Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

Analyzing the Traits and Anomalies of Political Discussions on Reddit

Anna Guimarães, Oana Balalau, Erisa Terolli, Gerhard Weikum
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
{aguimara, obalalau, eterolli, weikum}@mpi-inf.mpg.de

Abstract

Online communities like Reddit feature intensive discussions on political topics, engaging many users who give and receive votes for their posts and comments. Prior work has focused on detecting controversial discussions and analyzing the role of troll-like users. This paper provides a systematic analysis of the characteristics of user actions, like replies and votes, in discussion threads, identifying different patterns that refine the notion of controversies into disputes, disruptions and discrepancies. Most importantly, our study connects the dimension of user actions with a rich feature space comprising also the sentiments expressed in post contents and the topical variation across the posts and replies. We use this framework to gain insight on the traits of different archetypes of discussions, and to statistically test a variety of hypotheses on expected and abnormal behavior.

Introduction

Motivation: Discussions in online communities, such as Reddit, Quora etc., reveal people's opinions on many topics of societal importance. Moreover, it is often insightful to analyze the structure and dynamics of the discussion threads themselves. In this paper, we focus on Reddit-style discussions of political news. These include *harmonious* discussions where users agree on a certain stance (e.g., grief and anger about a school shooting), but also a large amount of *controversial* discussions with users strongly disagreeing (e.g., consequences regarding gun control). An interesting research objective in this setting is to identify such controversies and understand the role of individual posts in setting their tone and direction.

However, online discussions are more than this dichotomy of harmonies and controversies. In this paper, we take a broader and deeper look into different patterns of discussion. We propose four pattern groups to represent frequent and interesting conversational archetypes: Harmony, Discrepancy, Disruption, and Dispute.

Some discussions may lack any disturbances, constituting a *Harmony*, while others contain only isolated instances of disagreements, which stand out as *Discrepancies*. A *Disruption* may occur when the sentiment in the discussion shifts, or when there is an abrupt change in the topic. For example, consider the Reddit thread[1] about a news article from January 2016 on "Noam Chomsky on Clinton vs Sanders". At a certain point, a user chimes in with strongly negative opinions on Chomsky and steers the discussion away from the original topic into a more polemic debate on the person Noam Chomsky. Finally, *Disputes* represent conversations where users repeatedly disagree in their opinions about a particular topic, for example, when speculating about the winner of an election[2].

Understanding and characterizing these discussion patterns requires an analysis that goes beyond the level of user *actions* (posts, replies, votes) and also considers *topics* and *sentiments* jointly with the dimension of user actions.

Prior Work and its Limitations: There is abundant work on analyzing social media with regard to mining sentiments on specific topics (e.g., (Liu 2012)), predicting the popularity of individual posts (e.g., (Aggarwal 2011; Zhao et al. 2015)), identifying influential users (e.g., (Al-garadi et al. 2018)), and detecting abnormal or malicious behavior in terms of content (spam, fake, etc.) and users (trolls etc.) (e.g., (Jiang, Cui, and Faloutsos 2016; Cheng et al. 2017)). Much of this work has focused on Twitter as the underlying forum. Research on political discussions has largely focused on specialized topics such as migrant assimilation, and on adversarial debates between two parties, like election campaigns (e.g., (Rizoiu et al. 2018)).

Most related to this paper is the prior work on detecting controversial discussions and analyzing them. Recent studies by different groups devised pattern-based characterizations of discussion threads (Coletto et al. 2017; Garimella et al. 2018; Glenski and Weninger 2017; Zhang et al. 2018) or use post features, including controversiality, to predict post popularity(Zayats and Ostendorf 2018). However, as far as we know, this is the first study that characterizes Reddit discussions considering multiple meaningful facets of a conver-

[1] www.reddit.com/r/politics/comments/43m8lq/chomsky_on_sanders_vs_clinton/ czj9lwo/

[2] www.reddit.com/r/politics/comments/3yzek4/bernie_sanders/cyieq2l/

sation: users, sentiments, and topics.

Approach and Hypotheses: The unique element in our approach to understanding political discussions in broad online communities is to consider three dimensions jointly:

i) user *actions* like posts and votes,

ii) the *sentiments* expressed in post contents, relative to preceding posts and the root of the conversation,

iii) the variation of *topics* across posts.

To the best of our knowledge, prior work has not addressed all of these aspects in a joint manner. Our analysis is not specialized for specific themes like election campaigns, but covers a wide spectrum of political topics.

We approach this space by first identifying salient patterns expressed in user actions, most importantly, by positive or negative votes for posts in a discussion. Based on this action-centric model, we propose the conversational archetypes of Harmony, Discrepancy, Disruption, and Dispute. Each of these archetypes is then analyzed on its sentiments and topical variation based on the contents of posts.

We formulate hypotheses about each of the discussion archetypes and their characteristics, and use statistical tests to retain or refute hypotheses based on a large and thematically broad corpus of discussions from two prominent subreddits (www.reddit.com/r/politics, www.reddit.com/r/worldnews).

Key questions and hypotheses that we aim to gain insight on are the following:

- Are Harmonies representative of positive and on-topic conversations?
- Do Discrepancies occur when a single post expresses a negative sentiment or is off-topic?
- Is a Disruption a case of a sudden shift in topic or sentiment?
- Are Disputes predominantly negative in sentiment?

Contributions: The paper's salient contributions are:

- We introduce a pattern-based model of different archetypes of online discussions, refining the established notion of controversy into dispute, disruption, and discrepancy.
- We present the first study of these archetypes by jointly looking into user actions, post sentiments, and topical variations across posts.
- We report findings about the nature of controversial discussions and their refined facets.
- We statistically test a suite of hypotheses on a large and thematically broad corpus of Reddit discussions.

Related Work

Discussion threads. Much prior work on online discussions has aimed to predict the popularity of the discussion itself, via the number of comments or users it attracts, or of its underlying posts, via the ratings (scores, votes, likes) they receive. Thread popularity is often addressed under generative models for online discussions, which model the arrival of new replies based on the number of existing replies, novelty, and bias towards the initiators of the discussion (Gómez et al. 2013), structural properties of the comment tree (Nishi et al. 2016), or reciprocity between users (Aragón, Gómez, and Kaltenbrunner 2017).

(Liang 2017) studied the relationship between post scores, participating users, and thread structure in the Q&A subreddit, TechSupport. (Zayats and Ostendorf 2018) tackled comment score prediction on Reddit by modeling each post in a comment tree as a recurrent neural network, which learns features about the post content, local context, timing, and structural properties. (Glenski and Weninger 2017) monitored the browsing behavior of Reddit users to predict future interactions based on users' voting habits and page-browsing activities.

(Zhang et al. 2018) studied reply-trees on Facebook in combination with user-user interactions. The authors derived features to describe discussion evolution, including a summary of degree distributions, edge properties, and graph motifs. These features are then used to predict the growth of the discussion, and whether it will exhibit abnormal behavior that lead to participant blocking. Post content was not considered in this work at all.

(Zhang, Culbertson, and Paritosh 2017) developed a taxonomy of discourse acts in online discussions, proposing 9 categories, such as "agreement" or "answer", based on randomly sampled Reddit threads and crowdsourced annotation. This study noted patterns of disagreement chains, particularly in debate-oriented forums such as PoliticalDiscussion, but not so in the Politics subreddit.

(Weninger, Zhu, and Han 2013) studied the progression of topics in Reddit threads based on a hierarchical latent model.

Controversy and antisocial behavior. A prominent aspect of online social discussions is the presence of controversial topics and antisocial (troll-like) users.

(Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015) characterized antisocial behavior by studying the history of banned accounts in the comment section of three news sites. The resulting features were used to predict whether a user will likely be banned in the future. Subsequent work (Cheng et al. 2017) also investigated trigger mechanisms for antisocial behavior, or trolling.

Controversial topics are studied by (Coletto et al. 2017) as graph motifs in the network of user interactions on Twitter. Frequent motifs are coupled with structural, temporal, and propagation-based features from the graph in order to identify controversies. However, this work did not consider the contents of user posts.

(Garimella et al. 2018) also leveraged the network structure surrounding specific hashtags to quantify the degree of controversy for a given hashtag.

(Rizoiu et al. 2018) studied the influence of social bots in the diffusion of tweets containing partisan hashtags surrounding a political debate. (Vilares and He 2017) proposed a method for political stance classification with a hierarchical Bayesian model, where topics and stances are latent variables.

Data Modeling

A discussion starts on Reddit when a user posts an initial piece of content, such as a news article or a video, called a submission. Users comment on the submission, while also receiving replies of their own, and as users respond back and forth to each other, the discussion grows in a tree-like manner.

Submissions and posted comments alike may receive feedback in the form of upvotes and downvotes from users, which are combined to give a total post score. While voting behavior and the reasons for upvoting or downvoting a post are varied[3], we interpret scores as a measure of the community reaction to a post. Allowing for some noise, a post with a positive score can be seen as having been more well-received than a post with a negative score.

On Reddit, only the final scores resulting from the difference between upvotes and downvotes are displayed, and the total number of votes a post has received is hidden. Thus, posts that have been heavily downvoted may still have positive overall scores. In order to identify these posts, Reddit provides a "controversial post" flag. Posts which, in turn, have received significant negative feedback and have negative overall scores can become hidden in the discussion once their score falls below a certain threshold[4].

In this work, we denote these posts which have received a negative or mixed reaction from the community as **X-posts**. We consider a post as an X-post if it has been flagged as controversial or if it has a score equal to or below -4.

At the level of entire discussions, the presence of X-posts gives rise to several kinds of observable patterns. Our model considers these discussion archetypes by proposing four distinct groups: Harmony, Discrepancy, Disruption, and Dispute.

Definitions

We abstract the political discussions on Reddit into the following general concepts:

- A discussion is initiated by a **submission**, consisting of a piece of media or text, which attracts comments from users. These initial comments are called **top-level comments**.
- Comments may also receive comments, or replies, of their own. These chains of replies thus form **post trees**, where the root is a top-level comment made in response to the submission. When referring to these trees, we do not distinguish between top-level comments and replies, and simply refer to all user-provided content as **posts**.
- We consider all **paths** in a post tree rooted at a top-level comment and ending at a leaf node. Each path is a sequence of posts, where each post is a direct reply to its immediate predecessor. Note that in this model, paths might differ only in a suffix of nodes, by sharing a common prefix before the post tree branched out.
- Each post receives a number of **upvotes** and **downvotes** by the user community, and is then associated with an integer-valued **score**, which is a function of upvotes and downvotes. Our model assumes that $score = \#upvotes - \#downvotes$.
- Individual posts may be explicitly flagged as "controversial" (in Reddit jargon) when they have a substantial amount of votes and a roughly equal share of upvotes and downvotes. Posts are also subject to a visibility threshold and become hidden when they receive a sufficiently low score (≤ -4 as default). We denote both these hidden posts and "controversial" posts as **X-posts** to avoid the a priori connotation with semantic notions of disagreement and controversy. All other posts are called **normal posts**.
- Posts are further associated with **topics** and **sentiments**, which are expressed in the post's textual content.

Based on the dichotomy of X-posts vs. normal posts, we additionally define a path containing at least 5 posts to be labeled as:

- **Harmony**: a path where all posts are normal.
- **Discrepancy**: a path containing exactly one X-post.
- **Disruption**: a path that consists of two contiguous sequences: a sequence containing two or more normal posts and a sequence containing two or more X-posts, where the order of the two sequences is irrelevant.
- **Dispute**: a path where normal posts alternate with X-posts.
- **Others**: a path that does not follow any of the above patterns.

The intuition for this categorization is as follows. Harmonies represent general agreements, without any major disturbances. Discrepancies exhibit outlier behavior by one user but are otherwise harmonious conversations. Disruptions are discussions which abruptly shift, being composed of two opposing conversations, a harmonious one and a highly contentious one. Disputes would represent controversial discussions where users disagree.

Dataset

Our first dataset comes from the Politics subreddit[5], a forum for "current and explicitly political U.S. news." In an effort to promote serious discussions, the forum's guidelines ask that submissions be external links to recent political news articles, videos, and sound clips from reputable pre-approved sources, which include media outlets, polling and research centers, and government bodies[6]. This differs from many other subreddits, which also allow free-form text, questions, and images to be submitted.

We complement this dataset with posts from the WorldNews subreddit[7], where submission guidelines are similar to those in the Politics subreddit (external links to recent news articles), but specifically excludes US-related news.

[3] blog.disqus.com/here-are-the-reasons-why-people-downvote-comments

[4] www.reddit.com/r/AskReddit/comments/uxq79/what_does_comment_score

[5] www.reddit.com/r/politics/

[6] www.reddit.com/r/politics/wiki/index#wiki_submission_rules

[7] www.reddit.com/r/worldnews/

We collected all submissions and available comments posted to these communities in 2016 via the platform's API (accessed in February 2018), as well as the original news articles the submissions were referencing. We then discarded submissions linked to (currently) inaccessible articles and submissions which received fewer than 5 posts. An overview of our dataset is given in Table 1. This data is available at http://people.mpi-inf.mpg.de/∼aguimara/trait anomalies.

As comments and users may be removed from the discussion over time, some sequences of posts may have gaps. In these cases, we link the orphaned comment to its closest predecessor in the post tree.

Source	#Submissions	#Posts	#Users	#Paths
Politics	34,786	3,571,752	189,711	971,241
WorldNews	24,278	3,727,955	352,055	1,260,515

Table 1: Politics and WorldNews subreddit datasets.

Post Dimensions

In this section, we examine posts in terms of how they appear in the discussion, the sentiments they express, and their topical content. First, we revisit our notion of X-posts, which serve as the building block for the conversational patterns we later investigate. Then, we provide an overview of the sentiments and topical cohesiveness of posts in our dataset, which we later relate to each of our proposed pattern groups. Lastly, we derive the notion of X-users from our definition of X-posts and from observations about users' posting behavior.

X-Posts and Normal Posts

We introduced the notion of X-posts on Reddit. These posts stand out for having attracted a notable amount of negative attention from the community, manifest in terms of downvotes. In total, 13% and 12.3% of all posts are X-posts in the Politics and WorldNews subreddit, respectively.

While X-posts and normal posts differ principally in terms of their scores, with X-posts having lower overall scores due to the greater amount of downvotes they have accumulated, they differ also in the level of activity they generate. When comparing the number of replies received by each post, we find that X-posts get significantly more replies ($M = 1.78, SD = 1.69$ for Politics and $M = 1.87, SD = 2.14$ for WorldNews) than normal posts ($M = 1.11, SD = 2.10$ and $M = 1.13, SD = 3.09$)[8], ($p < 0.001$).

We also find that X-posts and normal posts can both be "controversial" with regards to their mentions of controversial issues. For this, we compiled a list of phrases related to controversial issues from Wikipedia[9], which contains "articles deemed controversial because they are constantly being re-edited in a circular manner, or are otherwise the focus of edit warring or article sanctions." From this list, we removed several categories, such as People, Languages and Philosophy, and we considered the titles of articles (or shortened versions) to be controversial phrases.

[8] M and SD denote the empirical mean and standard deviation, respectively.

[9] en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

On average, X-posts on the Politics subreddit contain more controversial terms ($M = 0.006$) than normal posts ($M = 0.005$), but only slightly so ($p < 0.001$). The opposite is true for WorldNews ($p < 0.001$), where X-posts feature fewer controversial terms ($M = 0.009$) than normal posts ($M = 0.012$). The most frequent terms in both types of posts are *women, crime, cult, god, rape, NATO, prison, racism, islam, drug*, several of which are often at the center of political and world-wide news. We leave it to further work to investigate if certain phrases in our list are more controversial in the context of discussions on political forums.

Sentiments

As a measure of the sentiments expressed throughout discussions, we evaluate the language used in each post in our datasets using VADER (Gilbert 2014). VADER is a human-validated sentiment analysis method created from a gold-standard sentiment lexicon, specialized for social media text. For each post, VADER assigns a sentiment intensity score from -1 to 1 and a sentiment polarity: posts with intensity scores in the range $[-1, -0.05)$ have negative polarity, posts in the range $[-0.05, 0.05]$ have neutral polarity, and between $(0.05, 1]$ positive polarity. Although this tool does not distinguish between opinions in text (i.e., positive or negative sentiment *towards* a topic), it still allows us to compare the use of positive and negative language and detect posts which differ from others in a conversation.

While we observe a similar proportion of X-posts and normal posts in both our datasets, there are differences in the distribution of sentiment polarities across the two subreddits. On Politics, we find a majority of positive posts (43.1%), followed by negative posts (38.3%) and a smaller amount of neutral posts (18.4%). Meanwhile, negative posts make up the majority on the WorldNews subreddit (38.2%), followed by positive (34.8%) and a significant amount of neutral posts (26.8%). These numbers indicate that discussions on the Politics subreddit tend to be more polarized, with relatively fewer neutral posts. In terms of the intensity of the sentiments being expressed, neither community tends toward extreme polarization, and sentiment scores are uniformly distributed.

Posts of different sentiment polarities do differ in terms of the attention they generate. Negative posts in both subreddits receive more replies on average ($M = 1.26, SD = 2.17$ for Politics, $M = 1.34, SD = 3.24$ for WorldNews) than positive ($M = 1.20, SD = 2.11$ and $M = 1.22, SD = 3.11$) or neutral ($M = 1.07, SD = 1.70$ and $M = 1.04, SD = 2.42$) posts ($p < 0.001$). These numbers may be explained by the nature of posts expressing a negative sentiment, which are likely to include hostile or inflammatory remarks designed to provoke a response from other users.

In addition, when examining sentiments at the path level, we find that the sentiment of the post at the root of a path (i.e., the top-level comment) tends to influence the sentiment of subsequent posts. On the Politics subreddit, the predominant sentiment polarity of a path matches the sentiment polarity of the root post in 71% of paths, and the same can be observed in 56% of paths in the WorldNews subreddit.

Topics

In order to evaluate the topic cohesiveness of a path, we consider both the topic similarity between posts and similarity of posts with the news article being discussed (i.e., the submission).

We transform posts and news articles to document embeddings using Doc2Vec (Chen 2017), an unsupervised method that learns fixed-length feature representations of words and documents. To capture language peculiarities of each community, we learn sentence representations from 5 years of Reddit text data, compiled from posts made to the Politics and WorldNews subreddits between 2012 and 2016.

To evaluate the similarity between two pieces of text, either two posts or a post and a news article, we consider the fact that users might respond to only a subset of the ideas stated previously, for example:

Person A: This 'article' smells of satire, but I could be wrong. Where do you guys find this stuff? The coin toss is for county delegates not state delegates. Its not a big deal.

Person B: what are county delegates?

To capture such situations, we consider the topical similarity of two posts p_i and p_j, $\textbf{sim}(\boldsymbol{p_i}, \boldsymbol{p_j})$ to be the maximum cosine similarity[10] of the embeddings of all text spans with consecutive sentences within p_i against the embeddings of p_j. We proceed in the same way when calculating the similarity between posts and news articles, $\textbf{sim}(\boldsymbol{news}, \boldsymbol{p_i})$.

Analogous to what we found when examining X-posts and normal posts, as well as posts of different polarities, there is also a difference in how "on-topic" and "off-topic" posts affect the activity in discussions. On average, posts which are highly similar to the news articles (with similarity scores above the $75th$ percentile) receive 50% more replies ($M = 1.44, SD = 2.64$ for Politics and $M = 1.53, SD = 4.03$ for WorldNews) than posts with low similarity (with similarity scores below the $25th$ percentile) ($M = 0.96, SD = 1.50$ and $M = 0.91, SD = 2.16$).

X-Users

Posts that show signs of being poorly received by the community, as we define X-posts to be, are often associated with trolls and ill-intentioned users, who deliberately antagonize other community members. However, even productive users are susceptible to occasional backlash. Off-topic content, biased opinions, and even bad jokes may come from any participating user over the course of a discussion, and all may be met with a mixed reaction from other users. Indeed, we find that there is a linear relation between a user's total number of posts and their number of X-posts, with a Pearson correlation coefficient of 0.825 for Politics and 0.999 for WorldNews.

We introduce the notion of **X-users** as users who make X-posts more frequently than others. To find these, we compute the number of posts per user, and for each set of users with the same number of posts we compute the average of X-posts. Given the distribution of the number of X-posts divided by the number of posts, we consider as X-users those with an X-posts-to-normal-posts ratio higher than the $95th$ percentile. In total, we label 17.3% of users on Politics as X-users, and 15% on WorldNews. These users are responsible for 14.7% and 12.9% of all posts (both normal and X-posts) in each respective community.

[10] We also experimented with the Word Mover's Distance presented in (Kusner et al. 2015), and we selected the cosine similarity as it produced better results.

Path Patterns

In this section, we turn our focus to the Harmony, Discrepancy, Disruption and Dispute conversational patterns, which we define according to how X-posts feature into different conversation paths.

As different paths belonging to the same post tree may share prefixes with the same posts, considering all paths would constitute data dependencies and would lead to non-iid[11] samples. Therefore, we perform our analyses on a subset of the data, containing one randomly sampled path from each post tree in the dataset (where each tree is rooted at a top-level comment). Table 2 lists the number of sampled paths that fall into each of the path pattern categories.

Pattern	#Paths in Politics	#Paths in WorldNews
Harmony	83,657	43,055
Discrepancy	54,562	30,801
Disruption	10,538	6,619
Dispute	8,565	4,167
Others	44,073	26,798
Total	201,395	111,440

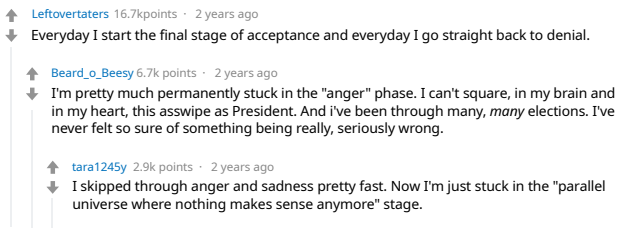
Table 2: Number of path samples for each pattern.

In the following, we express our expectations about each of these patterns as hypotheses and use statistical tests to evaluate how they are expressed in the sentiment, topic, and user dimensions. When examining the role of X-posts in specific path patterns, we employ Student's t-tests to compare them to normal posts in the same paths, with regard to their mean sentiments and topics. For these tests, we report the *t-value*, *p-value*, and effect size, which quantifies how pronounced the results are in the data, measured with Cohen's *d* (Cohen 1988). Cohen's *d* represents a very small effect size if $d \in [0.01, 0.20)$, small effect if $d \in [0.20, 50)$, medium if $d \in [0.50, 0.80)$, and large if $d \geq 0.80$. When analyzing the traits of each path pattern, we employ one-way ANOVA tests followed by Games Howell post-hoc tests, to compare post dimensions across different pattern categories. For these, we report the *F-test* statistic, *p-value*, and the effect size expressed as Eta-squared (η^2) (Sawilowsky 2009), which correspond to a small effect size if $\eta^2 \in [0.01, 0.059)$, medium if $\eta^2 \in [0.059, 0.138)$, and large if $\eta^2 \geq 0.138$. Table 3 shows a summary of our findings.

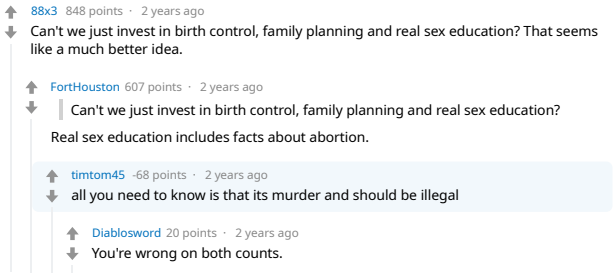
Harmony

Harmonies correspond to paths made up entirely of normal posts, that is, posts that have received no notable negative reaction from the community. Intuitively, such paths might

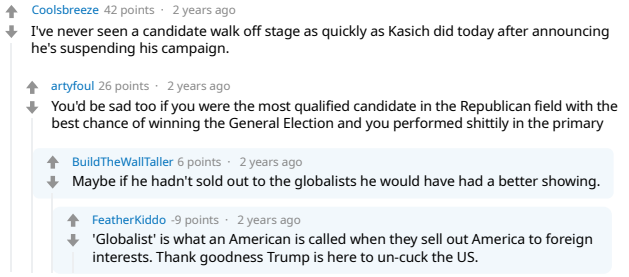
[11] iid = identically independently distributed



(a) Harmony.



(b) Discrepancy.



(c) Disruption.

(d) Dispute.

Figure 1: Examples of paths following each path pattern (X-posts are highlighted).

represent agreements, or at least balanced debates, without extreme sentiment polarization. Figure 1a shows an example of a path from the Politics subreddit which follows this pattern. In the following hypotheses, we assess the notion of Harmonies as positive and cohesive conversations.

H1: Harmony paths have the highest sentiment score.

To test this hypothesis, we compute the average value of the sentiment scores for all paths. We then compare these values for paths which follow the Harmony pattern against Discrepancy, Disruption and Dispute paths. Indeed there is a statistically significant difference for both datasets ($F(4, 188333) = 197.958, p < 0.001, \eta^2 = 0.004$ for Politics and $F(4, 110142) = 336.688, p < 0.001, \eta^2 = 0.012$ for WorldNews), indicating that Harmony paths are overall more positive, but this effect is subtle.

H2: Harmony paths have the highest topic similarity with the news article.

As a measure of how on-topic a path is, we compute the average similarity of the posts in each path to the news article originally referenced by the path. As in the previous hypothesis, these values are compared for Harmony and other path patterns. We find that there is a statistically significant difference for WorldNews with respect to topic similarity between different patterns ($F(4, 110142) = 75.968, p < 0.001$) and that Harmony has the highest topic similarity compared to other patterns. For the Politics subreddit, there was no statistically significant difference in topic similarity with the news between patterns ($p > 0.05$).

The above results, along with those in the previous hypothesis, demonstrate that while Harmony paths lack significant disturbances, they are not necessarily representative of uniform, cohesive discussions, nor of positive and uplifting exchanges between users. Instead, this pattern represents more relaxed conversations, where users may freely stray off-topic and express themselves positively or negatively. A prominent case of such a Harmony is humor, where humorous posts in a path often differ in content from its respective new article and might contain expletives or negative terminology. The posts in Figure 1a are examples of posts that would be considered negative and off-topic by our toolset, but which are highly upvoted by the community.

Discrepancy

Discrepancies represent paths where a single post has received a negative or mixed response from the community. Figure 1b shows an example of this pattern, where the highlighted post was heavily downvoted in comparison to other posts on the same path. While certain posts may simply be outliers in terms of their scores, we postulate that X-posts in Discrepancies may be singled out as such due to being off-topic or differing in sentiment from the remainder of the path.

H3: The X-post in a Discrepancy path expresses a different sentiment than the rest of the path.

For this hypothesis, we check the sentiment polarity (positive, neutral, or negative) of an X-post against the polarity of the mean sentiment of normal posts on the same path. We find that on 55% of paths on Politics, the X-post has a different sentiment polarity from the rest of the path, while on WorldNews this is true for 57% of paths.

In addition to this, we compare the average sentiment score of X-posts with the average sentiment score of normal

Hypothesis	Politics	WorldNews
H1: Harmony paths have the highest sentiment score	True	True
H2: Harmony paths have the highest topic similarity with the news article	Inconclusive	True
H3: The X-post in a Discrepancy path expresses a different sentiment than the rest of the path	True	True
H4: The X-post in a Discrepancy path has low similarity with the news article	True	True
H5: The X-post in a Discrepancy path is made by an X-user	False	False
H6: Disruption paths exhibit a sentiment shift between normal posts and X-posts	True	True
H7: Disruption paths display a topic shift between normal posts and X-posts	True	True
H8: Disruption paths contain the largest fraction of X-posts written by X-users	False	False
H9: X-posts have lower sentiment scores than normal posts in a Dispute path	True	Inconclusive
H10: Dispute paths have the highest topic similarity between posts	Inconclusive	Inconclusive

Table 3: Summary of hypotheses results. A hypothesis is marked as True or False when there is statistically significant evidence supporting or contradicting the claim, and Inconclusive when results are not statistically significant. We note that for H5, results are based only on descriptive statistics.

posts on Discrepancy paths. We find that X-posts in these paths have statistically significant lower sentiment values when compared to normal posts ($t(103134) = 12.35,\ p < 0.001, d = 0.077$ for Politics and $t(61600) = 13.971,\ p < 0.001, d = 0.116$ for World News), which may account for some of the negative reaction they receive.

H4: The X-post in a Discrepancy path has a low similarity with the news article.

Here, we compare X-posts and normal posts with regards to how similar they are to the news articles they originally referenced. For the comparison, we use the average topic similarity between X-posts and the news, and normal posts with the news. We find that the X-post in Discrepancies has lower similarity with the news article than normal posts in these paths, in both datasets ($t(103134) = -31.209,\ p < 0.001,\ d = 0.15$ for Politics and $t(61600) = -8.26,\ p < 0.001,\ d = 0.06$ for World News).

These results, as well as those in the previous hypothesis, indicate that the X-post in a Discrepancy does differ from normal posts in the path, either by straying off the original topic or by expressing a different sentiment.

H5: The X-post in a Discrepancy is made by an X-user.

We investigate also whether X-users are more often behind X-posts in Discrepancy paths. Such cases may correspond to instances of users attempting (and failing) to create a disturbance, or of community bias (Cheng et al. 2017) against known users. We find that Discrepancies are caused by X-users in 38.5% of cases in the Politics subreddit and 36.7% in WorldNews. While these may be cases of X-users intentionally trying to disturb the conversation, Discrepancies appear to be a more general result of posts which go against the predominant topic or sentiment.

Disruption

Disruption paths are made up of sub-sequences of normal posts followed by X-posts, or vice-versa. In both cases, these paths can be viewed as discussions that went through a sudden shift in terms of the community reaction to the conversation. An example of such a pattern is shown in Figure 1c. In the following hypotheses, we focus on the contrast between X-posts and normal posts in these paths to show whether there is indeed a change in the conversation, whether from the topic or sentiment perspective.

H6: Disruption paths exhibit a sentiment shift between normal posts and X-posts.

To test this hypothesis, we calculate the average sentiment value of posts in each sub-sequence (X-posts vs normal posts) of a Disruption path. A comparison of these averages finds that there is indeed a difference between the sentiment of both sub-sequences ($t(19866) = 5.944,\ p < 0.001, d = 0.084$ for Politics and $t(13236) = -6.931,\ p < 0.001,\ d = 0.12$ for World News). In particular, the sub-sequence of X-posts in these paths is more negative on average (mean sentiment score of $M = -0.011, SD = 0.39$ on Politics and $M = -0.11, SD = 0.36$ on World News), compared to the sub-sequence of normal posts ($M = 0.019, SD = 0.33$ and $M = -0.07, SD = 0.32$). Additionally, we find that on 54% of paths in the Politics subreddit and 53% of paths in the WorldNews subreddit there is a polarity shift from one sub-sequence to another, most frequently from positive to negative.

H7: Disruption paths display a topic shift between normal posts and X-posts.

For this hypothesis, we again rely on news articles as a point of reference for topic cohesiveness in paths and calculate the average topic similarity between posts in each sub-sequence of a Disruption path and the news articles they originally referenced. Comparing these two means reveals a statistically significant difference between topic similarities in the two sub-sequences ($t(19866) = -15.527,\ p < 0.001,\ d = 0.22$ for Politics and $t(13236) = -7.912,\ p < 0.001,\ d = 0.137$ for World News). Additionally, we find that the sub-sequences of X-posts have, on average, a higher topic similarity with the news article ($M = 0.57, SD = 0.127$ for Politics and $M = 0.53, SD = 0.15$ for World News), when compared to the sub-sequences of normal posts ($M = 0.55, SD = 0.13$ and $M = 0.51, SD = 0.15$).

H8: Disruption paths contain the largest fraction of X-posts written by X-users.

A possible explanation for the phenomenon of Disruption patterns is that a path is "highjacked" by an X-user. Given this, we would expect to find a larger fraction of X-posts written by X-users in Disruption paths than in Discrepancy and Dispute paths. There is indeed a statistically significant difference between these values. However, Disputes appear as the pattern containing the highest fraction of X-posts

made by X-users ($F(4, 911984) = 78036.47, p < 10^{-5}$, $\eta^2 = 0.247$ for Politics and $F(4, 198804) = 16573.25, p < 10^{-5}$, $\eta^2 = 0.25$ for WorldNews). Nonetheless, the majority of Disruption paths contain at least one X-post written by an X-user (65% on Politics and 68% on WorldNews), which demonstrates that these users are significantly involved in these conversations.

Together, these hypotheses confirm that there is a difference between the two portions of a Disruption path. More noticeably, we find that X-posts in these paths are both more negative and more closely related to the news article being discussed. As such, X-posts in these paths are likely to represent more polarized (and less popular) opinions about the subject matter of the news article, rather than user attempts at thread highjacking or "whataboutism", in which the discussion is shifted towards a new topic.

Dispute

Dispute paths alternate between X-posts and normal posts in their entirety. Intuitively, such paths might represent arguments or disagreements in which one side has the majority of the support from the community. Figure 1d shows an example of a Dispute. In the following hypotheses, we test whether these paths comprise opposing sentiments with regards to a specific topic, as would be typical in a contended debate.

H9: X-posts have lower sentiment scores than normal posts in a Dispute path.

For this hypothesis, we compare the average sentiment value of X-posts and normal posts in a Dispute path. A test of these values finds that there is a statistically significant difference between X-posts and normal posts in Dispute paths in the Politics dataset ($t(16202) = 3.155,\ p < 0.001,\ d = 0.05$), with X-posts being slightly more negative in sentiment ($M = -0.02, SD = 0.37$) than normal posts ($M = -0.002, SD = 0.34$), on average. However, no significant difference is found on the World News dataset ($p > 0.05$), where X-posts and normal posts are both negative, on average ($M = -0.095, SD = 0.35$ and $M = -0.093, SD = 0.32$ respectively). Therefore, X-posts are not necessarily the most "negative" side of a Dispute, and the high sentiment variance we find indicates that there may be a mix of sentiments expressed by both X-posts and normal posts throughout these conversations.

H10: Dispute paths have the highest topic similarity between posts.

To measure whether Dispute paths address a single issue from different perspectives, we compare the average topic similarity of posts in these paths against the post similarity in other path pattern types. However, we find no significant evidence to confirm this hypothesis ($p > 0.05$). One potential reason for this result is that opposite sides in a debate may use different arguments to back up their individual claims, so that post content between normal posts and X-posts may be highly varied.

In addition to the above, we also find that Dispute paths are shorter in length than other path types, with an average length of 5.7 posts (compared to 6.5 for Harmony, 6.98 for Discrepancies and 6.95 for Disruptions). This highlights the fact that disputed conversations are often short-lived.

Discussion of Findings

We studied several dimensions of conversations on two prominent sub-forums of the Reddit community. Using explicit cues like downvotes and the Reddit "controversiality" flag, we introduced X-posts to denote posts that have received a negative or mixed community reaction. Based on the pattern of occurrences of X-posts throughout conversation paths, we then proposed and analyzed four discussion archetypes: Harmony, Discrepancy, Disruption, and Dispute.

The Harmony pattern is intuitively supposed to represent positive conversations with high consensus on a topic. We found that, although Harmony paths tend to be slightly more positive than others, they often deviate from the topic brought up by the news article submission that started a discussion. This pattern is, therefore, more indicative of discussions without strong disagreements. Interestingly, although politics is often not associated with harmonious conversations, this is the most frequent pattern in our datasets. This reveals, to some extent, that the Politics and WorldNews subreddits mostly contain fairly civilized discussions.

The Discrepancy pattern represents conversations where a single post stands out from the rest by having received a markedly different community reaction. We found that this deviation is reflected across multiple dimensions of the discussions, with X-posts having a different polarity from the rest of the path and being more off-topic than normal posts in these paths.

The Disruption pattern indicates a strong shift in the discussion. We postulated that this shift is related to a sudden change in the sentiment or the topic of a conversation, and found that there is indeed a significant difference between the sentiments and topics expressed by X-posts and normal posts in Disruption paths. In particular, we found that X-posts tend to be more negative and more closely related to the news article. One plausible explanation for this is that X-posts discuss news articles in more detail and in a more negative light than normal posts in the same paths.

Finally, the Dispute pattern intuitively corresponds to disagreements over a given topic. We did not find significant evidence that these paths are topically more cohesive than others. This is likely a reflection of users posting different arguments to support their individual views on the same topic. The presence of mixed and negative sentiments also hints towards an exchange of polarized opinions, although this effect is subtle. We found, however, that X-users tend to participate more in writing X-posts in Disputes. This is interesting as it shows that X-users are less inclined to completely disturb conversations by creating Disruptions, and more likely want to have (healthy) arguments with other members of the community.

We highlight that content moderation also affects the discussions we observe in the Politics and WorldNews subreddits, particularly those that would, in principle, fit the Dispute and Disruption patterns: posts which contain very extreme statements or personal attacks are likely to be quickly

removed by moderators, and therefore would be absent in our datasets.

Conclusion

Discussions in online forums are very rich and complex regarding both the content and dynamics of conversations and the features of the underlying platform. Our proposed archetypes connect these important elements and give us insights into the relationship between sentiments, topics and user actions.

In future work, we plan to investigate whether these conversational patterns can be found also in other communities and whether similar cues regarding community reaction, sentiments, and topics can be used to characterize archetypical phenomena.

References

Aggarwal, C. C., ed. 2011. *Social Network Data Analytics*. Springer.

Al-garadi, M. A.; Varathan, K. D.; Ravana, S. D.; Ahmed, E.; Shaikh, G. M.; Khan, M. U. S.; and Khan, S. U. 2018. Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Comput. Surv.* 51(1):16:1–16:37.

Aragón, P.; Gómez, V.; and Kaltenbrunner, A. 2017. To thread or not to thread: The impact of conversation threading on online discussion. In *International AAAI Conference on Web and Social Media (ICWSM-17)*.

Chen, M. 2017. Efficient vector representation for documents through corruption. In *International Conference on Learning Representations (ICLR 2017)*.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17. ACM.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *International AAAI Conference on Web and Social Media (ICWSM-15)*, 61–70.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. 2nd.

Coletto, M.; Garimella, K.; Gionis, A.; and Lucchese, C. 2017. A motif-based approach for identifying controversy. In *International AAAI Conference on Web and Social Media (ICWSM-17)*.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying controversy on social media. *Trans. Soc. Comput.* 1(1):3:1–3:27.

Gilbert, C. H. E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *AAAI Conference on Web and Social Media (ICWSM-14)*.

Glenski, M., and Weninger, T. 2017. Predicting user-interactions on reddit. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17. ACM.

Gómez, V.; Kappen, H. J.; Litvak, N.; and Kaltenbrunner, A. 2013. A likelihood-based framework for the analysis of discussion threads. *World Wide Web* 16(5):645–675.

Jiang, M.; Cui, P.; and Faloutsos, C. 2016. Suspicious behavior detection: Current trends and future directions. *IEEE Intelligent Systems* 31(1):31–39.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, 957–966.

Liang, Y. 2017. Knowledge sharing in online discussion threads: What predicts the ratings? In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17. ACM.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Nishi, R.; Takaguchi, T.; Oka, K.; Maehara, T.; Toyoda, M.; Kawarabayashi, K.-i.; and Masuda, N. 2016. Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining* 6(1):26.

Rizoiu, M.-A.; Graham, T.; Zhang, R.; Zhang, Y.; Ackland, R.; and Xie, L. 2018. #debatenight: The role and influence of socialbots on twitter during the 1st 2016 u.s. presidential debate. In *AAAI International Conference on Web and Social Media (ICWSM-18)*.

Sawilowsky, S. S. 2009. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods* 8:597–599.

Vilares, D., and He, Y. 2017. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1573–1582. Association for Computational Linguistics.

Weninger, T.; Zhu, X. A.; and Han, J. 2013. An exploration of discussion threads in social news sites. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*.

Zayats, V., and Ostendorf, M. 2018. Conversation modeling on reddit using a graph-structured lstm. *Transactions of the Association for Computational Linguistics* 6:121–132.

Zhang, J.; Danescu-Niculescu-Mizil, C.; Sauper, C.; and Taylor, S. 2018. Characterizing online public discussions through patterns of participant interactions. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18.

Zhang, A.; Culbertson, B.; and Paritosh, P. 2017. Characterizing online discussion using coarse discourse sequences. In *AAAI International Conference on Web and Social Media (ICWSM-17)*.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.