Vague Comment Classifier

• • •

r/AmiTheAsshole

Introduction

Reddit

- The self-acclaimed front page of the internet
- a social news aggregation, web content rating, and discussion website.
- Registered members submit content such as links or multimedia to the site through subreddits that form communities which are then voted up or down by other members.

r/AmiTheAsshole

- Niche community that has arisen with a unique premise that aims to both help and entertain users.
- People post situations that they have found themselves in in which they feel they will be perceived as an 'asshole'.
- Commenters reply to the original posters situation and vote on them by using coded language designated in the subreddit rules. [YTA| NTA | ESH | NAH]

Problem Statement

Often times, people leave conditional or vague statements as votes, for example: "YTA if you do it this way, but NTA if you do it that way!"

This project aims to create a model that can predict on those comments in an attempt to classify if their language use makes them YTA or NTA.

Data Pulling & Handling

- Using PRAW (Python Reddit API Wrapper)
- Pulled 5 sets of posts & comments 1 week apart starting from 26/12/20(except for pull 5 due to complications)
- Data Merged
- Created columns YTA & NTA by scanning through the text
- Kept entries with only YTA or NTA
- Created the test set with both YTA and NTA

yta	nta	esh	nah	
0	0	0	0	224793
	1	0	0	137384
1	0	0	0	40631
0	0	1	0	7030
		0	1	5615
1	1	0	0	880
0	1	1	0	316
1	0	1	0	252
0	1	0	1	252
1	0	0	1	123
0	0	1	1	85
1	1	1	0	45
		0	1	21
0	1	1	1	13
1	0	1	1	12
	1	1	1	7

Spot the Difference

```
c5 = pd.read_csv('datasets/comments5.csv')
c5.head()
```
executed in 561ms, finished 00:24:04 2021-01-28

	post_id	comment_text	comment_distinguished	comment_score
0	koga1e	I feel like this subreddit has deteriorated ov...	NaN	73
1	koga1e	How many accounts of our "my stepmom sucks" tr...	NaN	71
2	koga1e	Can there be ban on "aita for reporting my [th...	NaN	122
3	koga1e	Something has to be done against the "sick bur...	NaN	122
	koga1e	People really need to be encouraged to use ESH...	NaN	50

```
c4 = pd.read_csv('datasets/comments4_1.csv')
c4.head()
```
executed in 495ms, finished 00:24:03 2021-01-28

	post_id	comment_text	comment_distinguished	comment_score
0	koga1e	I feel like this subreddit has deteriorated ov...	NaN	69
1	koga1e	How many accounts of our "my stepmom sucks" tr...	NaN	62
2	koga1e	Can there be ban on "aita for reporting my [th...	NaN	119
3	koga1e	Something has to be done against the "sick bur...	NaN	114
4	koga1e	People really need to be encouraged to use ESH...	NaN	47

```
c3 = pd.read_csv('datasets/comments3.csv')
c3.head()
```
executed in 521ms, finished 00:22:27 2021-01-28

	post_id	comment_text	comment_distinguished	comment_score
0	koga1e	How many accounts of our "my stepmom sucks" tr...	NaN	49
1	koga1e	I feel like this subreddit has deteriorated ov...	NaN	45
2	koga1e	Can there be ban on "aita for reporting my [th...	NaN	91
3	koga1e	Something has to be done against the "sick bur...	NaN	77
4	koga1e	People really need to be encouraged to use ESH...	NaN	37

title	id	date_created	text
Monthly Open Forum January 2021	koga1e	2021-01-01 18:33:45	Welcome to the monthly open forum! This is the...

"Welcome to the monthly open forum! This is the place to share all your meta thoughts about the sub, and to have a dialog with the mod team.\n\n#Keep things civil. Rules still apply.\n\nIt's 2021! Everything is fixed now!\n\nA couple notes.\n\n* Our bot is live, but definitely still in testing. **Please help us by reporting the ~~judgement bot comment~~ post when it doesn't actually explain why they think they may be an asshole.** Some people are using it like a TL;DR or just copying and pasting their post as a reply. [ETA - sounds like the report option doesn't work on all platforms for the bot comment, so you can just report the post. The option is bundled with the META report]\n\n* Please stop PMing mods. We spam the hell out of the [modmail link](https://www.reddit.com/message/compose/?to=/r/AmItheAsshole). When you PM us, it's super easy for things to get buried in our inbox and delay your response time.\n\n#As always, do not directly link to posts/comments or post uncensored screenshots here. Any comments with links will be removed. \nThis is to discourage brigading. If something needs to be discussed in that context, use modmail."

Text Processing

Created 4 different sets of data

- Cleaned_word : Simple cleanup of text, only alphabets without punctuation
- Nltk_lem : Using NLTK word lemmatizer
- Nltk_pos : Using NLTK pos tagger to help filter & Lemmatize text
- Spacy_lem : Using Spacy word lemmatizer

Text Processing

```
final['comment_text'][0]
```
executed in 9ms, finished 01:00:31 2021-01-28

'Thanks for this. When I was getting gaslit I came here a couple of times and got a lot of "YTA for staying with him/dating him in the first place". I think this happens more on the low traffic posts especially. Not particularly helpful, especially since I just took that quiz and the threshold for abuse is 5 and that relationship scored 33!'

```
final['cleaned_text'][0]
```
executed in 14ms, finished 01:00:31 2021-01-28

'thanks getting gaslit came couple times got lot staying dating first place think happens low traffic posts especially particularly helpful especially since took quiz threshold abuse relationship scored'

```
final['nltk_lem'][0]
```
executed in 14ms, finished 01:00:31 2021-01-28

'thanks getting gaslit came couple time got lot staying dating first place think happens low traffic post especially particularly helpful especially since took quiz threshold abuse relationship scored'

```
final['nltk_pos'][0]
```
executed in 14ms, finished 01:00:31 2021-01-28

'thanks get gaslit come couple time get lot stay date first place think happens low traffic post especially particularly helpful especially since take quiz threshold abuse relationship score'

```
final['spacy_lem'][0]
```
executed in 14ms, finished 01:00:31 2021-01-28

'thank get gaslit come couple time get lot stay date place think happen low traffic post especially particularly helpful especially take quiz threshold abuse relationship score'

Models

0	0.771874
1	0.228126

- Data: cleaned_text, nltk_lem, nltk_pos, spacy
- Vectorizers: CountVectorizer, TfidfVectorizer
- Sampler: SMOTE
- Classifiers: MultinomialNB, Logistic Regression

Production Model

Text: Cleaned_text

Vectorizer: TFidfVectorizer

Classifier: MultinomialNB

Accuracy: 0.846

	features	coef
325103	quotes needed	-13.7789
206883	keep decision	-13.7789
206891	keep dick	-13.7789
318200	probably feels utterly	-13.7789
393992	support structure	-13.7789
393994	support stupid	-13.7789
421539	transition whatever	-13.7789
206914	keep doubt	-13.7789
421534	transition need	-13.7789
421533	transition hard	-13.7789
206933	keep emergency	-13.7789
206942	keep engaged	-13.7789
32135	back folding	-13.7789
32136	back folding laundry	-13.7789
421520	transgender man	-13.7789
32138	back foot	-13.7789
421541	transitional period	-13.7789
318162	probably even know	-13.7789
32110	back feet one	-13.7789
421574	transphobic asshole	-13.7789
206687	keep apologizing	-13.7789
206702	keep awake	-13.7789
393968	support sisters	-13.7789
318248	probably gf	-13.7789
206754	keep blaming	-13.7789

	features	coef
226845	like	-6.75228
459722	would	-6.88085
153347	get	-6.91859
301833	people	-7.19734
130384	family	-7.23697
437842	want	-7.24865
414743	time	-7.27165
295630	parents	-7.28151
285727	one	-7.28863
118006	even	-7.34657
260593	mom	-7.37777
271295	need	-7.37784
167484	good	-7.39585
244168	make	-7.40369
409479	think	-7.40919
209762	kids	-7.42403
213387	know	-7.43354
12902	also	-7.44437
340543	right	-7.44695
162833	go	-7.45458
443318	way	-7.47081
262709	money	-7.47693
366066	sister	-7.49649
273857	needs	-7.50549
402700	tell	-7.51696

"YTA - you're happy to dump your kids on her, and you make an uneducated guess (you can't even name any of the reptiles) as to how dirty or dangerous they are over her advice as an educator.   Then you get mad when she tells your kids the truth!   There is not a single comment in your narrative that would make you NTA."

'YTA to your wife as you ARE stating that you do not trust her. It takes two to tango and your ex gf is at fault the same as your brother. You wife is not your ex girlfriend, but you are treating her as if she were. NTA to your brother as you have the right to want him out before you leave no matter the reasoning.'

"YTA if you don't walk away from such a selfish person that she thinks a WHEELCHAIR bound child should be upstairs solely because a room that MIGHT have guests in it is larger.\n\nNTA for insisting the downstairs room is rightly more suitable for her. \n\nIf your fiance is fighting about this now, do you really trust her to be a good mother to your daughter when you aren't around?"

"NTA but simply ask him to keep it down after X time especially on week nights (weekend nights you might have to compromise). If you bring up the fact he has people over when he's not home then YTA albc thats none of your business and a little creepy that you're keeping tabs on him."

"YTA - you're happy to dump your kids on her, and you make an uneducated guess (you can't even name any of the reptiles) as to how dirty or dangerous they are over her advice as an educator.   Then you get mad when she tells your kids the truth!   There is not a single comment in your narrative that would make you NTA."

YTA: 0.76

'YTA to your wife as you ARE stating that you do not trust her. It takes two to tango and your ex gf is at fault the same as your brother. You wife is not your ex girlfriend, but you are treating her as if she were. NTA to your brother as you have the right to want him out before you leave no matter the reasoning.'

YTA: 0.759

"YTA if you don't walk away from such a selfish person that she thinks a WHEELCHAIR bound child should be upstairs solely because a room that MIGHT have guests in it is larger.\n\nNTA for insisting the downstairs room is rightly more suitable for her. \n\nIf your fiance is fighting about this now, do you really trust her to be a good mother to your daughter when you aren't around?"

NTA: 0.81

"NTA but simply ask him to keep it down after X time especially on week nights (weekend nights you might have to compromise). If you bring up the fact he has people over when he's not home then YTA albc thats none of your business and a little creepy that you're keeping tabs on him."

YTA: 0.76