X-Posts Explained: Analyzing and Predicting Controversial Contributions in Thematically Diverse Reddit Forums

Anna Guimarães, Gerhard Weikum
Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{aguimara, weikum}@mp-inf.mpg.de

Abstract

Most online discussion forums capture user feedback in the form of "likes" and other similar signals, but limit this to positive feedback. A few forums, most notably Reddit, offer both upvotes and downvotes. Reddit posts that received a large number of both upvotes and downvotes receive an explicit "controversiality" marker, while heavily downvoted posts are hidden from the standard view of the discussion, and only shown upon explicit clicks.

This paper aims at understanding the nature and role of controversial posts in Reddit, considering four subreddits of very different natures: US politics, World politics, Relationships and Soccer. We design a feature space and devise a classifier to predict the occurrence of a controversial post given a prefix of a path in a discussion thread. Our findings include that these classifiers exhibit different behaviors in the four subreddits, and we identify key features for the respective cases. An in-depth analysis indicates that controversial posts do not arise as troll-like behavior, but are often due to a polarizing topic (mostly in US politics), off-topic content, or mentions of individual entities such as soccer players or clubs.

Introduction

Detecting, analyzing and characterizing sentiments, bias and controversy in online discussion forums has been a major research topic for years (see, e.g., (Kumar, Cheng, and Leskovec 2017; Garimella et al. 2018; Hutto and Gilbert 2014) and references given there). Prior work has largely focused on antisocial behavior, such as trolling (Zhang et al. 2018; Liu et al. 2018), hate-speech (Davidson et al. 2017; Mondal, Silva, and Benevenuto 2017), and other kinds of polarization (Garimella and Weber 2017; Joseph et al. 2019). These, however significant, represent severe instances of disturbances in a discussion, rather than regular characteristics. Work on understanding polarization in social media has mostly looked into limited kinds of sources like Twitter and Wikipedia (edit history and talk pages). There is little work on more elaborate discussion forums, like Quora or Reddit, exceptions being (Wang et al. 2013; Peddinti et al. 2014; Guimarães et al. 2019; Grover and Mark 2019; Chang and Danescu-Niculescu-Mizil 2019; Jhaver et al. 2019) where the focus is mostly on aspects like community structure and dynamics or privacy-sensitive topics.

In this paper, our goal is to understand the role and nature of controversial posts in Reddit discussions. We focus on Reddit for two reasons. First, it covers a wide spectrum of topical domains with in-depth discussions, with diverse sub-forums known as subreddits. We hypothesize that controversies have very different characteristics in subreddits as diverse as (US) Politics, (personal) Relationships, and Soccer. Second, Reddit is one of the few communities where users can give both positive and negative feedback on posts, in the form of upvotes and downvotes. We expect that this can give us a more informative signal about emerging controversies, compared to forums with likes only.

Specifically, we build on the notion of *X-posts* introduced by (Guimarães et al. 2019). These are posts that have attracted negative community feedback, despite not being necessarily associated with trolling. Such posts may instead represent unpopular opinions on controversial topics, strong sentiments, or off-topic content that does not contribute to a discussion. A particular point of interest is the fact that different communities may have unique notions of what constitutes an X-post in their specific contexts: a community strictly dedicated to political discussions may embrace controversiality and differences of opinion but discourage off-topic content, whereas a community focused on general interpersonal discussions may allow more room for tangential topics and be less tolerant of controversial content that may result in conflict.

Our goal in this paper is to understand the content signals that lead to an X-post within a discussion. Specifically, we investigate the following research questions:

- Which features in a discussion are indicative of the occurrence of X-posts?
- Are there specific topics that often incur X-posts, regardless of whether the discussion itself is controversial?
- Given a prefix of initial posts in a discussion path, can we predict whether the path will eventually have an X-post?

To address these questions, we design a feature space to describe various aspects of online discussions, including sentiments, cohesiveness, activity levels, and the presence or absence of X-posts. We use these features to learn logistic-regression classifiers trained on discussions from

four prominent and thematically diverse subreddits: Politics, World News, Relationships and Soccer. As X-posts may represent different types of posts depending on the community they appear in, we compare our findings on each of these subreddits and provide insight into the roles fulfilled by X-posts in different contexts. All of the data used in this work, as well as the features we derive, are made available for use (see Datasets section).

Our model has benefits along two major lines. First, it has potential to support the moderation of online debates. The X-post predictor may, for example, be used to alert moderators of discussions that require intervention. More strategically, our feature model can convey insights on the evolution of forum polarization and user behavior, while taking forum-specific traits into account. Second, longitudinal research studies on how content and behaviors differ across topics and forums, and how they change over time, may be supported by our model.

Related Work

Trolling and antisocial behaviour. The tasks of identifying, characterizing, and predicting malicious online behavior have received considerable attention in recent research.

(Zhang et al. 2018) devises a method to predict whether antisocial behavior will appear in Wikipedia discussion pages, based on linguistic cues reflecting politeness and rhetorical prompts. Follow-up work in (Chang and Danescu-Niculescu-Mizil 2019) extends this theme to Reddit discussions, using neural-network models for prediction. The work exclusively focuses on the special case of personal attacks in user posts, independently of topics and the nature of the discussion. In contrast, our work aims to understand a broader spectrum of controversial posts and the signals that lead to flagging them.

(Addawood et al. 2019) investigates troll behavior on Twitter during the 2016 US election campaign. The authors identify several linguistic features in tweets made by Russian troll accounts, and uses random forest and gradient boosting classifiers to predict troll behavior from deceptive language cues. (Liu et al. 2018) employs a logistic regression classifier to predict the occurrence and intensity of hostile comments on Instagram, based on linguistic and social features of earlier comments. (Cheng et al. 2017) argues for a broader definition of trolling, by investigating comments that were reported for abuse in the comment section of CNN.com news articles. The authors use a logistic regression classifier to show that comments may be considered "trolling" based on factors such as user mood and context, rather than a repeated history of malicious behavior.

(Hine et al. 2017) studies an extreme case of anti-social behavior, in the form of the 4chan board /pol/, a community specifically centered around hateful content. The authors provide insight into typical activity associated with extremism and how it carries over into other platforms. (Flores-Saviaga, Keegan, and Savage 2018) also analyzes the mobilization of "trolls" from the Reddit community The_Donald, highlighting the usage of inflammatory language that led to users engaging in trolling activity.

Controversy. Related to the issue of disruptive behavior on social media is the problem of recognizing and handling online controversy. (Gao et al. 2014) proposes a collaborative filtering method to estimate user sentiment, opinion, and likelihood of taking action towards controversial topics on social media. (Garimella et al. 2018) builds a domain-agnostic framework to identify controversial topics. The method proposes the use of a social graph of agreements between users in a conversation, which can be partitioned to represent opposing viewpoints, and allows for controversy to be quantified by network metrics like betweenness and connectivity.

In the opposite direction, (Napoles, Pappu, and Tetreault 2017) develops a pipeline to identify productive discussions in comment sections of Yahoo News articles. The proposed method relies on both textual features, like part-of-speech tag and entity mentions, and post features, like length and popularity, and a combination of ridge regression, CRFs, linear regression, and a convolutional neural network to automatically determine whether a comment thread is engaging, respectful, and informative.

Reddit discussion threads. Prior research on Reddit has looked into its voting system, moderation, and thread organization. (Jhaver et al. 2019) performs a detailed study on the role of moderators and automated moderating tools ("automods") on Reddit, examining how these tools impact content regulation on the platform and providing an overview of posting behavior, comment etiquette, and community-specific guidelines in different subreddits. (Liang 2017) analyzes the voting behavior in the Q&A TechSupport subreddit. The author uses negative binomial regressions and negative binomial mixed models to investigate the relationship between users, thread structure, and voting in determining post quality. (Fiesler et al. 2018) analyzes rules for community governance and self-organization across a large number of subreddits. (Grover and Mark 2019) presents a systematic study of early indicators for political radicalism in the alt-right subreddit. (Datta and Adar 2019) investigates antagonistic interaction between different subreddits (e.g., leading to the closure of an entire subreddit).

(Zayats and Ostendorf 2018) models the structure of Reddit discussions as a bidirectional LSTM. The authors show how the model can be used to predict the popularity of individual comments in terms of their scores, and how it may be used in conjunction with textual features to predict controversial comments. (Guimarães et al. 2019) proposes four archetypes of Reddit discussions built around the concept of X-posts, i.e., posts that received negative community feedback. These archetypes are then characterized via a series of statistical tests expressing expectations about their overall sentiment and topical cohesion. This work does not, however, address how X-posts might emerge in a discussion.

Data Modeling and Analysis

On Reddit, discussions are based on a user submitting a piece of content or media to a community (subreddit), for example, a news article or an advice-seeking question or statement. A *discussion thread* originates from a *submission* by having one or more community members posting *initial*

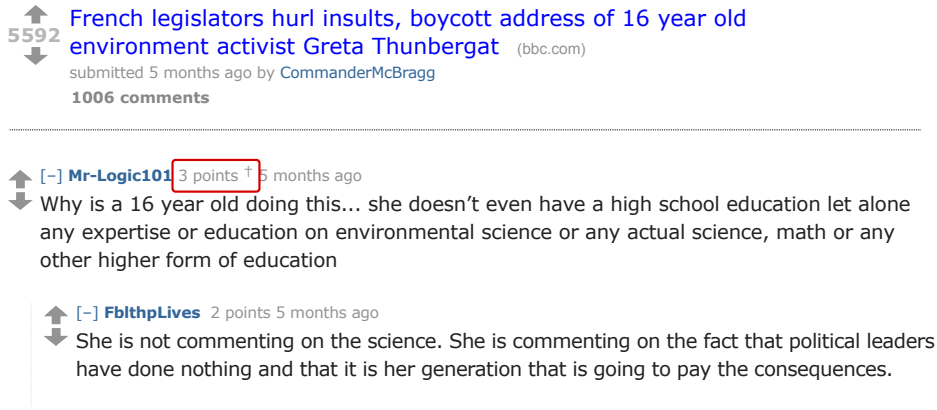

Figure 1: Submission and posts from the World News subreddit, with X-post marked by the typographical dagger.

Figure 2: Submission and posts from the World News subreddit, with X-post indicated by upvote/downvote difference of -24.

comments. As users reply to these comments, entire discussion *trees* unfold, sometimes comprising a large number of user posts (100s or more) and going into considerable depth (10 or more). Each submission can thus lead to a set of trees of posts, one tree per initial comment.

Unlike most social media platforms, Reddit allows users to give both positive and negative feedback in the form of *upvotes* and *downvotes*. Each submission and each post on the platform is associated with a score, representing the difference of upvotes and downvotes it has accumulated.

While scores from voting are mostly used for guiding readers through discussions in the Reddit UI, posts that have attracted negative attention are handled in specific ways. Posts that have received a substantial amount of votes and a roughly equal share of upvotes and downvotes are explicitly flagged as "controversial"[1]. This allows users to easily find and distinguish these posts in a discussion, as their overall scores may be positive or negative as usual. Posts may also be automatically hidden from the UI view if they have received a majority of downvotes, resulting in negative scores (by default, posts are hidden once they have a score equal to or below -4). Such posts may still be accessed, but doing so requires additional user interaction. Figures 1 and 2 show examples: the first case has a post explicitly flagged as controversial, symbolized in the UI with a typographical dagger, and the second case includes a post with a notably negative difference of upvotes and downvotes of 24 points.

In this work, we focus on these posts that have attracted significant negative attention, which we refer to as **X-posts**. In particular, we are interested in the context in which these posts appear and the elements of the discussion that are associated with their occurrence.

Definitions

We build on the definitions in the recent work of (Guimarães et al. 2019) to describe Reddit discussions:

- A **submission** refers to the starting point in a discussion, and consists of an initial piece of media or text submitted to a community by one of its users.
- Users post initial comments on the submission, which are referred to as **top-level comments**. Further posts are later made in reply to existing comments.

[1] www.reddit.com/r/announcements/comments/293oqs/new_reddit_features_controversial_indicator

- The result of these chains of comments and replies, rooted in a top-level comment, are referred to as **post trees**. As shorthand to describe user-posted content, top-level comments and replies are both referred to here as **posts**.[2]
- A **path** in a post tree denotes a sequence of posts, where each post is a direct reply to its immediate predecessor.
- **X-posts** denote posts which have attracted notable negative feedback from the community. A post is considered an X-post if it has been explicitly flagged as controversial on the Reddit interface, or if its score ($\#upvotes - \#downvotes$) is sufficiently negative (≤ -4). All other posts are referred to as **normal posts**.

Datasets

Our datasets comprise content from four prominent subreddits: Politics[3], World News[4], Relationships[5], and Soccer[6]. On the first two communities, posting guidelines dictate that all submissions must be links to external news articles of reputable sources and thematically appropriate (US politics and non-US news, respectively), while Relationships calls for text posts, and Soccer allows a mix of both free-form text submissions, links and media related to soccer. Thus, the four subreddits differ not only in terms of their content, but also in how their discussions are initiated, structured, and regulated. We chose these four so as to study this variety.

We collected all submissions and available comments posted to each of these communities in 2016 and 2017 using the PSRAW wrapper for the Reddit API[7] (last accessed in January 2019). We removed posts and submissions that had their text deleted or which linked to inaccessible external sources. As we are interested in discussions, rather than single posts that received little interaction or follow-up, we additionally discarded very short paths from the data, keeping only those that had a minimum of 5 posts.

From the remaining data, we created our datasets by randomly selecting one path from each post tree, where a post tree is rooted at a top-level comment made to a submission. We employed this one-path-per-tree restriction to ensure statistically independent samples in our study. In other words, we excluded overlapping paths that share a prefix.

The resulting datasets are summarized in Table 1. The distribution of posts that fall under the definition of an X-post in each of the datasets is shown in Table 2. All data is available at `people.mpi-inf.mpg.de/~aguimara/xposts`.

Properties of Paths and Post Trees

Building on the definitions of post trees and paths, we distinguish three categories of paths, according to the presence or absence of X-posts in a path and its surrounding tree:

- N: paths from trees containing only normal posts

[2]Note that this definition may differ from varying Reddit terminology, where submissions are sometimes called "posts".

[3]www.reddit.com/r/Politics

[4]www.reddit.com/r/WorldNews

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

[6]www.reddit.com/r/Soccer

[7]psraw.readthedocs.io/en/latest/

Source	Year	Submissions	Replies	Users	Paths
Politics	2016	34,785	1,350,866	114,970	201,395
	2017	19,477	468,383	54,799	71,067
WorldNews	2016	24,277	743,542	133,118	111,440
	2017	28,733	873,954	143,977	129,750
Relationships	2016	26,773	327,564	44,528	53,437
	2017	34,261	395,464	51,055	64,486
Soccer	2016	34,358	772,998	51,048	124,599
	2017	23,797	475,686	35,186	71,510

Table 1: Subreddit datasets.

Source	Year	Controversial	≤ -4 Points	Both
Politics	2016	150,456	95,841	19,574
	2017	23,642	34,917	3,570
WorldNews	2016	86,839	60,155	12,787
	2017	95,556	69,985	15,904
Relationships	2016	16,718	27,973	2,992
	2017	21,767	20,983	3,317
Soccer	2016	21,727	20,882	2,901
	2017	34,478	38,833	5,981

Table 2: Number of posts that satisfy each criterion for the definition of an X-post.

- NX: paths that contain only normal posts but are part of a tree containing at least one X-post
- X: paths that contain at least one X-post

The intuition for this categorization is that post trees with X-posts may address contended topics or have a bigger potential for disruptions compared to trees containing only normal posts, even if such disruptions are not present in every individual path in the tree. These differences would be particularly notable on those paths which themselves contain an X-post.

To determine whether the textual content of paths in these categories reflects notable differences, we computed frequently mentioned named entities in each of the categories N, NX and X. We identified the 50 most frequent entities per category, using the named entity recognition component of the AIDA tool (Hoffart et al. 2011). To highlight the differences across categories, we calculated the ratio of frequen-

Source	Year	N	NX	X
Politics	2016	71,898	129,497	117,738
	2017	33,507	37,560	32,375
WorldNews	2016	33,130	78,310	68,385
	2017	40,461	89,289	77,419
Relationships	2016	28,859	24,578	22,106
	2017	39,443	25,043	22,606
Soccer	2016	27,265	29,505	22,273
	2017	23,860	47,650	34,797

Table 3: Number of sampled paths belonging to the N, NX, and X categories.

Source	Year	Top Entities (X/N)
Politics	2016	Jill Stein (26774), November (14265), ISIS (11233), the Supreme Court (11229), Islam (11000), BLM (10552), ID (10531), Mexican (10420), Gore (10325), the Clinton Foundation (10239), Bernie (1.89), Reddit (1.67), Hillary (1.60), TPP (1.58), Comey (1.56), Clinton (1.52), Democrats (1.37), FBI (1.31), Muslims (1.24), 2008 (1.24)
	2017	Nazi (8451), Perez (6185), Ellison (5425), 2008 (3924), Islam (3543), Jews (6525), MSM (3457), Gorsuch (3383), Syria (3124), Milo (3106), State (3092), Jewish (3025), Bernie (6.31), Hillary (3.10), Clinton (2.82), Democrats (1.89), Muslim (1.85), Reddit (1.73), CNN (1.63), Obama (1.51)
WorldNews	2016	Hamas (27048), Jesus (19004), Gaza (18071), Quran (17909), Christianity (16952), Nazi (16081), Kurds (15339), Crimea (15305), Merkel (15154), DNC (15071), Palestinians (8.70), Israel (2.81), the Middle East (2.46), Jewish (2.23), Clinton (2.06), Hillary (1.98), Trump (1.88), Ukraine (1.77), Islam (1.67), Obama (1.66)
	2017	Democrats (47097), Nazi (37917), FBI (28121), Jerusalem (19231), Bush (18468), Hamas (16495), the Middle East (15456), Crimea (15264), Venezuela (14993), Poland (14857), Palestinians (6.85), Israel (2.61), Christian (2.60), Clinton (2.48), Jewish (2.31), Hillary (2.18), Obama (1.92), Republicans (1.88), Muslim (1.86), Ukraine (1.85)
Relationships	2016	Callie (2007), OP (753), Japanese (734), Indian (684), Japan (597), STD (592), NYC (485), Vegas (471), Christian (451), Reddit (1.91), Asian (1.536), America (1.47), Jesus (1.41), American (1.16), Christmas (1.12), FWB (1.07), US (1.04), English (1.04), Europe (1.03), CPS (0.92)
	2017	OP (798), NYC (569), PPD (409), Asian (1.84), GF (1.53), Reddit (1.48), SIL (1.40), America (1.37), FWB (1.29), American (1.24), Europe (1.17), BPD (1.12), IUD (1.11), Jesus (1.06), Christmas (1.06), US (1.02), STD (1.01), CPS (1.00), English (0.93), Google (0.89), Christian (0.85)
Soccer	2016	La Liga (3015), Messi (2871), Real Madrid (2378), Bale (2375), Klopp (2358), Atletico (2123), Ozil (1685), Zidane (1682), Wenger (1660), America (1660), Costa (1656), Guardiola (1620), Spurs (1595), Giroud (1558), China (1532), USA (1530), Iniesta (1516), American (2.80), Ronaldo (2.90), Suarez (2.37)
	2017	Hazard (5820), Ozil (5792), Qatar (5601), Southampton (4189), Celtic (4085), UEFA (4004), Spurs (3954), Zidane (3723), Atletico (3720), Kante (3605), UK (3562), Griezmann (3538), Cristiano (3492), Bundesliga (3487), Pogba (2.86), Messi (2.51), Ronaldo (2.49), Mourinho (2.40), United (2.39), Suarez (2.10)

Table 4: Top 20 entities with highest X/N ratio of occurrence frequencies.

cies of the top entities in category X and in category N, as $freq_{entity}\_X/max\{freq_{entity}\_N, 1\}$. The entities with the highest X/N ratios in the 8 datasets are shown in Table 4.

While popular entities are frequent across both X and N categories, the ratios bring out some notable differences.

For the Politics datasets, the most interesting observations come from contrasting the two years 2016 and 2017. For example, in 2016, Jill Stein, who was the Green Party's nominee for the US presidential election, was ranked highest in terms of X/N ratio with substantial controversiality, but was almost entirely absent in the 2017 data.

The frequent entities in the World News community mostly pertain to countries and leaderships. Religion and ethnicity are more frequent in paths containing X-posts. Among countries, Israel is among the ones most related to X-posts, appearing more than twice as often in the X category than in the N category. In contrast, countries like China and Turkey appear with roughly the same frequency in both X and N.

The Relationships datasets show the least amount of differences when comparing frequent entities between the X and N categories. A portion of the entities retrieved refer to acronyms, such as MIL (mother-in-law) and OP (original poster), rather than real-world named entities. Mental illness, ethnicity, and online platforms (Facebook, Reddit) also featured prominently in all categories. Given the personal nature of the community, it is natural that real-world entities would be infrequent.

For the Soccer datasets, similar to what we see in the two political communities, common themes appear across both X and N paths. Nonetheless, certain entities stand out as being closely associated with X-posts: several prominent figures, like the team manager Jürgen Klopp and player Mesut Özil, are frequent only in paths containing X-posts, while others, like Ronaldo and Suarez, are twice as frequent in category X paths than in N paths.

These findings highlight the fact that, although there are interesting differences in the entities and topics that discussions center on, these topics and entities alone are not sufficient to determine the presence and influence of X-posts. In the next section, we introduce additional features of discussions, which we then use as the basis for a classifier to predict future occurrences of X-posts.

Features of Discussions

We propose a feature space containing three main axes, each of which captures a different aspect of discussions: i) the *sentiments* expressed in posts, ii) their topical *cohesiveness*, and iii) the *activity level* and types of posts (X-posts and normal posts) in a path. A summary of the features is shown in Table 5.

$frac\_pos$, $frac\_neg$, $frac\_neu$	$\frac{\#\{positive,negative,neutral\}posts}{\#posts}$
avg_sent, var_sent	$\frac{\sum postsentiment}{\#posts}$
avg_pos, var_pos_sent	$\frac{\sum pospostsentiment}{\#positiveposts}$
avg_neg, var_neg_sent	$\frac{\sum negpostsentiment}{\#negativeposts}$
$diff\_sent$	$\frac{\sum_{i=1} sent_i \neq sent_{i-1}}{\#posts}$
$post\_sim$, var_post_sim	$\frac{\sum_{i=0,j=1} sim(p_i,p_j)}{\#posts}$
sub_sim, var_sub_sim	$\frac{\sum_{i=0} sim(p_i,s)}{\#posts}$
$root\_sim$, var_root_sim	$\frac{\sum_{i=1} sim(p_i,p_0)}{\#posts}$
$contains\_entity_{i,K}$	$= \begin{cases} 0 & \text{if } entity_{i,K} \notin path \\ 1 & \text{if } entity_{i,K} \in path \end{cases}$, $\forall K \in \{N, NX, X\}$
$prior\_X$	$= \begin{cases} 0 & \text{if } X \notin path \\ 1 & \text{if } X \in path \end{cases}$
$avg\_replies$, $var\_replies$	$\frac{\sum replies}{\#posts}$
avg_delay, var_delay	$\frac{\sum_{i=1} timestamp_i - timestamp_{i-1}}{\#posts}$
$frac\_X$	$\frac{\#X-posts}{\#posts}$
$uniq\_users$	$\frac{\#users}{\#posts}$

Table 5: Feature summary.

Sentiment Features. For each post in our dataset, we calculate its sentiment score using VADER (Hutto and Gilbert 2014), a sentiment analysis method created from a gold-standard sentiment lexicon, specialized for social media text. The sentiment scores range from -1 to 1, where a score of -1 indicates extremely negative polarity, and a score of 1 indicates maximum positive polarity. Posts with a score in the range $[-0.05, 0.05]$ are labeled as neutral.

To describe the overall sentiment expressed over a series of posts in a path, and how the sentiment fluctuates, we calculate the following metrics for each path:

- Fractions of posts in the path with negative, neutral, and positive sentiment scores.
- Average and variance of the sentiment scores across all posts in the path.
- Average and variance of the sentiment values across all positive posts in the path.
- Average and variance of the sentiment values across all negative posts in the path.
- Fraction of posts that have a different polarity than their immediately preceding post in the same path (polarity shifts).

Textual Features. To capture the textual content of posts, we transform them into sentence embeddings using Doc2VecC (Chen 2017), an unsupervised method that learns a fixed-length vector representation of sentences. For each pair of consecutive posts in a path, we consider the text similarity of two posts p_i and p_j, $\mathbf{sim(p_i, p_j)}$ to be the maximum cosine similarity of the embeddings for the sentences in p_i and p_j. Similarly, to account for the initial posts in the discussion, we compute the text similarity between the top-level post in the path and each subsequent post as $\mathbf{sim(p_i, p_0)}$, as well as the similarity between the original submission and posts in a path, $\mathbf{sim(p_i, s)}$.

To quantify the topical cohesion between posts in a path and how the posts relate to the initial topic of the submission, we calculate the following metrics per path:

- Average and variance of the text similarity between consecutive posts in the path.
- Average and variance of the text similarity between the original submission and the posts in the path.
- Average and variance of the text similarity between the top-level post at the root of the path and subsequent posts in the path.

Additionally, we capture the influence of individual terms that appear prominently in different categories of discussion paths. For this, we consider the top 50 most frequent entities in each of the N, NX, and X categories, as described in the previous section, with the following features:

- Binary flags that denote whether each frequent entity from categories N, NX, and X is present in at least one post in the path.

Post Features. Direct signals from the posts themselves can also describe the development of discussion paths. The presence and prevalence of X-posts, for example, may indicate intense disagreements. In addition, the time between successive posts, the number of replies received by each post, and the number of unique users participating in a path all constitute signals about its overall activity level.

To capture these features for each path, we calculate the following metrics:

- Binary flag that denotes whether the path contains an X-post or not.
- Average and variance of the number of replies received by each post in the path.
- Average and variance of the timespan between consecutive posts (post delay).
- Fraction of posts in the path that have been flagged as an X-post.
- Fraction of distinct users in the path.

Predicting X-Posts

In this section, we investigate whether it is possible to predict the occurrence of X-posts based on features of a discussion during its initial stages. We formulate this as the following prediction task: given a set of features derived from a path prefix, will the path suffix include an X-post?

For this task, we devise a binary logistic regression classifier where the predicted output variable is the presence of an X-post in the path suffix ("X-post" or "No-X-post"), and where the features of the previous subsection are computed for the path prefix only. As paths in our data have a minimum length of 5 posts, we consider the first 4 posts as the prefix of the path, and the remaining posts as its suffix.

We trained the classifier on each of our eight datasets. As X-posts are relatively rare, making up less than 15% of posts in our datasets, we balanced classes with oversampling using SMOTE (Chawla et al. 2002), using 70% of the resulting observations as training data and the remaining 30% as test data. Across all datasets, instances that contained an X-post in the path suffix were underrepresented, hence the need for balancing. The number of instances prior to oversampling are shown in Table 6. Note that without addressing this class imbalance, a classifier may learn to simply assign the dominant class label to any input and still achieve high overall accuracy. To underline this point, we also trained a classifier with the original class-imbalanced data for comparison.

The prediction results for each of the datasets are shown in Table 7. For each dataset, we present precision (true positives/(true positives+false positives)), recall (true positives/(true positives+false negatives)), F1-score (harmonic mean of the precision and recall), and AUC (area under the receiver operating characteristic curve).

Overall, the classifiers achieved F1-scores between 65 and 75 percent. This is a decent result, in line with values observed for other kinds of predictors over social media. Note that it is unrealistic to expect very high precision and recall, say with F1 around 90 percent, for our setting. Even more restricted tasks, like the neural classifier for predicting personal attacks in discussions (Chang and Danescu-Niculescu-Mizil 2019) with well-curated training data, did not exceed 70 percent in F1.

When comparing results for subsequent years in the same community, we find only small differences in prediction results. The only drop comes for the Soccer datatset, where predictions also had the lowest F1-scores, at 0.67 and 0.65 for 2016 and 2017, respectively. We refer back to Tables 1 and 2 to note that despite a drop in activity in this subreddit from 2016 to 2017, the amount of X-posts increased, revealing a significant shift in the community's posting behavior.

Source	Year	No-X-post	X-post
Politics	2016	164,073	26,713
	2017	63,751	7,018
WorldNews	2016	93,282	18,158
	2017	107,775	21,002
Relationships	2016	45,964	7,413
	2017	57,696	6,760
Soccer	2016	49,046	7,055
	2017	54,924	14,978

Table 6: Number of instances in the No-X-post and X-post classes prior to balancing.

Source	Year	Precision	Recall	F1-score	AUC
Politics	2016	0.67	0.81	0.73	0.79
	2017	0.70	0.81	0.75	0.77
World News	2016	0.63	0.74	0.68	0.72
	2017	0.64	0.76	0.70	0.73
Relationships	2016	0.73	0.74	0.73	0.79
	2017	0.75	0.74	0.75	0.80
Soccer	2016	0.69	0.65	0.67	0.74
	2017	0.68	0.62	0.65	0.71

Table 7: Prediction results for the X-post class.

Politics and Relationships exhibit the best prediction scores, with F1 at 0.73 for the 2016 data and 0.75 for 2017. We recall that the latter is the only community among our datasets where submissions are exclusively text posts by users, i.e., there is no outside content being brought in for discussion, which may reduce the amount of variance in topic cohesiveness and sentiments across paths. Compared to the other communities, (US) Politics, with its strong focus on the two main parties Republicans vs. Democrats during the election year of 2016, is presumably the one with the most narrow topical focus, which results in more topically cohesive discussions overall.

In contrast, the World News dataset shows comparatively worse results, with F1 scores at 0.68 and 0.70 for 2016 and 2017, respectively. We attribute this to the much larger diversity of topics and consequently wider range of opinions in the discussion about world-wide politics. Thus, the classifier for this community faces a more difficult task than the one for US politics.

We also conducted this evaluation with classifiers trained on the original class-imbalanced data. These predictors achieved good overall accuracy, between 0.72 (for Soccer 2017) and 0.89 (for Politics 2017). However, this was at the total negligence of the minority class of X-posts, with recall at or near 0% for the X-class. Consequently, both F1-score and AUC were very poor as well, and far inferior to the classifiers trained with re-balanced data.

Feature Influence

To understand the influence of specific features on the classifiers' prediction performance, we show the most significant features for each dataset in Table 8. The table gives the weights as learned by the logistic regression models for each of the three highest-weighted, and thus most influential, features.

Across all datasets, the fraction of controversial posts and the presence of an X-post in the path prefix were among the top predictors. Another important feature across all datasets was the topical cohesiveness of posts within a path, represented by the average similarity with the root post. This shows the importance of the initial topic for the subsequent discussions. Features representing the similarity with the submission and among the posts in the path were also weighted highly.

An interesting observation for the Relationships datasets is that the fraction of sentiment-wise neutral posts, which is an indicator for the absence of X-posts in the other communities, is among the high-weight features for future X-posts in 2017. This suggests that posts with a neutral tone about personal relationships are viewed as a deviation from the more emotional nature of this community's usual posts.

In World News, two predictors of future X-posts stand out: the fraction of consecutive posts with alternating sentiment polarities, and the fraction of unique users in a discussion. Together with the high weights for features relating to cohesiveness, these suggest that the community is less tolerant of arguments.

Source	No-X Predictors	X Predictors
Politics 2016	$post\_sim$ (-0.395) $root\_sim$ (-0.336) $frac\_neu$ (-0.271)	$prior\_X$ (1.553) $frac\_X$ (1.406) $avg_replies$ (0.145)
Politics 2017	$uniq\_users$ (-0.298) $root\_sim$ (-0.240) avg_pos (-0.210)	$prior\_X$ (1.732) $frac\_X$ (0.934) avg_neg (0.109)
WorldNews 2016	$root\_sim$ (-0.431) $frac\_neu$ (-0.338) $post\_sim$ (-0.315)	$frac\_X$ (1.344) $prior\_X$ (1.082) $uniq\_users$ (0.224)
WorldNews 2017	$root\_sim$ (-0.332) $post\_sim$ (-0.308) sub_sim (-0.228)	$frac\_X$ (1.347) $prior\_X$ (1.264) $uniq\_users$ (0.169)
Relationships 2016	$root\_sim(-0.339)$ $frac\_neg$ (-0.271) $post\_sim$ (-0.261)	$prior\_X$ (1.888) $frac\_X$ (1.075) $avg\_replies(1.86)$
Relationships 2017	$sub\_sim(-0.330)$ $frac\_pos$ (-0.321) $root\_sim$ (-0.313)	$prior\_X$ (2.086) $frac\_X$ (0.879) avg_neg (0.273)
Soccer 2016	$frac\_neu$ (-0.518) $frac\_pos$ (-0.210) $root\_sim$ (-0.205)	$prior\_X$ (1.466) $frac\_X$ (0.979) $post\_sim$ (0.279)
Soccer 2017	$frac\_neu$ (-0.172) $root\_sim$ (-0.153) $uniq\_users$ (-0.144)	$prior\_X$ (1.148) $frac\_X$ (0.463) $avg\_replies$ (0.082)

Table 8: Feature weights.

X-post Entities

The presence of specific entities in a path often features as a good indicator of the future of discussions, as most of the communities we examine highlight.

In the Politics dataset, while several political figures are more frequent in paths containing X-posts, they are less significant in predicting their occurrence in the 2016 dataset. Instead, entities like Israel, ISIS, and TPP (Trans-Pacific Partnership), feature more prominently.

Interestingly, Hillary, Bernie and Obama are among the top predictors of future X-posts in the 2017 dataset, during a time when these figures received less attention in the political landscape. The explanation is that their total popularity in 2016 was orders of magnitude higher. In 2017, the normal posts about these entities dropped drastically, but the amount of polarizing posts stayed relatively high, so that their X/N ratio increased substantially.

For World News in both years, Palestine and Israel had the highest feature weights among the top frequent entities, and are good indicators of future X-posts. Interestingly, mentions of religions, like Christianity and Islam, are inversely related to future occurrences of X-posts, despite being more frequent in paths that contain them (see Table 4). This result indicates that discussions involving religious topics often evolve in a fairly civilized manner – a good sign that this subreddit community welcomes healthy disagreement without acting negatively.

For the Soccer dataset, we again find heavily debated players and teams, like Messi, Ronaldo and (Manchester) United, as good predictors of future X-posts, whereas national teams and locations are indicators for the absence of X-posts. The results for this community largely echo our observations from Table 4.

For the Relationships datasets, as expected from the nature of this community, named entities play a minor role. While they are not entirely insignificant, even terms like STD (Sexually Transmitted Diseases) and PPD (Post-Partum Depression), which are potentially controversial, contribute little to the model when compared to other textual, structural, and sentiment features.

Robustness to Changing Topics

As the topics and associated entities in forum discussions change over time, the question arises as to what extent our model and method can gracefully handle such evolution. In the previous subsection, we notice how the same features often appear as top predictors for both 2016 and 2017 data, which indicates that past activity may be used to predict X-posts even farther into the future. To test this hypothesis, we apply the models trained on 2016 data to 2017 data. Results are shown in table 9.

The prediction results here are comparable to those achieved when the model is trained and applied to data from the same year, with F1-scores above 0.70 for all but one community. This indicates that despite potential changes in the community's topic of interest, discussions tend to follow similar patterns, such that the learned models remain viable over a longer time horizon.

Source	Precision	Recall	F1-score	AUC
Politics	0.70	0.81	0.75	0.80
World News	0.63	0.79	0.70	0.73
Relationships	0.75	0.73	0.74	0.80
Soccer	0.67	0.61	0.64	0.71

Table 9: Prediction results for the X-post class on 2017 data, with the model trained on 2016 data.

We highlight that the worst result is found for Soccer, the community in which we observed the largest shift from 2016 to 2017, particularly in terms of top entities and posting behavior, as previously discussed. We offer more discussion on evolving community interests and behaviors in the next section on model Limitations and Extensions.

Limitations and Extensions

Our model and its supporting framework are designed to be modular enough to be altered and extended as needed for other settings. In particular, it is easy to replace the components for entity detection and for sentiment features with alternative models and tools. To validate that our results do not unduly rely on specifics of our choices, we varied the predictors to replace AIDA with the popular spaCy[8] tool and VADER with LIWC[9].

While the alternative for NER did not lead to any major difference, we observed some degradation on the sentiment features when not using VADER. Naturally, several configuration and tuning issues may be at work here, and we did not investigate these issues to full extent. Rather, we believe that sentiment features are a generally challenging aspect that may require further extension, along the following lines.

Contextual Sentiment. VADER, like other tools for sentiment analysis, is built from a lexicon where terms were evaluated independently of context. This means that nuances in a community's use of language, which come as a result of its central theme, are largely ignored. For instance, while "war" is assigned a negative sentiment value in VADER, it may not necessarily convey a negative sentiment in the context of news or political discussions. Therefore, a specialized dictionary that reflects a community's vocabulary, or is otherwise sensitive to the context in which a term appears, would lead to more refined insights about the role of sentiment in how discussions progress.

Online training. Our results on robustness to changing topics show that despite changes in a community, its core behavior remains fairly consistent. This holds both for entities under discussion and for the language style of posts and replies. Nevertheless, it is conceivable that some forums undergo rapid shifts in what entities are of interest and even in the vocabulary and style of user posts. This raises the question of if and how a feature-based model for analysis and training predictors can keep up with the pace of changes.

Our approach to this end would be to frequently re-build the model and re-train the classifiers. This could be done on a weekly or even daily basis, as none of our components is prohibitively expensive. Feature extraction, including entity detection, can be performed in a few hours on a commodity machine, and training a logistic regression classifier takes only seconds. Still, proof of practical viability remains as future work.

[8] spacy.io/

[9] liwc.wpengine.com/

Conclusion

In this paper, we investigated the phenomenon of X-posts in discussions of four major Reddit communities. We devised a feature space that captures key aspects of discussion threads, including sentiment variation, topical cohesiveness, frequent entity mentions and activity levels. We leveraged these features for prefixes of discussion paths to learn classifiers for predicting if the initial path later leads to the occurrence of an X-post.

Our analysis of feature influence reveals that the topical cohesiveness across posts and the existence of an X-post early in the discussion are most informative across all four communities. In contrast, sentiment variation, as expressed, for example, by strong language, does not play a major role in triggering downvotes and controversiality flagging. Overall, these four Reddit communities seem to be very healthy in terms of tolerating disagreements and argumentation, as long as the user posts stay on topic.

The varying performance results for the dataset-specific classifiers also bring out key differences between the four subreddits, Politics, World News, Relationships, and Soccer. In particular, it appears that the prediction of X-posts is easier for US Politics than for World News, probably because of the highly polarized nature of the US political system with two major parties that are strongly opposing each other. Entities that appear in the submissions or root posts play a major role in leading to X-posts, except for the Relationships community. For Soccer, it is often the case that fans of debated players or teams get into emotional disagreements, leading to X-posts. These differences highlight the fact that X-posts are contextually defined by the communities in which they appear, rather than adhering to a single definition of controversiality.

References

Addawood, A.; Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Linguistic Cues to Deception: Identifying Political Trolls on Social Media. In *Proc. of the 13th International Conference on Web and Social Media, ICWSM*, 15–25. AAAI Press.

Chang, J. P.; and Danescu-Niculescu-Mizil, C. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, 4742–4753. ACL.

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, W. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16: 321–357.

Chen, M. 2017. Efficient Vector Representation for Documents through Corruption. In *5th International Conference on Learning Representations, ICLR*. OpenReview.net.

Cheng, J.; Bernstein, M. S.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proc. of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW*, 1217–1230. ACM.

Datta, S.; and Adar, E. 2019. Extracting Inter-Community Conflicts in Reddit. In *Proc. of the 13nth International Conference on Web and Social Media, ICWSM*, 146–157. AAAI Press.

Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proc. of the 11th International Conference on Web and Social Media, ICWSM*, 512–515. AAAI Press.

Fiesler, C.; Jiang, J. A.; McCann, J.; Frye, K.; and Brubaker, J. R. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proc. of the 12th International Conference on Web and Social Media, ICWSM*, 72–81. AAAI Press.

Flores-Saviaga, C.; Keegan, B. C.; and Savage, S. 2018. Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community. In *Proc. of the 12th International Conference on Web and Social Media, ICWSM*, 82–91. AAAI Press.

Gao, H.; Mahmud, J.; Chen, J.; Nichols, J.; and Zhou, M. X. 2014. Modeling User Attitude toward Controversial Topics in Online Social Media. In *Proc. of the 8th International Conference on Web and Social Media, ICWSM*. AAAI Press.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying Controversy on Social Media. *ACM Trans. Social Computing* 1(1): 3:1–3:27.

Garimella, V. R. K.; and Weber, I. 2017. A Long-Term Analysis of Polarization on Twitter. In *Proc. of the 11th International Conference on Web and Social Media, ICWSM*, 528–531. AAAI Press.

Grover, T.; and Mark, G. 2019. Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit. In *Proc. of the 13th International Conference on Web and Social Media, ICWSM*, 193–204. AAAI Press.

Guimarães, A.; Balalau, O. D.; Terolli, E.; and Weikum, G. 2019. Analyzing the Traits and Anomalies of Political Discussions on Reddit. In *Proc. of the 13th International Conference on Web and Social Media, ICWSM*, 205–213. AAAI Press.

Hine, G. E.; Onaolapo, J.; Cristofaro, E. D.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *Proc. of the 11th International Conference on Web and Social Media, ICWSM*, 92–101. AAAI Press.

Hoffart, J.; Yosef, M. A.; Bordino, I.; Fürstenau, H.; Pinkal, M.; Spaniol, M.; Taneva, B.; Thater, S.; and Weikum, G. 2011. Robust Disambiguation of Named Entities in Text. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 782–792. ACL.

Hutto, C. J.; and Gilbert, E. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proc. of the 8th International Conference on Web and Social Media, ICWSM*. AAAI Press.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput. Hum. Interact.* 26(5): 31:1–31:35.

Joseph, K.; Swire-Thompson, B.; Masuga, H.; Baum, M. A.; and Lazer, D. 2019. Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-Based Measures. In *Proc. of the 13th International Conference on Web and Social Media, ICWSM*, 290–301. AAAI Press.

Kumar, S.; Cheng, J.; and Leskovec, J. 2017. Antisocial Behavior on the Web: Characterization and Detection. In *Proc. of the 26th International Conference on World Wide Web Companion*, 947–950. ACM.

Liang, Y. 2017. Knowledge Sharing in Online Discussion Threads: What Predicts the Ratings? In *Proc. of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW*, 146–154. ACM.

Liu, P.; Guberman, J.; Hemphill, L.; and Culotta, A. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proc. of the 12th International Conference on Web and Social Media, ICWSM*, 181–190. AAAI Press.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017. A Measurement Study of Hate Speech in Social Media. In *Proc. of the 28th ACM Conference on Hypertext and Social Media, HT*, 85–94. ACM.

Napoles, C.; Pappu, A.; and Tetreault, J. R. 2017. Automatically Identifying Good Conversations Online (Yes, They Do Exist!). In *Proc. of the 11th International Conference on Web and Social Media, ICWSM*, 628–631. AAAI Press.

Peddinti, S. T.; Korolova, A.; Bursztein, E.; and Sampemane, G. 2014. Cloak and Swagger: Understanding Data Sensitivity through the Lens of User Anonymity. In *2014 IEEE Symposium on Security and Privacy*, 493–508. IEEE.

Wang, G.; Gill, K.; Mohanlal, M.; Zheng, H.; and Zhao, B. Y. 2013. Wisdom in the social crowd: an analysis of quora. In *22nd International World Wide Web Conference, WWW*, 1341–1352. ACM.

Zayats, V.; and Ostendorf, M. 2018. Conversation Modeling on Reddit Using a Graph-Structured LSTM. *Trans. Assoc. Comput. Linguistics* 6: 121–132.

Zhang, J.; Chang, J. P.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, 1350–1361. ACL.