6.806 Final Project Report

Justine Jang, Anderson Wang, Patrick Yang

December 2015

1 Abstract

We analyzed comments on the internet discussion forum at reddit.com/r/changemyview, a location where people can submit an opinion and ask for rebuttals or opposing viewpoints. Selected viewpoints can be awarded 'deltas,' signifying that they successfully changed one's opinion. We built a crawler to generate our own dataset, since this space did not have an existing corpus for use. We then built multiple classifiers that use information both about the comments themselves as well as their context in the entire comment tree to attempt to predict what posts would be awarded 'deltas'. Our best classifier achieved an F1 score of 0.147651 on our test set of approximately 40,000 comments, improving a significant amount on a baseline. The numbers reflect the difficulty of the task, because a variety of difficult-to-quantify features factor into the ability of a post to be convincing, such as personal opinions and long multi-comment discussions. The github directory for the project can be found at github.mit.edu/pbyang/6806-project.

2 Introduction

The forum /r/changemyview on the site reddit.com is structured similarly to a debate forum. A user (the *original poster* or *thread creator*) will post a submission, which begins a new thread, stating a personally held belief and requesting a dissenting opinion. Users can then post comments that either reply to the submission or to other comments. If the original poster deems a comment as having changed their opinion, they can note that by replying and awarding a 'delta' (denoted on the forum with the Δ symbol). This provides an opportunity to observe a space of argumentative posts which have an interesting structural component and are naturally annotated.

Our goal in this study was to attempt to automatically classify comments, using both textual data - information solely from the text of the comment - and contextual, structural data - ranging from the comment's placement in the overall reply tree to the reputation of the author, to determine whether it was

likely to be awarded a delta. We wanted to see whether there were any textual or structural markers that would cause the comment to more likely convince a human reader, whether overtly or subconsciously.

The inspiration for our study is one of the suggested projects, Joint Models of Disagreement in Online Debate [5]. That study also used a combination of textual and structural methods. However, many of their methods are inapplicable for our project, because the comments in our space have very different characteristics and the ultimate goal is to identify a strongest argument, not discern stances of individual authors.

In the remainder of this report, we will refer to comments awarded a 'delta' as *positive*, with all others being denoted *negative*.

3 Experimental Setup

3.1 Corpus

We wrote our own crawler, interfacing with the Reddit API via PRAW, to generate two different datasets, one small and one large. The small dataset was used for debugging and to verify the functionality of the models we were testing. The large dataset was used for actual evaluation.

The small dataset consists of the comments from 50 threads, divided in approximately a 4:1 ratio into a training subset and development subset. This dataset can be found in the project GitHub repository at `https://github.mit.edu/pbyang/6806-project/tree/master/data`. The small dataset assisted in rapid development, including debugging the model and optimizing major hyperparameters (in particular the weight ratio of the positive to negative samples).

The large dataset consists of the comments from 500 threads, divided in approximately the same ratio as the small dataset. This dataset is available on Dropbox.

The large dataset contains approximately 200,000 total comments, of which only about 800 are positive. This large skew in tagging presented a significant challenge when training the model.

3.2 Evaluation Metrics

We tested our models with two related tasks that aimed to provide different insights into the performance of the model:

Metric 0 (tag-all) required the model to tag each comment individually with whether the comment was expected to receive a delta. These predicted tags were compared against the gold-standard tags, and the score on this task was the F1 score on positively tagged comments only. Given the approximate 0.4% rate of positive comments, random guessing at would produce an expected score

of 0.004. This is the more standard evaluation task.

Metric 1 (1-of-n) required the model to predict, given a random selection of n comments under the stipulation that exactly 1 was positive, which of the comments was the positive. Put another way, this task asked the models to determine which of a given collection of comments was the single most likely to be awarded a delta. Score is simply the percentage of correct predictions over all rounds. In our experimental runs, n was set to 10 (so, random guessing would produce an expected score of 0.1). We devised this metric to try and capture scenarios where the models scored a comment highly, but below the threshold for being classified positive. The tag-all task doesn't provide any insight into how close a model is on a false negative, but this task provides some visibility into that facet of 'correctness.'

The first task proved to be harder by far, both for humans (via informal survey and our own experiences) and the models.

4 Models Used

4.1 Baseline

Our baseline model was a simple two-feature maximum entropy classifier. The two features are the length of the comment's text and its depth in the comment tree: one text-based feature and one structural feature. These were chosen based on a cursory manual inspection of the positively tagged comments to find simple features that seemed likely to be useful. The brief intuition behind choosing these features as a simple baseline is that a longer comment is more likely to be both well-researched and contain a component that is convincing, while a comment that is deeper in a tree is more likely to be the culmination of a long back-and-forth discourse. This classifier weighted the positive comments 10 times higher, a rate determined experimentally using the small dataset to best alleviate the extreme bias towards negative comments.

4.2 Final Model

The final model that performed the best was also a maximum entropy classifier, albeit with many more features than the baseline. For a discussion of the most performant and least performant features, see the section dedicated to them below. This model, like the baseline framework it was built on, also weighted the positive comments 10 times higher.

Some features use NLTK, the Natural Language Toolkit [1].

4.3 Other Models Investigated

One other model we investigated was an ensemble classifier featuring several maximum entropy classifiers, each of which was trained on all of the positive comments but a small random selection of negative comments, which would perform majority vote to predict tags. We thought this might be an alternative way to reduce the learned bias towards negative comments. However, this ensemble actually performed worse than the single maximum entropy classifier with weighted samples. We theorize that this is due to a combination of the voting system negatively impacting the model's performance and the ensemble classifier having less data about common trends among negative samples (due to the random selection).

5 Results

5.1 Model

For the tag-all-comments metric, the F1 scores were:

	Baseline	Final model
F1 score	0.118467	0.147651

For the 1-of-n metric, the accuracy rates were:

	Baseline	Final model
Accuracy	0.320	0.387

Note that even our baseline classifier did much better than randomly guessing, which as previously mentioned would have produced an expected F1 score of .004 for the first metric and an accuracy of .1 for the second metric. The final model performed modestly better than the baseline in both metrics.

5.2 Features

After brainstorming around 20 features, we normalized the feature values to be between 0 and 1 and ran a χ^2 test to calculate how much each feature's values varied from expected. The higher the χ^2 values, the more likely we were to reject the null hypothesis of independence and take the feature to be a positive indicator for whether a post received a delta. A χ^2 score of 6.635 or higher indicates a statistical significance of 0.01. We found 6 features that reached this level of significance, which can be found along with their χ^2 scores in the table below, in decreasing order.

Feature	χ^2 Score
Length	22.368
Parity Depth	13.715
Quotes Another Post	12.453
Author Reputation	7.923
Is Sourced	6.612
Contains Link	6.431

Our most effective feature was the length of the post in characters, followed by the parity depth (the parity of the post's depth in the tree of posts, where the root is the original submission). That was closely followed by 'Has Quote', an indicator of whether the post quoted another post or not. Next was 'Author Reputation', the number of times the author has received a Delta in their entire history, and 'Is Sourced', an indicator of whether the post included 'source', 'paper', and other similar words. Finally, we have 'Contains Link', which is an indicator for whether the post contains a link or not.

Many of these features make a lot of intuitive sense as to why they are correlated with the effectiveness of a post. For example, it is easy to see that the features 'Length', 'Quotes Another Post', 'Is Sourced', and 'Contains Link'all positively correlate with the amount of effort put into the post as well as the perceived legitimacy by the reader. Furthermore, the number of Δs earned in total, openly represented as $n\Delta$ in the flair next to the author's username, is positively correlated with a history of writing compelling arguments as well as an increased amount of reputation in the forum. Interestingly, one of the most effective features was the parity of the depth of the comment in the tree of posts. We guess that this feature was a good indicator because people tend to respond to posts that they disagree with. Thus posts that have a depth-parity of 1 tend to be countering the original submission and therefore have a greater chance of receiving a Δ.

Two of our most effective features were also our simplest, namely the length and depth of our features. Because our baseline classified using these two features, our baseline ended up performing much better than a randomly guessing classification model. Our more complicated features, like a cosine similarity between the tf-idf vectors [3] of the posts and a sentiment analysis using the VADER corpus [2] did not make much improvement over the baseline.

5.3 Qualitative Survey

We sent out an informal survey to a handful of people in which we gave them a screenshot of a post on CMV along with screenshots of several comments replying to that post, and asked them to try to identify which one of the screenshots was awarded a delta. We also asked them to optionally provide the reasoning for their decision. There were a few goals of this survey:

1. We chose comments that received deltas but our classifier thought would

not receive a delta. This was to see if humans could correctly identify these trickier posts.

2. We also chose comments that our classifier identified as receiving deltas but did not actually receive deltas, to see if these options might be picked.

3. By asking them to provide reasoning, we wanted to get a wide variety of opinions on what sorts of features are important when analyzing the convincingness of these comments.

We had three questions in our survey, corresponding to three different threads on CMV. In one of the questions, the majority of people correctly identified the comment that received the delta, and in the other two, about half of the people were correct, with most of the other half picking the incorrect comment that our classifier also picked. Although this survey had a fairly small sample size, it does seem like that the problem of identifying which comments receive a delta is fairly difficult, but humans can still perform quite a bit better.

The biggest thing that people mentioned in their reasoning was how well the comment addressed the issue, especially whether it directly responded to one of the points in the original post. This makes sense, and could be a reason for why this task was so difficult for computers: it's easy to quantify word similarity or even sentiment, but calculating whether a point was addressed is a different beast.

6 Conclusions and Further Work

Ultimately, we improved over the baseline more on the harder task than the easier task, though our score on the hard tasks demonstrates that there are still many complications left to solve when tagging comments of this nature. A manual inspection into our false negatives revealed difficulties such as sarcastically or whimsically awarded deltas that might have made training even more challenging, given the low volume of positive posts to begin with. Furthermore, because of the subjective nature of argument strength, knowledge of the author's beliefs would be useful in discerning how likely they are to accept a given counterargument; this knowledge is difficult to obtain via reading the comments in a single thread. Another factor that makes this task difficult is long chains of multiple back-and-forth comments: many times, such a chain will end in a delta being awarded, but predicting where the chain ends is very difficult, as many comments in the chain will be very short and only clarify previously stated opinions.

There are many avenues for further exploration which were beyond the scope of this project. One such direction might be to use semantic analysis to fact-check individual arguments, and assign them a 'correctness'score as a feature. Another extension would be to span multiple threads to build a profile of the original poster, and extract some features from that profile to determine, for

example, how likely that user is to be convinced, and the general leaning of their political and social stances. These explorations are enabled by the relative complexity of this space, both textually and structurally.

Another approach we can take is adding lots of features and employing a feature selection algorithm to determine the highest-performing ones, the method used by Persing and Ng in [4] to make improvements in automated argumentative essay grading. Most of the features described in their paper would not be very applicable to this problem, however, because student essays are completely different from internet comments, in format, tone, and style, so some additional research would have to be done into possible kinds of features.

Finally, we might want to try models other than a maximum entropy classifier. In particular, neural nets seem like they could perform well because they can find combinations of features. In this case, we would still have to deal with the problem of the negative tags hugely outweighing the positives.

Acknowledgements

We would like to thank all the staff of 6.806, and particularly Prof. Barzilay, for providing lots of pointers and assistance throughout the course of the project.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[2] C.J. Hutto and Eric E. Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: (2014). URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Scoring, Term Weighting, and the Vector Space Model*. 2008, pp. 118–130. URL: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

[4] Isaac Persing and Vincent Ng. "Human Language Technology Research Institute". In: (2015), 543–552. URL: http://aclweb.org/anthology/P15-1053.

[5] Dhanya Sridhar et al. "Joint Models of Disagreement and Stance in Online Debate". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. 2015, pp. 116–125. URL: http://aclweb.org/anthology/P/P15/P15-1012.pdf.