Predicting Who The Asshole Is

By Robert Malka

(With thanks to Ben Bell & Kenneth Gil-Pasquel!)

Who's the Asshole, Here?

How do people perceive what we're saying and why? Are there any qualities about our tone, word choice, and phrasing that would make someone like us less, think better of us, or carry our message across with further clarity? If we could find answers to this question by sifting through aggregate data, we might be better able to convince a judge/jury that we're not at fault, create a more effective business pitch, or, for the field of self-help, help others be more persuasive.

Shifting away from the narrow perspective of business value-add, we can also examine the limits of machine learning's understanding by means of this dataset – examining where machine learning, and the current state of AI generally, falls (far) short in its assessment of e.g. a person being (or not being) an asshole. Insofar as this can be shown through this project, we validate the value and uniqueness of being human.

We examine the subreddit, Am I The Asshole (AITA), which receives thousands of unique visitors a month. Its goal is to serve as a forum for people to ask the crowd whether or not they, within their self-reported experience, can be understood as an asshole. We investigate features that might serve to better predict whether or not an individual is behaving like an asshole, and experiment with Decision Trees, Random Forests, and Logistic Regressions to predict assholery based on features related to, among other things, the author's post and post title.

The business problem may not have a specific timespan or clear numerical benchmark attached to it, but that does not stop it from being potentially useful to a consulting firm or other entity. We will examine the presence or absence of features relative to the chance of receiving an NTA ("Not the Asshole) or YTA ("You're the Asshole) and, in spite of philosophical limitations, exhibit statistically significant improvements in ML benchmarks.

Caveats

Reddit's AITA subreddit decides whether or not someone is the asshole, given a situation they describe from their own experience, in their own words, based on a majority vote. This majority vote relies on numerous factors, namely:

- The people who are judging are self-selected rather than randomized (they choose to click on the thread).
- People who comment often give nuance in their posts – cultural markers for what is and isn't appropriate, ask clarifying questions, are sometimes thoughtful and unthoughtful – which isn't reflective of the final binary decision given (YTA/NTA). Upon individual inspection of these comments, a reader (or dare I suggest, a majority of readers?) might note a comment in the

minority opinion that is overall more persuasive, and more well-thought-out, than the inclinations of the majority judgment.

- The audience making these decisions may be grouped according to time of day, positioning of the post, controversy of the times (related news, fiery memes, something in the culture wars) that leads to one opinion or the other. There is no way, based on the data, to account for these inevitable movements of the national (or international) zeitgeist. We are assessing the smallest snapshot of the particular crowd X's overall, on-balance judgment.

1. Data

My data was received, largely clean, from:

- https://github.com/iterative/aita_dataset

2. Data Cleaning

Fortunately, not much cleaning was required. I changed the "edited" column, signifying whether or not the original post was edited, to "True" or "False" as opposed to epoch time; I changed the epoch time of the post's publication to datetime; and I took out a series of symbols that would have rendered feature analysis inaccurate.

3. EDA

For a full examination, please see the EDA report.

The hypotheses I examine can be grouped under three broad categories:

- The things that are said (such as how many of a certain pronoun are used versus another pronoun);
- The way they're said (such as whether people use "Am I the asshole" versus "Would I be the asshole);
- What the posts are about (familial disagreement, professional concerns, or romantic relationships).

Some interesting observations:

- Not having a question mark at the end of your sentence – leaving it as a period, say – correlates slightly with not being an asshole.
- The valence of one's tone is not correlated with assholery.
- There is a greater likelihood of being deemed an asshole if the issue is romantic (particularly if it involves sex).
- The younger the individual being discussed relative to the poster (and the lower in the power structure), the more likely the poster is to be an asshole.

- Conversely, the higher the authority (boss, parent), the less likely the poster is to be an asshole.

Basic facts about the dataset

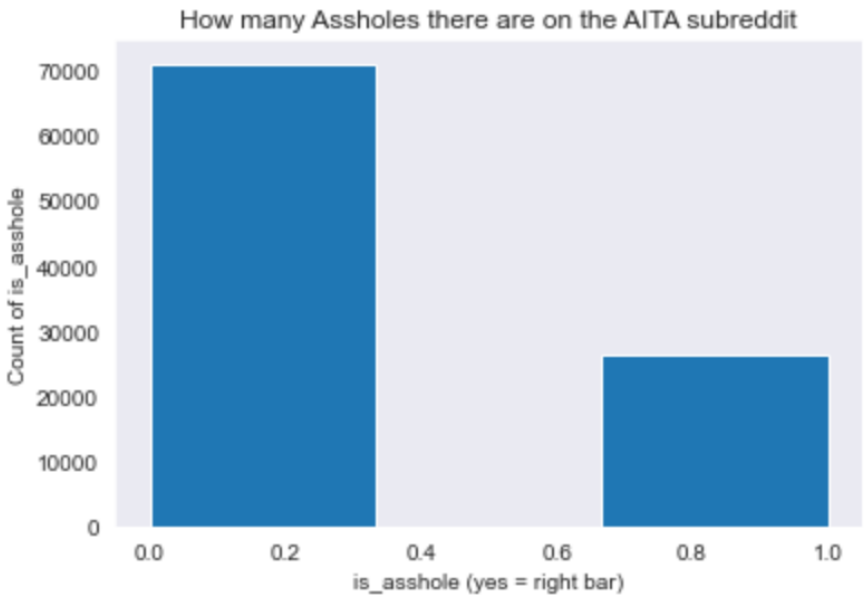
The basic features of the dataset are:

- id: The unique post ID
- timestamp: The timestamp of the post in datetime (originally in epoch time)
- title: The subject line of the post
- body: The full body of the post
- edited: Whether it was edited or not (T/F)
- verdict: The community's decision ("You're the Asshole" and "Everyone Sucks" both count as being an asshole)
- score: The number of "upvotes" the community gave to that post (roughly, how much interest the community had in that post)
- num_comments: How many people commented on the post
- is_asshole: 1 for yes, 0 for no

Of the *body* and *title* features, I developed core features to examine:

- countI: How many times I is said in the body of the text. (Does number of "I"s correlate with narcissism/being an asshole?)
- countHeSheThey: How many times He/She/They is said in the body of the text. (Does number of times OP (original poster) mentions others correlate in either direction to being an asshole (being thoughtful, or placing blame on others)?
- IvsHeSheThey: countI divided by countHeSheThey. Does any proportion between the se two numbers suggest being/not being an asshole?
- post_word_count: The length of the post.
- questionmarklast: Whether or not the post ends in a question mark (does asking versus saying, in this case, the last sentence, suggest a humility or arrogance?).
- WIBTA_AITA: Whether the title asks "Would I be the asshole", versus "Am I the asshole" -- are there any differences in decision based on choosing one or the other?
- bodyreadinglevel: Estimating the reading level of the body text.

The dataset was ~97,000 rows, and, as you'll see below, imbalanced:



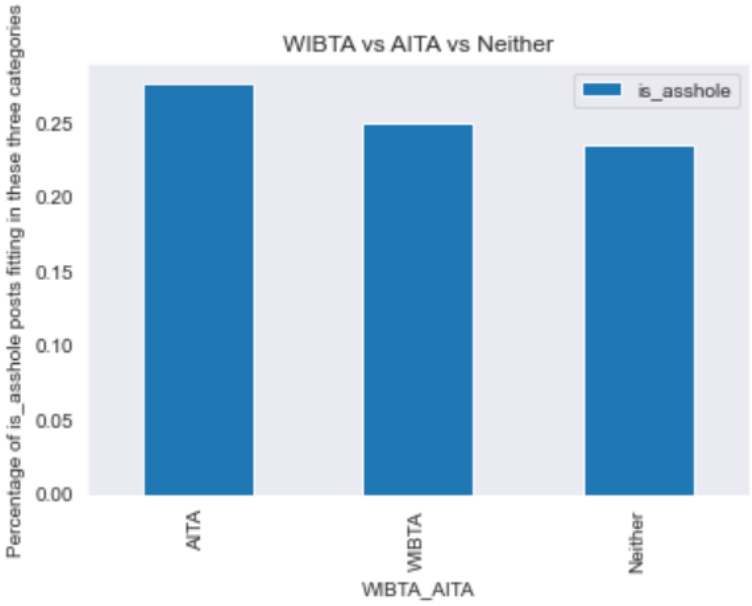
Most of the time, people are NOT considered assholes ('0' (the left side) are not deemed assholes) -- it's actually only about 25,000 of 97,000 people -- or a little more than 25% -- of people who are deemed an asshole at all. Not sure if this indicates a generosity in the community, since we have no broad standard. Suffice it to say: More than a quarter of the people who decide to self-reflect on their assholery are confirmed to have behaved like an asshole by the Reddit community.

The dataset was scraped from Reddit as recently as Feb 2020.

I also did a heatmap of the initial features, which is displayed below. I opted to examine other features by t-test and frequency relative to the is_asshole global average, which is why the features integral to a number of hypotheses are not included in this heatmap.

	edited	score	num_comments	is_asshole	countI	countHeSheThey	IvsHeSheThey	post_word_count	questionmarklast	bodyreadinglevel
edited	1.000000	0.070000	0.100000	0.090000	0.140000	0.120000	-0.000000	0.150000	-0.380000	0.000000
score	0.070000	1.000000	0.840000	-0.010000	0.050000	0.060000	-0.020000	0.060000	-0.030000	-0.000000
num_comments	0.100000	0.840000	1.000000	0.040000	0.060000	0.060000	-0.010000	0.070000	-0.030000	-0.000000
is_asshole	0.090000	-0.010000	0.040000	1.000000	-0.020000	-0.020000	-0.000000	-0.030000	-0.050000	-0.010000
countI	0.140000	0.050000	0.060000	-0.020000	1.000000	0.640000	0.200000	0.800000	-0.070000	-0.020000
countHeSheThey	0.120000	0.060000	0.060000	-0.020000	0.640000	1.000000	-0.350000	0.900000	-0.060000	0.060000
IvsHeSheThey	-0.000000	-0.020000	-0.010000	-0.000000	0.200000	-0.350000	1.000000	-0.150000	-0.000000	-0.080000
post_word_count	0.150000	0.060000	0.070000	-0.030000	0.800000	0.900000	-0.150000	1.000000	-0.080000	0.070000
questionmarklast	-0.380000	-0.030000	-0.030000	-0.050000	-0.070000	-0.060000	-0.000000	-0.080000	1.000000	-0.010000
bodyreadinglevel	0.000000	-0.000000	-0.000000	-0.010000	-0.020000	0.060000	-0.080000	0.070000	-0.010000	1.000000

One of these variables is WIBTA_AITA, or the question of whether or not someone is more likely to be an asshole using conditional versus present language ("Would I be the asshole" vs "Am I the asshole"). The difference between WIBTA, AITA, and neither is visible in the graph below:



Next, we look at the words that are most likely to predict is_asshole versus the words most likely to predict "not the asshole."

4. Wordcloud

A wordcloud of the titles of each post show the frequency of words outside prepositions, indefinite articles, and so on.

The most common words are nouns – girlfriends, friends, parents. A lot of the verbs are fundamentals of our modern experience: Wanting, telling, asking, letting, refusing. This suggests that many of the conflicts in our modern lives really do come down to clashes of what people want, what people say, what people consent to (or don't). (We recall that there were once worlds in which conflicts centered around beliefs, needs, sacrifices, enemies.) Interestingly, emotions aren't as present (upset, getting mad, and annoyed are tiny). Tragically, cutting is a word that occurs with enough frequency to be present in the wordcloud (it's bigger than "cheating", and only slightly smaller than "making").

5. Interesting Words

In order to figure out which words were indicative of being an asshole or not, I vectorized every word from *body*, excluding conjunctive, prepositional, and other related words. I organized them into a training set against is_asshole, and then used a naïve bayes model to examine how accurately they reflected is_asshole in any given post. From there I created a new dataframe with only the words and the percentage of the time they reflected is_asshole, and examined the top 2,000 and bottom 2,000 words.

While there are plenty of words I personally found interesting (the appendix of this report contains the top and bottom 20), I'd like to point out a few, from which I allow the reader to write his own story:

First, no words definitively indicated that someone was NTA. The scale was 1 for NTA, and 0 for YTA, on a 100-point scale. The ones I noted were heavily on the NTA side and particularly interesting were *revenge* (0.79), *insecure* (0.70), *included* (0.65), and *unsolicited* (0.65). Apparently, these words suggest contexts within this dataset favorable to the poster. (Note, surprisingly, that *include* is 0.01 – on the opposite side of the scale!)

More neutral, but leaning NTA, we notice *judging* (0.60), *addiction* (0.59), *force* (0.59), *attracted* (0.56). It seems safe to suggest that with judging, there may be a denial ("I'm not judging"), or perhaps a more passive statement ("He's judging me"). Particularly notable is *force* (0.59), which has a very different statistic from *forced* (0.35) and *forcing* (0.39) – which is fascinating.

To be *pc* is pretty neutral (0.51), which is interesting so long as we're not talking about computers, but political correctness; *disagreeing, nsfw, std, lose,* and *cheat* all go right down the middle (0.50), particularly surprising with std, but perhaps more understandable if one is describing whether or not they gave context surrounding their possible std. Finding two random samples, one described calling karma on a friend who had 2 STDs (NTA), and another who failed to properly split up with his FWB (YTA). Makes sense.

Cheated (0.41) and *cheating* (0.32) are both similarly (and strangely) worse than *cheat*, each successively nine points worse than the other.

We slowly enter into firm YTA territory when we see *affair* (0.49), *daddy* (0.46), *shaming* (0.43), *lazy* (0.43), and *sex* (0.39). You probably shouldn't talk to *strangers* (0.42). Even if *god* isn't one of us, He remains as controversial as the *bedroom* and trying to *forgive* (0.38). Whether you *apologize* after an *argument* or not, it ends roughly as badly for you (0.37).

Don't be *racist*… or *vegan* (0.32).

Consent is exactly as fiery as you might expect (0.30), but not as bad as whatever people report happens on or around *Christmas* (0.25). And if you like to talk about demanding *respect*, others being *controlling,* or about who's the *asshole*… well, it's probably you (0.24).

If it deals with a *newborn,* you're (basically) always wrong (0.13). *Abuse* is a *massive* no-no (0.01). *Minor*, be it the noun or the adjective, shafts the storyteller (0.01). Don't even try to be *accommodating* or *noisy* (0.01). And if you have to get to the point where you're *unfriending* or you *block* your friends, and then tell everyone about it… keep it to yourself (0.01).

And, finally, the scale you never knew you needed – the platforms + games where people are most likely to be affiliated with assholery, in descending order:

Tinder 0.42
Reddit 0.38
Instagram 0.36
Minecraft 0.30
Snapchat 0.23
YouTube 0.09
PS4 0.07
Netflix 0.06

You're welcome.

6. Thresholding

Thresholding gives us the level at which to determine, in our classification problem, whether or not someone is an asshole.

The optimal threshold for the decision tree classifier came out to 0.409.

For the random forest, the best threshold came out to 0.725, though the value of the F-score at that value was NaN. I opted therefore for the F-Beta score (0.5 threshold).

The Logistic Regression showed an optimal threshold of ~0.55.

Gradient Boosting gave a best threshold of .391.

7. Modeling

While I played around with accuracy as a metric (able to get as high as 72%), It seems to me that accuracy is the wrong metric for this dataset. Our goal is an optimal F-score, since we want a sense of the optimal precision and recall, or an understanding of how many true positives we have over the entire dataset tit predicts how many true positives there are over the entire set of assholes – and that's exactly what we want to know: How many assholes did we properly predict would be assholes, based on our features?

I grid-searched for each model and applied the ideal results, given my limited search. The model metrics and models I opted for are in the following table:

Model	Gridsearch Parameters	Recall	Precision	Accuracy	F1
Decision Tree Classifier	Criterion = Entropy, max depth = 4	0.577	0.555	0.551	0.669
Random Forest Classifier	Criterion = Gini, n_estimators =97, max depth = none (chose 12 from prev. gridsearches)	0.573	0.580	0.577	0.577
Logistic Regression	C = 0.01, penalty = 12	0.551	0.571	0.566	0.561
Gradient Boosting	Criterion = mae, learning rate = 0.1, loss = deviance, max depth = 3	0.549	0.559	0.555	0.669

We see that the best model for F-score is tied between the Decision Tree and the gradient boosting, although superior recall goes to the Decision Tree classifier, at 0.577 - I opted for a max depth of 4. The Random Forest Classifier had superior Precision and Accuracy, and recall in second place. These are mild but notable gains. All models fared only somewhat better than chance.

More investigation into the dataset is needed, to continue searching for superior features.

8. Most Predictive Features

For the Decision Tree, the most helpful features were *edited* (58.2%), and *parents* (41.7%).

For the Random Forest, it was *post_word_count* (8.1%), *IoverPostCount* (8.1%), and *bodyreadinglevel* (7.8%).

For Gradient Boosting, we see it was *edited* (25.7%), *parents* (16.9%), and *questionmarklast* (14.5%).

9. Summary

The business problem we focused on was seeing how we could best convey our concerns to juries, therapists, bosses and coworkers in the workplace, let alone all of the personal ways in which such information could be applied to make us better understood, and ideally better people.

While we found only a few features that came to be useful, we find them to be descriptive rather than prescriptive. Still, noticing some interesting suggestive (cultural) patterns and tendencies, we could infer with some confidence who the asshole was more or less likely to be, up to several percentage points above the global average of asshole or not.

Some descriptive features of note: The person we presume to be in a lesser state of power (a child) is less of an asshole than their parents; the boss is more likely to be an asshole than the coworker; the inlaws more so than the spouse. Question marks – presuming to ask a question about who the asshole is – suggests some humility in the person; when the issue is about sex, tensions are high. A number of words are highly suggestive of being an asshole (and so one hopes the topic is less about those subjects), while others suggest innocence.

So, in all frankness, the business problem is somewhat limited with respect to the scope of this project. But it is not irrelevant. Future improvements, below, could be gateways to new, usable insights for our business problem.

In terms of predicting assholery, the best model was tied (unusually) between the decision tree classifier and the gradient boosting, at an identical F-score of 0.669. However, it is worth noting that the gridsearch for the gradient boosting was done on a very small sample due to processing time taking at least five days on the whole dataset, so one suspects that a better f-score could be found on the GB with a TPU.

10. Future Improvements

We can group individual words by their part of speech, which I have left as a separate Dataframe for others to fool around with. Certainly other ways to examine tonality are interesting, as posters considered arrogant might get less sympathy from the community. Being able to examine and weigh comments based on their scores, and the awards they get, might help us put YTA/NTA on a spectrum, allowing us to see it as the more fluid subject that it is. Grouping the titles of posts by more nuanced commonalities than the three headings I created will likely bear more fruit. One could also run handpicked posts through the model and see if the model agrees with them. It's always cool to have a model useful to an individual's moral standards.

11. ***Huge Optional Aside:*** *On the Limitations of Our ML Analysis*

You're certainly wondering why I mention something so grandiose as examining the limits of machine learning's analysis within an otherwise-straightforward search for features predicting assholery.

The answer is, simply put, that the features we would like to have in examining this dataset, to accurately guess at the crowd's decisions, are not possible through machine learning, and that the features which do have significant predictive power, and which are wonderfully creative and helpful, only reveal the chasm between AI as it is, currently, and the human being, and which show clearly what problems ought to be worked on with ML and which ought better to be left to humans and their inherent strengths – problems, for example, like predicting who's an asshole (or, better yet, deciding what an asshole is).

Let me explain.

At the heart of this conversation live two famous philosophers: Plato and Heidegger. Plato asserted that there exists a form of things, an ideal, which nothing ever truly is, but always aspires to be. For example,

a four-legged chair, while coming in many different shapes and sizes (including three- and two-legged chairs), 'borrows' in some way from the form of a chair. That's how we know it's a chair. We adopt this philosophy when we use supervised learning: Every picture of a chair gets labelled a "chair," and gradually some "form," some estimation of a chair, exists in the network. It thus predicts whether or not this or that picture is of a chair (does it have features such that it belongs in this group?), or something else. (Plato applied this philosophy also to The True, The Good, The Beautiful, and so on, but we won't go into that here.)

Heidegger, by contrast, insisted that things do not reach upwards, into idealized forms: They are best understood by means of their functions. How do we know whether a chair is a chair? By means of the fact that we sit on it. This is why a beanbag chair and a three-legged chair are both chairs. They are still capable of being wielded by a human being for their decided-upon function: Sitting. Heidegger understood objects as being not a "what" but a "what-for." Yet – and this is key – chairs do not have to be understood as made for sitting. They could be used as tables, such that humans sit on their knees. They could be sacred emblems representing the great Chair Goddess. They could be building blocks used to build the world's largest blanket fort.

The only way we can understand a chair – really understand it – is if we are embodied beings, able to interact with, participate in, and be inseparable from the World, as we are. AI is a tool that is part of this world, but it is not able to come to grips with the world because, as of yet, it is not encapsulated within a body/ies able to seamlessly interact with chairs, or beanbags, or assholes.

So, why does this matter?

Because what we most want to know, in this project, is whether person X is an asshole, according to the community – but there are innumerable cultural grounds through which to understand this or that person as an asshole (and what the community might even mean by designating someone as an asshole or not) – and almost all of them are inferred 'features,' not explicitly stated in the dataset. I'll share an example:

I use reading level as one feature to see whether people might be more or less of an asshole, given a certain writing ability. The ratings system I use is Flesh-Kincaid**.** I noticed at least one negative (!) reading level, right off the bat. This means the post was too basic – too short – for the algorithm to properly assess it (the math behind it can be found here). Yet the community seemed to have understood: The poster was "Not the Asshole." A cursory review clarified things. The post in full is:

**Title:* I told a goth girl she looked like a clown.*

**Body:* I was four.*

So we see immediately that not only is it not "confusing", but remarkably clear, in context. (When title and body were included, the reading level remained negative, as a slightly-improved -1.4.)

Notice how much isn't included in this post, how much we bring to it, to understand why the crowd determined this person was not an asshole. We introduced the archetype most people hold for "a goth girl," and all of the nuanced qualities a goth girl has; the quality of the insult "clown" and how it fits (or doesn't) with being a goth; the fact that this could have been understood as an insult or presumed to be so; and how a human being can broadly come to understand and value (or not) the remark as it's made

by a four-year-old. There is nothing reducible to "data-crunching" in the interplay of each of these pieces. Even if, for example, we established a rule dictating that remarks made by children should be labelled as “Not The Asshole”, that is a culturally-contextual understanding. Is that still the case in China?, Papua New Guinea?, the United States in 1953 or 2031?

Even within the context of this post, a whole story can be developed, perhaps even one in which the author is a less sympathetic character. We can imagine several responses from the goth girl: laughter, maybe, or, if she felt insecure, perhaps she *was*, in some indirect way, insulted; perhaps the child’s tone was unacceptably inappropriate, indicative of a budding future asshole. But the response of the community (a community that, remember, is ever-changing – reliant on who saw and decided to respond to the post) seems to have decided here that a kid's innocence excuses him from what is otherwise a crass remark.

The bottom line: culture, be it American or any other, is not reducible to atomic elements. It is an ecosystem with an interplay of nearly infinite parts, to which we, and not AI, have access. We are part of a culture that understands the significance of being an ‘asshole,’ all the ways in which one can be one, *why* this or that person was called an asshole, whether we agree with that assessment, and so on.

By this project I’m reminded that the words on a page or images on a screen are referential, pointing towards things that are in the world. They are not the world itself. It’s therefore impossible to be able to communicate the significance of any event, thing, or experience the data is pointing towards using only the data, since we fill it in with our embodied experience. Word mappings, as one example, capture the relationship between words as they are said, but they do not reveal what those words *embody*. Word mappings will always be a cursory explanation (at best) of the embodied understanding.

There is one caveat here: Embodiment, it seems to me, can be manifested in ways outside of possessing an explicit body *vis a vis* robotics or brain implants. The analyzing of complex, nuanced human creations such as culture can be understood in virtual space, where bodies navigate using technical proxies (eventual full-body suits, for example). This seems to me a possible opening to kind of being-in-the-world currently out of the realm of machine learning.

Until this happens, though, the reader will notice in our analysis that these features of the dataset fundamentally cover only the shallowest inferences of assholery.

I hope we humans fearlessly assert what is human about us – our ability to make meaning of things, and invent and create those meanings for us to share and celebrate – and not presume AI has that power over us, and that we are helpless before it. It is, after all, stupid, as existentially stupid as can be.

For a full discussion on the limitations of AI, investigate Hubert Dreyfus’ *What AI Can’t Do*.

12. Acknowledgements

Major thanks to Benjamin Bell, my Springboard mentor, for his vision and help with this project; to Kenneth Gil-Pasquel for his help in cleaning up my work; to the AITA community for this wonderfully-rich data, from which we discovered some tremendous insights.

Appendix: Predictive Words (top 20, bottom 20)

Top 20 words		**Bottom 20 words**	
revenge 0.79	intentionally 0.64	fwb 0.13	abuse 0.01
ass 0.78	failed 0.63	snitching 0.12	noisy 0.01
vs 0.75	policy 0.62	relatives 0.12	carry 0.01
insecure 0.70	wont 0.62	vasectomy 0.11	minor 0.01
failing 0.69	supervisor 0.61	stepmom 0.11	accommodating 0.01
purpose 0.68	judging 0.60	doctors 0.10	adopting 0.01
cussing 0.68	please 0.60	baseball 0.09	include 0.01
embarrassed 0.66	addiction 0.59	youtube 0.09	unfriending 0.01
dumb 0.66	force 0.59	miscarriage 0.09	block 0.01
included 0.65		ps4 0.07	
unsolicited 0.65		netflix 0.06	