Evaluating Identity and Arguments Online

Amanda Baughan
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98103
baughan@uw.edu

Elena Khasanova
Department of Linguistics
University of Washington
Seattle, WA 98103
ekhas1@uw.edu

Erik Tomasic
Department of Linguistics
University of Washington
Seattle, WA 98103
etomasic@uw.edu

1 Introduction and Motivation

An enormous amount of interpersonal communication now occurs online, as the number of people on social media platforms is higher than ever [1]. Through messaging platforms, social media, dating apps, comment sections, and more, users can collaborate, debate, and engage on virtually any topic with a vast array of other people. These interactions can have enormous benefits for users, including stronger interpersonal bonds [11] and opportunities to fulfill needs of socialization [14].

However, not all online conversations are productive. Users report that heated conversations with friends and connections are routine [5], and in some instances, online arguments with loved ones lead to permanently severed relationships [15, 19]. Despite this, completely eradicating online conflict may also be undesirable. Arguments are an important feature of close interpersonal relationships [18], and they play a critical role in people's negotiations with one another [12]. Prior work has shown that people leverage digital platforms as a tool for carrying out these essential disagreements [12], and that it is possible for users to argue productively online [17].

Thus, this works seeks to evaluate if and how people argue differently based on whether they share values. We draw from work in philosophy [20], social computing [16, 23], and machine learning [17, 22] to evaluate the hypothesis that *people who share values are more likely to argue in ways that support a restorative justice framework, including appeals to shared values, civility, and reconciliation than people who don't share values, which will be likely more susceptible to personal attacks and arguments without resolutions* [20].

Restorative justice is characterized as a "repair of justice through reaffirming a shared value-consensus in a bilateral process between victim and offender" [20]. This is in direct contrast to retributive justice, which is a "repair of justice through unilateral imposition of punishment" [20]. Much of current social media's justice process skews towards retributive justice, with features such as blocking, censoring, and account suspension/deletion. In contrast, restorative justice emphasizes striving for consensus (for instance, on harm, responsibility, and values) as necessary to restore justice. The goal is to reaffirm what are considered *shared* and *identity-defining* values of a community.

Thus, this work contributes:

- Analysis of Twitter users based on salient shared-value groups, including a comparison of sentiment and discourse types within and between these groups.

Preprint. Under review.

- Analysis of discourse types in two different subreddits: /r/changemyview, an example of a community with a shared value of tolerance, and /r/PoliticalDiscussion, a community with less prevalent shared values.

We find that there is cautious evidence that sentiment within groups is more positive than between groups, as 4/5 clusters analyzed had a positive or neutral average sentiment polarity, whereas between group polarity was negative on average. Additionally, we find that the most common reaction to disagreement is more disagreement, regardless of shared values. We discuss our findings and future work to study and design for constructive argumentation online.

2 Related Works

2.1 Identity Groups: Clustering Social Media Users

Stewart et al. [16] used structural data (network ties) on Twitter to cluster social media users based on a similarity metric of shared audiences. The "shared audience" metric was the Jaccard similarity of the follower lists of each pair of users present in the data and comprised the edge weight of their graph network of users. They clustered users with the Louvain modularity [2] algorithm, where each community initially consists of one node (user) and modularity compares the intracommunity edge weights (determined by shared audience in this case) with intercommunity edge weights. Stewart et al. [16] conducted grounded qualitative analysis of clusters to develop findings characterizing the shared audience clusters and their interactions.

2.2 Structural Properties of Discourse and Interaction Patterns

A large body of research in Natural Language Processing is focused on finding the language indicators of constructive, healthy conversations or predictors of conversation failures followed by destructive behavior and classifying these features. The technical component of these papers is not particularly remarkable (many rely on linear or neural classifiers), however, the feature sets and annotation schemata are informative.

Structural discourse research models discussions as dynamically developing sequences of discourse acts seen as a classification or sequence-to-sequence problem. Ghosh et al. [4] follow Pragmatic Argumentation Theory to annotate online discussion entries as Callout (reaction to a message that includes attitude and ideally a rationale) and Target (a part of the previous message that is called out) and train a hierarchical clustering model to group discourse units into types such as Agreement and Disagreement. This model achieved relatively low accuracy, however, it supported argument analysis by evaluating subsets of messages, rather than treating messages as a whole. The approach presented in Zhang et al. [23] is focused on the high-level interaction patterns organized in a reply-tree, with the nodes representing the volume of responses received, edges representing response types (reactions, replies).They evaluated public discussions on roughly 9,000 Facebook pages with high traffic. One of the most relevant findings is that *commenter*'s were more important for setting the tone of arguments (expected outcomes, such as blocking, and length of argument) than the *content* of the post.

Zhang et al. [22] produced an annotated dataset based on interaction patterns and content of online discussions on the platform Reddit. Zhang et al. [22] used iterative crowdsourced annotation to define Reddit thread discourse types and used this to train a supervised model to predict discourse types. They found that the discourse types that may indicate or lead to an argument or its resolution, are AGREEMENT, APPRECIATION, DISAGREEMENT, NEGATIVE REACTION, HUMOR. These are in addition to several other discourse types, such as QUESTION, ANSWER, and ELABORATION. This dataset takes into account both discourse acts and discourse relations (linking to a particular comment in the tree). The classification is based on a rich feature set combining information about language, structure, author, thread, and community for each discourse act. The ablation study showed that content and structure are essential to the classification of discourse acts while removing author, thread or community is less harmful. Thus, the study finds that the discourse acts involved in arguments receive the lowest inter-annotator agreement (measured with Krippendorf's alpha across 3 annotators of each entry), which has a strong positive correlation with the lowest F1 score. This emphasizes the complexity of argumentative conversations and the need to further research. The analysis of the "chains of opposition" provided in the paper gives some insight on conflict resolution common to Reddit discourse: 61% of DISAGREEMENT acts were followed by nothing, 18% were

continued with elaboration, and only 7% ended in agreement. This provides cautious support that retributive justice is more widespread in online communities than restorative justice.

The studied body of literature suggests the importance of both language features characterizing the content and the higher-level interaction patterns to offer a more holistic view of discussions. We were surprised to find that the research on argumentation mainly discusses negative outcomes of online disputes and consider markers such as "a comment removed by a moderator" label [7], blocking of the participants [22], or antisocial behavior that followed the conversation, which are all signs of retributive justice. Some attempts to evaluate conflict mediation were made in Konat et al. [8], however, this framework is fairly complex and is hard to extrapolate. These findings suggest that a closer look at conflict resolution patterns and possible ways to scaffold restorative justice is a promising and much needed avenue of research.

3 Data Collection

A prerequisite to this work is to use a large social network platform with potential for opposing viewpoints. To this end, we decided to experiment with two approaches across Reddit and Twitter.

3.1 Reddit

We attempted to analyze user upvote data in order to construct edges for a graph network, however, only 1% of users make that data public. To overcome this limitation, we used subreddit as a proxy for shared values, and collected around 350 posts and 30,000 comments from `/r/changemyview` and `/r/PoliticalDiscussion`. We hypothesize that the shared value of tolerance will be more salient on `/r/changemyview` than `/r/PoliticalDiscussion`, and therefore use `/r/changemyview` as the shared value group and `/r/PoliticalDiscussion` as the non-shared value group.

3.2 Twitter

We collected data from Twitter on #Coronavirus as we expect some division and controversy around this topic. We used a subset of the dataset on #Coronavirus [10], which contains Tweets published between May 02, 2020 10:18 AM and May 03, 2020 09:57 AM. These Tweets were collected in real time and supplied with sentiment polarity scores obtained from TextBlob[1]. The data was filtered to only include English language, and keywords “corona”, “coronavirus”, “covid”, “covid19” and “sarscov2”. This initial dataset contained 2,216,553 Tweet IDs due to the Twitter Developer Policy, from which we "hydrated" (obtained metainformation for) 153,600 Tweets. To hydrate the Tweets, we used DocNow Hydrator[2]. This led to a dataset that included the count of how often each Tweet was favorited (liked), but it did not contain information about the users who liked the Tweets. This information is also not available through any standard Twitter API. This information is, however, essential for our research on shared values, therefore we used the `request` library to scrape the *likes* from Twitter web pages. We further filtered our dataset to only the Tweets that were posted by users who liked other Tweets. The final dataset consists of 35,950 Tweets posted by 33,916 users. This dataset was used for defining identity groups reported in Section 4.1. Further, in an attempt to source more chains of conversations, we used twarc command line interface to collect an additional set of 49,230 Tweets that are responses to the Tweets posted by 7,103 users present in our largest clusters. To increase the likelihood of querying only the Tweets with responses (critical for managing request rate limits), we searched only the Tweets that were liked at least 30 times. The majority of these additional Tweets were, however, authored by the users outside our clusters making it insufficient for the analysis of disagreement resolution. Applying more filters in data sourcing has to be carefully considered as it may introduce implicit biases and skew distributions. This dataset was supplied with the ’root’ label for the id of the Tweet that started the conversation converted to ConvoKit[6] format to be conveniently analysed with this toolkit.

[1] https://textblob.readthedocs.io/en/dev/

[2] https://github.com/DocNow/hydrator

4 Methods

4.1 Clustering Identity Groups

We used the methods set forth in Stewart et al. [16] to conduct network analysis. We represented users as graph nodes, and edge weights were the sum of likes in either direction. (e.g. edge weight between users A and B = count of times user A liked user B's Tweets + count of times user B liked user A's Tweets). We then ran Louvain's algorithm for network modularity [2] and used Gephi to build the graph visualization. This resulted in the 33,916 users being grouped into 6,978 clusters, with a modularity of 0.93 and resolution 1.0. This graph had a diameter of 29 and density of < 0.001. The average path length was 8.55, meaning, the average graph-distance between all pairs of nodes was approximately 9 nodes. The average degree was 1.96. The top 5 clusters represent 11,396 (33.6%) total users, connected by 13,630 edges (see Figure 1). The remaining clusters had an average group size of 3.23 users per cluster.

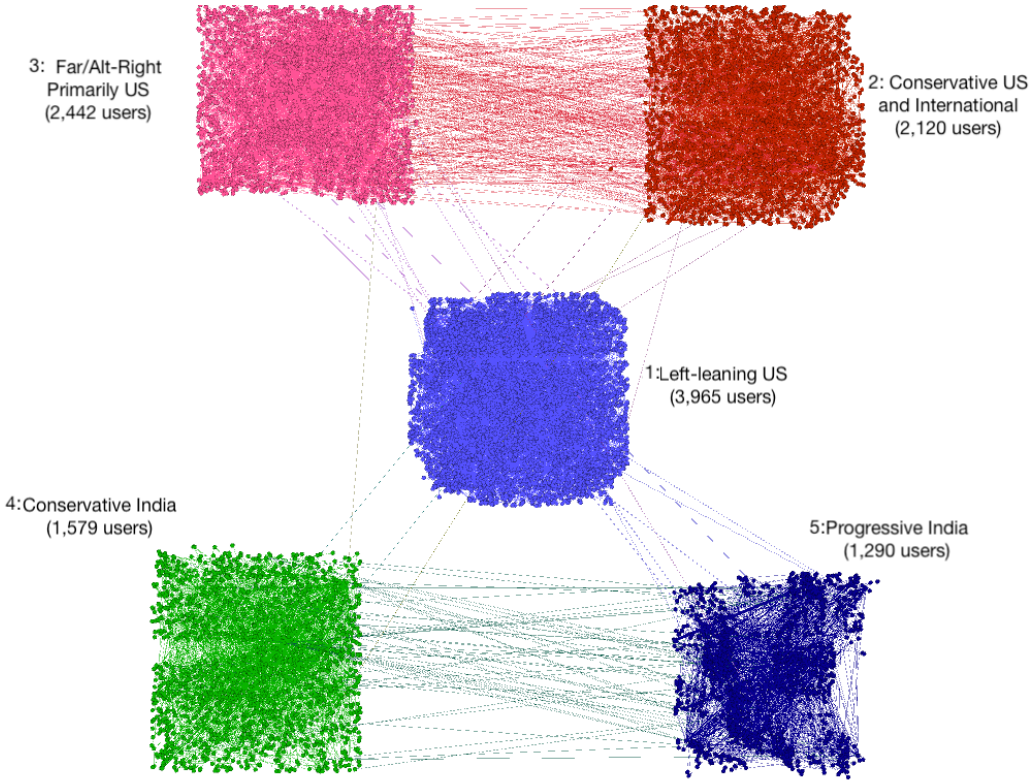


Figure 1: The top five clusters from Twitter data. Nodes are users, edges are number of liked Tweets between two users.

4.2 Evaluating Argumentation

Using the CRFsuite Python library, a conditional random field (CRF) model was built from the annotated dataset used by [22] to analyze discourse acts within `/r/PoliticalDiscussion` and `/r/changemyview`. An F1 score of 0.782 was achieved, using features such as n-grams and post depth.

Twitter data was analysed with three methods: comparison of sentiment polarity scores supplied with the initial dataset, evaluating differences between clusters using FightingWords Transformer from ConvoKitJ. P. Chang [6] implementing a modeling approach for lexical feature evaluation from Monroe et al. [13] based on z-score of logg-odds ratio, and manually annotating discourse patterns in the larger Twitter dataset. Other methods offered in ConvoKit such extraction of politeness patterns

or coordination scores between the groups were attempted but not reported as they require further investigation and adaptation of the data.

5 Results

5.1 Twitter: Characterizing Clusters of Users in #Coronavirus

- **Cluster 1: Left-leaning US accounts**. Cluster 1 is comprised of 3,965 unique users. They have 4,269 followers on average ($SD = 69,093$). 1% ($n = 56$) of these users have verified accounts, many of which are journalists for left-leaning news sources. These accounts tend to be journalistic and factual in nature, with some critiques of Republican governments' reactions to the pandemic. "*TN has 1156 new confirmed COVID-19 cases in one day and still plans to re open :(*" from Dr. Hana Ali, a county Democratic Party chair in Tennessee.
- **Cluster 2: Moderate Right-leaning accounts.** Cluster 2 is comprised of 2,442 unique users. They have 6,239 followers on average ($SD = 66,309$). 0.2% ($n = 7$) of these users have verified accounts, many of which are journalists for right-leaning news sources. This cluster is the most diverse in terms of nationality, including primarily US accounts, along with India, Canada, Europeans, and New Zealanders. However, some moderate and left-leaning journalism is also present in the accounts in this cluster. One of the most favorited Tweets from these users include "*Children should all be back in school nationwide—we have done a total disservice to children and their parents*" from Laura Ingraham at Fox News. However, Tweets from CNN and moderate US news sources were also popular.
- **Cluster 3: Alt-Right-leaning accounts.** Cluster 3 is comprised of 2,120 unique users. They have 6,753 followers on average ($SD = 15,672$). 0.4% ($n = 9$) of these users have verified accounts, all of which are journalists or influencers for right-leaning news sources, including the Washington Examiner (@dcexaminer) and Tim Young (@TimRunsHisMouth), a Fox News contributor and personality on SiriusXM Patriot, which amplifies Breitbart News. These accounts tend to be characterized by critiques of democratic and socialist governments and left-leaning news sources, and support of Donald Trump's presidency. They also contribute to questioning of verified news sources. One such Tweet from @TimRunsHisMouth reads: "*HOLY S**T: Did I read this wrong or did the CDC just revised* [sic] *the national COVID-19 deaths to 37,308?!?!*" This was followed with "*They've now segmented the deaths to show that only 37,308 are directly from COVID-19.*"
- **Cluster 4: Moderate and Conservative Journalism from India.** Cluster 4 has 1,579 unique users. Each user has 8926 followers on average ($SD = 155,361$. 2% ($n = 24$) of accounts are verified. These accounts include the Indian Chamber of Commerce and Industry, the Hindustan Times, Andaman & Nicobar Police, and various reporters and politicians in India. Topics include coronavirus and how to identify fake news regarding coronavirus, such as this Tweet from the Hindustan Times: "*Remdesivir gets US emergency approval for treating Covid-19 patients: Know all about the antiviral drug...*"
- **Cluster 5: Left-leaning Journalism from India** Cluster 5 has 1,290 unique users. Each user has 169,318 followers on average ($SD = 924,489$). 11% ($n = 78$) of accounts are verified. Accounts include CNBC India, feminist Indian influencers, and various news outlets. Indian Twitter account @SheThePeople Tweeted, "*Delhi High Court has directed the central and the state government of Delhi to ensure women from COVID-19 hotspots, needing maternity care, be given immediate attention during the lockdown. Reports @PoorviGupta08 Coronavirusoutbreak*".

As seen in Figure 1, there is relatively little liking between the left and right-leaning clusters (top three clusters). However, the two right-leaning clusters often like each other's Tweets. The US accounts do not often like the Indian accounts' Tweets, but the Indian accounts do often like each other's Tweets.

5.2 Twitter: Sentiment and Discourse Analysis Between and Within Groups

We used the sentiment polarity scores present in the initial Twitter dataset [10] to describe the language styles of five largest clusters and the between group communication. Fig. 2 shows the distribution of positive, negative and neutral scores. As can be seen, neutral Tweets prevail within each cluster, and the distribution of positive, negative and neutral Tweets is fairly uniform across

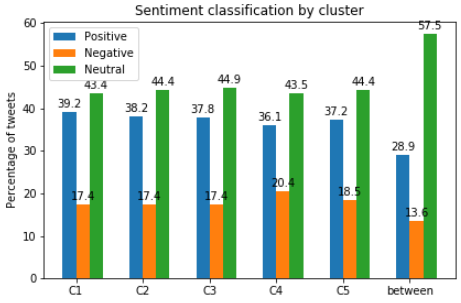


Figure 2: Sentiment classification of Tweets in the top five clusters from Twitter data.

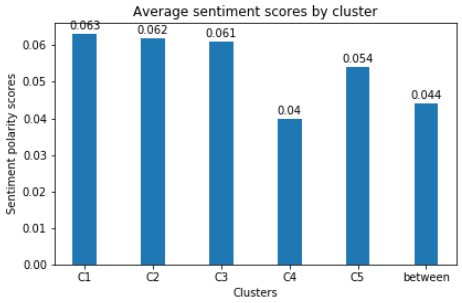


Figure 3: Average sentiment scores in the top five clusters from Twitter data.

clusters. "Between" represents conversations occurring between the users from different clusters, including those outside the five clusters.

The average sentiment polarity scores for each cluster and between groups are given in Fig. 3. Sentiment polarity scores are in the (-1,1) range and the commonly used thresholds are the following: Tweets with scores below -0.05 are treated as negative, scores above 0.05 are positive indicators, and the rest are neutral. Clusters 1, 2, and 3 have an average positive polarity, cluster 5 is close to neutral, and cluster 4 and between clusters have average negative sentiment. This provides cautious evidence that shared values may correlate with more positive communication than in the absence of shared values, however, it is inconclusive.

To study interaction patterns in the five major clusters, we used the second, larger Twitter dataset, from which we extracted only the chains of conversations of at least two Tweets between the users in 5 clusters and manually annotated for discourse acts, using the set of labels from Zhang et al. [22] (with an addition of "opinion" label). The number of conversation chains between the users of different clusters is insignificant for analysis: only sixteen exchanges were found across pairs of clusters. With such small number of interactions no conclusions can be made with statistical confidence. Within-cluster chains of communications were more prevalent, but still rare. From the entire dataset only 255 conversations occurred within clusters, with the majority of them being just two Tweets. These conversations were manually annotated for discourse acts. The results are summarised in Figure 1. Label "other" includes question, answer, appreciation, humor discourse types. This analysis reveals the type of communication in each cluster: clusters 1 and 4 are majorly fact-reporting and include majorly journalism and retweeted journalism, while cluster 2, 3 and 5 introduce more opinions and reactions, mostly negative.

% of Tweets	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
announcement	**42.7**	**33.3**	3.12	30.4	42.9
elaboration	**45.3**	**25.9**	**25**	**39.1**	0.0
negative reaction	5.1	**22.2**	**25**	14.5	**38.1**
agreement	2.6	0.0	6.3	1.4	0.0
disagreement	1.7	7.4	0.0	0.0	4.8
opinion	1.7	7.4	6.3	10.1	9.5
other	2.5	3.7	6.2	4.3	4.8

Table 1: Discourse annotation of Tweets in 5 major clusters

Further, we analyzed the differences in language via pairwise comparisons of the clusters using FightingWords Transformers from the ConvoKit library [6]. This method compares the languages of two groups based on the usage rates of n-grams with a smoothing Dirichlet prior showing how much certain words were used by one side more than the other as described in Monroe et al. [13]. This method is reported to work well on the datasets with disproportionate samples and accounts for different word frequency in common language [13]. We performed the analysis of bigrams and unigrams with default uniform prior, setting a different prior and more sophisticated preprocessing may reveal more distinct trends of the language use. See Tables 2 and 3 for comparisons.

C1: Left-Leaning US accounts	C2: Moderate Right-Leaning, Primarily US accounts
'covid19'	'indian'
'total deaths'	'indiafightscorona'
'total confirmed'	'rt pti'
'new cases'	'medical professionals'
'new deaths'	'aircraft'
'odisha'	'air force'
'coronavirus'	'police'
'live updates'	'health department'
'reports'	'express gratitude'
'confirmed'	'positive cases'

Table 2: FightingWords: Top 10 terms for Cluster 1 and Cluster 2

C2: Moderate Right-Leaning, Primarily US accounts	C3: Alt-Right Leaning US accounts
'indiafightscorona'	'black'
'health department'	'you'
'the independent'	'killing'
'secpompeo'	'victims'
'natashabertrand'	'white'
'bcpoli'	'qasimrashid'
'uspoli'	'the poor'
'positive cases'	'potus'
'fbi'	'coronadon'
'police'	'don'

Table 3: FightingWords: Top 10 terms for Cluster 2 and Cluster 3

Cluster 1 is characterized by reporting the factual information about the pandemic. Notably, Cluster 2 (Moderate-Right) compares differently to Cluster 1, emphasizing Indian issues, health concerns and expression of gratitude towards medical personnel, and Cluster 3 (Alt-Right US accounts), introducing more concerns with security and health (e.g. mentions of FBI, Secretary of State Mike Pompeo, and National Security correspondent Natasha Bertrand). The discussion in cluster 3 is obviously politicized and differs from cluster 2 in their use of words associated with hate speech[3]. Clusters 4 and 5 (Indian accounts) share common language (the FigtingWords algorithm returned only 5 words from the order of 10 for cluster 4). The prevailing terms in Cluster 4 are @realdonaldtrump, enforcement, and incompetence. Cluster 5 is characterized by mentions of news resources such as @xhnews, @cnn, @ndtv, and @reuters. The findings of language analysis aligns with the characteristics of users presented in Section 5.1.

5.3 Reddit: Sentiment and Discourse Analysis Between and Within Groups

A conditional random field (CRF) model was trained on the annotated dataset from [22]. The results of training on the data can be seen in Table 4. Of particular note is the difficulty it has with negative reactions, disagreement and humor, each with an F1 score lower than 0.4. Humor can be hard to detect even by humans, so this result is understandable; Zhang et al. [22] also reported low F1 scores in humor relative to other discourse types. However, the low scores for negative reactions and disagreement are concerning, given that this is the focus when evaluating argumentation. However, the precision scores for all categories are relatively high (each greater than 0.5), meaning that there are fewer false positives than false negatives.

The model was then used to predict labels for the comments pulled from `/r/changemyview` and `/r/PoliticalDiscussion`. The proportions of each label can be seen in Table 5, with the ranges based on probabilities outputted by the model. Overall, the proportions of types of acts is relatively similar across the two subreddits, which makes sense given that they are meant to foster discussion. This is especially seen in the possible high proportion (up to ≈45%) of elaboration comments. The main observation that seems important is the ratio of questions to answers. In `/r/changemyview`, the proportion of questions (up to ≈ 15%) is close to that of answers (up to ≈ 17%). However, in

[3]https://www.adl.org/resources/reports/the-online-hate-index

	precision	recall	f1-score	support
negativereaction	0.6538	0.0914	0.1603	2233
agreement	0.7087	0.4072	0.5172	5504
disagreement	0.6996	0.2618	0.3810	4225
elaboration	0.5384	0.7792	0.6368	22225
announcement	0.9373	0.9452	0.9412	14005
answer	0.8443	0.9510	0.8945	47760
appreciation	0.7725	0.6542	0.7084	9461
other	0.5596	0.3320	0.4168	18433
question	0.9260	0.9579	0.9417	58005
humor	0.5539	0.1300	0.2106	2922
weighted avg	0.7939	0.7998	0.7820	184773

Table 4: Overview of model performance on types of discourse in Reddit data.

% of comments	CMV	Political Disc.
negativereaction	0.060-0.472	0.000-0.105
agreement	1.339-3.818	1.050-3.520
disagreement	4.213-13.266	2.481-10.367
elaboration	8.701-46.156	7.195-47.010
announcement	1.141-1.587	0.033-0.039
question	**9.241-14.982**	**6.283-10.693**
answer	**7.826-16.990**	**20.220-27.780**
appreciation	0.815-1.570	0.420-0.685
other	0.721-14.347	0.713-6.957
humor	0.000-0.017	0

Table 5: Probabilities of discourse types present in `/r/changemyview` and `/r/PoliticalDiscussion`.

`/r/PoliticalDiscussion`, there is a much larger proportion of answers, compared to questions. (up to $\approx 28\%$ and up to $\approx 11\%$, respectively).

Next, the proportions of discourse acts were reanalyzed after filtering to only comments posted in reply to a disagreement. The results can be seen in Table 6. `/r/PoliticalDiscussion` still maintains its higher proportion of answers, but only by $\approx$1%, so this is likely not significant. There is also slightly more elaboration in `/r/PoliticalDiscussion`, possibly indicating that users try to drive their point home more clearly after encountering a difficulty. Otherwise, there is a slightly higher proportion of negative reactions and disagreements in `/r/PoliticalDiscussion`, which could reflect lower community standards and/or moderation. Of note across both subreddits is the high occurrence, more than half the time, of chained disagreements, where a disagreement is followed by another disagreement.

Finally, to see whether user similarity had an effect on interactions, we looked at the Jaccard similarity of subreddits that users have commented in. Only pairs of users that had a positive interaction (appreciation or agreement) or negative interaction (disagreement or negative reaction)

% of comments after a disagreement	CMV	Political Disc.
unresolved	12.437	8.968
negativereaction	0.000-0.101	0.000-0.377
agreement	0.708-2.629	0.678-1.884
disagreement	**55.106-68.756**	**53.580-73.775**
elaboration	2.224-12.538	1.356-17.483
answer	0.000-0.101	0.226
appreciation	0.404-0.708	0.000-0.151
other	1.011-3.134	0.075-1.583
question	3.438	3.693-4.823

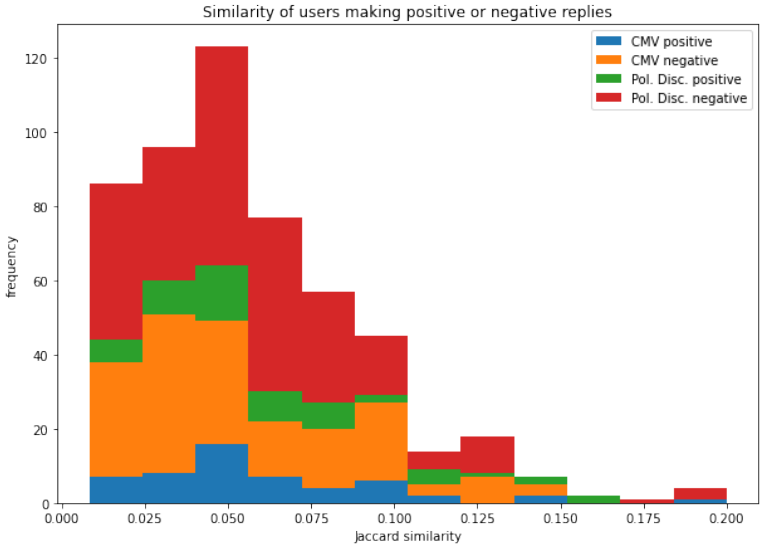
Table 6: Types of discourse in reply to a disagreement. Unresolved indicates that no further discussion is made.

	CMV	Political Disc.
Positive interaction	0.060	0.064
Negative interaction	0.055	0.057

Table 7: Jaccard similarity of users based on subreddits they have been active in.

were considered. From this, the data was filtered to only pairs of users that both had at least 50 comments. The results can be seen in Table 7.

A histogram showing the distribution is shown below. There are two possible conclusions to draw from this, though it is possible they are not statistically significant. First, the users in `/r/PoliticalDiscussion` are more similar than those in `/r/changemyview`. This can be seen by the higher similarity in the second column, compared to the first. This would negate our hypothesis used to describe and analyze the two groups. Second, negative interactions happen more between less similar users. This can be seen by comparing row values. However, to draw conclusions on both, it would be necessary to run statistical analysis.



6 Discussion

6.1 Prevalence of Echo Chambers Impeding Restorative Justice Research

Our Twitter analysis revealed that participants far more often like, retweet, and reply to users who share their values as compared to those who do not. The sparsity of data available for analysis between different clusters of users was remarkable: there were only 16 strings of comments in which users of different clusters communicated with each other. Colleoni et al. [3] have shown that political homophily is very common online (and specifically on Twitter), and our results confirm this.

These echo chambers also bring the relevance of our original research question - restorative justice online - into sharp focus. Perhaps large, public social media networks are not the ideal environment for investigations of restorative justice, as prior work has shown that Twitter as a platform typically leads to polarization [21]. Or, perhaps political discussions are not conducive to diverse perspectives, whereas a different topic such as entertainment or sports may yield more disagreement. Future work could explore arguments and ruptures on other platforms, topics, or simply incorporate more data.

6.2 Disagreements Lead to Disagreement

The most prevalent difference between discussions on `/r/changemyview` and `/r/PoliticalDiscussion` was that `/r/PoliticalDiscussion` had approximately three times more answers per question. In contrast, `/r/changemyview` had roughly similar ratios of questions and answers. `/r/PoliticalDiscussion` was our proxy for lack of shared identity, and it is curious that there would be so much more engagement regarding answering questions

than on /r/changemyview. It raises the question that user motivations for engaging in the /r/PoliticalDiscussion may drastically differ from /r/changemyview. A variety of social and cognitive factors have been found to significantly influence engagement and satisfaction with online communities [9], and these may change over time as users interact with the community. Thus, it could be interesting to couple large-scale quantitative data with survey data or qualitative data on why users began engaging with the community, what value they currently get from the community, and what motivates them to continue engagement with the community in the future. This may help us understand why these discourse patterns exist across the different communities.

Perhaps one of the most interesting results from the Reddit data is that disagreements tend to lead to more disagreement (in approximately 60% of the follow-up from an initial disagreement, there was more disagreement). There was not a large difference across the /r/changemyview and /r/PoliticalDiscussion subreddits in terms of prevalence of follow-up disagreement, however, it would be interesting to further evaluate language style and sentiment in these two scenarios. For instance, it could be possible that in /r/changemyview, users employed more hedging language and appeals to shared values while disagreeing, or used more neutral or positive sentiment. Understanding if differences exist in how these two communities disagree requires further research, and may yet elucidate more information on our hypothesis of restorative justice and shared identity.

7 Contributions

Elena gathered the Twitter data, hydrated the Tweets, and captured likes and replies to subsets of Tweets using alternate methods. Elena wrote most of the related works on argumentation, did the analysis on differences in discourse content and structure within and across Twitter groups. Erik was the main contributor for analyzing the Reddit data, using it to build the CRF model and apply it in various ways, looking at discourse chains and user similarities. Amanda introduced the topic and structured the project, pulled the initial Reddit data, did the identity clustering for Twitter, and qualitatively analyzed clusters. All group members contributed to the writing and presentation of the final project and contributed equally to the project.

References

[1] Social media fact sheet, 2019. URL https://www.pewresearch.org/internet/fact-sheet/social-media/.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008.

[3] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.

[4] D. Ghosh, S. Muresan, N. Wacholder, M. Aakhus, and M. Mitsui. Analyzing argumentative discourse units in online interactions. 06 2014. doi: 10.3115/v1/W14-2106.

[5] B. Group. Why people fight online, 2017. URL https://www.barna.com/research/people-fight-online/.

[6] L. F. A. W. J. Z. C. D.-N.-M. J. P. Chang, C. Chiam. Convokit: The Cornell Conversational Analysis toolkit, 2019. Retrieved from http://convokit.cornell.edu.

[7] C. D.-N.-M. Jonathan P. Chang. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of EMNLP*, 2019.

[8] B. Konat, J. Lawrence, J. Park, K. Budzynska, and C. Reed. A corpus of argument networks: Using graph properties to analyse divisive issues. 01 2016.

[9] C. Lampe, R. Wash, A. Velasquez, and E. Ozkaya. Motivations to participate in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1927–1936, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589299. doi: 10.1145/1753326.1753616. URL https://doi.org/10.1145/1753326.1753616.

[10] R. Lamsal. Corona virus (covid-19) tweets dataset, 2020. URL http://dx.doi.org/10.21227/781w-ef42.

[11] D. Levin, J. Walter, and J. Murnighan. Dormant ties: The value of reconnecting. *Organization Science*, 22:923–939, 08 2011. doi: 10.2307/20868904.

[12] M. Madianou and D. Miller. Polymedia: Towards a new theory of digital media in interpersonal communication. *International Journal of Cultural Studies*, 16(2):169–187, 2013. doi: 10.1177/1367877912452486.

[13] B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

[14] Z. Papacharissi and A. Mendelson. Toward a new(er) sociability: Uses, gratifications and social capital on facebook. *Media Perspectives for the 21st Century*, pages 212–230, 01 2011. doi: 10.4324/9780203834077.

[15] C. Sibona. Unfriending on facebook: Context collapse and unfriending behaviors. In *Proceedings of the 2014 47th Hawaii International Conference on System Sciences*, HICSS '14, page 1676–1685, USA, 2014. IEEE Computer Society. ISBN 9781479925049. doi: 10.1109/HICSS.2014.214. URL https://doi.org/10.1109/HICSS.2014.214.

[16] L. G. Stewart, A. Arif, A. C. Nied, E. S. Spiro, and K. Starbird. Drawing the lines of contention: Networked frame contests within blacklivesmatter discourse. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), Dec. 2017. doi: 10.1145/3134920. URL https://doi.org/10.1145/3134920.

[17] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883081. URL https://doi.org/10.1145/2872427.2883081.

[18] R. Trapp and N. Hoff. A model of serial argument in interpersonal relationships. *The Journal of the American Forensic Association*, 22(1):1–11, 1985. doi: 10.1080/00028533.1985.11951297.

[19] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. "i regretted the minute i pressed share": A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450309110. doi: 10.1145/2078827.2078841. URL https://doi.org/10.1145/2078827.2078841.

[20] M. Wenzel, T. G. Okimoto, N. T. Feather, and M. J. Platow. Retributive and restorative justice. *Law and human behavior*, 32(5):375–389, 2008.

[21] S. Yardi and D. Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327, 2010. doi: 10.1177/0270467610380011. URL https://doi.org/10.1177/0270467610380011.

[22] A. Zhang, B. Culbertson, and P. Paritosh. Characterizing online discussion using coarse discourse sequences. 2017.

[23] J. Zhang, C. Danescu-Niculescu-Mizil, C. Sauper, and S. J. Taylor. Characterizing online public discussions through patterns of participant interactions. *Proc. ACM Hum.-Comput. Interact.*, 2 (CSCW), Nov. 2018. doi: 10.1145/3274467. URL https://doi.org/10.1145/3274467.