Comment Bandwagoning on Inspirational Reddit Posts

David Van Anda
Indiana University at Bloomington, Luddy School of Engineering
Fall 2021 ILS-Z639 with Dr. Ali Ghazinejad
dvananda@iu.edu

Abstract

This study sought to find a relationship between similarities in Reddit comments and whether or not the post they commenting on is inspirational. The hypothesis is that Redditors, having been exposed to inspirational content, are inspirationally primed and will be more likely to 'follow the leader' in the comment section. The study did not yield statistically significant results, however, there is room for improvement in methodology; particularly, in the dataset and the classification of inspirational posts.

1. Introduction

The Bandwagon Effect was first described by Sundar et al. in 2008.[1,2] In their studies, they show that feedback about products from other people will influence an individual's decision to purchase. This usually comes in the form of ratings and/or reviews. Prior to the formalization of the concept in 2008, there was a study in 2006 that found that music selection and enjoyment is socially influenced.[3] These ideas relate to much broader concepts about increasing returns in the economy and winner-take-all environments, sometimes known as the Matthew Effect.[4,5]

Relatedly and more recently, Ognat et al. have a working paper that uses machine learning to identify inspirational posts on social media.[6] In the paper, they classify posts as being Inspiring or Not-Inspiring. They discover important insight into characteristics that might make a post more Inspiring. In the paper, they describe inspiration as having two separate stages. The first is a feeling of motivation. The second is an urge to act. Using a RoBERTa Model, they are able to identify Inspiring posts with approximately 88% accuracy. Depending on the outcome of this study, this can be an important tool in combating toxic feedback loops on social media.

1.1. Research Questions

In my Content Analysis Paper, I observed a Bandwagon Effect in the comment sections of political subreddits; that is, for top-level commenters to echo the sentiment of the top-voted, top-level comment. The first comment displayed under the post itself was indicative of the tone of the rest of the top-level comments. My primary hypothesis for the purpose of this paper is that Redditors, having already seen an Inspiring post, are "inspirationally primed" to follow the comment leader and that I will observe a stronger bandwagon effect on inspirational posts.

The Ognat paper provides a dataset of Reddit posts that are labeled Inspiring and Not-Inspiring. I use this dataset and examine the comments on these posts to determine whether or not the Bandwagon Effect is stronger on inspirational posts. Comments are evaluated by computing the cosine similarity of word embeddings as well as a vector of the normalized sentiment scores returned by the Empath text analysis package.

2. Data

The dataset for this study comes from the Ognat paper. They've made available a csv file that contains almost 11,592 post IDs which are each assigned to one of two columns: Inspiring or Not-Inspiring. The assignment was made either by professional annotators on Amazon's Mechanical Turk or labeled as such by a fine-tuned RoBERTa Model.[7]

The researchers were rigorous and did what they needed to do in order to conduct the study, but there are some limitations. The initial set of posts did not yield nearly enough Inspiring posts, so they had to select posts and posts with comments that contained specific substrings such as "inspir" or "uplift". Another possible limitation is that subsequent sets of posts were selected by training a machine learning model on the human annotated posts. Though they did seem to confirm the accuracy of the model by having humans annotate some of these posts selected by the model, this method seems potentially circuitous since they are training a model on data selected by a similar model.

The dataset, unfortunately, only contained the post IDs and no actual content. The authors note that they were using the Pushshift Reddit API, however I opted to use the Python Reddit API Wrapper or PRAW. It wasn't necessary for this study, but in the accompanying notebook, I replicate (nearly) the authors prediction accuracy results. I trained 3 models (fastText, BERT, and RoBERTa)[8,9] on post titles and post content. My best result was, likewise, RoBERTa, which was 85% accurate. Just shy of the authors' 88%. Having these models trained will prove useful if I expand this research to include posts from outside the provided dataset.

I used PRAW to extract top-level comments. I stored the top-voted, top-level comment and compared all subsequent posts to this one. The similarities were the summed and averaged and stored in a dictionary where the keys were post IDs and the values were the total number of comments, average similarity, and total similarity. There was one dictionary each for Inspiring and Not-Inspiring.

Due to the fact that cosine similarity can only be computed with two non-zero vectors, the number of comments in the Empath sentiment section is limited. If Empath does not detect a sentiment it will return a dictionary where all values are 0 and zero vectors cannot be used. Many comments on Reddit are very short or may contain slang or shorthand that do not express obvious sentiments. The number of top-level comments in the embedding similarity section is unlimited since all comments could be embedded as non-zero vectors. An additional challenge was that a significant number of posts in the dataset have been removed, deleted, or are otherwise no longer on Reddit. This presented a challenge in replicating the results of the prediction models as well as in collecting comment data for this study.

3. Measurements

3.1. Cosine Similarity

The cosine similarity of two non-zero vectors is a number between -1 and 1 that describes the similarity or dissimilarity of the two vectors. 1 is exactly similar and -1 is exactly dissimilar. The raw counts in the Empath sentiment vectors were normalized before calculating the cosine similarity and so the measurement "is equivalent to the Pearson correlation coefficient."[10] In this way, the two sections are measured differently. However, within each section the Inspiring and Not-Inspiring posts are measured the same way. So, the intra-section comparisons are valid while inter-section comparisons of the scores are not valid. Cosine similarity is derived below.[11]

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Equation 1

3.2. Sentence Embeddings

Comments were embedded using the SentenceTransformers Python framework.[12] SBERT (Sentence-BERT) is ideal for comparing sentence embeddings using cosine-similarity. I used an out-of-the-box pre-trained model to embed the comments and compared the top-voted comment to all subsequent top-level comments. Figure 1 at the left is from the original paper and shows the architecture for comparing two sentences.

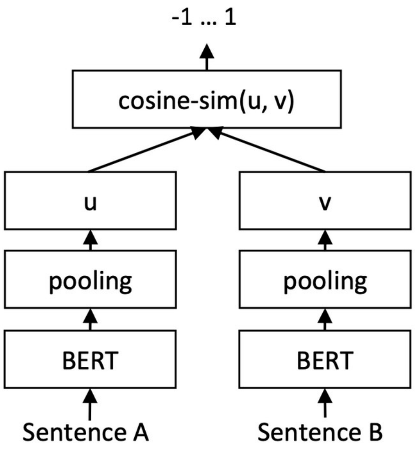

Figure 1: Architecture for comparing two sentence embeddings.

3.3. Sentiment

The Empath text analysis package has 200 categories that are assigned a raw count of matches found within the text being analyzed.[13] Empath allows researchers to create new categories using different models. One of the models is Reddit, however this process requires the researcher to choose specific categories to create. The topics of the Reddit posts that were analyzed in this study were broad, so I didn't think it would be appropriate to subjectively choose categories to be analyzed. For this reason, I used the 200 pre-built categories which were constructed by analyzing works of fiction. If the topics of the comments being studied were more narrow, I think that it would surely be beneficial to create new categories using Empath's Reddit model.

I originally planned to narrow the topics to only positive and negative and then simply compare the distribution of sentiment in the corpus to the sentiment distributions within the comment threads of Inspiring and Not-Inspiring posts. I abandoned this idea for two reasons. The first is that the majority of comments did not have any emotional sentiment as detected by Empath. The second is that I thought that cosine similarity of large vectors of categories/sentiments is a better way to answer my questions.

3.3. Comparisons

Within each section, I compared the results of the Inspiring posts and Not-Inspiring posts by simply comparing the average similarity scores across the dataset. Each key (post ID) in the dictionary has an average similarity score as a value. Because the cardinality for each post ID is different, I needed to calculate a weighted average in order to find the true result.

In order to describe the difference between Inspiring posts and Not-Inspiring posts, I calculated a p-value which tells me how much confidence I can have in the fact that the difference in the results is not due to chance.

4. Results

4.1. Sentence Embeddings

The weighted average of cosine similarities for word embeddings on Inspiring posts was 0.254. The weighted average for Not-Inspiring posts was 0.263. The p-value is 1.82e-10, so we can feel quite confident that this difference is not random. So in this case I have to reject the hypothesis that users are inspirationally primed to follow the comment leader after reading an inspirational post.

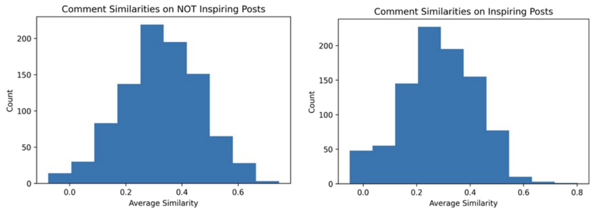

Figure 2: Histograms for comment similarity on each type of post. Not inspiring posts on the left and inspiring posts on the right.

Figure 2 above shows histograms of comment similarity from both Inspiring and Not-Inspiring posts. They are both approximately normal, though the Inspiring post dataset did have a higher maximum.

Figure 3 shows a scatter plot that relates the average similarity in comment sections to the number of comments on a post. This plot includes both datasets and shows a very wide range of similarities when the number of comments is small and a narrowing of this range as the number of comments increases.

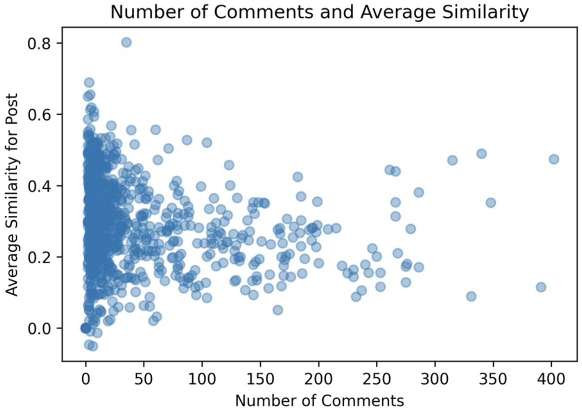

Figure 3: Scatterplot showing average similarity relating to number of comments.

Since the number of comments in this section was unrestricted, I was able to plot this and found that the number of comments on a post possibly follows a power law distribution. Using the python powerlaw package, I found that for the inspirational posts and for the Not-Inspiring posts, where is the scaling parameter. [14] These alphas are a bit low though and the idea that there is a power law distribution here should be confirmed with a goodness of fit test.[15] Figure 4 shows a kernel density estimation of this distribution.

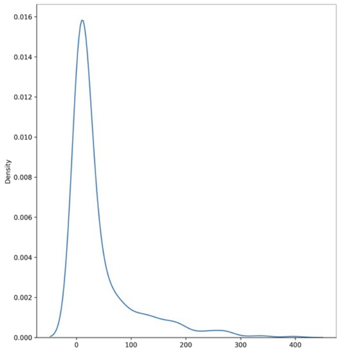

Figure 4: Kernel density estimation of comment distribution. Number of comments on the x-axis.

4.1. Sentiments

The average similarity in sentiment in comments on Inspiring posts was 0.828. The average similarity on Not-Inspiring posts was 0.803. By simply looking at these two numbers, this dataset would appear to confirm my hypothesis. The similarities were not normally distributed and so I was not able to conduct the same t-test as was done on the sentence embedding data. Instead, I used the Wilcoxon signed-rank test.[16] This gave me a p-value of 0.54, which is quite high. This method is more likely to produce a significant p-value when the data is not normally distributed, but 0.54 is certainly high enough to, once again, reject the hypothesis that the Redditors are inspirationally primed to follow the leader after reading an inspirational post.

Figure 5 below shows corresponding histograms for average similarities of comments on Inspiring and Not-Inspiring posts.

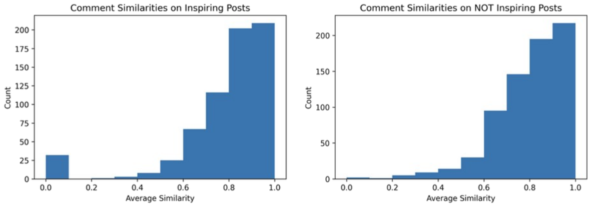

Figure 5: Histograms for comment similarity on each type of post. Inspiring posts on the left and not-inspiring posts on the right.

5. Conclusion

This study did not produce results that show that Redditors are inspirationally primed to follow the comment leader. In fact, the sentence embedding data showed, with statistical significance, that comments are more similar on Not-Inspiring posts. There was not a statistically significant difference in the Empath sentiment/category data, though comparison of the means did favor the hypothesis.

The sentiment similarities were, for both datasets, quite high on average. A possible explanation for this is the way that Reddit presents comments to users. Normally and unless a user changes this preference, the comment at the top of the page is the one with the most votes and not the first temporally. Possibly then, this top comment is merely a reflection of the sentiment of the other users viewing the post and not an influencer of them. I have a large dataset of WhatsApp messages that are organized temporally and possibly this might be a better dataset to examine the influence of messages on subsequent messages. Though using this dataset would lose the inspirational priming component of this study.

The statistically significant difference in the embedding similarities must, I think, be due to the initial data collection. As I noted earlier in the paper, Ognat et al. used somewhat subjective methods to compile enough Inspiring posts. For instance, a lot of these Inspiring posts were from the r/AskReddit where a Redditor will ask a question and receive different answers from potentially thousands of people. The Not-Inspiring posts were collected from a broader range of subreddits where topics can often be quite narrow and therefore the comments might be more likely to share similarities.

Another possible explanation for the closeness in similarities might be that once a consensus forms in a comment section, those with divergent opinions are less likely to comment. Doyle et al. observed that opinions have a quality of ‘stickiness’ that determines the difficulty in overturning the majority opinion.[17] Given how Reddit orders comments by number of votes, this may increase the stickiness of opinions and create a very strong positive feedback loop. This would follow some of the winner-take-all themes of this study, but at the same time make it difficult to measure the effect if it is itself dissuading Redditors with divergent opinions from commenting.

It is still my intuition that comments influence other comments, though this study failed to prove it. I think that inspirational priming is an important component of the idea, so while the WhatsApp dataset might be interesting to examine, it’s not ideal. Another way to conduct this study would be to revert to the original idea of comparing the sentiment in comment threads to the distribution of sentiment in comments more broadly. Though with the size of the sentiment vector |V|=200, this would get computationally quite complicated. In my first paper, I set up simple conditional probabilities to determine that commenters on political subreddits followed the comment leader. A solution might be to use my trained RoBERTa Model to select Inspiring and Not-Inspiring posts from a more narrow set of subreddits and then use Empath's Reddit model to build a limited set of custom categories. In this way, I might be able to use conditional probabilities to demonstrate inspirational priming and the Bandwagon Effect.

5.1. Limitations

The primary limitation of this study is that I used lists of Inspiring and Not-Inspiring Reddit posts that were collected with some subjectivity and may not be globally representative. Secondly, cosine similarity may not be sufficient on its own to measure Bandwagon Effect. It may be preferable to use conditional probabilities on a smaller set of topic categories and sentiments. Finally, my experience is limited with real-world, self-generated, non-parametric datasets and so this study would certainly benefit from a review by a more experienced researcher.

10. References

[1] Sundar, S. Shyam, Anne Oeldorf-Hirsch, and Qian Xu. "The bandwagon effect of collaborative filtering technology." CHI'08 extended abstracts on Human factors in computing systems. 2008. 3453-3458.

[2] Sundar, S. Shyam. The MAIN model: A heuristic approach to understanding technology effects on credibility. MacArthur Foundation Digital Media and Learning Initiative, 2008.

[3] Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. "Experimental study of inequality and unpredictability in an artificial cultural market." science 311.5762 (2006): 854-856.

[4] Arthur, W. Brian. "Competing technologies, increasing returns, and lock-in by historical events." The economic journal 99.394 (1989): 116-131.

[5] Merton, Robert K. "The Matthew effect in science: The reward and communication systems of science are considered." Science 159.3810 (1968): 56-63.

[6] Ignat, Oana, et al. "Detecting Inspiring Content on Social Media." arXiv preprint arXiv:2109.02734 (2021).

[7] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[8] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.

[9] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[10] "Cosine Similarity" Wikipedia, Wikimedia Foundation, 12Nov.2021, https://en.wikipedia.org/wiki/Cosine_similarity.

[11] "Pearson Correlation Coefficient" Wikipedia, Wikimedia Foundation, 12 Nov. 2021, https://en.wikipedia.org/wiki/Pearson_correlation_coefficien.

[12] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[13] Fast, Ethan, Binbin Chen, and Michael S. Bernstein. "Empath: Understanding topic signals in large-scale text." Proceedings of the 2016 CHI conference on human factors in computing systems. 2016.

[14] Alstott, Jeff, Ed Bullmore, and Dietmar Plenz. "powerlaw: a Python package for analysis of heavy-tailed distributions." PloS one 9.1 (2014): e85777.

[15] Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." SIAM review 51.4 (2009): 661-703.

[16] Wilcoxon, Frank. "Individual comparisons by ranking methods." Breakthroughs in statistics. Springer, New York, NY, 1992. 196-202.