body data had 17,868 words. I checked for any outliers in word occurrence, and there seemed to be only one in the title data. The acronym AITA(Am I the Asshole) occurred 3319 times, and the next highest was the telling with 588 occurrences. Many posts start with AITA, so this word was excluded to reduce redundancy.

Exploratory Data Analysis

Let’s begin with getting an idea of what words occur the most in both the body data and the title data. We see that the numbers are significantly higher for body data than for the title data. This makes sense as the body of posts is almost always longer than that of the post title, meaning words will occur more frequently. Figure one shows data counts for title data and figure two data for body data.

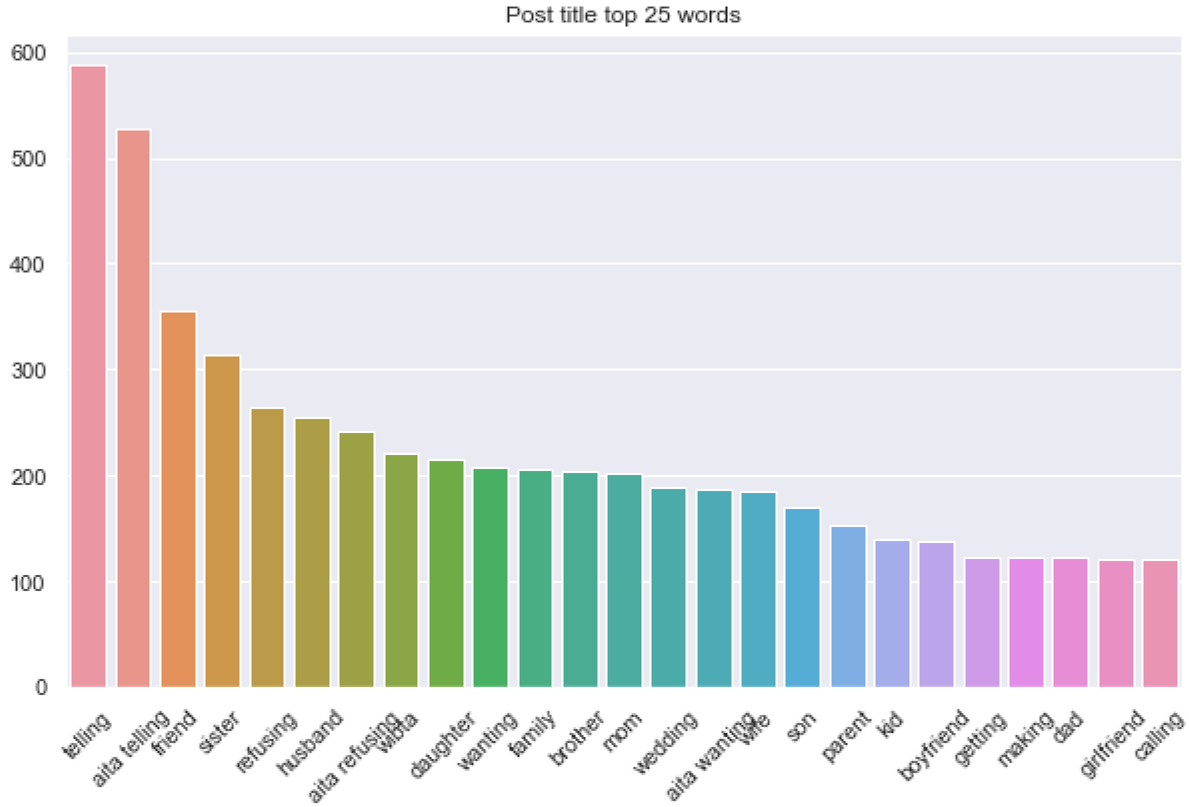

Figure 1 Word counts for title data

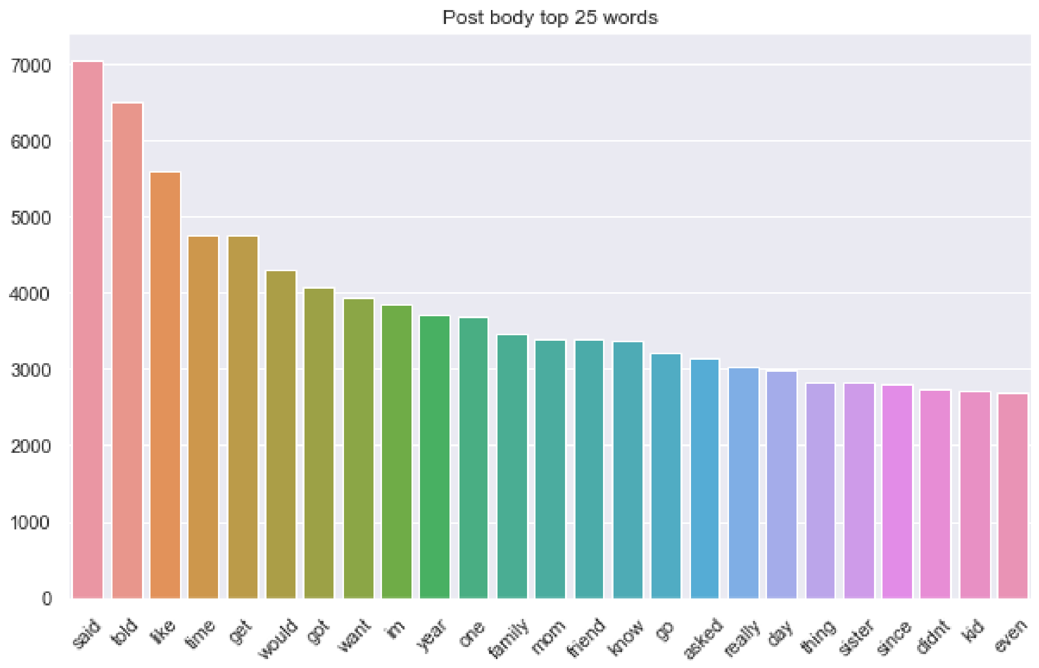

Figure 2 word count for body data

We see that words dealing with conversation and talking appear many times in both data sets. Words about friends and family also seem to pop up quite a bit. It will be interesting to see which of these words and themes pop up when feature importance is examined. This next figure, Figure 3, shows the percentage of assholes by word count within equal sized buckets.

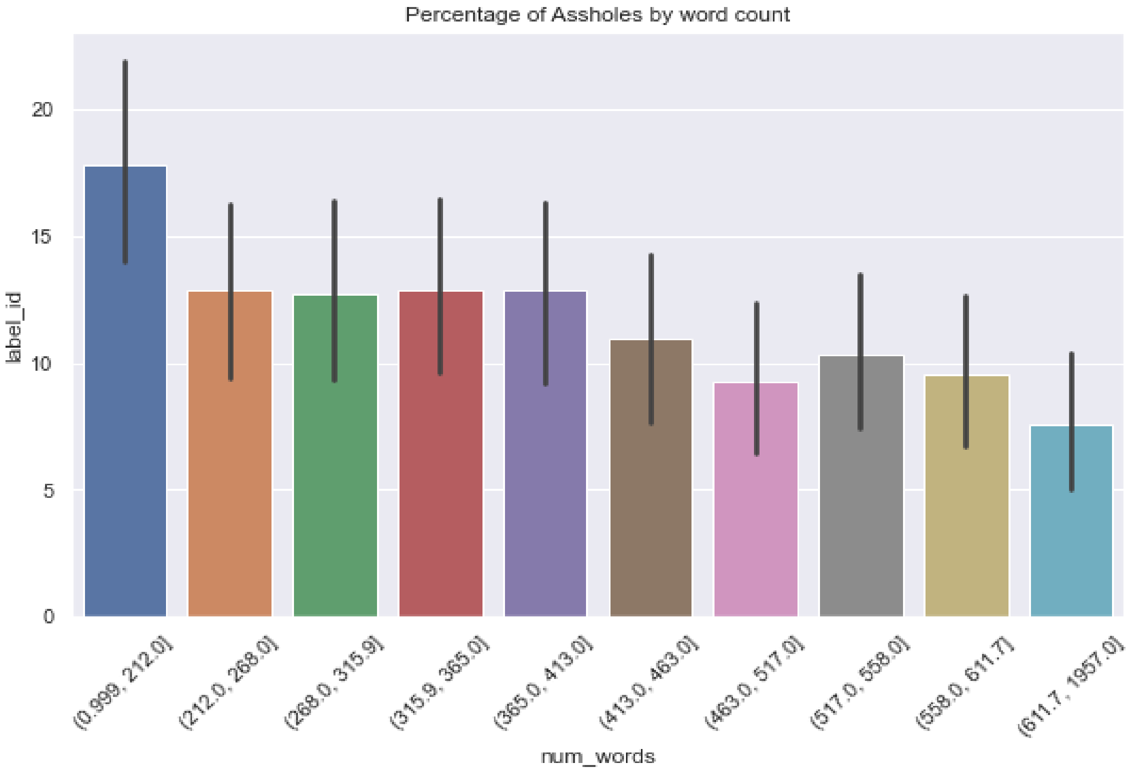

Figure 3

I find Figure 3 quite interesting. While the change isn't huge as the number of words increases, there is a clear trend of fewer assholes. This distribution also has a p-value of <.01, showing us that this trend is not due to chance. This trend, for some reason, makes intuitive sense to me. I would think that an asshole would want to use fewer words for their story to hide facts and twist the story into their favor. On the other hand, an non-asshole would want to elaborate more and provide more facts and details to ensure the whole story was understood correctly.

Most Predictive Words

Now that we have examined some distributions of our data, let's examine which of these words and features are most important for predicting assholes and non-assholes. To do this, I started with identity matrices, meaning square matrices with one's along the diagonal and zeros everywhere else, the size of the vocabularies of the body data and text data. Each column in these matrices represents a document of one word, and each word has only one document that contains it since each column will only have one non-zero value. I then used a Logistic Regression model to predict on these matrices. This returned the probabilities of each document being the asshole or not. Since each document contained one unique word from the vocabulary, these probabilities also are the probabilities of each word being the asshole or not. Therefore, these probabilities show which words correlate with being the asshole or not.

The Title Text

Let's begin by looking at the top predictive words of the title data. Figure 4 will show words for assholes and Figure 5 will show words for non-assholes.

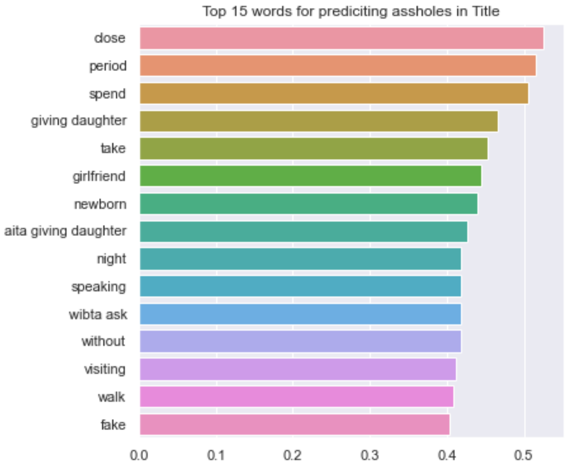

Figure 4

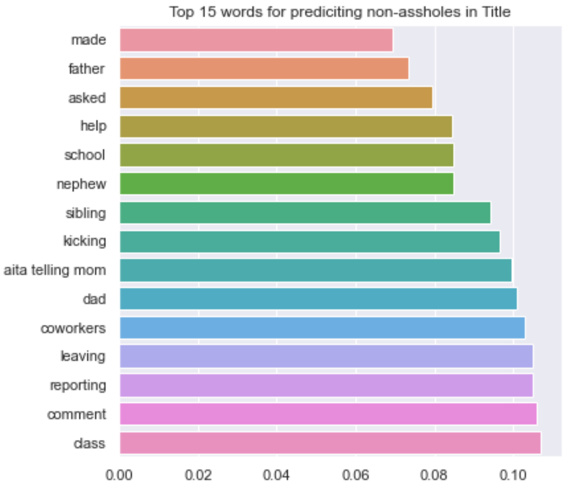

Figure 5

We can see that 'close', 'period' and 'spend' are the most important words for finding assholes while 'made', 'father' and 'asked' are the most important for finding non-assholes. Let's examine some titles containing these words and see if we can see how they are used. We will begin with the asshole words. Please note these titles and text are being taken verbatim from the source, so errors will be plentiful.

Predictive "Asshole" Words:

Close:

- AITA for being too *close* with my friend's boyfriend.
- AITA for still being *close* with a man I once loved

It seems that 'close' is typically being used to refer to relationships as opposed to being used to denote physical distance to something. So a takeaway based on the responses in this subreddit: if you are asking if you are too close to someone, you likely are.

Spend:

- AITA for wanting to *spend* Christmas with my brother and his new baby even though my fiance has to work and can't come?

- AITA for locking myself and my newborn in the guest bedroom so that I could finally *spend* some time wither away from my wife?

Again we have an interesting result here. I thought that 'spend' might refer to money, but we can see that it is referring to time. In fact, of the 7 asshole titles containing 'spend', only one is referring to money. The rest refer to time. This word also appears 11 times in non-asshole titles, and over ⅔ of those are referring to time as well.

Let's now examine some non-asshole words such as 'made' and 'help'. While 'father' is the second most predictive word for non-assholes, this word is less ambiguous. The word 'help' on the other hand leaves a lot more up in the air, so this word will be examined instead.

Predictive "Non-Asshole" Words:

Asked:

- AITA for tell my sibling that my gf was not the person they *asked* to babysit?
- WIBTA if I *asked* a tattoo artist to tattoo someone else's art?

"Asked" appears only once in asshole titles. This isn't all too surprising, as generally "asking" isn't considered to be an offensive activity, but it certainly can elicit negative reactions. However, it seems that in these cases the reddit hivemind tended to agree that simply "asking" doesn't make you an asshole.

it seems that if someone is asking something, they aren't being an asshole. '

Help

- AITA for expecting my husband to *help* me when I am grieving
- AITA for refusing to *help* my ex pay rent for an apartment I no longer love at?(guessing they mean live here not love)

I thought that help would be used in terms of someone offering help, but it seems that isn't typically the case. It seems to be used more often when someone is refusing help or executing help. In fact, of the 49 non-asshole titles that contain 'help' 19 start with the phrase 'AITA the for refusing'. In retrospect, this shouldn't be too surprising. People aren't going to ask if they are the asshole for offering help. They are going to ask when they are refusing or expecting it, and typically they aren't the asshole.

The Body Text

Now that titles have been explored, let's explore the post bodies now. These words were found using the same identity matrix trick explained above. Figure 6 contains the asshole words and Figure 7 the non-asshole words.

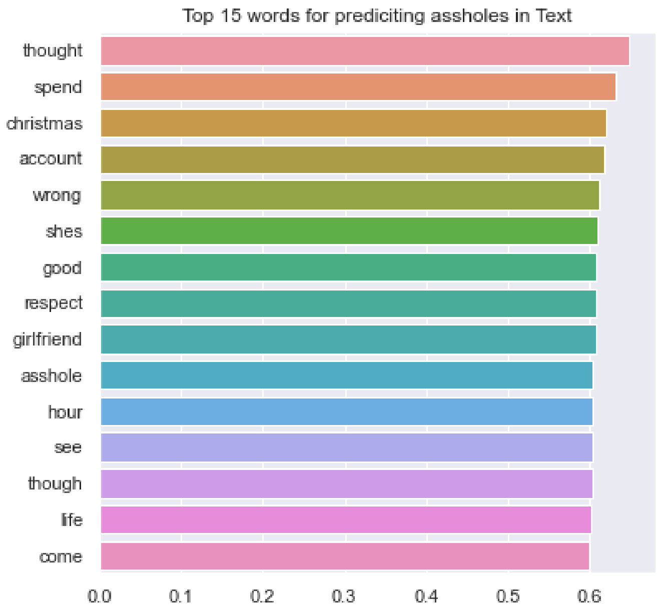

Figure 6

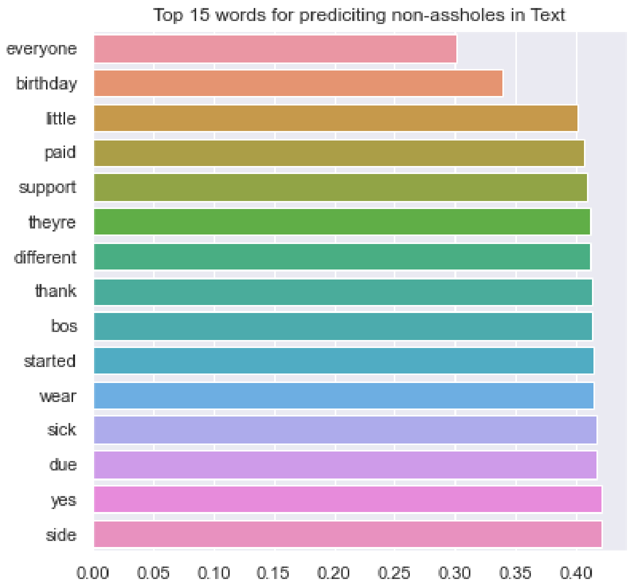

Figure 7

We see that 'thought', 'spend' and 'Christmas' are the top words for finding assholes while 'everyone', 'birthday', and 'little' are the top words for finding non-assholes. It is interesting that 'spend' and 'girlfriend' show up on both asshole charts. Let's look at a few examples of 'thought', 'wrong' and 'account' among assholes in the text data. Well will skip 'Christmas' for similar reasons that we skipped 'father' in the title data. We will skip 'spend' as we already covered this word previously. The titles will be shown for these posts as well as a snip of the body text to help provide context.

Predictive "Asshole" Words:

Thought:

- AITA for wearing revealing clothes around my BFs dad?
  - "Once again, I *thought* I looked really nice…"
- AITA for downplaying my friend's trauma without even knowing it?

    - "I didn't know I was making her uncomfortable and like I said I should have *thought* through what I said…"

I wasn't really sure why this word was appearing for the assholes, but these examples really shed some light. Seems that this word is being used to portray some lack of understanding or awareness. Maybe even that something is being assumed, and we all know the saying about assumptions. It may go without saying, but this alludes to assholes not being as aware of their surroundings as they should be.

Wrong:

- AITA for helping my girlfriend's maid clean?
    - "I'm not sure if I was in the *wrong* for helping her clean or if I'm in the wrong now for not helping her."

- AITA for telling my daughter to apologize after she said that she hates her sister?
    - "I said it would be *wrong* to exclude Jane but Sarah said didn't want her to come at all."
    - "Was I in the *wrong* for telling Sarah she needs to apologize…"

I was expecting 'wrong' to appear in places where the author was telling someone that they were wrong. Instead it seems to be used to denote moral correctness. While these examples show how 'wrong' is being used, it doesn't shed much light on why the word is associated with assholes. Interestingly enough, 'wrong' appears in 91 asshole posts and in 548 non-asshole posts. There may be some combination of words or number of appearances within posts that correlates this word with assholes.

Account:

- AITA for telling my fiance she is not allowed to take out a loan for her brother?
    - "I have a large savings *account*, which is going towards our wedding…"
- AITA For kissing my friend's brother without his consent during a truth or dare?
    - "I used a throwaway *account* for privacy reasons…"

The reference to a bank account is what I was expecting from this word. For some reason, I had not taken into account(pun intended) that it would reference a throwaway account. A throwaway account, is as it sounds, an account that is thrown away, usually after serving one or a few purposes. It is common for people to use these when they want to protect their privacy. It makes sense to me that an asshole may want to use a throwaway. Though, that doesn't mean they are the only ones that use them.

Predictive “Non-Asshole” Words:

Now that we have seen the assholes in the body data, let’s check out the non-assholes. We will look at the words ‘everyone’, ‘little’ and ‘paid’. We will skip ‘birthday’ here as that word isn’t very diverse or interesting as its use cases are slim.

Everyone:

- WIBTA for making my kid give me a gift she really wants for her bday to the person she stole from?
    - “I just wanted to let *everyone* know who is suggesting therapy that we have been taking advantage of that option.”
- AITA For arguing with my brother’s wife after she announced her pregnancy at my daughter’s birthdays party?
    - “My sister in law announced that she and my brother were expecting and *everyone* started congratulating her…”
    - “Maya seemed upset until the party was over she didn’t even open some gifts as *everyone* else was sitting and talking to my sister in law.”
    - “*Everyone* please understand that I am not in the USA.”

The biggest thing that I see in these two posts is that both posts have ‘everyone’ appear in an edit. The first example only has ‘everyone’ in an edit and the second example’s last ‘everyone’ is from an edit. It is possible that non-assholes have more interaction with the comments through edits and feel justified enough to come back and make edits.

Little:

- AITA for not offering to split costs on food I was invited to cook?
    - “I’m amazed my *little* issue got so many replies!”
- AITA for talking to a friend’s ex?
    - “So it’s a *little* more complex than that, but that is the gist.”

Both of these examples see ‘little’ being used to describe the issue at hand. The first one is an edit replying to the comments and the second is the first sentence in the body referring to the title. While the second is being used in a way to almost mean the opposite of little, the fact that it refers to the issue is still important. It shows that non-assholes may downplay their issue and make it not a big deal. This shows some humility to me, which is not a typical asshole trait.

Modeling

The modeling for this analysis is pretty straight forward. There were no missing values that had to be worked around and the data had already been formatted for machine learning previously. Text was preprocessed and fed into a CountVectorizer with a min_df of 6, the resulting term-document matrices were combined with the grade level scores for reading difficulty, the number of words and the readability scores for each title or body text. The data was split, leaving 15% for validation, and scaled using Robust Scaler. I chose three different algorithms and grid searched them: Logistic Regression, Random Forest and MLP Classifier. The metric I chose to go with for this modeling was the ROC AUC score. This is because there are many more non-asshole posts than asshole posts. This would cause the accuracy score to be much too misleading as labeling everything as non-asshole would result in high accuracy. The ROC AUC score ensures that this unequal distribution of posts will not affect our metric. The ROC AUC has the added advantage of being threshold independent so models can be evaluated across different business scenarios. The title and body data was kept separate for EDA, and they will also be analyzed separately during modeling.

The Title Data

Grid searching three models on the title data resulted in the scores and hyperparameters shown in Table 1. Logistic regression yielded the best result with a ROC AUC score of 0.56, with the other models not far behind. It appears that determining the asshole may be a difficult thing to predict or there may be another flaw somewhere in the analysis. We can further see the results of this modeling in Figure 8. This figure shows the ROC curves for the three models discussed in this section.

Classifier	ROC AUC Score	Hyperparameter Values

MLP Classifier	0.54	Activation = relu
		Alpha = 0.05
		Hidden_layer_size = (100,)
		Learning_rate = adaptive
		Solver = adam

Logistic Reg.	0.56	C = 0.559
		Penalty = l2

Random Forest	0.52	Max_depth = 3
		Max_features = log2
		n_estimators=200

Table 1

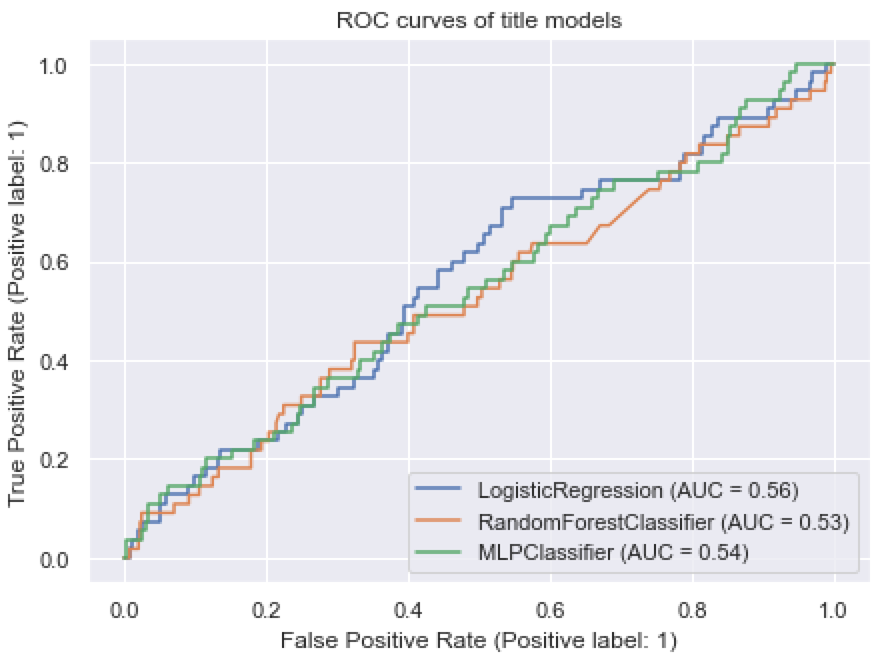

Figure 8

The Body Data

The same three models were grid searched for the body data as for the title data. The scores and hyperparameters for these models can be seen in Table 2. Likewise, the ROC curves can be seen in figure 9. This time the MLP Classifier yielded the best result with an ROC AUC score of 0.60.

Classifier	ROC AUC Score	Hyperparameter Values

MLP Classifier	0.60	Activation = relu
		Alpha = 0.0001
		Hidden_layer_size = (100,)
		Learning_rate = constant
		Solver = sgd

Logistic Reg.	0.59	C = 0.00599
		Penalty = l2

Random Forest	0.54	Max_depth = 5
		Max_features = log2
		n_estimators=100

Table 2

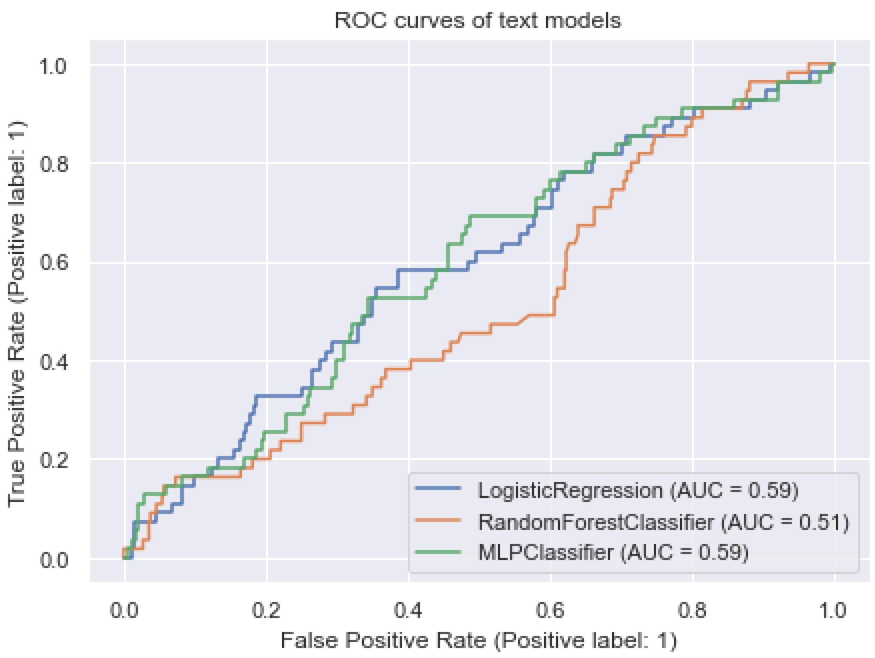

Figure 9

Conclusion

r/AmITheAsshole is a place containing thousands of accounts and questions of people trying to make sense of their stories. We can see that topics vary widely from dads trying to figure out how to handle their kids to people worried about splitting the cost of food. This model seeks to help writers and storytellers verify their stories to ensure that they are coming off in the way that they want. We can see from the predictive word section what words could portray what feelings. Even better, these words help to show what situations to avoid and how to convey stories without being the 'asshole'.

How could these models and problems be improved? My first thought goes to the data. While the source of the data is solid, the amount of data that was gathered was insufficient to say the least. With only ~4,500 posts the models had a difficult time finding patterns in the data and therefore making predictions. Also the skew in the asshole and non-asshole posts may be contributing to this problem. To fix this, more data could simply be pulled using the same methods but with more time. For example, one of the pulls made was top posts from the week. This data could be repulled every week for several months and every post would be new. Additionally, other API's or methods could be researched or created to help solve the amount of data.

More data would be a great way to help these models, but extra ideas can be considered. The first thing that comes to my mind is using the models together to make better predictions. Are the title and body models predicting the same things for each post? This could be analyzed to see if the models could be combined somehow to make better predictions. Another thing that could be explored is setting up rules based on things that were found during EDA. Many different trends were found during the predictive word analysis. These trends could be turned into steadfast rules for prediction to improve modeling. Overall, this analysis helps give insight into how people can express themselves in a way most likely to be received well by their intended audience.