Hamza Baccouche
hbaccouc@umich.edu
University of Micigan
Ann Arbor, MI

ABSTRACT

This study was conducted with the objective of constructing a classifier to predict relatively reliably whether a post was sourced from r/askReddit or r/askScience based on its linguistic properties. The classifier was built primarily using tokenization, stemming, and logistic regression methods that associated root words with the likelihood of originating from a given subreddit. The model succeeded, generating precision and recall scores of 87% and 96%. The most strongly correlated words were as expected, as scientific terms were most strongly associated with r/askScience and casual terminology such as profanity was strongly associated with r/askReddit. The takeaways from this were the adhering nature of conversation on r/askScience and the strictness of their moderators, as opposed to the open-ended nature of r/askReddit. The most prominent opportunity for application of this model is auto-moderation with further training and model reliability.

ACM Reference Format:
Hamza Baccouche. 2019. . In ,. ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

The model used in this study is attempting to predict which subreddit a given text post on reddit belongs to, either r/askReddit or r/askScience, using sets of test data and training data.

There are a variety of potential takeaways to be had from the study. One potential takeaway lies in an extension of the research currently being done of classifying reddit posts as civil and uncivil: rather than pulling posts from a single subreddit and categorizing them, posts from multiple subreddits can be categorized and then rates of incivility can be compared across subreddits should the model be able to reliably predict the subreddit of origin. Another takeaway from the data will be a general sense of how people seek information based on different topics; the terminology used and their respective correlations to subreddits will tell a lot. Things like vulgarity, politeness, and expression of opinion vs. fact (i.e. terms such as "I think" being associated with one subreddit over another) will all be telling of the nature of interactions within the two subreddits.

There aren't any very clear potential uses of this experimentation, but one thing that I see as having the most potential is auto-moderation, but it would require further experimentation. The classifier will sort posts by subreddit by examining the vocabulary content of its posts, and a lot of the posts will simply be [removed]. With further experimentation and access to the contents of removed posts, the model could learn to identify posts that should be removed based on previously removed posts, and if it turned out to be reliable enough it could eventually eliminate the need for manual moderators and instead auto-moderate the subreddits on its own.

2 RELATED WORK

2.1 Bag of Communities:

Chandrasekharan et al. 2018[1] explore a new approach to auto-moderation in online communities through the "BoC," or "bag of communities," concept. This approach is unique in that goes outside the scope of the community that is being moderated. It instead utilizes pre-existing data on moderated content from nine different communities across reddit, 4chan, Voat, and MetaFilter. The primary benefit of this is that it is much more objective in judging content, thus removing explicit bias in situations like Reddit or YouTube where designated moderators of an online community dictate what is appropriate and what is not.

The model utilized 10 million posts of training data and was tested on 200,000 posts from the target community. Using only training data from the nine communities (aka the "Mixed Bag") and with no training data from the target community, the model was 75% accurate in removing posts that should be removed according to guidelines. After being trained on 100,000 of the target communityâĂŹs posts, the model improved to over 91% accurate. The research team thus concluded that the Bag of Communities method would be relatively effective in doing most of the work of moderation.

2.2 Language Style in Subreddits:

Tran et al. 2016[2] conducted at the University of Washington sought out to categorize posts in various subreddits with argumentative natures (changemyview, politics, atheism, etc) by the language styles used in the subreddit and identify how new content was received in thos communities. They found that the language style of a subreddit to be much more indicative of how well new content is received than the actual subreddit topic. This means that even if debating a touchy subject, such as politics or religion, a community whose discourse is typically civil will handle controversial content much better than a less controversial subreddit with more uncivil discourse would, such as fitness.

3 DATASET

The dataset used for this study was a randomized sample of all of the posts from all of the subreddits used in the study from the month of December 2018, but my portion only required the data from the r/askReddit and r/askScience subreddits. Having such a large dataset (the entire month's posts) allows for a very representative sample of what a typical post looks like within a subreddit so the model can be as accurate as possible. Each post's data set contained the author of the post, when it was posted (UTC standard time), which subreddit it originated from, the link id and parent id, the

Classifier performance:

Test set classification report:

		precision	recall	f1-score	support
AskScience	(0)	0.96	0.87	0.91	1028
AskReddit	(1)	0.87	0.96	0.92	966

upvote 'score' for the post, and the post's text body. When running the model, the dataset was siphoned down to 10,000 posts for the model to use for the sake of running code in a timely manner.

4 RESULTS

4.1 Model Performance

The following data was reported when the model was run, where 0 was r/askScience and 1 was r/askReddit. As seen from the data, r/askScience had very high precision but not as high recall, and r/askReddit was the opposite, meaning if the model predicted a post to be from r/askScience then 99% of the time it was in fact from that subreddit, whereas if the model predicted it to be from the latter subreddit it was only correct 95% of the time. This brings us to some of the most strongly correlated words in the model:

Top coefficients in model:
Positive values indicate a strong correlation with r/askReddit
Negative values indicate a strong correlation with r/askScience

- remov: -5.99
- askreddit: 2.48
- thank: -2.26
- fuck: 2.13
- chemic: -1.86
- shit: 1.77
- explan: -1.75
- sub: -1.72
- surfac: -1.54
- astronomi: -1.52
- bodi: -1.51
- crow: -1.48
- rule: 1.45
- rain: -1.44
- particl: -1.36
- bad: 1.36
- medicin: -1.34
- field: -1.33
- god: 1.31
- whiski: -1.29
- longer: -1.29
- asksciencediscuss: -1.29
- hotter: -1.29
- anecdot: -1.28
- away: -1.28
- infinit: -1.26
- moon: -1.25
- patient: -1.24
- water: -1.22
- energi: -1.22

4.2 Feature Analysis

There are a few takeaways to be had from the aforementioned data. The first, as seen from the most strongly correlated words, is that there are certain words that can make the model much more certain that a post belongs to r/askScience than there are available for r/askReddit, which is what led to the precision and recall scores of the model.

The most notable of terms is remov, the stem of [removed], whose model coefficient has a magnitude almost three times that of the next strongest stemmed word. This is because r/askScience has much stricter response guidelines that moderators enforce due to the need for factual, evidence-based responses. Any responses that don't follow the guidelines are heavily moderated and likely to be removed. r/askReddit, on the other hand, is much more open-ended and typically opinion-based and thus isn't as heavily moderated. Typically, only responses that violate community interaction guidelines surrounding keeping things civil will be removed and that is a relatively rare occurrence. This difference in removal frequency is what leads to remov being so strongly associated with posts from r/askScience.

Another observation that doesn't require much explanation is the association of scientific terms with the askScience subreddit. Stemmed terms such as chemic, surfac, and astronomi, are much more likely to come up in the context of a scientific question and answer as opposed to everyday inquiries. On the opposite end, casual conversation in the askReddit subreddit is likely to be much more informal, which is likely why shit, bad, and fuck are some of the terms most strongly associated with r/askReddit.

The comparison between these two subreddits is fairly specific, so it's tough to project this to a large-scale application. One way it could be used, however, is for auto-moderation. The model is by no means perfect but is closer than I expected it would be, and perhaps with more sample data or experience it can be made to be more reliable. If the model itself can predict whether a post belongs on r/askScience or any other science forum on the internet, then it can moderate forums on its own and take down posts that do not belong.

4.3 Error Analysis

False Negatives:

- Using scientific or measurement terms: 2,3,6,8,9
- Removed 1,4
- Just outweighed by word values, no clear problem: 5,7,10

False Positives:

- Outweighed by positive word values: 11,14,15,16,19
- Should be negative judging by word weight values: 12,20
- No training data despite clearly being negative: 13
- Informal Language 17,18

The majority of incorrectly labeled posts came as a result of simply being outweighed by word values in the wrong direction. There isn't much that can be done to fix this – perhaps a larger training data set, but nonetheless it will always be an inevitable problem that contexts of a given subreddit will lead to posts that mimic the language of the other. That is also the case with informal and formal language restrictions – while the two tones apply to askReddit

and askScience, respectively, there will always be crossover and it cannot be fixed.

The only other thing of note is that, due to the stricter moderation guidelines of r/askScience, any removed posts are by default classified as r/askScience. This cannot be relied on and to improve the model's accuracy, consider removing [removed] posts from the data set to begin with as they do not provide much tangible benefit.

5 CONCLUSION

The model created for this experiment turned out to be quite reliable; much more so than I expected given how similar the two subreddits appear to be on the surface. The model's breakdown of how posts were classified highlighted the differences in terminology, tone, and adhesion to guidelines between the two subreddits that I initially didn't take note of.

The model would definitely see improvement both from larger training data sets as well as cutting out content that has a high risk of incorrect labeling, such as [removed] posts. The larger training set will increase the strength of highly correlated words to reduce error from 'crossover content,' and removing high-risk content should remove a noticeable amount of avoidable mis-labels from the testing data set.

In the future, if strengthened to a significant degree, this model's most beneficial application would be for auto-moderation or at least moderation assistance on large platforms such as reddit, where the scale of user content makes large-scale reliable moderation an almost impossible task.

REFERENCES

[1] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, Denver, Colorado, USA, 3175–3187. https://doi.org/10.1145/3025453.3026018

[2] Trang Tran and Mari Ostendorf. 2016. Characterizing the Language of Online Communities and its Relation to Community Reception. *arXiv:1609.04779 [cs]* (Sept. 2016). http://arxiv.org/abs/1609.04779 arXiv: 1609.04779.