Analyzing Reddit Posts to Predict Society's Verdicts on Morals

Sahil Farishta
University of Michigan
sahilf@umich.edu

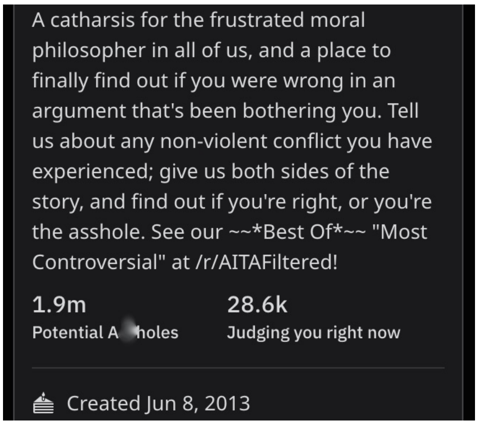
Rupal Nigam
University of Michigan
rupaln@umich.edu

Abstract

Reddit is an increasingly popular social news website. The online forum has hundreds of communities within it called subreddits. A popular subreddit is AITA, where a user posts a morally ambiguous situation they have experienced and the community comments and votes on whether or not the user made the morally correct decision. In this paper, we discuss methods to predict whether or not society finds a member to be in the wrong and what factors contribute to this verdict. The results of this project will help better understand societal norms. This has applications in a variety of situations, such as psychology, crime, and political scandals.

1 Introduction

This project will work with data from the AITA subreddit on the social media platform, Reddit. This community within Reddit consists of posts written by a user who has been in a complex situation and is reaching out to other members to determine whether or not the user was in the wrong. The AITA subreddit is arguably one of the more popular and active ones on Reddit, with about 1.8 million people subscribed to it, as shown in the figure below.



A catharsis for the frustrated moral philosopher in all of us, and a place to finally find out if you were wrong in an argument that's been bothering you. Tell us about any non-violent conflict you have experienced; give us both sides of the story, and find out if you're right, or you're the asshole. See our ~~*Best Of*~~ "Most Controversial" at /r/AITAFiltered!

1.9m
Potential A[redacted]holes

28.6k
Judging you right now

Created Jun 8, 2013

Figure 1: Description of the AITA subreddit

The subreddit allows readers to vote on various situations indicating whether they thought the original poster is morally wrong. Understanding what people look for in a post as they judge whether it is morally right or wrong can shed light on societal values. Thus, we propose a system that will attempt to predict whether a given post would be found to be morally right or wrong.

1.1 Problem Statement

The problem we are trying to solve is whether or not we can predict whether a person in a morally ambiguous situation was correct in their actions or not. This is important because these types of situations can fuel intense debates and can arise in all facets of life. Determining what factors contribute to these situations can help prevent them from escalating.

Additionally, analyzing what features lead to a post being perceived as morally right or wrong will help us understand what society values in terms of morals along with how society perceives content in terms of the structure it is presented in. It may be the case that content that is presented briefly in a polite manner with no grammatical errors is perceived to be morally right more often than content presented in a verbose manner with many grammatical errors. These insights can lead to valuable conclusions about society's morals, values, and decision making.

1.2 Related Work

Previous work has been done in getting models to make decisions in moral dilemmas. In the artificial intelligence and autonomous vehicles realm, a team from MIT has used data from Moral Machine to describe an individual's values using a Bayesian network (Kim et al., 2018). Moral Machine is an application that collects judgements on ethical dilemmas regarding autonomous vehicles. The team found that the model performed well in situations where the dilemma was not too difficult and was unable to predict more difficult dilemmas. The team was interested in expanding their work to applications outside of autonomous vehicles.

Another team from Duke questioned whether AI systems have to use a set of ad-hoc rules or can make moral decisions like humans do (Conitzer et al., 2017). They propose a framework that uses a game theory approach with two players and discuss extending it to more players and broader situations. The team concludes that using machine learning with this framework is a good options because the features and results can be explained and analyzed. That way conclusions can be drawn about what factors contribute to a decision in a moral dilemma.

1.3 Proposed Method

To address this problem, we created a system that will take in various features from a Reddit post and run them through a Random Forest Classifier in order to determine the label of the post. This involved balancing and preprocessing the data. We then determined the optimal hyperparameters by performing a gridsearch over the various hyperparameter options using cross-fold validation on the training data using weighted F1-score as our metric. Once the model was trained, it was run against the test data and the model was analyzed to determine its performance and find what features were important in predicting the label.

1.4 Summary of Results

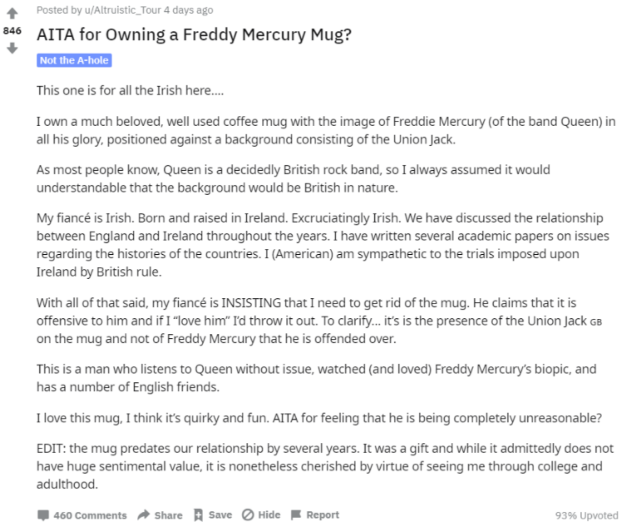
We found that the model had an accuracy and F1-score that surpassed that of a model with random guessing and a model that choose the label with the highest frequency. We were able to analyze the model coefficients to determine what features the model found to be the most important in determining a posts label. The analysis revealed that the Reddit features, sentiment analysis of the post, the number of errors in the post, and the post length to be the most important features. We also found that the two labels with significantly fewer data points, ESH and NAH, were rarely classified correctly by the model, indicating that a more balanced dataset with larger amounts of data is necessary for work in this area.

1.5 Report Organization

In this report, we discuss the data used to train our model in greater detail. We analyze the data that we collected and how it was processed before being fed into the model. We then discuss the way the model was trained and how hyperparameters were selected. We discuss the results and what we learned from the model based on its performance on the test data. Finally, we analyze the ethical concerns associated with this work and future considerations.

2 Data

Data from the AITA subreddit is used to extract features that will be used to train the model. This subreddit is a place for users to present a moral dilemma they were involved in and ask for a judgment of their actions from the community. Below is an example of a post that can be found in this subreddit.



AITA for Owning a Freddy Mercury Mug?

This one is for all the Irish here....

I own a much beloved, well used coffee mug with the image of Freddie Mercury (of the band Queen) in all his glory, positioned against a background consisting of the Union Jack.

As most people know, Queen is a decidedly British rock band, so I always assumed it would understandable that the background would be British in nature.

My fiancé is Irish. Born and raised in Ireland. Excruciatingly Irish. We have discussed the relationship between England and Ireland throughout the years. I have written several academic papers on issues regarding the histories of the countries. I (American) am sympathetic to the trials imposed upon Ireland by British rule.

With all of that said, my fiancé is INSISTING that I need to get rid of the mug. He claims that it is offensive to him and if I "love him" I'd throw it out. To clarify... it's is the presence of the Union Jack GB on the mug and not of Freddy Mercury that he is offended over.

This is a man who listens to Queen without issue, watched (and loved) Freddy Mercury's biopic, and has a number of English friends.

I love this mug, I think it's quirky and fun. AITA for feeling that he is being completely unreasonable?

EDIT: the mug predates our relationship by several years. It was a gift and while it admittedly does not have huge sentimental value, it is nonetheless cherished by virtue of seeing me through college and adulthood.

Figure 2: Screenshot of a post in the AITA subreddit

We used Python scripts that utilize Reddit's API to scrape posts from the AITA community from the past week. This was done through the Python Reddit API Wrapper (PRAW) library and the Python Pushshift.io API Wrapper (PSAW) library.

2.1 Data Volume and Criteria

Using the aforementioned Python libraries, 25k posts were scraped from the subreddit. It was decided that all data selected should be tagged as ”Safe For Work” as many posts on this subreddit can contain ”Not Safe For Work” content. Additionally, we only look at posts that have a flair, which is what we were using as our labels. After applying these criteria, 7k posts remained, which will be down sampled and split into training and testing data.

2.2 PRAW

The data from Reddit was mined using the Python Reddit API Wrapper (PRAW) library. The python script written goes through the posts on the subreddit in chronological order. Posts that have a flair and are 'safe for work' are stored in a dataframe, along with the . The data was then exported as a csv file.

2.3 PSAW

Using PRAW alone we can only scrape 1k posts due to request limitations set by the library. This amount was not nearly sufficient enough, so the PSAW library was necessary. This library allowed us to request far more posts and 25k was set as the limit.

2.4 Labels

In this subreddit, members of the community can express their opinion on the situation detailed in the post by voting in comments. There are five votes possible:

1. NTA: Not the A-hole. This indicates the opinion that the poster is morally correct
2. YTA: You're the A-hole. This indicates the opinion that the poster is morally wrong
3. ESH: Everyone sucks here. This indicates the opinion that the poster and everyone involved in the situation described is morally wrong
4. NAH: No A-holes here.This indicates the opinion that the poster and everyone involved in the situation described is morally correct
5. INFO: This is a request for more information from the poster before making a final decision

We are using a classification scheme for determining the label. There are four types of posts- NTA, YTA, ESH, and NAH. Posts that have had sufficient votes on the subreddit are assigned a flair, indicating the overall opinion of the voters, using the most voted option. We use this flair as the label for the post, and as such, only consider posts that have been assigned a flair.

3 Processing

We used Doc2Vec to provide continuous bag-of-words model to pre-processs the post content and extract the text related features. This allowed for contextualization of words. Previous research papers analyzing sentiment in Reddit posts have had success using this pre-processing model (Shen and Rudzicz, 2017). The title was also pre-processed into a Doc2Vec array, but no additional features were extracted from the title. Doc2Vec was used to create 300 dimensional vector representations of the text. 300 was chosen as the vector length based on previous findings indicating that a 300 dimensional vector resulted in the best performance while not being overly complex (Pennington et al.).

3.1 Features

We used post length, sentiment, grammar and spelling, and post content as the main features. Additionally, we look at some of the post metadata, including, title, post score, upvote ratio, number of awards received, number of comments, and number of cross-posts. Using features such as the post length and sentiment in Reddit posts has previously been done in research, with a team using sentiment, politeness, and post length in their research on whether a request for Pizza in the subrredit Random Acts of Pizza would be fulfilled or rejected (Althoff et al., 2013).

Initially, we anticipated extracting features from the mined Reddit posts to be a substantial amount of work. However, the aforementioned Python libraries vastly simplified this task. We were able to extract the desired features using the library in the python script itself. The extracted features are part of the csv data file.

For politeness, we are using the NLTK SentimentIntensityAnalyzer package. This gave us positive, negative, neutral, and compound sentiment scores. The compound score is the norm of the other computed scores. To determine the grammar and spelling of a post, we use Python's language-check library. The post content was in the form of the output from Doc2Vec, a numerical representation of each post. These features were used to train the model. As development continues and results are analyzed, some of these features may be found to not be relevant while other features not considered at this time will be added.

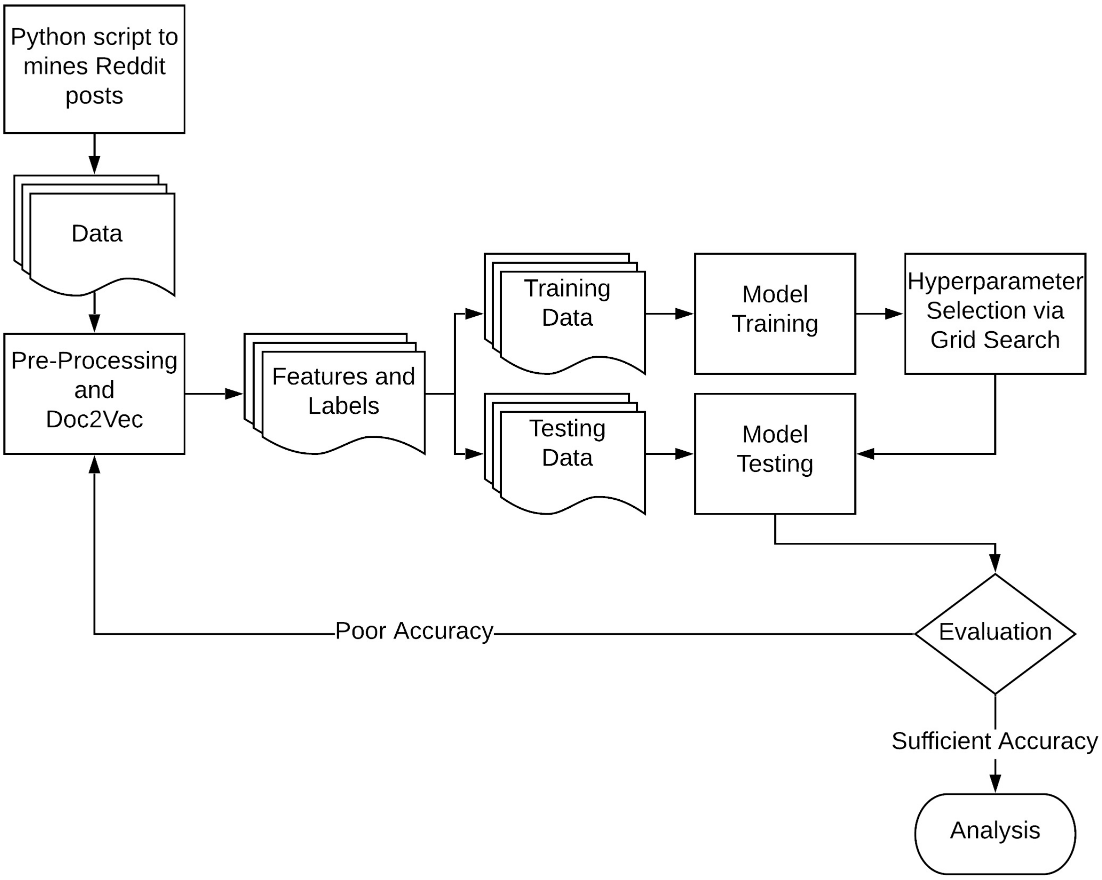


Figure 3: Overview of the steps taken to train and evaluate the model

The data mined from Reddit was read from the resulting csv file in order to pre-process the text. The post and title content were used to train a Doc2Vec Continuous Bag of Words model. Each title and post text were converted into a feature vector using Doc2Vec and stored with the data. Additionally, we analyzed the sentiment of each post and checked for grammar and spelling errors. The sentiment, number of errors, and length in words of the post were saved with the data as well.

3.2 Method

The aspects of the project detailed above form the pipeline for this project. This entire process is shown in Figure 3 and highlights the iterative process of designing and training the model until satisfactory results are achieved. Following this, analysis on the results will be done to determine which features contribute to a verdict of morally wrong or right.

4 Algorithms

We began by downsampling the data obtained in order to balance the collected data. We used a Random Forest Classifier as our model as its ability to have multiple trees voting on a problem closely resembles how commentators vote on posts on the subreddit. To select the model hyperparameters, we performed cross-fold validation on the training data. The training data comprised of a random sample of $\frac{5}{6}$ of the downsampled data. The remaining $\frac{1}{6}$ of data was used for testing the model afterwards.

4.1 Downsampling

The raw data was incredibly unbalanced, with 61% of the data having a label of NTA. To downsample, we brought the number of posts with label NTA down to 1500. This resulted in 39.4% of posts being labeled NTA, 35.4% of posts being YTA, 15.4% of posts being NAH, and the final 9.8% being ESH. This balanced dataset helped prevent the model from developing a bias.

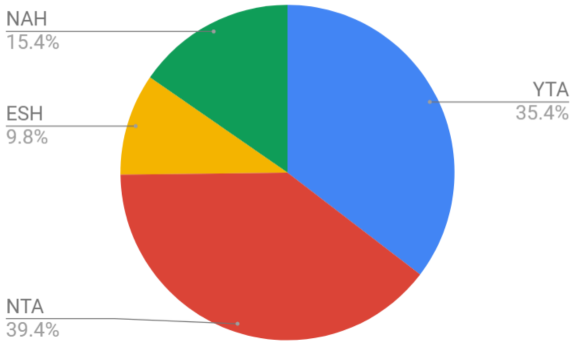


Figure 4: Overall label distribution after downsampling

4.2 Performance Parameters

When training the model we used weighted F1-score as the performance metric. The reason we used this above a metric such as accuracy is it penalized the model for simply picking the most occurring label every single time. This was incredibly important when dealing with an imbalanced dataset. The F1-score is calculated as seen below:

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

Precision and Recall rely on the number of True Positives (TP), the number of False Positives (FP), and the number of False Negatives (FN). The precision is calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

And the recall is calculated as follows:

$$Recall = \frac{TP}{TP + FN}$$

For a model that guesses randomly every single time based on the label frequency in the training data, the true positive rate is the label frequency squared. This is because the probability of guessing any label is the frequency it appears at in the data and the probability that a given post belongs to that label is also the same frequency. The false positive rate is the label frequency times (1-label frequency) since that is the probability of guessing the label times the probability that a given post does not belong to that label. Finally, the false negative rate is (1-label frequency) times the label frequency since that is the probability that the model picks any other label times the the probability that a given post belongs to that label. Assuming f_i is the label frequency for the ith label, we can calculate the precision and recall if we guess randomly as follows:

$$Precision = \frac{f_i^2}{f_i^2 + (f_i \cdot (1 - f_i))} = f_i$$

$$Recall = \frac{f_i^2}{f_i^2 + ((1 - f_i) \cdot f_i)} = f_i$$

This gives an F1-score as follows:

$$F_1 = 2\frac{f_i^2}{f_i + f_i} = f_i$$

Finally, the weighted F1-score is the F1-score for each label multiplied by the we frequency at which the label appears. We calculate the weighted F1-score as follows:

$$Weighted\ F_1 = \sum_{i=1}^{4} f_i \cdot F_1 = \sum_{i=1}^{4} f_i^2$$

Which for our dataset is 0.3139.

For a model that simply predicts the label with the highest frequency every time, the precision and recall is 0 for every other label since there are no true positives. This results in an F1-score for those labels of 0 as well. For the label that we predict we have a true positive rate of the label frequency since the model always predicts this label but only gets this prediction right at the frequency this label appears. The false positive rate is (1-label frequency) since the model always guesses this label but is wrong at the frequency that every other label appears. Finally, the false negative rate is 0 since the model always predicts this label, leading to no false negatives. This allows us to calculate the precision and recall as follows where f_i is the frequency this label occurs in the data:

$$Precision = \frac{f_i}{f_i + (1 - f_i)} = f_i$$

$$Recall = \frac{f_i}{f_i + 0} = 1$$

This gives us an F1-score as follows:

$$F_1 = 2\frac{f_i}{f_i + 1}$$

For our data, the highest occurring label is NTA which occurs at frequency of 0.394. This gives us an F1-score of:

$$F_1 = 2\frac{0.394}{1.394} = 0.5652$$

Finally, the weighted F1-score is the F1-score for each label multiplied by the we frequency at which the label appears. We calculate the weighted F1-score as follows:

$$Weighted\ F_1 = \sum_{i=1}^{4} f_i \cdot F_1$$

For our dataset, the F1-score for every label other than the most frequent is 0, so we get a weighted F1-score of $0.394 \cdot 0.5652 = 0.2227$.

Comparing this to accuracy, where guessing randomly would simply yield an accuracy equal to the sum of the true positive rates which yields an accuracy of 0.3139 for our dataset. Similarly, for accuracy for a model that simply outputs the most common label every single time, the accuracy is simply the frequency of that label, which for our dataset is NTA, which occurs at frequency of 0.394. Thus, the accuracy of a model that simply outputs the most common label every time has an accuracy of 0.394 for our dataset.

From this, we can see that F1-score penalizes a model that simply outputs the most common label every single time much more than accuracy does. When dealing with an unbalanced dataset, this is a huge concern, as a model that outputs the most common label every time could have a high accuracy but would not have very strong predictive power. Thus, we choose F1-score as our training metric.

4.3 Hyperparameter Selection

During training, we used cross-fold validation for training and hyperparameter selection. We used weighted F1-score as the performance metric as described earlier in Section 4.2. The hyperparameters considered were the number of estimators, the minimum number of samples required to be at a leaf node, the minimum number of samples required to split a node, and the maximum depth of each tree. These hyperparameters were evaluated through a grid search where the model with the highest weighted F1-score was selected. For cross-fold validation, the training data was split up into 5 folds, and the average weighted F1-score across the folds was used as the score for the model.

4.4 Testing

To test our model, we trained a Random Forest Classifier using the hyperparameters that were found to create the model with the highest F1-score. This model was trained on all of the training data, and then run against the test data. The predictions were compared to the test labels in order to analyze the model performance.

5 Results and Discussion

After testing the model we were able to calculate the F1-score and accuracy of the model on the test data. We found our model had an F1-score of 0.4496 and an accuracy of 0.4945. We were able to construct a Confusion Matrix based on the predictions. This is shown in Table 1.

		Predicted			
		YTA	NAH	ESH	NTA
Actual	YTA	131	4	5	96
	NAH	22	2	2	39
	ESH	21	0	4	56
	NTA	60	3	13	177

Table 1: Confusion Matrix for Test Data

We can look at the model coefficients to evaluate what the model found to be important in terms of predicting the labels. The model found upvote ratio, number of comments, post score, compound sentiment, post length, number of grammar and spelling errors, and positive sentiment to be the most important features in determining the label. The least useful parameters were found to be number of awards and the number of crossposts. The rest of the parameters, including the values for the two 300 dimension Doc2Vec vectors that represent the post content and the title, were distributed in between the features aforementioned.

Looking at the features the model found to be important, we can make sense of why the model might be valuing them. Posts where the original poster is not in the wrong may have many comments and upvotes to show the poster support in the dilemma they are facing while a poster who is clearly in the wrong in eyes of many in the community will get downvotes or ignored. Additionally, posts that get upvoted are more likely to be seen by other users based on Reddits algorithms for determining top or hot posts. This would result in posts with a lot of upvotes getting more views and comments. The number of upvotes and downvotes on a post are used in determining the posts score, where downvotes count as negative and upvotes count as positive.

In terms of sentiment, posts where the user portrays their language in a positive manner may garner more public support and shape the view other users have of the dilemma that makes the original poster seem not at fault. Similarly, a post that gives a negative sentiment within the language may influence the viewers opinions in a negative manner, causing them to feel that the original poster was in the wrong in their actions. Posts where the original poster was found to be in the wrong may be inherently longer, as the poster may try to defend themselves from the criticism or explain themselves better. A post with many grammar and spelling errors may seem like its coming from a person with poor literary skills, which could trigger biases in the commentators. This may lead to the original poster facing prejudice and being unfairly voted as being morally wrong.

Two features that were found to be less useful were the number of awards and crossposts. Reddit awards are given to posts that are found by a user to be funny, interesting, high quality, or any other time an award giver feels a post is worthy of an award. Crossposts refer to the number of times a post was shared on another subreddit as well. Both of these metrics can be very high for viral posts, and are usually 0 for more posts. Since all types of posts can become viral, especially when the situation described in the post is quite comical, it makes sense that the model did not find these features to be useful in distinguishing who was at fault in the post.

We also see that the model struggles on the labels ESH and NAH which makes sense since these two labels had very little data associated with them compared to the other two classes. The model was biased against these two labels, with very few predictions in general made for these classes. In future work, it would be important to ensure we have a balanced dataset in order to have a model with high predictive power. Additionally, analyzing what features influenced the model towards a certain decision could shed some light on some of the more complicated features, such as neutral sentiment. With features like this, while the model values it, we are unsure if a high neutral sentiment results in a higher likelihood of the poster being classified as morally wrong or morally right. Determining these tendencies would help us understand better how people are influenced while making judgements on morally ambiguous situations.

6 Ethical Considerations

Below are three major ethical considerations for this project. Since the project uses data from a social media platform, there are concerns of invasion of privacy as users post personal and sensitive anecdotes. Additionally, often posts include the author's age and gender, so bias is also a potential issue. Lastly, since one of the model's prominent features was grammar, there is concern about bias against authors that speak English as a secondary language.

6.1 Personal and Sensitive data

The AITA subreddit centers around authors providing details of a morally ambiguous situation they were in. As a result, they often write and post about personal and sensitive stories. Although Reddit is a public forum, using the authors' stories to train the model is not something they gave explicit consent for.

6.2 Gender and Age Bias

Often, in Reddit posts, the author includes their age and/or gender by saying 19F, for example. Here, the reader would learn that the author is a 19 year-old woman. The model could potentially learn this information about age and gender and become biased either against certain age ranges or gender. For instance, if the data used to train the model happens to have younger authors labelled as YTA more frequently, the model may learn this behaviour and thus be biased against younger authors. A way to prevent this is to remove such information from the posts.

6.3 Dialects and ESL

Reddit is online community and thus attracts users worldwide and has a large international presence. The AITA subreddit is in English but has users from across the world, and thus sometimes posts are written with different grammars and spellings. People who have not learned English as their first language will write posts in a different style. Additionally, even within America, there are different words used to convey the same thing depending on region. The model does use grammar and spelling as a feature, so there is concern that it could be biased against such users.

7 Conclusion

In summary, the question we set out to answer was whether we could predict which side the majority of society would side with in a morally ambiguous situation. In a nutshell, our model is able to predict this about half of the time. This is better than guessing randomly (since in our data there were 4 possible labels), and also better than guessing the most frequently occurring label. However, it is not a very high accuracy in general. This can be reconciled with the fact that we are attempting to predict the majority vote, which is not necessarily the majority by a large margin.

The factors that were most important in predicting a verdict were post popularity, sentiment, grammar, and length. From this we can infer that the style in which the post is written carries the greatest weight. Additionally, many features specific to Reddit were favoured over post content. In the future it will be interesting to train the model solely on post content as that would give us a better understanding of when society judges a situation in either direction. This would also allow the model to be applied in situations outside of Reddit more easily.

References

Tim Althoff, Niloufar Salehi, and Tuan Tu Nguyen. 2013. Random acts of pizza : Success factors of online requests.

Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *AAAI Workshop on AI, Ethics, and Society*.

Richard Kim, Max Kleiman-Weiner, Andres Abeliuk, Edmond Awad, Sohan Dsouza, Josh Tenenbaum, and Iyad Rahwan. 2018. A computational model of commonsense moral decision making. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics.