Visualizing Subreddit Similarity Using Natural Language Processing

Jourdan DeVies, Tess Leggio, Rachel Shah, Alex Sitaras, Lauren Skorb, Darsh Thakkar

Introduction - Motivation

In today's society, we are plagued with political bias. People tend to seek news from a single source which frequently serves to further polarize and reinforce held beliefs. Social media is a popular outlet for political debate and news consumption. For our project, we leverage Reddit data to expose individuals to all sides of a given topic and make inferences about groups of online communities. By being able to make comparisons about communities in relation to given topics, it allows us to uncover the nuances of political discourse and how information is dispersed on Reddit.

Problem Definition

We grouped similar subreddits based on the choice of words used within comments and submissions for a particular topic. Specifically, we were interested in discussions taking place within the most populated politically-related subreddits about some of the most notable events that occured in 2018. The events we chose to explore were the Singapore Summit (North Korea meeting), Hurricane Florence, the Kavanaugh hearing, and the government shutdown. We provide a visual tool that can be used to see how different subreddits engage with these topics, so that people can explore a diverse set of perspectives. We believe that access to nuanced opinions and disparate viewpoints is important because our society is so polarized. We want to portray opinions as part of a spectrum to show radical, mild, and neutral thoughts regarding our selected topics. If people become aware of the larger picture and distance themselves from online echo chambers, we believe that people can start forming opinions more nuanced than the opinions they presently hold. Our project represents our attempt to take a step towards a less biased society.

Literature Survey

General

Pang and Lee provide a general background in sentiment analysis and the importance of topic material on degree of positivity (2008). Discussions of how to analyze "big" unstructured data more broadly were informative, but not directly relevant to our approach (Gandomi & Haider,

2015)

Topic/Sentiment

We used a topic cluster model, hLDA, as proposed by Weninger, Zhu, and Han, but apply it between subreddits to get a subreddit clustering as opposed to within to get a hierarchy of topics for each subreddit(2013). Others have used LDA and hierarchical clustering models to group similar news articles based on detected bias (Hamborg et. al., 2018). We considered a joint model which performs sentiment analysis and topic modeling in a single, unsupervised model, but decided to perform sentiment analysis separately (Lin & He, 2009). An approach to analyzing Twitter data used a combination of a hierarchical feature subset selection algorithm and sentiment analysis to visualize thought "bubbles" (Diehl et al., 2018). However, since Twitter has certain measurable user interactions that Reddit does not, it is difficult to implement on Reddit data. We considered using the sentiment of child comments in relation to parent comments on a topic thread, but ultimately did not do so, instead using a bag-of-words (corpus) for each subreddit with hLDA to capture hierarchical structures (Zayats & Ostendorf, 2018). An alternate implementation is a neural network for short texts to determine comment sentiment (Dos Santos & Gatti, 2014). Topic-Based Latent Semantic Analysis can be used to quantify topic similarity between noisy social media texts; however, this is beyond the scope of our approach (Dang et. al., 2016). Topic models and Jaccard coefficients can be used to find community pairs with similar userbases but dissimilar content, however this is the reverse of our goal (Hessel et. al, 2015).

Community Detection

Similarity of subreddits have been historically determined through Latent Semantic Analysis (Dumais, 2004) and cosine distance, clustering, and topic models. In determining community structure, Bedi and Sharma provide a survey of algorithms such as agglomerative hierarchical clustering and graph partitioning (Bedi & Sharma, 2016) . These compute edges based on user connections which we did not implement in our project, but if we did, we would have had to adapt these methods to create weighted edges based off community similarity (Olson & Neal, 2015). Subreddit recommendation systems were described using graph partitioning, but is primarily done based on user-user similarity rather than subreddit similarity (Sundaresan, Hsu, & Chang, 2014). We could have used classification to find subreddit similarity using TF-IDF occurrences and Jaccard's coefficient, but this method might not detect true differences (Alquaddoomi & Estrin, 2018). Studies show that style of language used performs better in community detection than topic alone, so adding sentiment analysis is an asset of our project (Tran & Ostendorf, 2016).

Visualization

Interest-based visual maps of subreddits have been created, but with edges constructed based on

user crossover (Olson & Neal, 2015). Instead, we used sentence trees to convey relationships between words and sentences that we could visualize (Hu, Wongsuphasawat, & Stasko, 2017). Visualizing semantic spaces derived from Latent Semantic Analysis reduces dimensionality, but can be slow (Landauer, Laham, & Derr , 2004). FluxFlow demonstrates a unique way to visualize anomalous information and sentiment in Twitter data, which, if our project were developed further, could be helpful in detecting subreddit bots and other spamming attempts (Zhao, et. al. 2014).

Proposed Method

Intuition

Various techniques exist for performing sentiment analysis on social media text and, separately, for detecting communities within social media platforms. Our approach is novel in that we found clusters of communities within Reddit by clustering subreddits based on the language unique to each subreddit relating to a given topic. To our collective knowledge through our research, using this approach to visually inform users about clusters of similar subreddits has never been done.

Data Collection

Over 1.7 billion Reddit comments and submissions are available as public datasets through Google BigQuery. We limited our choice of subreddits to 30 political- and news-related subreddits. We queried submissions and comments using SQL queries for subreddits for each month of 2018, resulting in 12 datasets of 2 tables for each month. The resulting tables were too large to export directly to a csv, so we exported these tables from Big Query into a Google storage bucket as gzip compressed csv files. In total, the compressed files are about 15GB. After data collection, we uncompressed the files in a python script and combined each month of comments and submissions files into separate dataframes for each subreddit.

Filtering Our Data

Given that our collected data is not filtered by topic, we had to determine keywords related to each topic to search within our data for. If a comment or submission contained a match, then we added the selection to word banks that we assembled for each subreddit by topic.

Algorithms

What is our approach? We are implementing hierarchical Latent Dirichlet Allocation (hLDA), a probabilistic topic model, to conduct hierarchical topic clustering on subreddit post and comment threads within a constrained subset of polarizing political topics. In addition to our hLDA algorithm, we integrated sentiment analysis by subreddit and topic. Using Python's

Natural Language Toolkit (NLTK), we computed average positive and negative valence for subreddit by topic using Vader's SentimentAnalyzer (Bird, Klein, and Loper, 2009). We also used the Python package TextBlob to compute subjectivity of subreddits by topic, where subjectivity represents more emotive and judgemental language, as opposed to more fact-based.

How does our approach work? The hLDA algorithm identifies shared words among subreddits, and at each node, all of the subreddits under that node share a set of unique words. Words are defined after first iterating through all of the subreddits with initial text analysis to extract lemmas, allowing for more precise comparison across different word forms. Importantly, this algorithm does not specify a number of clusters, but we can specify the level at which we would like to cluster. In terms of sentiment analysis, NLTK Vader's Sentiment Analyzer uses a rules-based approach to assign positive, negative, neutral, and compound scores of valence. Also, TextBlob subjectivity is based on pre-determined subjectivity scores for the entire vocabulary, which is multiplied by intensity words and averaged across comments and submissions.

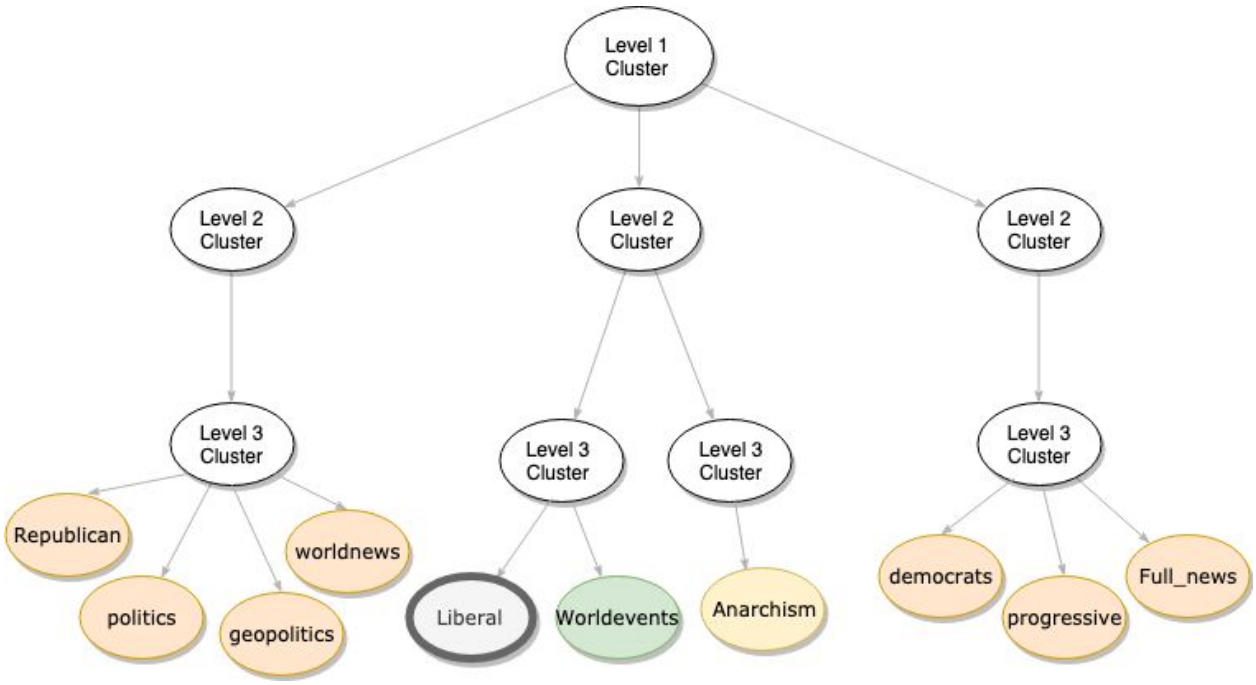

Figure 1. Example hLDA output of three-level subreddit clustering. If we are analyzing Liberal, it's word usage makes it most closely related to Worldevents, then Anarchism, and finally the orange colored subreddits.

How does our approach effectively solve our problem? By implementing hLDA across subreddits for each topic, we are able to identify commonalities in the language used to discuss

various polarizing topics (Hamborg et. al., 2018). This will allow us to identify more granular differences in language than, for example, binary positive/negative sentiment analysis.

What is new in our approach? Various techniques exist for performing sentiment analysis on social media text and, separately, for finding communities within social media platforms. Our approach is new in that we will find communities within Reddit by clustering subreddits based on overall topic analysis within each subreddit for a given topic.

Visualization

Our interactive visualization, created using D3, will show how subreddit clusters change based on political topic as well as depict word clouds and sentence trees for each subreddit pertaining to each topic. For subreddits that do not contain many submissions or comments related to a given topic, a word cloud will not be produced for the given subreddit and the sentence tree will be limited in functionality. However, clustering, sentiment analysis, subjectivity, etc will still be able to be evaluated.

The centerpiece of our visualization is an interactive scatterplot that allows viewers to engage with our results and allows axis zooming, panning, and clicking of data points (each subreddit represents a data point) to reveal word cloud and sentence tree analysis corresponding to the data point. The scatter plot also enables the viewer to choose the metrics they wish to present on the x- and y-axis. The metrics we chose to include as part of our scatterplot are positive sentiment, negative sentiment, neutral sentiment, compound (overall) sentiment, subjectivity, number of subscribers, number of submissions (related to the chosen topic), and number of comments (related to the chosen topic).

Link to our visualization: https://cse6242-subredditsimilarity.github.io/

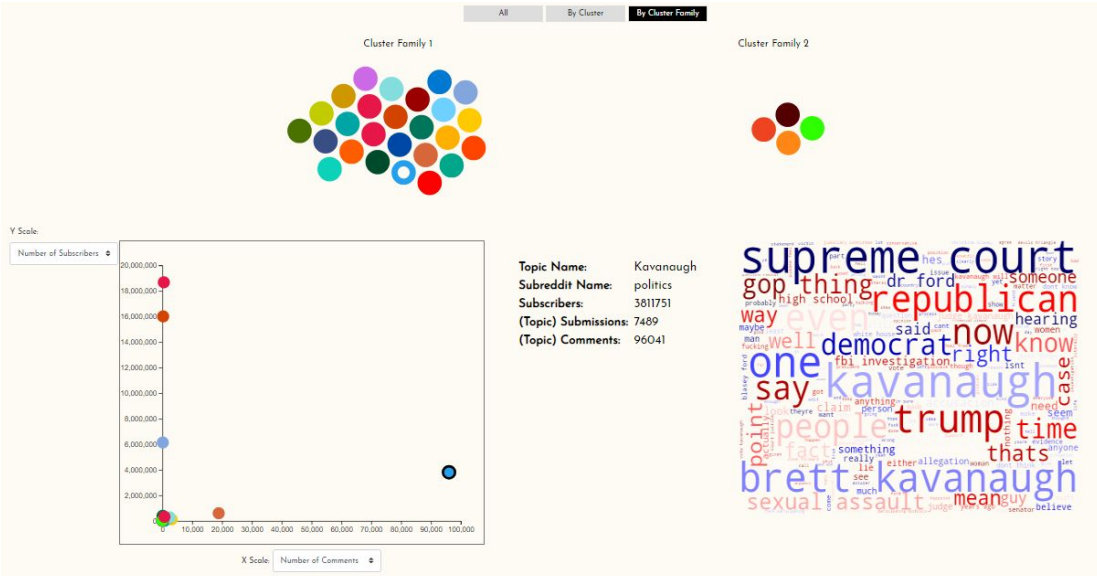

Figure 2. A screenshot of our visualization that reveals comparably high discussion of the Kavanaugh hearing in the politics subreddit compared to other subreddits.

Experiments and Evaluation

There are two primary components of our tool that must be evaluated: the quality of clustering performed by the hLDA analysis and of the interactive visualization. We have collected and cleaned our data, and created hLDA models using subsets of our data (subsets are chosen given the runtime required to use the hLDA code) based on tuning parameters that we found to work best for our data. We tested several different tuning parameters, which determined the level of smoothing required in our clustering algorithm, and we evaluated the face validity of each set of unique tuning parameters. We tested eta (smoothing over topic word distributions) values between 0 and 1 and number of cluster levels or the depth of the tree structure, and we then evaluated the output of these models. We considered how many clusters were created and if the clusters created matched our understanding of the subreddits. We chose to run hLDA with an eta value of 0.7 and three clustering levels.

To assess the effectiveness of our interactive visualization, we conducted user surveys to rate the following sentences on a scale from one to five:

Q1. It was easy to interact with this website.
Q2. It was easy to understand the information conveyed in the website.
Q3. I learned something new from interacting with the website.
Q4. I would recommend this website to my friends.
Q5. I thought that the design was visually appealing.
Q6. I would use this website in the future.

Q7. The website presents new information than what I have seen before.

We used the online forum NextDoor and social media to sample survey respondents. We found the following results:

	1	2	3	4	5	Average Score
Q1	5%	5%	15%	20%	50%	4.10
Q2	5%	15%	40%	15%	25%	3.40
Q3	0%	10%	5%	60%	25%	4.00
Q4	5%	10%	30%	20%	35%	3.70
Q5	0%	10%	10%	20%	60%	4.30
Q6	5%	25%	20%	25%	25%	3.40
Q7	5%	0%	0%	30%	65%	4.50

We received the lowest scores on Q6 (*I would use this website again in the future*) and Q2 (*It was easy to understand the information conveyed in the website*). We received the highest score on Q7 (*The website presents new information than what I have seen before*). Our survey results show that users find the website interactive and that it provides new information, but we also find evidence that our visualization and results may require more explanation because the information and format is novel in method for displaying Reddit information. We had users interact with the tool on their own, so we could have seen different results had we walked the users through how to use our visualization with a video or in person.

Conclusions and Discussion

Our use of hLDA has allowed us to obtain and visualize insights into Reddit communities in a clear manner, contrasting other existing visualizations. Furthermore, hLDA allows for sentiment analysis on chosen topics, an area of inquiry not often explored in research related to subreddit data. Finally, users could potentially be introduced to communities with contrasting views.

Despite our success in using hLDA, certain limitations exist. The process of running the model is time-consuming especially if additional topics are implemented or if the number of subreddits explored increases. With present day computing power, classification by a hLDA model in real time is challenging, if not impossible, and this represents a limitation in our work and in obtaining inferences about enormous communities in real time.

From our visualization, a plethora of insights can be made about how Reddit users engage with the political topics we explored. For instance, we're able to validate our intuition that people, as a whole, speak more negatively on the internet than positively, but we're also able to make less intuitive findings such as the fact that the NeutralPolitics subreddit held less comments expressing negative sentiment (as opposed to positive sentiment) towards the government shutdown and that the Kavanaugh hearing had a lesser subjectivity score on average, meaning that language use was less emotive, than the also controversial Singapore Summit and government shutdown (the Kavanaugh hearing did have a higher subjectivity score than Hurricane Florence, an event regarded as generating less controversy, and Hurricane Florence's lesser subjectivity score is therefore intuitive).

Distribution of Team Effort

All team members have contributed a similar amount of effort.

References

1. Alquaddoomi, F., & Estrin, D. (2018). Ranking subreddits by classifier indistinguishability in the Reddit Corpus, presented at eKnow 2018: The Tenth International Conference on Information, Process, and Knowledge Management, Rome, 2018.
2. Bedi, P., & Sharma, C. (2016). Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *6*(3), 115-135.
3. Dang, A., Moh'd, A., Islam, A., Minghim, R., Smit, M., & Milios, E. (2016). Reddit temporal n-gram corpus and its applications on paraphrase and semantic similarity in social media using a topic-based latent semantic analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3553-3564).
4. Diehl, A., Hundt, M., Haussler, J., Seebacher, D., Chen, S., Cilasun, N., ... & Shreck, T. (2018, October). SocialOcean: Visual Analysis and Characterization of Social Media Bubbles. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)* (pp. 1-11). IEEE.
5. Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69-78).
6. Dumais, S. T. (2004), Latent semantic analysis. Ann. Rev. Info. Sci. Tech., 38: 188-230. doi:10.1002/aris.1440380105
7. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.
8. Hamborg, F., Donnay, K., & Gipp, B. (2018). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 1-25.
9. Hessel, J., Schofield, A., Lee, L., & Mimno, D. (2015). What do vegans do in their spare time? Latent interest detection in multi-community networks. *arXiv preprint arXiv:1511.03371*.
10. Hu, M., Wongsuphasawat, K. & Stasko, J. (2017). Visualizing social media content with SentenTree. *IEEE Transactions on Visualization and Computer Graphics*, *vol. 23*(1), pp. 621-630. https://doi.org/10.1109/TVCG.2016.2598590
11. Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5214-5219.

12. Lin, C., & He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 375-384). ACM.
13. Olson, R. S., & Neal, Z. P. (2015). Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, *1*, e4.
14. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.
15. Sundaresan, V., Hsu, I., & Chang, D. (2014). Subreddit Recommendations within Reddit Communities.
16. Tran, T., & Ostendorf, M. (2016). Characterizing the language of online communities and its relation to community reception. *arXiv preprint arXiv:1609.04779*.
17. Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* (pp. 579-583). IEEE.
18. Zayats, V., & Ostendorf, M. (2018). Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association of Computational Linguistics*, *6*, 121-132.
19. Zhao, J., Cao, N., Wen, Z., Song, Y., Lin, Y. R., & Collins, C. (2014). # FluxFlow: Visual analysis of anomalous information spreading on social media. *IEEE transactions on visualization and computer graphics*, *20*(12), 1773-1782.
20. Bird, Steven, Ewan Klein, and Edward Loper (2009), *Natural Language Processing with Python*, O'Reilly Media.