Predicting the Changing of Views on a Reddit subreddit

Assignment 2 for CSE190A, Fall 2015

Jonathan Pham
A12491003
jhp021@ucsd.edu

1. INTRODUCTION

1.1 Reddit and the voting system

Reddit (www.reddit.com) is an online message board with a voting system and numerous communities. These communities (named **subreddits**) have a focus on a particular interest, such as sports, world news, personal stories.

The foundation of any subreddit's content begins with **Submissions**. Submissions are either text or links to other content on the web. Users of Reddit are able to vote on these Submissions with either an **upvote** or **downvote**. To put it simply, an upvote signifies approval and a downvote means disapproval. Upvotes and downvotes affect the visibility of Submissions, where Submissions with many upvotes tend to appear closer to the top of a subreddit's page.

A **Comment** is a reply to a Submission or another Comment and can be voted on with upvotes and downvotes as well. A Comment that is a direct reply to a Submission forms a thread, which also consists of Comments that are replies to the original Comment.

With all of these systems, Reddit aims to order and organize discussion. Upvotes and downvotes let the users have an influence on what a subreddit's attention is on. Subreddits and Comment threads organize related discussions together. My particular predictive task involves one particular subreddit that makes good use of this organization, but strays away from some of Reddit's traidtions.

1.2 Change My View (CMV)

Change My View (www.reddit.com/r/changemyview) is a subreddit that, instead of presenting news or sharing content like other subreddits, allows users to post Submissions that articulate a view they hold. Other users are then challenged to change it. If anyone, including the original user (called the **OP**), believes that their view has been changed by a Comment, they can award the Comment's author with a Δ. This is done by including the Δ symbol in a reply to that Comment.

There are some special rules with CMV that make it unlike other subreddits.

1. You are unable to downvote Submissions or Comments.
2. Upvotes are hidden for 24 hours after a Comment has been posted.
3. (ou muFor Submissions) Yst personally hold the view and be open to it changing.
4. Direct responses to a CMV post must challenge at least one aspect of OP's stated view (however minor), unless they are asking a clarifying question.

As I explore this subreddit, I hope to look at how and what it takes to receive a Δ.

Table 1: A General Look at CMV

# of Comments	892871
# of Submissions	19865
# of users	47923
Avg # Comments per Submission	44.95
# of Δs given	6970
# of OPs who gave a Δ	2922
# of Δs given by OPs	4553
Avg # of Δs given by OPs	1.56
% of Submissions with a Δ given	14.7%

1.3 Analysis of CMV

The dataset I'll attempt to analyze comes from Submissions and Comments from the entirety of last year, 2014. Before I begin, to prune the dataset, I removed Submissions that were not actual views, such as posts from Moderators and discussions about the subreddit itself. Basic information is contained within Table 1.

What is especially interesting is that the number of Δs given is tiny compared to the number of comments. Also, on average, the average number of Δs given by an OP of a Submission is closer to 2 than 1.

1.3.1 Submissions with Δs Given

Perhaps most importantly, we must look the differences between Submissions that do and do not result in a Δ given. Here are some basic information on this data in Table 2. There are huge differences between the average length of a Submission's text, the number of Comments, and the number of upvotes of Submissions that result in a Δ and those that do not. Perhaps as users explain their views more, they express more enthusiasm for the ensuing discussion and are more willing to change their view. And due to the general attitude of the subreddit, users of CMV may upvote Submissions that had a view changed more than those that did

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Table 2: Submissions that give/don't give a Δ

	Overall	Yes Δ	No Δ
Avg length of Submission text	200.13	344.28	175.27
Avg # of Comments	44.95	84.24	35.21
Avg # of upvotes	28.05	66.64	21.39

Table 3: Commenters do/don't have at least 1 Δ)

	Overall	Yes Δ	No Δ
Avg length of one's Comments	88.58	103.87	75.59
Avg # of one's Comments	42.42	84.24	35.21
Avg # of upvotes on Comments	3.42	5.23	3.33

not.

1.3.2 Users

I then attempted to analyze all the users of CMV. First, I look at those who receive Δs, in Table 3. To extrapolate meaning from these numbers, it seems like the longer one's comments are, the more likely they will be awarded a Δ. And those who participate often in CMV Submissions are much much more likely to get Δs, with their experience contributing to earning their Δs. Upvotes may also indicate whether or not a Commenter would receive a Δ, with more upvotes being better.

On the flip side, let's look at those who give Δs, in Table 4. It seems as though the length of comments from those who give Δs differs adequately from those who do not. This seems like it contradicts how Submissions with more text correlate with Δs given, but it's important to understand why users would Comment on their own Submission in the first place. Perhaps those who are reluctant to give Δs are those who feel they must more strongly defend their view in response to Comments. Thus, they are more likely to express their views using more words. They continue the discussion. As for upvotes, there is a slight correlation between the number of upvotes an OP receives and the likelihood that OP gives Δs.

1.3.3 Number of Comments in Submissions

Perhaps the biggest predictor of whether a Submission will have its view changed is the number of replies it receives. Figure 1 shows the distribution of the number of comments submissions receive. The graph resembles power law, so Submissions with few Comments are more frequent than those with more Comments. Some information about these Submissions are in Table 5.

The median being 16 is surprisingly small. This may come from the strict moderation of the subreddit, as Submissions and Comments have very strict rules on what is allowed to be posted. There are thousands of Submissions that have 0 or 1 Comments. To discount these throwaway Submissions, the newly calculated median rises up to 32, which seems like it would lend to more productive conversations.

Table 4: OPs that give/don't give Δs

Field	Overall	Yes Δ	No Δ
Avg length of Comment	88.58	72.28	97.62
Avg # of upvotes on Comments	2.08	2.55	1.83

Table 5: OPs (Yes meaning they gave a Δ)

# of Submissions	19865
# Submissions with 0 comments	2416
# Submissions with 1 comment	4240
# Submissions with >500 comments:	159
# Submissions with >1000 comments:	20
Average # of Comments	44.95
Median # of Comments	16
Median # of Comments (Filtered)	32

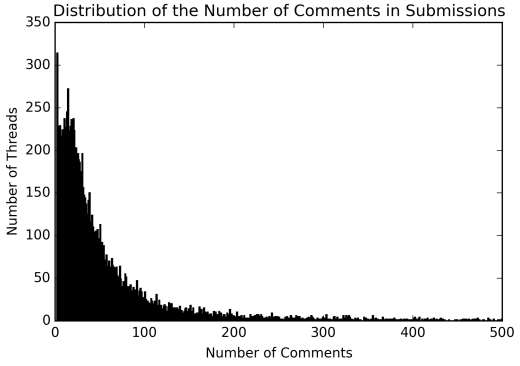

Figure 1: Ignoring Submissions with 0, 1, or over 500 Comments

In addition, Submissions with well over a thousand Comments exist. Variance with this feature is high, so it may lend itself well to our predictive task.

1.3.4 Topic Modeling/Text Classification

To get a general idea of the topics in CMV, I ran a python package called gensim's implementation of Latent Dirichlet Allocation. LDA nondeterministically clusters together closely related documents to find a distribution of topics for that document. To prep the corpus and dictionary of LDA, for every Submission I collected all of its words and the words of its Comments, converted them to lowercase, removed punctuation, and removed those in the nltk stopwords list.

Topics come from weighing each word in the corpus. The number of topics is set to 15 and number of passes equal to 15. The topics identified and the words associated with them are in Table 6. The topics identified by LDA are not comprehensive, it seemed. Quite a few threads in CMV involve social issues such as abortion, legalization of marijuana, and philosophy. Perhaps with more iterations it could converge on a more representative list of topics for the dataset.

Nonetheless, it was able to identify some important and frequent topics for users of CMV, topics such as US foreign affairs, the cost/reward of education, the relationship between government and business, and human relationships.

As another approach, I read through about 200 Submissions and classified them by hand with the following categories: Political, Social, Economic, Philosophical, Cultural, Moral. Each Submission needed to belong to a category. The distribution is as shown in 7. There seems to be quite a few posts related to popular culture, which LDA did not find unfortunately. This may be due to the variety of proper

Table 6: OPs (Topics found by Latent Dirichlet Allocation)

ID	Words	Interpretation
0	would money work school job pay college get students system	Education and Income
1	us would government country war states world countries police vote	Politics and Foreign Affairs
2	would could use technology new space also used much water	Space Travel and Water?
3	reddit posts comments think post word use like would subreddit	Reddit
4	dont think im like would know get want even one	Desires?
5	like game games time one play even get really dont	Time for playing video games
6	car drugs alcohol drug driving cars food get smoking health	Drugs and Driving
7	white black animals race racist human humans animal racism species	Racism/Animal species
8	women men sex gender sexual relationship man gay male woman	Sexual relationships
9	government business money companies company public would internet market free	Governments and Businesses
10	please like dont one see things feel first look gt	Feelings?
11	child children would parents life believe women society think rape	Children & Women in Society?
12	religion believe god human religious one moral rights society would	Religion and Morality
13	culture science social different language use many think music art	Culture
14	time past look say remember photos internet travel celebrity wars	Culture?

Table 7: Topics identified by me

Category	Number
Political	46
Social	65
Economic	33
Philosophical	15
Cultural	41

Table 8: Data Split

	# of Submissions
Training set	12713
Validation set	3179
Test set	3973

nouns people use when talking about pop culture (names of people, movies, shows), which are commonly rarer than other words.

2. THE PREDICTIVE TASK

The predictive task I am aiming to look at is given a Submission with all its various Comments, predict whether or not the OP will award a Δ to any Comment.

Originally, I wanted to predict whether, given a certain Comment, would it receive a Δ. This can prove to be extremely difficult as very few Δs are given compared to the number of Comments that CMV receives. In addition, perhaps what is most important is that getting a view changed is a a team effort, where Comments on a Submission would all, hopefully, nudge the OP away from their view. Not only that, but it would be important to look at the context of Comments, as conversations would be important to look at.

Now, onto splitting the dataset. To start off, I did a split of 80:20 on the Submissions to construct my training and test data. On top of that, I split the training data again with a 80:20 split again for training:validation sets. See Table 8 for the numbers.

A simple baseline to use is just flipping a coin with a bias of the # of Submissions with a Δ given (good submissions) / # Submissions without a Δ (bad submissions) given. This turns out to be a bias of $p = .147$. The results of this baseline are in Table 9.

Table 9: Accuracy of this baseline

	# Right / # Wrong
On validation set	0.750865051903
On test set	0.754090108231
On all set	0.752529574629

Baseline performance is not bad. Given the fact that a small number of Δs are given anyway, it is easy to simply predict no Δ the majority of the time, which is what the coin flip does. Hopefully my model can do better than this!

3. BUILDING THE MODEL

The task is a classification problem, so I am choosing support vector machines as the model for the job. I will be using sklearn's svm module to create my SVMs, with a reguarlization parameter of $C = 1000$.

For picking the features in my SVM's feature matrix, I first started with features related to the Submission itself. This includes its number of upvotes, comments, and words in the Submission text. These were discovered to be important features in Task 1.

Accuracy on Validation Set: 0.834853727587.

A nice improvement over the baseline! We can do better. To try and improve this, I add features based on the user who submitted the Submission, the OP. OPs who have Δs and who give Δs may be more likely to give more Δs. Sadly, though, adding the number of Δs an OP has actually decreased accuracy on the validation set. So, it will simply be the number of Δs they has already given.

Accuracy on Validation Set: 0.836111984901

And not only that, we can add the number of comment upvotes the OP has accumulated while processing our training data. It seems that those who genuinely had their view changed will have more upvotes as a result.

Accuracy on Validation Set: 0.841459578484

Now we must look into the Comments of a Submission. We first add the a feature based on the number of times an OP replies to their own Submission. This indicates that the

OP continued to discuss their view and addressing counter-arguments to the view. With enough conversation, maybe even the staunchest of staunch can have their view changed.

Accuracy on Validation Set: 0.843976093111

Next, we can add the average Comment length of Comments within a Submission as well as the number of total upvotes a Submission's Comments receives. If a Comment had an intellectual "quality", then the length of a submission and the number of upvotes it receives could be indicators of this quality.

Accuracy on Validation Set: 0.846492607738

And finally, this is where we end our model. Features include the number of Comments, upvotes, words, and replies from the OP a thread has; the number of Δs given and Comment upvotes the OP has; and the average length, total number of upvotes, and total number of Δs the Comments of a Submission are. We can run our SVM on the Test Set and see how we did.

Accuracy on Test Set: 0.863830858293

Accuracy of Baseline: 0.754090108231

I attempted to fit features based on the average upvotes on Comments in a Submission, average length of Comments in Submissions, and other information based on Comment data. Unfortunately, these features did not lead to an increase in accuracy on the Validation Set.

Other models include those that only examine the OP of a Submission. Since this model relies on the fact that these Submissions have ended long ago, in practice it may only do well weeks after a Submission has been posted. Instead, a good online model would be looking at what would be usually available at the time of a Submission's posting: the user who posted it and the Submission itself. It may be difficult to build an accurate model based on just this information, so the model would really have to focus on knowing a lot about the user.

And due to time constraints, I did not get a chance to look at using topics found by LDA or determined by myself. This might have lead to a great performance boost. Going more on the by-hand approach, I would have definitely made use of a logistic regressor to train a model that would classify a Submission/Comments based on my by-hand classification. I am not sure if that would have lead to any performance improvement, but it would have been fun to use a model to build the feature matrix of another model!

4. LITERATURE

I found a dataset that contained all of Reddit's Comments and Submissions on a subreddit called datasets. I downloaded these datasets and filtered out all the ones pertaining to CMV. Due to the sheer volume of the data, and the age of CMV, I chose to only analyze data from 2014 for this assignment. The datasets were line-delimited json blobs, much like in class.

In general, there are numerous studies on Reddit, such as Bruno JakiÄĞ's MSc Thesis "Predicting sentiment of comments to news on Reddit", available at

$dare.uva.nl/cgi/arno/show.cgi?fid = 451648$

There is also Stanford student Katyaini Himabindu Lakkaraju's "Predicting Content Popularity on Reddit"

$snap.stanford.edu/class/cs224w-2012/projects/cs224w-025-final.v01.pdf$

I don't believe there is any literature regarding this kind of predictive task. There was a discussion on the Library of Economics and Liberty forum. Users on this discussion debated the reasons a person's views may change.

$econlog.econlib.org/archives/2015/07/change_my_view.html$

As for analyzing text, there are numerous tools that can come from analyzing the data. LDA is pretty recent (2003) and gensim's implementation is an online model that can adapt to new data coming in. As computers get faster, it will be even easier to transform bags of words into usable features for a model.

No doubt there are psychological studies on how humans change their views. Not only that, I imagine certain topics or beliefs are hard to change. This may come from how controversial the topic is.

5. CONCLUSIONS

This was an interesting topic. Not only that, but I was able to beat the baseline model with a simple selection of features that were "given" to me by the data set. Unfortunately, perhaps the biggest gains in performance would have come from interpreting the topics of CMV Submissions and turning them into features. And what's more unfortunate, I was not able to locate any literature on this kind of task. Nevertheless, the constructed model works well with tools I learned from this class. Perhaps when I am more capable, I would love to revisit this problem, with an even bigger dataset and more tools at my disposal.