Analyzing behavior spread through content quality in Reddit.com/r/politics

Tom Kremer
Stanford University
tkremer@stanford.edu

Vein Kong
Stanford University
vskong@stanford.edu

Ben Krausz
Stanford University
bkrausz@stanford.edu

1 Introduction

In light of the recent US election, it has become increasingly clear that social media has had a tremendous effect on how Americans consume news. In April 2016, for example, 243 million people visited reddit [1]- and a recent study revealed that a whopping 70% of reddit users get news from the site [2]. In an age rife with memes and fake news, it is vital that our social media platforms are able to facilitate intelligent discussion. With the increasing popularity of online forums and the pressing relevance of politics in the United States today, it's only natural that political discussion forums like Reddit's r/politics explode with user created content. The aim of the forum, according to Reddit's rules, is to provide a forum for thoughtful political discussion, but due to the sheer number of active members, well thought out comments can get lost in a sea of jokes, one-liners, memes, and name calling. We were curious if the community actually participates in and rewards any form of thoughtful political discussion, and if "good" behavior can be spread. In this paper, we explore the traits of the most popular comments in the subreddit and the effects of more thoughtful comments on the community to judge reception and reaction. We then present a final answer to the question: "How does positive behavior spread on r/politics?"

2 Related Work

The topic of assessing community activity in online forums is not a new one. Anderson et al [3] focused on the collection of responses on question answering sites like Stack Overflow and Quora, which emphasize community through reputations and rewards for good behavior. The paper tries to use the metric of voting as a measurement of answer quality, but there is an inherent flaw in this approach because it can easily confuse quality with exposure. It's possible (and likely) that there are countless helpful responses that were either posted too late after the question was asked, did not receive enough initial votes to be seen by voters sooner, or were lacking in some other way that isn't related to post quality. Many well thought-out and helpful answers could have been deemed as bad behavior, even though a forum that was comprised of these comments is most definitely the wanted type of forum behavior.

Wang, Ye, and Huberman improved on this concept of good online behavior by focusing instead on post frequency and the likelihood of online conversations [4]. They found that user activity follows a power law distribution, such that there is a heavy tail of users that post very often and considered this behavior good for the community. Though this helps to negate a bias toward exposure, unlike Anderson's paper, it does not look at actual content. It's possible that a "troll" user who posts nonsensical or unrelated content very often will be deemed as practicing good community behavior even though he or she isn't helping to facilitate discussion. Though post frequency is a legitimate measurement for judging a user's community activity, it doesn't necessarily correlate with "good" users for the community overall.

Lastly, Damon Centola et al looked at how behavior spread through an online community [5]. They found that users who had more exposure to their immediate network's actions were more likely to adopt that same behavior. This is applicable to how good online behavior can spread through the network; if more users engage in a certain type of behavior, it encourages other users to do so as well. Unfortunately, this was dealing with the very simple action of registering for a certain health website, and is probably not as prevalent in more complex actions like posting thoughtful content in a political forum. Furthermore, it's possible that a user can be exposed to more "bad" comments that do not promote discussion and conform to the community's poor comment standards. If comments that encourage discussion inspire other comments that further encourage this good behavior, then we can measure the effectiveness of good behavior in the community.

3 Methods

3.1 Data Collection

We used the Reddit.com/r/politics 2014 comment data as posted on the CS224w.stanford.edu website [4]. The dataset contains each comment in the subreddit community in 2014, including the comment's identification string, the parent comment's identification string, the time of posting, the author's username, the net score (through Reddit's upvote system), the thread's identification string, and the content of the post. We developed a class that read through the data and organized the information into instances of a custom Comment class for ease of use down the line.

The dataset contained data for 2,333,826 comments belonging to 70,856 discussion threads. Because the dataset only contained comments, threads were not directly represented. Threads were identified purely by their identification strings, which were represented in each comment's metadata. This means that any threads that had no comments were not represented in our dataset or in our analyzations. Luckily, we are interested purely in comment traits, behavior, and reception, so even though we are missing data on which threads spark the most discussion, it does not affect our search for which comments do so.

Of those comments, 476,380 were "root comments," meaning that they were not direct replies to other existing comments but rather were the initial comments within a thread that sparked other replies. These were identified by the fact that they all had parent comment ID strings that were also thread ID strings because they were direct replies to a thread rather than to another comment.

3.2 Network Organization

In our network, we chose nodes to represent comments or threads (i.e. posts) and edges were directed from each parent post to its child post. To avoid having a disjoint network for each thread, we constructed a master root node, which has to data directly associated with it. However, there is a directed edge from this root node to each of the 70,856 thread nodes, which also do not have any data directly associated with them. Each of these thread nodes act as roots for the tree structure that represents all the comments in that thread. Thus, every root comment is a direct child of its appropriate thread node, and the total number of comments in a thread is the size of the thread's tree minus one (to account for the thread node that isn't technically a comment). With this organization, we can analyze all the threads in a single graph. It also allows us to visualize this network in a radial pattern, with the master root node in the center, a sphere of thread nodes surrounding it, and comment trees extending outward from each thread node.

4 Comment Classification

4.1 Identifying "Good" Comments

We wanted to find comments that provoked the most discussion on reddit. We decided to use the following heuristics to train a word-to-vector classifier:

- Degree (number of direct replies to a comment)
- Max Depth (length of longest "reply chain" to a comment)
- Tree size (total number of comments existing in the tree spawned by a comment)
- Score (Number of upvotes minus the number of downvotes)

All of these heuristics were normalized by the number of comments in their respective threads to account for the bias of thread popularity.

We also used two other measurements to determine if a comment was "good:" a comment's length (number of words) and it's Flesch-Kincaid Ease of Reading Level (calculated through the Python textstat package[4]. Rather than use a classifier with a training set and test set, we simply identified the appropriate score for a comment for both heuristics and identified it as long or of advanced reading level if its scores were more than a standard deviation above the mean among all comments in the dataset.

4.2 Classifying the Dataset

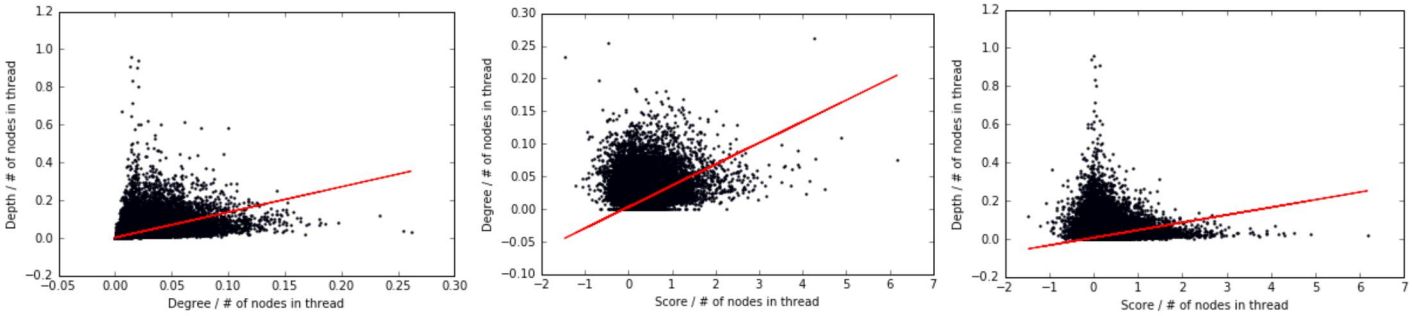
We attempted to classify the dataset using a few different machine learning algorithms. We first attempted to use NB (Naive Bayes) to classify the data and treat it as a sentiment analysis problem by training it with data that we perceive as "good" and "not good." However we found an issue with the approach because the classifier classified over 99% of the comments as good. This may be due to the fact that it was better at identifying extremely "bad" comments than good comments.

We then attempted to use word2vec [6] with logistic regression. This gave us more sensible results. Since word2vec preserves the syntactic and semantic relationship of the word, we think it did a better job of classifying the data compared to a boolean word approach of NB.

In the NB case, for each of the scoring heuristics, for each comment we created a set of words, we then used the NLTK library to train the NB classifier. Note: since this is a set it does not take frequency of the word into account.

In the word2vec + logistic regression case, we created the word vectors using gensim word2vec with the CBOW model. We then scaled the vectors to move it to a standard normal distribution. We trained the data using logistic classifier and the SGDClassifier in scikit-learn. We set the penalty to l1 norm because the dimension of the dataset is sparse.

There was a noticeable correlation between degree and depth (correlation coef = 0.64) and lesser correlations between degree and upvote score (0.45) and depth and upvote score (0.26). Since these correlations were not extreme, we decided that evaluating the popularity of comments based on multiple different heuristics was worth the effort.



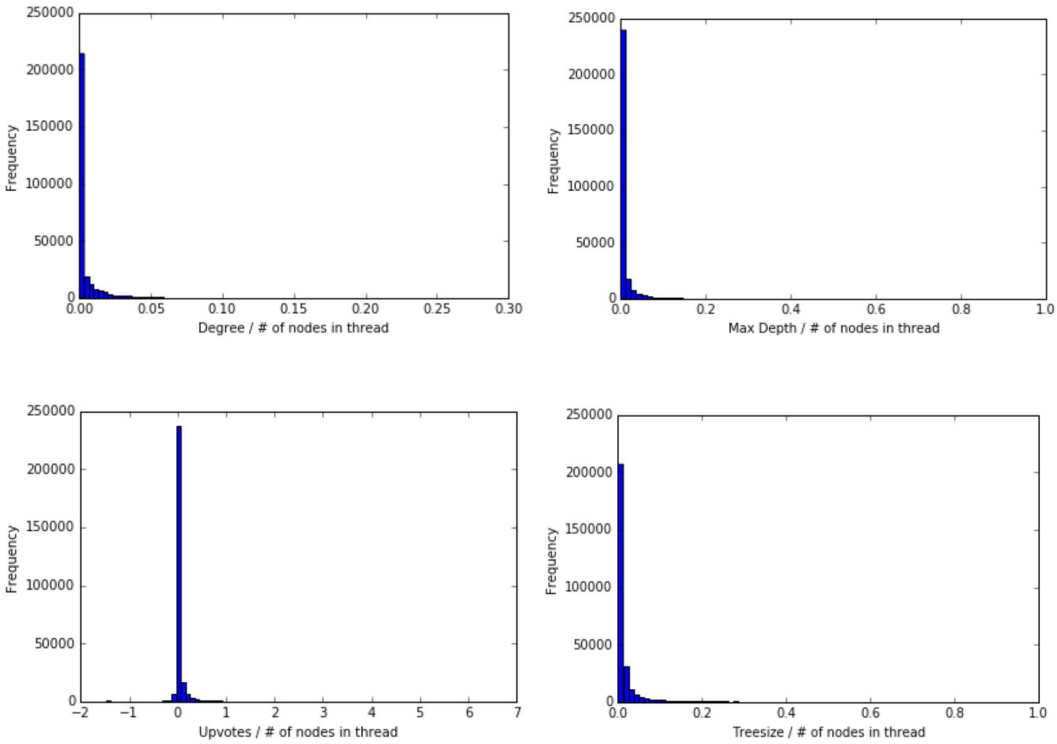
Figures 1a-1c: Graphs comparing the normalized heuristics of each root comment in the dataset. Note that, though its slope does not appear to be the steepest, degree and depth are in fact the most correlated

4.3 Constructing the Training Set

We used a word2vec and logistic regression classifier that was trained on each of these traits. For the classifier's training set, we only used root comments so that the classifier wouldn't create a bias against leaf nodes. Since comment trees are not infinite, there is a huge collection of leaf comments that have no replies. These would be classified as comments that did not spark discussion, but at that point, it's more likely that these comments don't have children because of poor exposure and not because of poor content quality. We also limited the training set to root comments in threads that had at least 60 comments in the thread's subtree, since we found the average thread size in our dataset to be around 26 comments. This way, a root comment's low performance on a heuristic wouldn't be solely because the thread was unpopular.

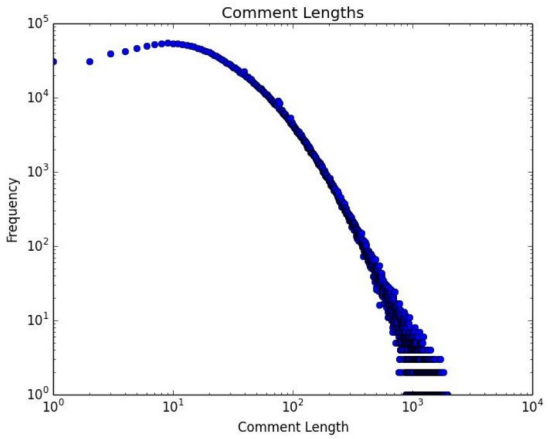
In this way, we could analyze the content of comments that resulted in the most relative direct replies, the longest relative reply chain, the largest relative subtree, and the highest relative upvote score. It is important to note that although relative post-time can greatly affect some of these classifications negatively, but the general training data should still give us a clear idea of the content of comments that perform the best and the worst for each heuristic.

Our approach resulted in a total of about 280,000 root comments to be judged based on our heuristics.We sorted all the comments in the dataset by each heuristic and used the top 15,000 comments (~5%) as positive and the bottom 15,000 comments (~5%) as negative to train the classifier. Though a normal classifier usually has a training set that is about 20% of the total dataset, we could not use such generous numbers because of the way the comments were distributed based on these heuristics. For most heuristics, at least 75% of the comments were placed in the worst category. For example, when judging comments by tree size, we found that 76% of the root comments had a relative tree size of less than 0.0125, meaning that they had less than 1.25% of the comments in that thread. The only exception was for relative upvotes, where 84% of the root comments had a score barely above 0, which would translate to a post having fewer upvotes than the number of nodes in the entire thread. Thus, a more usual training set size could encapsulate too wide of a range of scores according to our heuristics, meaning that the comments could have a hugely varying range of content quality.

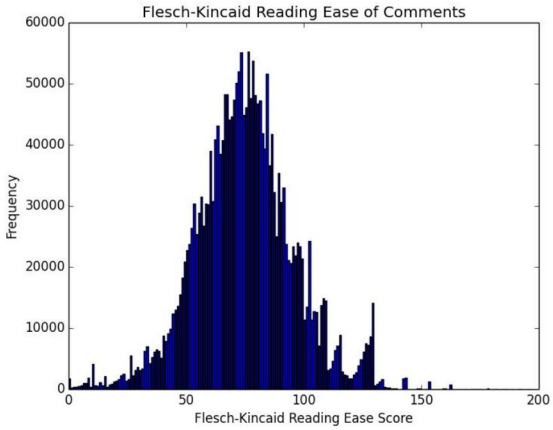


Figures 2a-2d: The histograms for root comments categorized by our normalized heuristics. Notice how most of the comments are classified in the same category associated with low visibility. Most comments received very few votes or replies of any kind.

Interestingly, the two other measurements for content quality showed different distributions. The word count histogram follows a power-law distribution, while the Flesch-Kincaid Ease of Reading Score looks more like a normal distribution. Note that the reading ease score is based on a 0-100 scale, where anything below 50 is university or post-graduate level, 50-60 is late high school student level, and subsequently higher scores represent lower grade level reading ease.



Mean: 48.8, Standard deviation: 71.3



Mean: 74.5, Standard deviation: 21.0

Figures 2e, 2f: The histograms for comment length and reading ease follow different distributions. Comment length follows a power-law distribution (with alpha to be just under 3), while the reading ease follows a normal distribution.

5 Analyzing the Classification Patterns

5.1 Analyzing Good Comment Influence

After classifying the entire data set according to our word-to-vector classifier and according to word count and ease of reading score, we calculated the probability of a certain comment having that trait given that different sets of its ancestors did. For example, if a parent node had a high word count, was a direct reply more likely to be longer? Through how many generations of comment replies can a trait be passed? Which traits trickle down with the most influence? We can find this by comparing the probability of any trait occurring in a comment and compare that to the probability of its ancestors possessing that trait. We'd hope to find that certain attributes are infectious and will induce higher quality responses. We theorized that each comment might possess the ability to "influence" it's children- similarly to how we studied the spread of "infections" in class. Except, in this case, the "infection rate" of a parent comment would depend on its distance from each child. To study this, we developed the following algorithm:

1) Choose an attribute 'A' that we wanted to investigate the spread of.
2) For each comment c with at least 1 ancestor (defined as a comment c' where c is in a chain of comments spawned by c', construct a vector v of length 9 defined in the following manner:
    a) For each v_i such that 0<= i <= 7, v_i = 1 if the i'th most direct ancestor of c has attribute A, v_i = 0 if the i'th most direct ancestor does not have attribute A, and v_i = 0.5 if the i'th most direct ancestor does not exist.
    b) v_8 = the total number of ancestors v has. We wondered whether attributes might be more or less likely to be
3) Also, record whether comment c has attribute A (1 if yes, 0 if no)
4) Perform logistic regression on all comments in the dataset- where are y value for each comment is 1 or 0 depending on whether the comment has attribute A, and the feature vector for each comment is v as defined above
5) Record the weight of each feature to determine the influence strength each parent has on each of its children for that attribute.

We studied two attributes in this method: comment length (comment is "long" if number of words was one standard deviation above the mean) and Flesch-Kincaid reading score (comment is "high reading level" if FK-score is one standard deviation below the mean). The coefficients from the linear regression are as follows:

	Parent	Grand parent	Great Grand parent	4th ancestor	5th ancestor	6th	7th	8th	# ancestors
Comment length	1.36	0.89	-0.08	0.2	-0.17	0.01	0	0.08	0.08
FK score	0.42	0.44	0.11	0.18	0.01	0.15	0.07	0.24	0

From this we can conclude that comments have a large influence on their children and grandchildren's length and style- but any ability to influence further descendents is dubious. Note that for both comment length and FK scores, every 2nd ancestor seems to experience a small coefficient spike. We theorize that this is due to the likelihood that in certain comment chains, every other comment was written by the same poster as the child in question- and thus is more likely to reflect the child's attributes. If we were to conduct this regression again, we might eliminate or otherwise account such posts. In addition, the total number of ancestors a comment possessed did not seem to have any bearing on whether an attribute was present. Since our linear suggests that certain attributes are only passed down by parents to their first 2 generations of children, we decided to empirically calculate the exact influence of parents on their first 3 generations of children. We also decided to throw our own classifier into the mix, and determine whether our own definition of a "good" comment was passed from a parent to its descendents.

5.2 The Empirical Data

Our classifications of the entire data set led us to some interesting data. We computed the probabilities of a child comment being classified as positive given that different combinations of ancestors also displayed the same trait. For example, if a parent comment's content was similar to the content of the comments with the most replies (highest degree), we want to learn how much more likely the child comment to mirror those same content traits.

To measure the influence of an ancestor node, we counted the number of child nodes that were positive for a certain trait, given that a specific ancestor was as well, and compared it to the number of child nodes that were negative for that same trait. We isolated the specific ancestor we were inspecting by ensuring we were including all possible middle generation permutations of positive and negative traits. We also made sure that the generation we were inspecting was a root comment, meaning it was a direct reply to a thread and not to another comment, so that no comments further back in the ancestry could further affect the child comment. It must be noted that this relationship is not necessarily causal- for example- different threads could encourage different types of discussion- and we did not normalize our attributes across all threads.

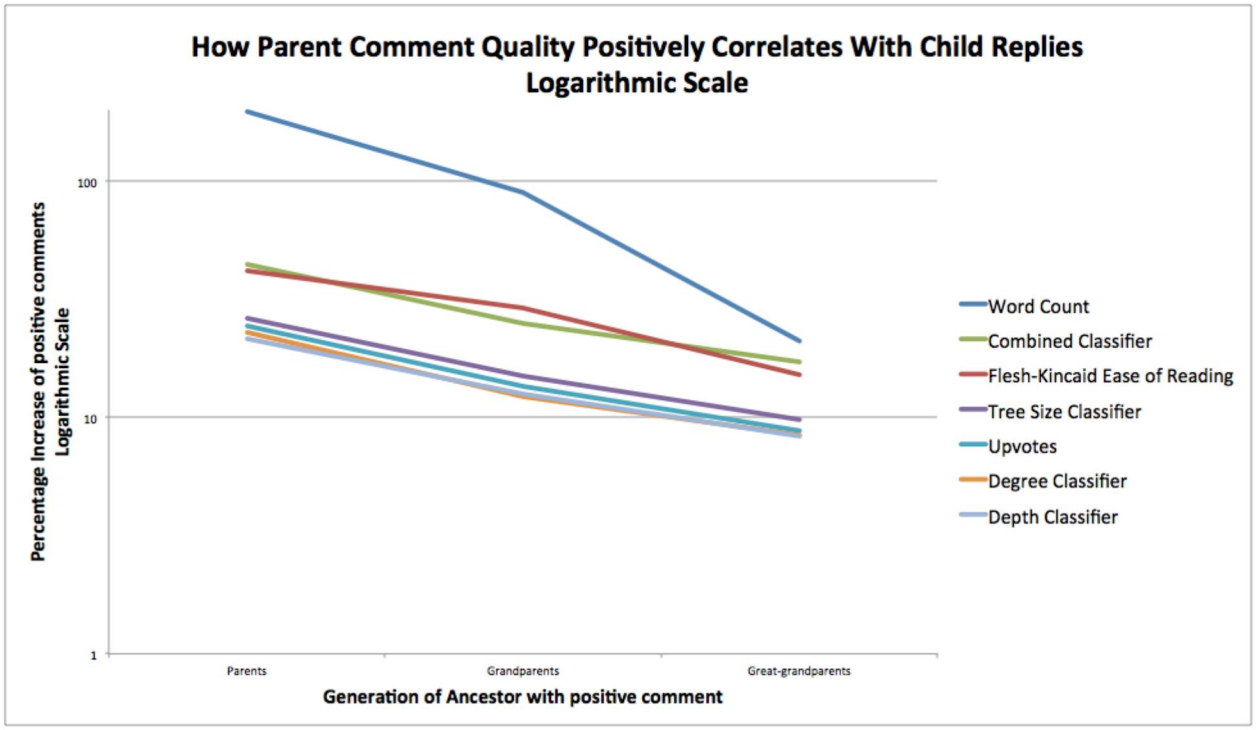


Figure 3: Percent increase in a child comment having a certain trait if its ancestor was classified positive compared to when its ancestor was negative.

Below is the table of our findings represented in figure 3, where each number represents the percent increase in the probability that a comment is classified positive for a trait if a specific ancestor is positive as well.

	Word Count	Flesch-Kincaid Ease of Reading	Degree Classifier	Depth Classifier	Tree Size Classifier	Upvotes Classifier	Combined Classifier
Parents	195.7	41.5	22.9	21.6	26.3	24.3	44.1
Grandparents	89.9	29.0	12.3	12.5	14.9	13.6	25.0
Great-grandparents	21.0	15.1	8.41	8.33	9.79	8.78	17.1

Based on the data that our different measurements and classifiers found, we can see a clear trend. According to every heuristic, there is a positive correlation between a child comment and up to a third generation ancestor having that same trait as well. Our theories about Word count and FK-ease of reading were largely backed up by our empirical findings- our word-count classifier saw a fairly large drop in influence between the parent and grand-parent- while the FK-ease of reading saw a smaller drop in influence. Not only that, but it seems to scale down at a steady rate (logarithmically), giving us reason to believe that the behavior trickles down steadily.

As we expected, the more distant the two nodes are in the graph, the less of an influence the ancestor has over the child. For example, a reply is 41% more likely to be a standard deviation above the

mean reading level if its parent comment is as well, but it is only 15% more likely when its great-grandparent is.

5.3 Conclusions

We found that "good" behavior spreads very well in r/Politics. According to each heuristic, a parent or grandparent being classified positive greatly increases the probability of a child comment displaying that trait as well. However, this isn't exactly an infection rate of good behavior, since there was influence across generations. For example, a grandparent's classification can correlate with a child comment's content to an extent, so it's not enough to only take into account a node's immediate neighbors in the network.

We were surprised to see that good behavior spread so pervasively in the forum. According to the prior research, a more difficult behavior (such as writing well-formulated comments) was predicted to have a very weak adoption rate, but across all metrics, the behavior spread very well. This was especially true for post length, where a child was 190% more likely (almost three times!) to have a length at least a standard deviation above the mean if it's parent did. We were not, however, surprised by the dampening of the effect from more distant generations. It makes sense that a comment further up in the reply chain has a weaker correlation with the quality of the next comment - it's a more distant voice in the online conversation.

In short, the Reddit community in r/Politics does reward intellectual content by replying and voting in higher numbers (as found by our classifier) and with higher quality (as found by the word count and reading ease scores). The community still provides a healthy platform to have intellectual discussion and debate about American politics if it is incited by a well-formed post.

6 Further Research

As stated previously, our models would benefit from considering whether descendents of comments were made by the same person. Also, one feature we neglected in this study was time of posting. Past research has proven that earlier comments are far more likely to receive more responses than later ones. Perhaps time between also plays a role in whether feature are passed on to their children.

We can also attempt to try modifying the classifier by running word2vec with the skip-gram model to see if it produces better results. In the NB case, we can take frequency of words into account to see if that would have improved the classification accuracy of the model.

Finally, the dataset we used was from 2014, before the political climate in the United States took some very interesting turns in 2016. It is quite possible that discussion on r/politics has changed significantly since then, and doing a similar study on a 2016 dataset might be valuable.

7 Group Contributions:

About 80% of the work was done in meetings with all 3 members present- so everyone was involved with each other's projects.

Ben: proposed the dataset, wrote the code to obtain our first heuristics (degree of each comment, maxdepth of each comment, etc.) from the dataset, graphed the correlation and distributions of these heuristics, developed and helped code logistic regression algorithm, wrote about these sections in both the milestone and final report.

Tom: crawled the dataset, defined the comment object, counted and graphed FK-reading level and comment length histograms and pre-computations, wrote sections of the milestone and report, measured probabilities for empirical data and graphed it, bought poster board.

Vein: wrote code for calculating thread size, naive bayes classifier, word2vec and logistic regression models, wrote sections of the milestone and report. Helped with analysis of data. Set up and taught other group members github. Designed and laid out poster.

References

[1] "Reddit Users: Unique Monthly Visits 2016 | Statista." *Statista*. N.p., n.d. Web. 10 Dec. 2016.

[2] Gottfried, Jeffrey, and Elisa Shearer. "News Use Across Social Media Platforms 2016." Pew Research Center's Journalism Project. N.p., 26 May 2016. Web. 10 Dec. 2016.

[3] Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. "Discovering Value from Community Activity on Focused Question Answering Sites." Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12 (2012): n. Pag.

[4] Wang, C., Ye, M., & Huberman, B. A. (n.d.). From User Comments to On-Line Conversations. SSRN Electronic Journal. doi:10.2139/ssrn.2012183.

[5] Centola, Damon. "The Spread of Behavior in an Online Social Network Experiment." *Science* 329.5996 (2010): 1194-197. *Snap.stanford.edu*. Stanford University, 2 Sept. 2010. Web. 18 Oct. 2016.

[6] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

[7] http://snap.stanford.edu/data/politics_subreddit.tar.gz
Note, clicking this link will begin the download process of the 0.5 gigabyte dataset.