# Am I the Asshole?

An important question turned into a data science project.

By: Cassie Buhler



Remember that argument you had the other day?

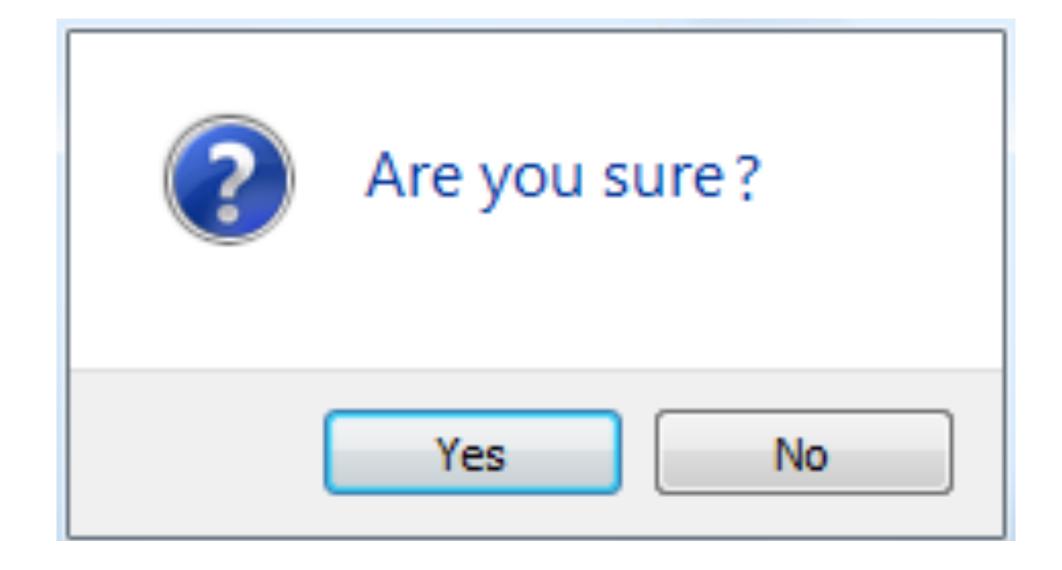


Yeah, they were the ones who messed up. Not you!

You did everything right. It's their fault!

Except... you *did* raise your voice and say some hurtful words...

But whatever— they were an asshole! Not you!



## r/AmITheAsshole

- r/AmlTheAsshole is a popular subreddit where you can post about a conflict (giving both sides of the story)
- Fellow Redditors will determine your judgement with the following votes:
  - Not the asshole (NTA)
  - You're the asshole (YTA)
  - Nobody is an asshole (NAH)
  - Everyone sucks/Everyone is an asshole (ESH)
- After 18 hours, moderator bots go through the comments and tally up the votes.
- The verdict is declared by the majority vote.

#### What Data Can We Collect?

- I took posts from the last 5 years and collected the following features:
  - The title/post content
  - The judgement (YTA, NTA, ESH, NAH)
  - Number of upvotes/downvotes
  - Number of comments
  - Number of awards

## Usage/Applications

- Using ML to detect assholes
  - By inputting a scenario, the algorithm will output the person who is in the wrong.
  - You could make it an app. Great for people who overthink things!
- Analyzing human thinking/behavior
  - Researchers could look at the different scenarios and their verdicts to understand how humans judge scenarios
  - Not only who is considered an asshole, but why? What about the situation makes them an asshole?
  - Do certain type of relationships (spouses, tenant/landlord) have more conflict?

## Distribution/Access

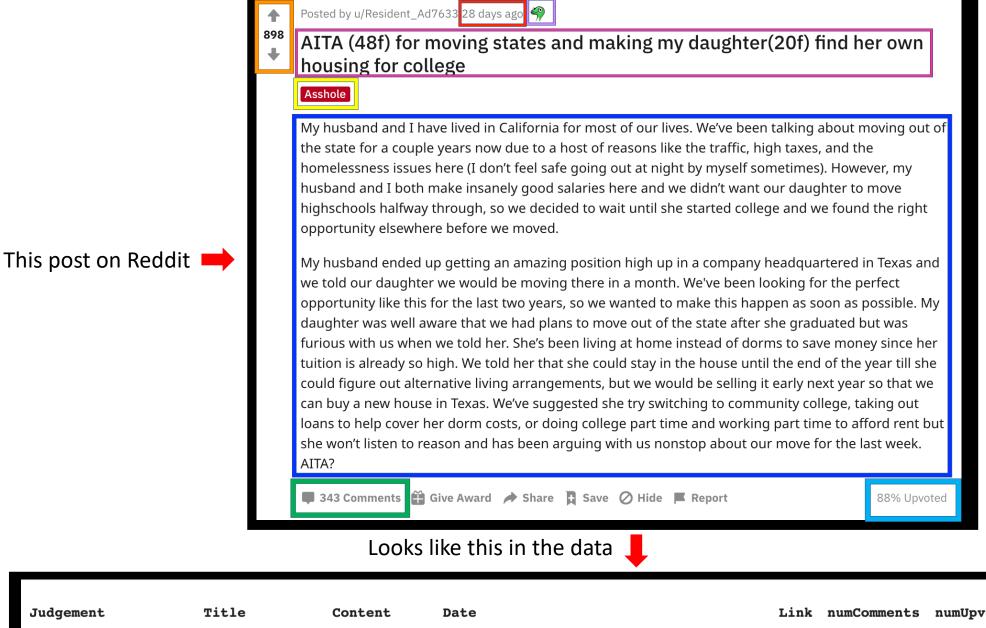
- To use Reddit's API, you need to create an app as a developer to get access keys
- Since this app is registered as a personal script, I didn't package up the code for distribution
- Reddit owns the data so if I were to create something with the data (e.g. an app that tells you if you're an asshole) I would need to get approval from Reddit.

#### Acquiring the Data

- 1. First, I was naïve and used BeautifulSoup to scrape the post links on the subreddit.
- 2. Then I used PRAW (The Python Reddit API Wrapper) to obtain features for each post.
  - I ended up with only 400 posts.
- 3. So I went back and used PRAW instead of Beautiful Soup to get post links.
  - This gave me **800 posts!**
  - But I was getting greedy and flew closer the sun.
- 4. I used pushshift.io to get post IDs between 01/01/15-11/01/20
  - This gave me over 700,000 posts- NOW we are talking!!!
- 5. However, collecting features for the posts uncovered a big problem.
  - Getting features for 700,000 posts would've taken my computer ~6 days to run
  - Therefore, I reduced the number of posts to 70,000
  - However, my computer had an issue and ended up only saving 114 posts:'(

## Cleaning the Data

- Posts were removed on the following conditions:
  - The post was deleted/removed/empty
  - There is no verdict
- There were also some minor renaming
  - Ex: not the asshole = not the a-hole = NTA
- Edited posts need to be dealt with
  - The author can edit their post with an update like "EDIT: yeah I was being an a-hole"
  - I found out after that you can filter edited posts out when collecting the links in PRAW
- I ended up with 1026 posts between the years 2015-2020.



JudgementTitleContentDateLinknumCommentsnumUpvotesupvoteRationumAwardsYTAAITA (48f) for moving states and making my dau...My husband and I have lived in California for ...2020-11-04 13:29:57https://www.reddit.com/r/AmltheAsshole/comment...3439050.881

#### Issues and Limitations

- Lost track of the big picture
  - I was so fixated on getting more post links
  - So I didn't think about how long it would take to process the links
  - This resulted in overwhelming my computer (and myself)
- Too much data for my little computer
  - This limited the number of features I could include.
  - Ex: I wanted to tally up the votes to see how close the margins are each verdict but there are thousands of comments for the thousands of posts days.

#### Additional Considerations

- There is additional data we can get from this
  - Demographics –e.g. 22F
  - Relationships- e.g. Tenant/landlord
- This data is biased—take it with a grain of salt
  - Based on opinions from strangers on the internet.
- There is no way to assure the validity of the posts
  - Did that conflict \*really\* happen? Or did the OP make it up? Does it matter if it's made up?

## The Team Behind The Project

#### Team Member:

1. Cassie

#### **Contributions:**

- Spent time on Reddit aka "researching"
- Wrote the code
- Made all the mistakes
- Fixed most of the mistakes

Thank you!