Data Science Methods for the Analysis of Controversial Social Media Discussions

by

Anna Christina de Carvalho Guimarães

Dissertation zur Erlangung des Grades des Doktors der Ingenieurwissenschaften (Dr.-Ing.) der Fakultät für Mathematik und Informatik der Universität des Saarlandes

Saarbrücken, 2022

Day of Colloquium	June 10, 2022
Dean of the Faculty	Univ.-Prof. Dr. Jürgen Steimle
Chair of the Committee	Prof. Dr. Anna Feit
Reporters	
First Reviewer	Prof. Dr. Gerhard Weikum
Second Reviewer	Prof. Dr. Pedro Vaz de Melo
Third Reviewer	Dr. Andrew Yates
Academic Assistant	Dr. Paramita Mirza

Abstract

Social media communities like Reddit and Twitter allow users to express their views on topics of their interest, and to engage with other users who may share or oppose these views. This can lead to productive discussions towards a consensus, or to contended debates, where disagreements frequently arise.

Prior work on such settings has primarily focused on identifying notable instances of antisocial behavior such as hate-speech and “trolling”, which represent possible threats to the health of a community. These, however, are exceptionally severe phenomena, and do not encompass controversies stemming from user debates, differences of opinions, and off-topic content, all of which can naturally come up in a discussion without going so far as to compromise its development.

This dissertation proposes a framework for the systematic analysis of social media discussions that take place in the presence of controversial themes, disagreements, and mixed opinions from participating users. For this, we develop a feature-based model to describe key elements of a discussion, such as its salient topics, the level of activity from users, the sentiments it expresses, and the user feedback it receives.

Initially, we build our feature model to characterize adversarial discussions surrounding political campaigns on Twitter, with a focus on the factual and sentimental nature of their topics and the role played by different users involved. We then extend our approach to Reddit discussions, leveraging community feedback signals to define a new notion of controversy and to highlight conversational archetypes that arise from frequent and interesting interaction patterns. We use our feature model to build logistic regression classifiers that can predict future instances of controversy in Reddit communities centered on politics, world news, sports, and personal relationships. Finally, our model also provides the basis for a comparison of different communities in the health domain, where topics and activity vary considerably despite their shared overall focus. In each of these cases, our framework provides insight into how user behavior can shape a community’s individual definition of controversy and its overall identity.

Zusammenfassung

Social-Media Communities wie Reddit und Twitter ermöglichen es Nutzern, ihre Ansichten zu eigenen Themen zu äußern und mit anderen Nutzern in Kontakt zu treten, die diese Ansichten teilen oder ablehnen. Dies kann zu produktiven Diskussionen mit einer Konsensbildung führen oder zu strittigen Auseinandersetzungen über auftretende Meinungsverschiedenheiten.

Frühere Arbeiten zu diesem Komplex konzentrierten sich in erster Linie darauf, besondere Fälle von asozialem Verhalten wie Hassrede und "Trolling" zu identifizieren, da diese eine Gefahr für die Gesprächskultur und den Wert einer Community darstellen. Die sind jedoch außergewöhnlich schwerwiegende Phänomene, die keinesfalls bei jeder Kontroverse auftreten die sich aus einfachen Diskussionen, Meinungsverschiedenheiten und themenfremden Inhalten ergeben. All diese Reibungspunkte können auch ganz natürlich in einer Diskussion auftauchen, ohne dass diese gleich den ganzen Gesprächsverlauf gefährden.

Diese Dissertation stellt ein Framework für die systematische Analyse von Social-Media Diskussionen vor, die vornehmlich von kontroversen Themen, strittigen Standpunkten und Meinungsverschiedenheiten der teilnehmenden Nutzer geprägt sind. Dazu entwickeln wir ein Feature-Modell, um Schlüsselelemente einer Diskussion zu beschreiben. Dazu zählen der Aktivitätsgrad der Benutzer, die Wichtigkeit der einzelnen Aspekte, die Stimmung, die sie ausdrückt, und das Benutzerfeedback.

Zunächst bauen wir unser Feature-Modell so auf, um bei Diskussionen gegensätzlicher politischer Kampagnen auf Twitter die oben genannten Schlüsselelemente zu bestimmen. Der Schwerpunkt liegt dabei auf den sachlichen und emotionalen Aspekten der Themen im Bezug auf die Rollen verschiedener Nutzer. Anschließend erweitern wir unseren Ansatz auf Reddit-Diskussionen und nutzen das Community-Feedback, um einen neuen Begriff der Kontroverse zu definieren und Konversationsarchetypen hervorzuheben, die sich aus Interaktionsmustern ergeben. Wir nutzen unser Feature-Modell, um ein Logistischer Regression Verfahren zu entwickeln, das zukünftige Kontroversen in Reddit-Communities in den Themenbereichen Politik, Weltnachrichten, Sport und persönliche Beziehungen vorhersagen kann. Schlussendlich bietet unser Modell auch die Grundlage für eine Vergleichbarkeit verschiedener Communities im Gesundheitsbereich, auch wenn dort die Themen und die Nutzeraktivität, trotz des gemeinsamen Gesamtfokus, erheblich variieren. In jedem der genannten Themenbereiche gibt unser Framework Erkenntnisgewinne, wie das Verhalten der Nutzer die spezifisch Definition von Kontroversen der Community prägt.

CONTENTS

1

INTRODUCTION

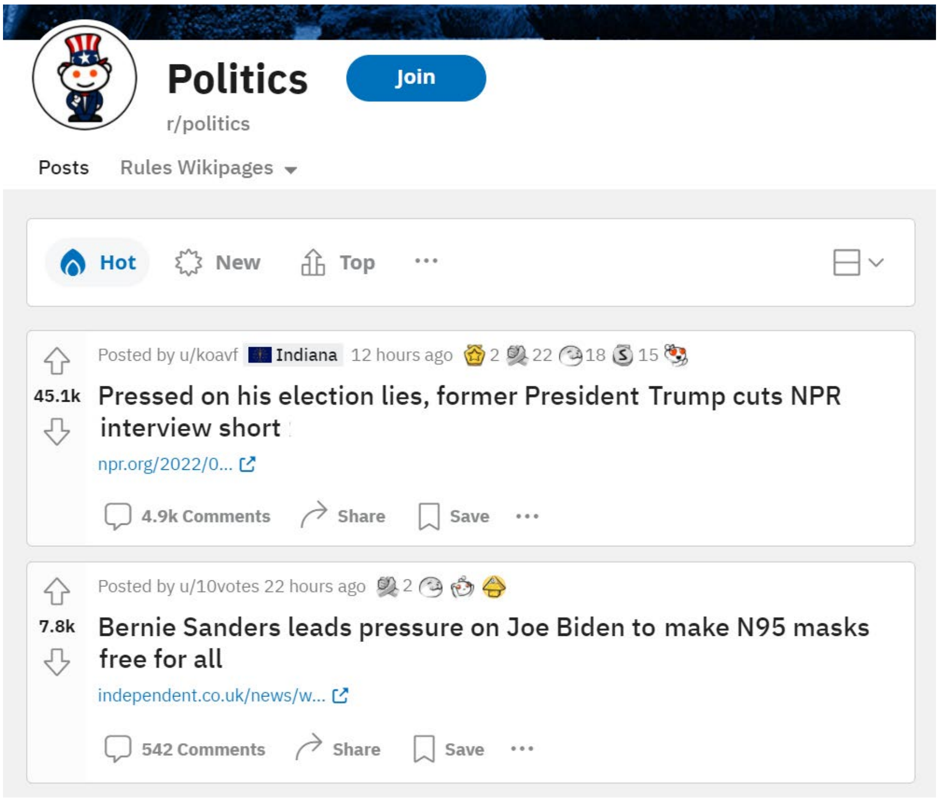
Contents

1.1 Motivation

Online discussion forums, including QA platforms and social media networks, allow users to engage with topics of their interest and with other users who may share or oppose their opinions. On Twitter (`twitter.com`), users can express an idea by the means of hashtags and easily seek out others who have used the same hashtag terms, while Facebook (`facebook.com`) and Reddit (`reddit.com`) users can organize themselves in large communities dedicated to specific interests. As one example, Figure 1.1 shows the front page of a Reddit community centered on US Politics (`reddit.com/r/Politics`), containing posts that have attracted thousands of replies from users over only a brief period of time.

In these settings, prolonged exchanges between users, via posts and replies, give way to full-fledged discussion threads. An initial post made to the community poses an initial topic, and users gradually add to it with their own thoughts, opinions, and related content. Side-conversations may develop as new topics arise, and as subsets of users respond back and forth to each other as a result of finding a particular common link.

In certain circumstances, discussions may include strong opinions, disagreements, and controversial statements. An example is shown on Figure 1.2, which highlights two discussion threads that develop from the same initial post on the controversial topic of abortion legislation. While the first two users interact productively by sharing their concerns, the users in the second thread display clearly different opinions as they voice their strong stances on the topic.



Posted by u/koavf Indiana 12 hours ago 2 22 18 15

45.1k Pressed on his election lies, former President Trump cuts NPR interview short

npr.org/2022/0...

4.9k Comments Share Save

Posted by u/10votes 22 hours ago 2

7.8k Bernie Sanders leads pressure on Joe Biden to make N95 masks free for all

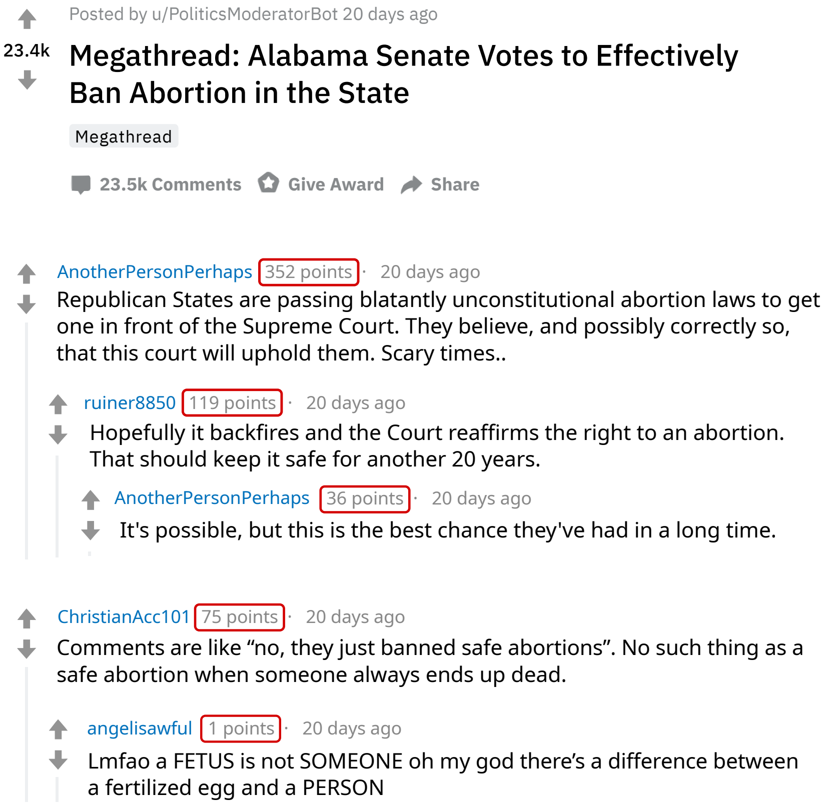
independent.co.uk/news/w...

542 Comments Share Save

Figure 1.1: Front page of the Politics subreddit.

As in the example, certain topics and domains are especially conducive to controversy and division. Political discussions, in particular, have become strongly associated with negative online phenomena such as trolling [Addawood et al., 2019] and echo chambers [Gillani et al., 2018]. These, however, do not preclude the development of productive conversations, and in some social media sites like Reddit, users are given explicit tools with which to curate discussions and thus preserve the quality of user interactions. Note that each post in Figure 1.2 is associated with a number of "points" (highlighted in red), which indicates the total of positive and negative votes it has accumulated from user feedback, and acts as a high-level representation of how well the surrounding community has received each post.

In this dissertation, our goal is to step beyond the established views about controversy in order to understand different types of interactions that take place in the presence of disagreements and polarization. This requires an analysis that connects several key aspects of online discussions, such as community feedback (e.g. via voting), textual content, and sentiment, all of which can influence how a discussion progresses. We focus primarily on political topics, where polarization features prominently, but also explore how controversy may appear in thematically diverse communities such as those centered around general news, sports, and personal relationships.



Posted by u/PoliticsModeratorBot 20 days ago

23.4k

Megathread: Alabama Senate Votes to Effectively Ban Abortion in the State

AnotherPersonPerhaps 352 points · 20 days ago

Republican States are passing blatantly unconstitutional abortion laws to get one in front of the Supreme Court. They believe, and possibly correctly so, that this court will uphold them. Scary times..

ruiner8850 119 points · 20 days ago

Hopefully it backfires and the Court reaffirms the right to an abortion. That should keep it safe for another 20 years.

AnotherPersonPerhaps 36 points · 20 days ago

It's possible, but this is the best chance they've had in a long time.

ChristianAcc101 75 points · 20 days ago

Comments are like “no, they just banned safe abortions”. No such thing as a safe abortion when someone always ends up dead.

angelisawful 1 points · 20 days ago

Lmfao a FETUS is not SOMEONE oh my god there’s a difference between a fertilized egg and a PERSON

Figure 1.2: Example of a controversial Reddit discussion on abortion legislation.

1.2 Challenges

Our study of controversial social media discussions faces several challenges:

- **Noise and Sparseness in Social Media Data**: Though social media offer an enormous amount of user-generated content, this content is not easily interpretable. Intended for an informal setting, social media post text is non-standardized and often contains misspellings, abbreviations, slang and “netspeak”, and emoticons. While state-of-the-art text processing tools are capable of handling many of these, the shortness of social media posts, the specificity of their language as a result of community-specific jargon and terminology, and their sheer volume pose challenges [Khalid and Srinivasan, 2020, Hutto and Gilbert, 2014]. These issues are further exacerbated in social media *discussions*, where posts do not express self-contained and complete ideas, and instead are pieces of a conversation that requires its full context in order to be understood.

- **Explicit and Implicit Signals**: A key component of social media is that users may interact with each other’s content to provide exposure and feedback, or as a way to initiate a personal connection. The way in which this is done varies from platform to platform:

while Twitter posts can receive “likes” and “retweets”, a Reddit post can receive “upvotes” and “downvotes”, and both can receive replies. Each of these interactions has its own meaning, and users may have different expectations and motivations for their usage, which can make these signals difficult to interpret, even in context.

- **Dynamics and Evolution**: Even online communities that are dedicated to specific topics of interest are seldom static. A constant flux of users joining and leaving a community means that new ideas, habits, and interests are always emerging and reshaping the discussions taking place within it. This is especially true for communities centered around real world events, like sports or politics, which see abrupt changes in their topical foci in response to new developments. This dynamic nature of online communities therefore makes discussion topics and style difficult to pin down, particularly when they are viewed in the long-term.

1.3 Prior Work and Its Limitations

Much of the prior research on online discussions has focused on their structure and growth. Such efforts typically involve modeling how a discussion expands and attracts new posts over time [Gómez et al., 2013, Nishi et al., 2016, Aragón et al., 2017a, Zayats and Ostendorf, 2018] or how popular it will ultimately become, in terms of the number of posts and users it involves, and how much visibility and positive feedback (likes, upvotes, retweets) it gets [Weninger et al., 2013, Cheng et al., 2016, Liang, 2017].

Recent work has attempted to investigate these discussions in greater depth, with a particular focus on discussion dynamics and evolution. [Zhang et al., 2018a] models the evolution of Facebook discussions, with a particular focus on interaction patterns correlated with disruptive behavior. Their methodology is centered strictly on the network structure surrounding a discussion, and leaves aside any consideration for its contents, in an effort to remain topic-agnostic. In the opposite direction, [Zhang et al., 2017] surveys and annotates Reddit post content corresponding to a limited set of interpretable discourse acts, like questions, answers, and disagreements, in order to describe a discussion as a whole. This work prioritizes textual elements of the discussion, and does not take into consideration is structure or elements such as post popularity and community feedback.

As these suggest, notable interaction patterns arising from conflict have been of particular interest in recent literature. These include phenomena such as the emergence and impact of hate speech [Davidson et al., 2017, Mondal et al., 2017, Chetty and Alathur, 2018, Liu et al., 2018] and trolling behavior [Cheng et al., 2015, Kumar et al., 2017, Coles and West, 2016, Flores-Saviaga et al., 2018, Garimella et al., 2018]. While a majority of the effort here has gone into detecting and counteracting isolated instances of these phenomena, some work has also examined how they manifest in discussions at large. [Cheng et al., 2017] examines the impact of the topic of news articles on trolling behavior in the comment section of news websites, while

[Coletto et al., 2017] models the propagation of controversial content on Twitter.

This existing work, however, has a narrow definition of controversy which often focuses on exceptionally egregious behavior. This overlooks the fact that controversy can also represent contended debates, differences of opinions, and otherwise turbulent content that does not go so far as to breach community guidelines or compromise the health of the discussion. Therefore, there is room to investigate a broader definition of controversy that includes such posts that deviate from the expected community behavior, taking into account the specific community context in which they appear.

1.4 Contributions

To overcome the challenges outlined in Section 1.2 and address the limitations of prior work described in Section 1.3, this dissertation offers the following contributions:

- **Framework for the study of controversial discussions.** The key contribution of this dissertation is a framework for the systematic analysis of controversial social media discussions. We design a detailed data modeling approach, where posts and discussions are described in terms of their textual content (including topics and similarity to other posts), the sentiments they express (whether positive, negative, or neutral), and the activity they receive (in terms of replies and feedback). Our scope is initially restricted to adversarial Twitter discussions with well-defined opposing sides, then expanded to general political discussions on Reddit, and finally to Reddit discussions surrounding a variety of topics where controversy may be present. Our methodology likewise evolves as our scope increases, gradually including additional and alternative features necessary to describe more complex discussions. Our first iteration of this framework was published as a workshop paper in ICDM [Guimarães et al., 2017].

- **Notion of X-posts.** As outlined in the previous section, prior research has focused on specific, heightened instances of negatively-labeled controversy. In this dissertation, we introduce the novel concept of *X-posts*, a general representation of posts that attract mixed responses from the community they are found in. Our definition relies on explicit feedback signals from the community and is validated with an in-depth characterization of post text, sentiment, and response. This allows us to explore the role of controversial posts in ongoing discussions and how they are impacted (or even dictated) by community context.

- **Categorization of discussion patterns.** Building on our definition of X-posts, we identify and characterize different discussion patterns that further refine the notion of controversies into *disputes*, *disruptions* and *discrepancies*, all of which exist in contrast to *harmonious* discussions. Similar to our approach in [Guimarães et al., 2017], we design a feature space that describes these patterns in terms of the sentiments they express, the feedback they receive, and their topical variation. These conversational archetypes, along with the definition of X-posts, were introduced in a full paper in ICWSM 2019 [Guimarães et al., 2019].

- **Predictive models for X-posts in different communities.** Taking advantage of the rich feature space we developed to describe X-posts and the discussions in which they appear, we devise a classifier to predict future occurrences of X-posts given the initial posts in a discussion. Our classifiers exhibit different behaviors across communities centered around different themes, and allows us to identify key elements that define controversy in each one. An analysis of our results also confirms our initial assumption that X-posts do not correspond only to extreme or antisocial behavior, and instead correspond to polarizing topics, off-topic content, or divisive entities, such as sports teams, politicians, and celebrities. The results of this work were published as a full paper in ICWSM 2021 [Guimarães and Weikum, 2021].

- **Comparison of online health-related communities.** To gain further insight into social media discussions, including those that do not necessarily entail controversy, we also perform a characterization study on three prominent online communities centered on health. Drawing from our established methodology, we design a feature space that expresses the user engagement, topics, and level of detail expressed in discussions. As in [Guimarães et al., 2017] and [Guimarães et al., 2019], we use these features to characterize and contrast these communities, finding notable differences in the user activity and the topical focus of each one. This comparison was published as a poster paper in CSCW 2021 [Guimarães et al., 2021].

- **Large annotated collections of social media discussions.** With this work we also publicly release the full datasets we collected, processed, and annotated for our analyses. Our Twitter dataset comprises over a million tweets made in response to political stakeholders in the US and the UK throughout 2016. Our Reddit datasets comprise over 5 million posts made to four communities in 2016 and 2017, and include annotations regarding post sentiment scores, post similarity, feedback given by upvotes and downvotes, and the presence of X-posts, among other features. Our health datasets comprises over 3 million posts made to Reddit, the Patient forums, and Health Boards, from their inception until 2020, and include annotations on post topics, activity, and medical detail. This data can be found at `https://socialdiscussions.mpi-inf.mpg.de`.

1.5 Publications

The research presented in this dissertation has been published in the following:

- Guimarães, A., Wang, L., Weikum, G. (2017). **Us and Them: Adversarial Politics on Twitter**. In *IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017*, New Orleans, LA, USA, November 18-21, 2017, pages 872-877.

- Guimarães, A., Balalau, O., Terolli, E., Weikum, G. (2019). **Analyzing the Traits and Anomalies of Political Discussions on Reddit**. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media, ICWSM 2019*, Munich, Germany, June 11-14, 2019, pages 205-213.

- Guimarães, A., Weikum, G. **X-Posts Explained: Analyzing and Predicting Controversial Contributions in Thematically Diverse Reddit Forums**. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021*, held virtually, June 7-10, 2021, pages 163-172.

- Guimarães, A., Terolli, E., Weikum, G. (2019). **Comparing Health Forums: User Engagement, Salient Entities, Medical Detail**. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media, CSCW 2021*, held virtually, October 23-27, 2021, pages 57-61.

1.6 Organization

The rest of this dissertation is organized as follows. Chapter 2 presents an overview of past research on social media discussions and online controversy. Chapter 3 explores these themes in the context of political discussions on Twitter and presents our work on identifying patterns in discussion topics and user behavior. Chapter 4 shifts our focus to political Reddit communities and introduces a novel notion of controversial posts, called X-posts, from which we derive and characterize different conversational archetypes that exist in these communities. Chapter 5 steps beyond the confines of politics and further explores the presence of X-posts in thematically diverse communities, additionally presenting a method to predict their occurrence in ongoing discussions. Chapter 6 experiments with applying our methodology to non-controversial discussions, presenting a framework to compare different online communities centered on health-related topics. Finally, Chapter 7 concludes this dissertation and gives directions to future work.

2

Related Work

Contents

This chapter presents an overview of the existing literature on modeling and analyzing social media discussions and their key elements. Section 2.1 examines research on the structural aspect of these discussions and its impact on their development across different platforms. Research on the typical roles of users driving online discussions is presented in Section 2.2, along with exceptional user behavior that may cause disruptions. Lastly, Section 2.3 examines prior work on how controversy, in its various forms, may manifest in online discussions.

2.1 Social Media Discussions

Social media provides a unique structure for users to discover and engage with content they are interested in. This process is formalized by [Lerman, 2007] with the concept of social information processing: users can establish their online presence by creating or sharing content on social media, they can engage with and evaluate existing content by means of “voting”, “liking”, or further sharing the content to other users, and they can form social networks, for instance by following or by directly interacting with other users. Due to the wealth of created content, the ability to organize and curate it becomes a crucial matter to the productive development of online communities.

Content curation and user interaction are both at the heart of the work by [Agichtein et al., 2008] on identifying quality content on social media. The authors highlight the importance of community feedback and the social activity between users in question-answering forums, demonstrating that the quality of posts are not attributed only to their semantic and syntactic content, but also to the interactions surrounding them.

On a wide scale, this points to how users may influence the visibility and reach of online content, and to their ability to shape the way in which content is received by a community of users. The tutorial by [Leskovec, 2011] surveys several techniques to identify temporal patterns arising from these user-content interactions and to model phenomena such as content diffusion [Aggarwal, 2011], popularity growth [Backstrom et al., 2013], and user influence on information-sharing [Cha et al., 2010].

2.1.1 Discussion Threads

More closely related to the work in this dissertation is the analysis of social media *discussions*, which arise as a result of direct user to user interactions. These discussions can take many different shapes according to the supporting platform in which they appear. In traditional message boards, posts mark the start of a discussion and their subsequent replies are displayed sequentially and in chronological order, forming a single linear discussion [Gómez et al., 2011, Samory et al., 2017]. Other social media networks, like Reddit and Twitter, allow posts to be made in direct response to existing posts and replies, with the resulting discussion being presented in a hierarchical view: a reply is shown closer to its parent post, and parallel branches of the discussion may emerge.

Much of the work analyzing the structure of these discussion has focused on modeling their growth, both in terms of the arrival of new posts and in terms of their expected total number of posts. The early work of [Backstrom et al., 2013] explores a probabilistic approach to predicting the length of discussions and the participation of users in Wikipedia and Facebook threads, incorporating features relating to the social network of participating users, the pattern of user arrival to the discussions, and the rate of replies. Several other probabilistic and generative models for growing discussion trees have followed this work, and are surveyed in [Aragón et al., 2017a].

Among the pioneering work into the tree-like thread structure of Reddit is [Weninger et al., 2013], which examines the topical hierarchy of a discussion and how it evolves, for instance by branching into multiple disjoint topics over the course of several replies. There the focus is on the progression of topics within a thread, which the authors extract with a latent topic model. Comparing the hierarchy of topic assignment to thread structure reveals that posts that share a common ancestor in a discussion tend to be, on average, closer in topic.

A deeper analysis of the Reddit thread structure is presented in the later work of [Choi et al., 2015], which characterizes discussions in one hundred subreddits in terms of their volume (i.e., the number of posts they comprise), their responsiveness (i.e., how quickly replies are made to existing posts), their structural virality (i.e., how likely it is that replies will be made to the post at the root of a discussion's post tree), their semantic content, and the rate of user participation.

[Liang, 2017] adds to this prior research into Reddit discussions by exploring the role of user feedback as a proxy for post and thread quality, and its relationship to thread structure. User feedback is represented by post "scores", given by the difference between upvotes (positive feedback) and downvotes (negative feedback) a post received from users. These scores, along with features describing the network structure of post trees and of participating users, are then used by the author to build negative binominal and regression models to predict post ratings in the Q&A subreddit TechSupport (reddit.com/r/techsupport). The model reveals a correlation between high post ratings and thread depth (i.e., posts that generate longer reply threads tend to be better rated) and user diversity (i.e., posts that attract a wider audience are better rated).

While many of the models reviewed here attempt to capture only the growth of online discussions, [Zhang et al., 2017] focuses instead on the dynamics of their contents. This work presents a classification of posts in Reddit discussions based on a categorization of coarse discourse acts displaying a common structure, and demonstrates how they can be used to identify interaction patterns and specific community behavior. Reddit posts are randomly sampled and crowd-annotated according to these pre-defined discourse acts, which include questions, answers, announcements, agreement, appreciation, disagreement, negative reactions, elaboration and humor. Notably, the authors find a prevalence of question posts, which are naturally accompanied by a majority of answer replies. Another interesting finding was the existence of chains of disagreements, particularly in debate subreddits such as Change My View (reddit.com/r/changemyview) or PoliticalDiscussion (reddit.com/r/PoliticalDiscussion). The proposed discourse act categories leave room for the definition of more fine-grained categorization for particular discourse acts that may arise in expanded datasets or from detailed annotations.

This collection of prior work reflects the importance of the structure surrounding a discussion when understanding how it develops, with incoming posts building off of the context provided by prior posts, and the possibility for branching, parallel discussions.

2.2 User Behavior in Online Communities

Driving online discussions are users themselves, who can simultaneously act as conversation starters who promote activity by contributing new content to their communities [Cha et al., 2010, Bakshy et al., 2011, Tinati et al., 2012], community leaders who take charge of moderating and organizing existing content [Zhu et al., 2011, Matias, 2016, Seering et al., 2019], and topical authorities who can answer questions from other users and help shape the focus of a discussion [Jurczyk and Agichtein, 2007, Weng et al., 2010, Bamakan et al., 2019].

2.2.1 Typical Roles and Behavior

The seminal work of [Cha et al., 2010] introduces the idea that every user has the potential to act as an influencer, promoting or stemming the propagation of content and ideas. Their empirical analysis of Twitter, focused on features such as a user's number of followers, retweets, and mentions, reveals that popularity (i.e., a high follower count) and name recognition (as in the case of real-world celebrities and brands) do not correlate with topical authority, and that ordinary users can therefore gain influence by posting original and topically-focused content.

Influential Twitter users are also the subject of [Weng et al., 2010]. This work creates a ranking algorithm that quantifies user influence, measured both by the follower structure surrounding a user and by the topical focus of their tweets. [Pal and Counts, 2011] develops this concept further by proposing an algorithm to identify topical authorities on Twitter, particularly in the presence of general authorities (e.g. news outlets with high visibility and follower count) and of specialized but lesser known authorities (e.g. accounts specifically created to report on an emerging event). Instead of relying solely on network metrics, which might be skewed towards established and influential accounts, the proposed methodology also takes into account the volume and nature of interactions between users and the topical coherence of their activity. Their characterization highlights that due to the highly dynamic environment of Twitter, the lifetime of topics can be short-lived, and that while a typical user's interest in a topic may be transient, true authorities display a more consistent topical focus over time.

Examining user roles beyond that of topical influencers, [Welser et al., 2011] explores social roles among Wikipedia editors and how they affect the quality and coordination of participants' contributions. Following the previous research of [Gleave et al., 2009], they define a set of potential roles, as well as potential fingerprints associated with them, linked to behavioral regularity (consistently performing a set of actions), network attributes (interactions with other contributors), and self-identity (information from editors' profile pages). Based on the edit histories of both long-term and new editors, the authors identify patterns in how users perform and adapt to these roles, finding marked differences in both their activity distribution and their network of interactions. [Zhu et al., 2011] later focuses specifically on the role of community leaders on Wikipedia, characterizing leaders as users who regularly provide positive and negative feedback for other editors, direct others to work on a particular task, or exchange social information to foster a community environment.

[Buntain and Golbeck, 2014] are among the first to study user behavior on Reddit with the goal of identifying the roles a user performs in a community. The authors focus on the prevalent role of the "answer-person", first highlighted in [Welser et al., 2007] in the context of the now-defunct Usenet news groups, which denotes users who predominantly reply to questions posed by others and seldom engage in further discussions. These users are identified according to network metrics derived from their corresponding user interaction graphs, built from their contributions to multiple subreddits. Using these metrics, the authors then build a supervised classifier to predict user roles, which reveals that users tend to take on different roles across the different communities they contribute to, though they behave consistently within each one.

Arguing for the possibility of mutable user roles on Reddit is the work of [Das and Lavoie, 2014], which examines how Reddit users change their posting strategies according to the community feedback they receive. Here, the authors adopt a reinforcement learning strategy to predict a user's future posts given the upvotes, downvotes, and replies they have received in past posts. This work stresses the notion that users adapt their behavior according to their interactions and surrounding context.

2.2.2 Disruptive Behavior

Just as users can contribute positively to the development of discussions, they can also deter and derail them. The most notable of these users are the so-called "trolls", who aim to cause disruptions by starting arguments [Buckels et al., 2014], attacking other users with insults and inflammatory remarks [Xu et al., 2012, Chatzakou et al., 2017], disseminating hate speech [Davidson et al., 2017, Mondal et al., 2017, ElSherief et al., 2018, Chetty and Alathur, 2018], and luring users into off-topic discussions [Shachaf and Hara, 2010, Dimitrov et al., 2021]. These behaviors typically exist outside of community guidelines and often lead to user bans [Geiger and Ribes, 2010], content deletion [Liu et al., 2018, Chandrasekharan et al., 2018], and even the shut down of entire communities [Newell et al., 2016, Chandrasekharan et al., 2017]. Of particular interest to our research is user behaviour which, while not so extreme as to fall into the categories of hate speech and trolling, can still create a turbulent environment for online discussions.

Under this scope, the prominent work of [Cheng et al., 2015] characterizes antisocial online behavior by investigating the posting and interaction history of banned accounts in the Disqus (`disqus.com`) comment section of three major news websites. Strong indicators of users with a tendency for antisocial behavior include language features and low readability, the concentration of replies to few individual threads, and a high rate of interactions. Additionally, the work finds that the behavior of such users tends to worsen over time and that the community refutes them faster the longer they remain active (i.e. by having their posts removed more quickly). Based on these observations, the authors use signals associated with antisocial behavior (post content, user activity, community response, and actions from community moderators) to develop a classification model that predicts whether a user will be banned in the future.

Subsequent work [Cheng et al., 2017] further explores the trigger mechanisms to antisocial behavior online, attempting to measure the impact of mood and influence on users' propensity for displaying negative behavior. The authors design an experiment simulating an online discussion in which participants are exposed to an unrelated positive or negative prior stimulus, and are then shown positive or negative posts in the discussion thread they are to participate in. The experiments indicate that negative prior mood and negative context both increase trolling behavior. The authors validate their findings by analyzing discussions in the CNN news website comments and confirm that users are more likely to engage in trolling when they have been exposed to it in the recent past, reinforcing the notion that mood is highly correlated with antisocial activity and that context greatly impacts the direction of the discussion. The topic of news articles likewise has an effect on the quality of discussions, with politics being fairly uncontroversial and sports being more so. Based on these findings, the authors then develop a logistic regression model to predict whether a user will display antisocial behavior in a given post. Unlike previous work, which relies on deleted posts as examples of trolling, this model makes use of flagged posts (i.e., posts marked by the community as potentially problematic), textual cues, and downvotes.

2.3 Controversy

Rather than focusing on disruptive individuals, research on online controversy focuses on content and discussion dynamics. These include the study of partisan networks which form in response to a divisive topic [Barberá et al., 2015, Conover et al., 2012], community responses to controversial content [Matamoros-Fernández, 2017, Choi et al., 2010], and the development of conflict and disagreements [Coletto et al., 2017, Stromer-Galley et al., 2020].

The work of [Adamic and Glance, 2005] is among the first to study the emergence of bipartisan networks on social media, in the context of political blogging during the 2004 US presidential campaign. An analysis of the links between conservative and liberal blogs reveals two distinct communities with few cross-links between them and little overlap between the topics and news they cover. Later research identified a similar partisan structure of political users on Twitter during the period leading up to the 2010 US congressional midterm elections [Conover et al., 2011]. Users were found to be more likely to retweet users matching their political leaning, and tweets of users with the same leaning were found to be more similar. Unlike the previous work, there was no significant segregation of users when it came to mentions (i.e. direct user interactions).

The long-term evolution of political partisanship on Twitter is comprehensively studied in [Garimella and Weber, 2017], which points to a significant increase in polarization over time. Compared to 2009, users in 2016 were less likely to follow and even engage with users of a different political leaning. [Joseph et al., 2019] adds to these findings by examining the political landscape of Twitter between 2017 and 2018 and analyzing the support for tweets made by then-president Donald Trump. This study finds that though there is significant polarization regarding absolute support of Trump's tweets, left and right-leaning users showed some agreement with regard to relative support (i.e. both groups show the least and most amount of support for the same tweets).

Moving away from strictly political discussions, [Garimella et al., 2018] develops a method for detecting controversy in diverse social media networks, drawing from both the content and the user structure surrounding it. Topics are initially determined by keyword queries and their associated discussions are represented in a conversation graph, where users are connected through their interactions with a given keyword. The resulting graph is then partitioned into two (mostly) disjoint sets of users representing opposing points of view on the same topic, if they exist. To measure topic controversy, the authors use random walks, measuring the probability that a random walk starts from the partition it ended in, and expected hitting time (i.e., the number of expected steps to hit the high-degree nodes on either partition). This approach is tested on several datasets, with the finding that political discussions on Twitter are among the most controversial, while political blogs tend to be less so. In order to compare their structural measures of controversy to those that rely on content analysis, such as [Choi et al., 2010], the authors also extract textual and sentiment-related features, finding that controversial topics tend to have a higher variance in sentiment and tone, even when they are not intrinsically associated with a negative or positive sentiment.

Another important work on the emergence of controversy in online discussions is [Zhang et al., 2018a], which investigates the role of politeness (or lack thereof) and other rhetorical devices as a predictive measure of conflict in Wikipedia "talk pages", the specialized discussion forums for Wikipedia editors. Unlike the antisocial behavior discussed in the prior section, here the focus is on continued interactions with the potential to become toxic, rather than isolated problematic posts or users. The authors are able to identify several conversational markers of emerging toxicity, such as the use of an accusatory "you" at the start of sentences and overly direct questions. Markers related to civil discussions include greetings and expressions of gratitude.

While the previous work focuses on detecting and learning from discussion derailment after the fact, the recent work of [Chang and Danescu-Niculescu-Mizil, 2019] develops a method to predict them as they happen. Their method relies on a generative dialogue model that learns conversational patterns from post text, and is then fine-tuned to forecast future events. In a similar line, [Hessel and Lee, 2019] proposes a method of early controversy detection on Reddit, in which controversy is given by the ratio of upvotes and downvotes a post has received over an observed period of time. Using a latent topic model, the authors extract and evaluate the topics of posts labeled as controversial, finding that they are strongly associated with controversy but also highly community-specific.

This prior research highlights the prominence of controversy, which emerges to varying degrees across several online communities, and points to the wide range of effects it can have on discussions at large. In the following chapters, we revisit these notions as we focus on controversy as not just disturbances, but also as building blocks for online discussions.

3

Adversarial Politics on Twitter

Contents

Both offline and online, political discussions take the form of debates, where opposing groups advocate their individual stances on a set of issues. Politicians themselves now have a growing presence on social media, with their campaigns extending to online spaces and including the active involvement of users who voice their support for one side while strongly opposing another. This constitutes a unique and important setting in which to observe how social media discussions evolve in the midst of diverging opinions, beliefs, and ideologies.

This chapter presents an analysis of political discussions on Twitter in the period leading up to two prominent events in 2016: the US presidential election and the EU referendum, widely known as "Brexit". Our methodology casts these discussions into a multi-faceted data space that captures their key topics and their factual and non-factual nature, described in Section 3.4, as well as the roles of participating users, described in Section 3.5. Our main findings, which include notable differences in the topical focus and user activity from supporters of each stance in the two campaigns, are summarized in Section 3.6.

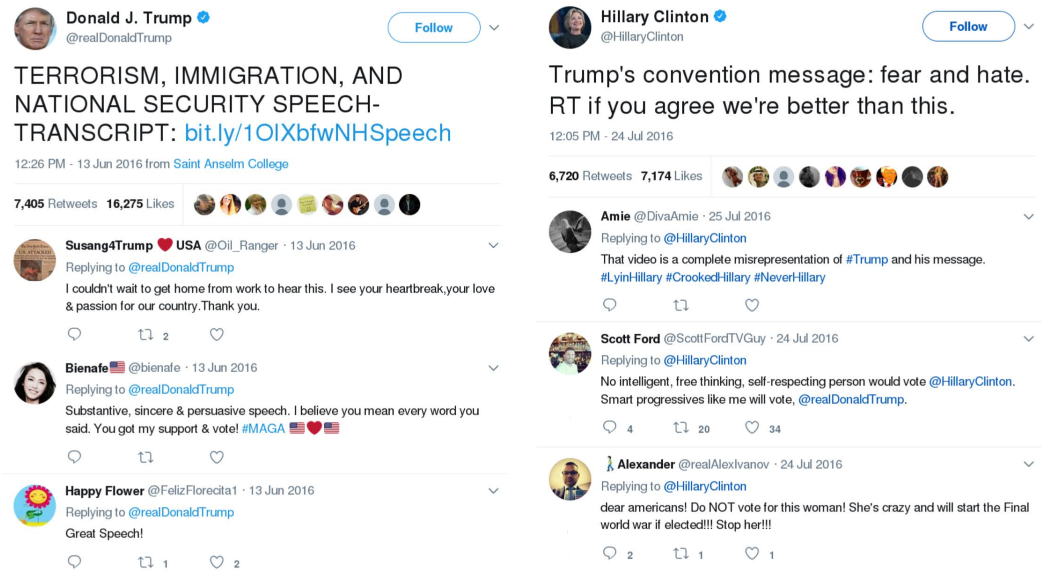
3.1 Introduction

Motivation. Social media, such as Twitter and large online forums, reflect grassroots opinions on controversial topics. Often, the resulting discussions take highly polarized forms where people either strongly support one stance or heavily oppose it. Politics is a prominent case: users inclined with either one of two parties engage in *adversarial discussions* over many months. Examples are the 2016 US Presidential Election campaign and the UK "Brexit" referendum.

A recent trend is that discussions also include original posts by the political stakeholders themselves, for example, Donald Trump and Hillary Clinton in the US, or Nigel Farage and Jeremy Corbyn in the UK. Figure 3.1 shows an example of tweets made by two of these politicians, along with the replies of support and opposition they garnered from other users.

Thus, regular users not only express their opinions, but also interact directly with political candidates and other leading figures. These interactions have distinct characteristics that have not been investigated in depth so far. Especially in light of the role of so-called "post-factual" statements (see, e.g., Wikipedia article on "Post Truth"), a fundamental study of these phenomena is needed.

Contribution. This chapter analyzes adversarial discussions on politics, as observed on Twitter over extended timeframes. We propose a general framework, based on latent topic models and user features, over a multi-faceted data space. The facets of interest are the *topics* of tweets, their *factuality* versus *sentimentality* (aka. post-factuality), the *inclination* of users with regard to the two involved stances ("us" and "them"), and the *roles of users* with regard to how they



Donald J. Trump @realDonaldTrump

TERRORISM, IMMIGRATION, AND NATIONAL SECURITY SPEECH-TRANSCRIPT: bit.ly/1OlXbfwNHSpeech

12:26 PM - 13 Jun 2016 from Saint Anselm College

7,405 Retweets 16,275 Likes

Susang4Trump ❤️ USA @Oil_Ranger · 13 Jun 2016
Replying to @realDonaldTrump
I couldn't wait to get home from work to hear this. I see your heartbreak,your love & passion for our country.Thank you.
2

Bienafe🇺🇸 @bienafe · 13 Jun 2016
Replying to @realDonaldTrump
Substantive, sincere & persuasive speech. I believe you mean every word you said. You got my support & vote! #MAGA 🇺🇸❤️🇺🇸

Happy Flower @FelizFlorecita1 · 13 Jun 2016
Replying to @realDonaldTrump
Great Speech!
1 2

Hillary Clinton @HillaryClinton

Trump's convention message: fear and hate. RT if you agree we're better than this.

12:05 PM - 24 Jul 2016

6,720 Retweets 7,174 Likes

Amie @DivaAmie · 25 Jul 2016
Replying to @HillaryClinton
That video is a complete misrepresentation of #Trump and his message. #LyinHillary #CrookedHillary #NeverHillary

Scott Ford @ScottFordTVGuy · 24 Jul 2016
Replying to @HillaryClinton
No intelligent, free thinking, self-respecting person would vote @HillaryClinton. Smart progressives like me will vote, @realDonaldTrump.
4 20 34

Alexander @realAlexIvanov · 24 Jul 2016
Replying to @HillaryClinton
dear americans! Do NOT vote for this woman! She's crazy and will start the Final world war if elected!!! Stop her!!!
2 1 1

Figure 3.1: Tweets by political stakeholders in the 2016 US Presidential campaign and subsequent replies they received.

affect activity in the discussions.

Within this framework we study two recent cases: the US Election and the UK Brexit. These cover more than a million tweets by several thousands of users over a period of ten and eight months, respectively. The two cases serve as examples to address the following general research questions about social media:

Question 1: What are the key topics of the adversarial discussion? Which topics are most polarized? Which topics are of factual nature, referring to political issues like jobs or immigration, and which ones are "post-factual", referring to subjective beliefs and sentiments?

Question 2: What are the roles played by different kinds of users? How strong is the influence of the leading figures themselves? Are there other, highly prolific, users who drive the adversarial opinions?

Although there is prior work on analyzing topic profiles and user influence in online communities, the outlined research questions address newly emerging phenomena that have not been studied before. The novel contributions of this chapter are 1) the methodology to systematically study adversarial discussions, and 2) insightful findings on the role of "post-factual" topics and the nature of influential "power users".

3.2 Related Work

Twitter Analyses. Social media like Twitter has been studied as a source for a wide variety of analyses. These aim to understand (and sometimes predict) the dissemination and virality of topics (e.g., [Ma et al., 2013, Grabowicz et al., 2016]), identify influential users (e.g, [Bakshy et al., 2011, Weng et al., 2010]) and characterize their behavior, study spatial and temporal patterns of trending topics and user activities (e.g., [Yuan et al., 2013, Choudhury et al., 2016]). Much of this prior work is based on latent topic models (e.g., [Zhao et al., 2011, Vosecky et al., 2014]), typically using variants of LDA [Blei et al., 2003] or word2vec [Mikolov et al., 2013] to analyze tweet content.

Polarized Topics. The task of identifying controversial topics and their polarized stances has received considerable attention in the literature. Prior work has largely focused on analyzing, modeling and predicting the political leaning of users (e.g., [Conover et al., 2011, Wong et al., 2016]). A fundamental approach to measuring the amount of controversy in social media discussions is presented in [Garimella et al., 2016] and further expanded in [Coletto et al., 2017] to cover the evolution of polarizing discussions. The work of [Vydiswaran et al., 2015] studied the role of echo chambers in biased discussions, and proposes countermeasures to polarization.

Political Campaigns. Closest in spirit to this study is the prior work on analyzing the 2012 US presidential election, based on Twitter data. [Wang et al., 2012] presents a tool for user sentiments in this context. Other studies on political campaigns or major incidents and their aftermaths have covered the 2008 German parliament election [Tumasjan et al., 2010], the 2012 US primaries [Mejova et al., 2013], the 2015 Scottish Independence referendum [Fang et al., 2015], and elections in developing countries [Ahmed et al., 2016]. [Le et al., 2017] presents an approach for gauging the slant of political news consumption on Twitter, according to the activity of Republican and Democrat-leaning users. Though general analytics such as [MonkeyLearn, 2016] have emerged, there are few in-depth analyses of social media discussions surrounding the 2016 US election campaign and the UK Brexit referendum.

3.3 Data Collection

Our datasets consist of discussions rooted on leading figures in the Brexit referendum and the 2016 US presidential election. For the first event, we identify politicians Nigel Farage and Boris Johnson as headliners of the "Leave" stance and Nicola Sturgeon and Jeremy Corbyn as main drivers of the "Remain" campaign. For the second event, we focus on then-candidates Hillary Clinton, the appointed candidate of the Democrat party, and Donald Trump, the candidate of the Republican party.

We collected all tweets posted to the official accounts of these politicians in 2016, as well as all the *replies* their tweets have received. Replies differ from the usual Twitter "mentions" in the sense that they are linked to a specific tweet, instead of linking to a user account. We consider only these reply threads (i.e., trees of tweets), and disregard tweets posted independently of the posts by the leading figures. An overview of our datasets is given in Table 3.1.

Note that the UK Brexit case had considerably fewer tweets, but still enough mass for an in-depth analysis. Also note that the notion of a user is syntactic: one user corresponds to one Twitter account. Some users, especially the leading figures themselves, may employ professional PR teams or pay other people to contribute on their accounts.

Stance / Leader	Clinton	Trump	Remain	Leave
#Posts	2,602	1,861	1,098	539
#Replies	586,335	549,799	101,193	72,190
#Users	153,786	146,255	35,504	27,941
Time Period	01-01-2016 to 15-11-2016		01-02-2016 to 01-10-2016	

Table 3.1: Twitter data on US election and UK referendum.

3.4 Factual and Post-Factual Topics

As a first dimension of the discussions, we start our analyses by looking into the topics brought up over the course of the UK referendum and US election campaigns. Here we are interested in

Topic	F/S	Salient Words
T0: pro Clinton	S	hillary, president, potus, imwithher, bernie, vote, berniesanders, love, clinton, trump, good, win, sanders, feelthebern, great, woman, hope
T1: contra Clinton	S	hillary, white, potus, house, liar, people, obama, black, lying, vote, clinton, woman, flotus, crooked, bill, corrupt, pandering, prison, billclinton
T2: contra Clinton	S	benghazi, neverhillary, hillary, liar, americans, crookedhillary, potus, hillaryforprison, maga, people, killed, die, america, lies, lockherup
T3: contra Clinton	S	hillary, trump, timkaine, lies, potus, usaneedstrump, lie, clinton, kaine, pence, truth, liar, video, lying, debate, mike_pence, crooked
T4: Social Issues	F	women, rights, care, health, pay, abortion, children, life, babies, hillary, woman, kids, support, change, gay, marriage, equal, healthcare, lgbt
T5: Gun Control	F	gun, vote, law, guns, potus, bernie, hillary, laws, berniesanders, party, voting, democrats, stop, nra, illegal, violence, control, amendment
T6: FBI	F	hillary, emails, fbi, clinton, potus, email, criminal, jail, wikileaks, server, investigation, classified, benghazi, lies, security, corruption
T7: Foreign Politics	F	money, hillary, clinton, foundation, wall, war, street, countries, millions, saudi, isis, foreign, iraq, russia, state, iran, obama, libya
T8: Economy	F	jobs, pay, money, taxes, tax, trump, people, class, debt, business, work, free, plan, middle, obama, raise, economy, wage, obamacare
T9: Bill Clinton	S	bill, women, hillary, rape, clinton, husband, trump, rapist, billclinton, child, victims, sexual, raped, victim, monica, girl, assault, wife
T10: Racism	F	trump, racist, hillary, people, hate, white, black, kkk, supporters, vote, support, donald, bernie, blacks, racism, anti, party, bigot, violence
T11: Hispanics	S	los, por, con, drudge_report_, hillary, para, una, presidente, usa, jillnothill, imwithher, clinton, pas, ser, pero, usted
T12: Trump Family	S	erictrump, melaniatrump, donaldjtrumpjr, ivankatrump, mike_pence, happy, love, donald, melania, laraleatrump, teamtrump, great, family
T13: Trump Scandal	F	tax, returns, account, trump, delete, release, taxes, show, donald, nevertrump, hiding, trumpdelete, fraud, records, money, liar
T14: Foreigners	F	muslims, muslim, wall, trump, illegal, country, isis, obama, islam, america, americans, build, immigrants, refugees, terrorists, illegals, border
T15: Media Bias	S	trump, cnn, media, hillary, polls, nytimes, poll, news, lies, people, truth, clinton, debate, donald, foxnews, facts, win, lie, rigged
T16: pro Trump	S	trump, cnn, foxnews, makeamericagreatagain, trump2016, megynkelly, trumptrain, fox, news, watch, debate, maga, donald, teamtrump, great
T17: pro Trump	S	trump, america, great, donald, president, vote, god, love, country, people, makeamericagreatagain, win, trump2016, bless, usa, good, maga
T18: Republicans	F	trump, cruz, ted, tedcruz, vote, rubio, gop, win, donald, jeb, people, jebbush, party, establishment, kasich, glennbeck, romney, bush, republican
T19: contra Trump	S	trump, man, donald, nevertrump, loser, good, people, nytimes, racist, cnn, big, sad, ass, president, tweet, stupid, hands, liar, orange

Table 3.2: Topics and top representative keywords identified by LDA for US Election data (F = factual, S = sentimental).

the thematic differences and similarities between issues addressed by proponents of either side.

To this end, we employ Twitter-LDA[1], an adaptation of the Latent Dirichlet Allocation model for topic discovery on tweets [Zhao et al., 2011]. For each dataset, we generate topics from the full corpus of tweets, with removal of stop words and embedded URLs. We set the model

[1] https://github.com/minghui/Twitter-LDA

Topic	F/S	Salient Words
T0: Referendum Day	S	leave, vote, brexit, nigel, ukip, remain, referendum, cameron, voted, country, hope, farage, campaign, voteleave, win, democracy, stay
T1: US Parallels	S	nigel, realdonaldtrump, good, hillaryclinton, brexit, farage, trump, ukip, boris, britain, country, luck, hope, day, love, god, independence
T2: pro Leave	S	boris, nigel, zacgoldsmith, brexit, london, farage, grassroots_out, ukip, daviddavismp, racist, change_britain, cameron, alllibertynews
T3: European Union	F	brexit, trade, leave, control, immigration, europe, free, ukip, world, borders, vote, britain, market, countries, deal, system, movement, economy
T4: Immigration	F	borders, europe, turkey, migrants, control, brexit, immigration, country, open, border, leave, countries, immigrants, free, british
T5: Foreign Politics	F	boris, foreignoffice, johnkerry, ukun_newyork, turkey, isis, syria, war, foreign, russia, stop, erdogan, assad, mfa_ukraine, ukraine, saudi
T6: Media Debates	S	david_cameron, nigel, cameron, brexit, itv, dave, truth, farage, man, head, bbc, people, debate, itvnews, ukip, dodgy, voteleave, lies
T7: Economy	F	tax, steel, david_cameron, money, industry, cameron, chinese, vote_leave, china, nigel, pay, tariffs, fishing, cheap, ukip, avoidance, labour
T8: UK	F	news, rights, human, year, foreign, aid, housing, law, article, nhs, account, build, homes, scotland, scotgov, money, labour, government, british
T9: Altruism	S	sharing, socialism, equal, virtue, failure, ignorance, envy, misery, philosophy, creed, gospel, tin, juice, women, edinburghpaper, snsgroup
T10: before Cameron	F	blair, war, tony, johnmcdonnellmp, hilarybennmp, lindamcavanmep, rcorbettmep, labour, emilythornberry, benn, karenpbuckmp, iraq, israel
T11: David Cameron	S	answer, question, cameron, david_cameron, corbyn, questions, jeremy, ireland, pmqs, northern, answers, scotland, david, north, labour, wales
T12: Healthcare	F	nhs, jeremy, minister, prime, labour, ttip, heidi_mp, doctors, great, corbyn, uklabour, junior, telegraphnews, david_cameron, support, health
T13: Public Services	F	money, public, nhs, labour, pay, steel, private, work, tax, government, train, rail, david_cameron, jeremy, energy, jobs, service, contracts
T14: Middle East	S	anti, labour, corbyn, ira, petermurrell, jeremy, uklabour, hamas, party, petition, israel, parliament, support, semitism, friends, jews, terrorist
T15: Khan election	S	sadiqkhan, happy, sad, jeremy, ruthdavidsonmsp, love, nicola, family, thesnp, hope, labour, thoughts, great, news, london, day, corbyn, peace
T16: Social Welfare	F	tax, labour, pay, money, workers, nhs, people, education, rights, tories, working, class, housing, poor, schools, rich, paid, work, disabled
T17: Scotland	F	scotland, thesnp, snp, nicola, scotgov, vote, scottish, independence, leave, scots, brexit, england, referendum, good, sturgeon, indyref2, scotparl
T18: pro Labour Party	S	labour, party, corbyn, vote, election, win, leader, uklabour, tories, tory, jeremy, resign, government, voters, voted, leadership, general, left
T19: pro Labour Party	S	jeremy, labour, corbyn, party, uklabour, leader, good, owensmith_mp, vote, support, members, resign, people, great, leadership, keepcorbyn

Table 3.3: Topics and top representative keywords identified by LDA for Brexit data (F = factual, S = sentimental).

hyperparameters as $\alpha = 2.5$, $\beta = 0.01$, $\gamma = 20$ and $N = 20$ topics. To evaluate the topic model for different choices of the dimensionality, we calculate the per-word perplexity for varying numbers of topics N. The lowest perplexity is found at $N = 11$, and only marginally increases for N up to 50. Thus, to tune N, we also consider the aspect of interpretability [Chang et al., 2009], based on human judgements. Feedback on our data shows that the choice of $N = 20$ topics leads to the clearest interpretation (while having near-minimum perplexity).

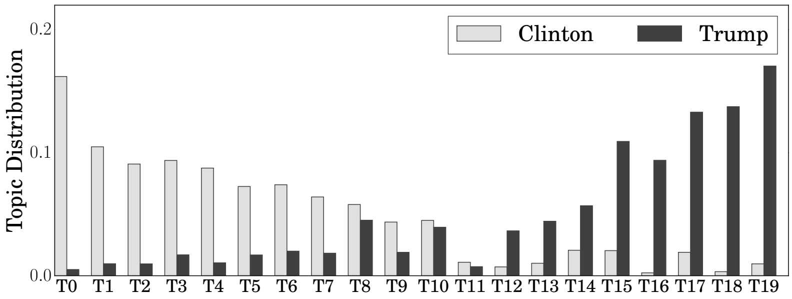
The discovered topics are displayed on Table 3.2 for the US Elections case and Table 3.3 for the UK Brexit case.

3.4.1 Factual vs. Sentimental Topics

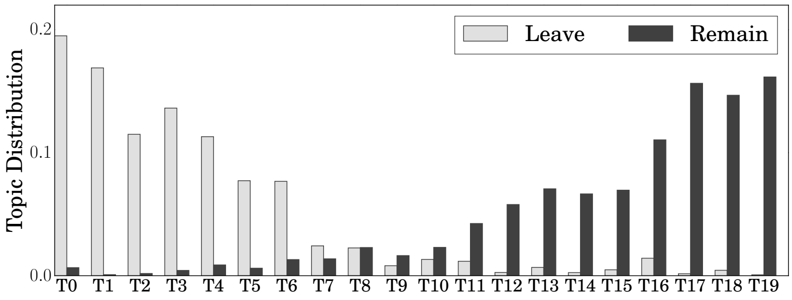
To derive further meaning from the topics, we employed the help of 10 judges to label them as *factual* or *sentimental*, where factual topics refer to concrete issues, facts, events and candidate agendas, while sentimental topics refer to personal opinions, emotional claims and speculation (aka. "post-factual"). Although some topics naturally include a mix of facts and opinions, we note a high agreement on their factuality, with 70% of topics receiving the same label from at least 8 out of the 10 judges, and an inter-annotator agreement (Fleiss' Kappa) of 0.42.

Contrasting the topical content of either side of the discussions, Figure 3.2 shows the distribution of topics across replies posted to Hillary Clinton and Donald Trump over the US Election campaign, and the Remain and Leave campaigners on the UK Brexit campaign.

In the US case, Clinton discussions display a wider topical spread, particularly across factual topics: while 16% of replies are sentimental messages of support (T0) and 29% are general criticism (T1, T2 and T3), factual topics such as those relating to gun control and foreign politics (T4, T5, T6, T7 and T8) each make up at least 5% of the replies. Meanwhile, replies to Donald Trump are largely sentimental: 10% of tweets are reactions to media coverage and preliminary



(a) US Election topics and replies to Clinton and Trump.



(b) Brexit topics and replies to Leave and Remain campaigners.

Figure 3.2: Distribution of LDA-generated topics over replies to leaders of either stance.

Label	Clinton	Trump	Remain	Leave
Factual	0.44	0.39	0.47	0.41
Sentimental	0.56	0.61	0.53	0.59

Table 3.4: Proportion of factual and sentimental replies to campaigners in the US Elections and Brexit.

poll results (T15), 22% express support (T16 and T17), and 17% criticism (T19). Topic T18, which incorporates terms relating to other Republican party members and the Republican primaries, is the main factual topic discussed, making up 18% of replies.

The Brexit case behaves similarly, with the Leave side displaying a narrower topical focus than its adversary. 48% of replies to the Leave side express pro-Leave sentiment (T0, T1, T2), while 25% address factual topics about aspects of the European Union and immigration (T3 and T4). On the Remain side, 30% of replies are devoted to pro-Labour party sentiment (T19 and T18), while 15% and 11% discuss Scotland (T17) and welfare issues (T16), respectively.

Replies on both sides of the campaigns are dominated by sentimental topics, and indeed more of such topics were detected for the US Election case. The overall distribution for each dataset is shown on Table 3.4.

3.4.2 Prominent Hashtags

The most popular hashtags in the discussions are primarily sentimental in nature and often among the salient words of the LDA-generated topics. Top hashtags #makeamericagreatagain, #trump2016 and #trumptrain, with over 30,000 combined uses, are captured in pro-Trump topics T16 and T17 of the US Election, while #crookedhillary and #neverhillary are picked up by contra-Clinton topic T2, and the #imwithher campaign motto features in pro-Clinton topic T0.

A potential exception to this pattern is #Brexit, which is picked up by factual topic T0, and may refer to the event itself rather than its endorsement. We find that the hashtag was nonetheless much more frequent on the Leave side, with 2,433 uses versus 685 on the Remain side.

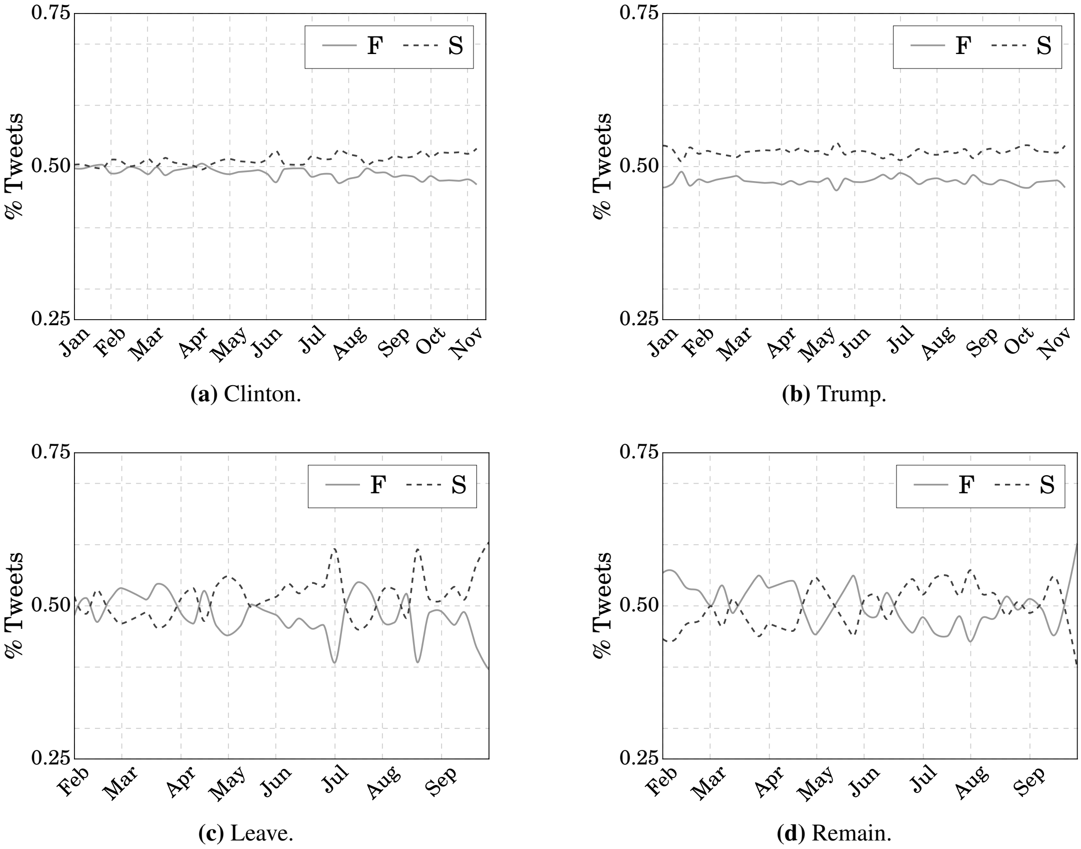
Interestingly, we find a frequent use of Trump-related hashtags in replies to Clinton, with hashtags #trump, #trump2016, #makeamericagreat and #trumptrain appearing more than 9,000 times. This phenomenon is not expressed in the opposite direction, i.e., there are hardly any Clinton hashtags in Trump threads).

This predominantly one-sided adoption of sentimental hashtags indicates that, though adversarial in nature, the opposing sides of the discussions are not often directly confrontational: topics referring to a particular candidate or stance, both favorably and unfavorably, are usually targeted at its stakeholders. This is particularly notable on pro and contra Clinton and Trump topics, as well as pro Labour Party topics. We also note that while unfavorable topics hint at a tendency to support the opposite stance, they do not necessarily convey this explicitly.

3.4.3 Evolving Topics

To understand the relationship between activity and topical focus, we also investigate the timeline for the LDA-based topics, grouped according to their factuality. Figure 3.3 shows the evolution of activity, in terms of the number of tweets, of factual topics (F) and sentimental topics (S), for both Clinton and Trump in the US Election case and the Remain and Leave sides of the Brexit case. Here we see a reflection of the overall topical distribution discussed previously, with a consistent predominance of sentimental topics throughout the campaigns. The Remain side of the Brexit discussion is again the exception, with a majority of factual topics on the weeks preceding and following the May 5 elections. The weeks following the announcement of the referendum (made on February 20) also saw an increase in the discussion of factual topics on the Leave side.

Both cases see an increase of activity for sentimental topics immediately after the end of the campaigns (i.e., the election on November 9 and the referendum on June 23). Even shortly before the election, discussions on Clinton's threads displayed a trend of growing sentimental content, following the reaction to new scandals surrounding the candidate. Meanwhile, Trump



(a) Clinton.

(b) Trump.

(c) Leave.

(d) Remain.

Figure 3.3: Timeline for factual (F) and sentimental (S) topic groups for Clinton (a) and Trump (b) on the US Election, and Leave (c) and Remain (d) sides of Brexit.

saw only a slight increase in sentimental tweets around the election itself.

For the Brexit case, Remain and Leave show spikes of sentimental activity which fueled the growing discussions following the decision. While this burst of activity quickly fades out on the Leave side, the Remain side exhibits a strong activity signal on sentimental topics for several weeks following the referendum. This reaction has been coined “Bregret”, for British regret, in the media.

3.5 The Power of Power Users

In this section, we turn our focus to the users involved in the discussions. In particular, we are interested in the role and influence of different kinds of users, as a function of their inclination towards either one of the two stances. We label each user according to:

- *Role*: user is either a leader (i.e., leading politician), power user, or regular user;
- *Inclination*: user leaning towards stance A or stance B.

In addition to the leading figures in the discussions (e.g., Clinton and Trump), we distinguish two other kinds of users, motivated by the observation that some accounts have a high activity level that makes them unlikely to be managed by single individuals. We suspect that some of these accounts represent entire teams, either professional PR teams or (paid or volunteering) workers.

To identify these, we obtained activity information from users’ Twitter profiles, including: i) account life time (in days) since its creation date, ii) number of tweets ever posted (not just within the discussion at hand), iii) number of users that the account follows, called *followees*. We manually inspected a random sample of the accounts and labeled 50 power users and 50 regular users as the training data with the above features. We then used libsvm[2] [Chang and Lin, 2011] to classify all other users. Using 5-fold cross-validation, we achieved an accuracy of 93% and 96% for the inclination and power user classification respectively in the US Election case, and 91% and 99% accuracy in the UK Brexit case. This high accuracy is in line with the significant disparity in the posting activity between regular and power users. Table 3.5 shows the break-down of users across these three roles.

These tables also show how the three user roles are distributed over the two inclinations. To determine these values, we again trained a binary classifier for user inclination with libsvm, using all original posts from leaders on both sides as positive and negative training examples. For each user, we concatenated all their tweets into a virtual document and fed this into the trained classifier. Interestingly, we see that the Remain side has twice as many power users as the Leave side, whereas in the US election case the number of power users is roughly the same for both sides.

[2] https://www.csie.ntu.edu.tw/ cjlin/libsvm/

Inclination	pro Clinton	pro Trump	Total
L	1	1	2
P	5,362	4,851	10,213
U	167,927	81,861	249,788
Total	173,290	86,713	260,003

(a) US Election

Inclination	pro Remain	pro Leave	Total
L	2	2	4
P	1,042	525	1,567
U	42,310	14,297	56,607
Total	43,354	14,824	58,178

(b) UK Brexit

Table 3.5: User roles and inclinations (L = leaders, P = power users, U = regular users).

3.5.1 Activity and Influence of Users

To assess the influence of users, we use two different metrics: i) their tweet activity in the scope of the adversarial discussion, and ii) the degree to which other users followed up on tweets by replying to them. Table 3.6 shows statistics for these metrics, for each of the US and UK cases. The first metric is given by the number of tweets made by each user category. The follow-up metric is given by #R2U: the number of replies from others in response to users in the different categories.

Table 3.6 shows that power users had a much higher share of activity in the Trump camp than in the Clinton camp. Trump-inclined power users were responsible for 12% of all replies to either candidate, whereas less than 3% of the replies were made by Clinton-inclined power users. The absolute numbers on the pro-Trump side are interesting as well: 134,000 tweets by power users and nearly 606,000 by regular users. This should be interpreted against the fact that Trump-initiated threads include a total of 550,000 tweets. This means there was a large number of pro-Trump tweets among replies to Clinton threads, and a substantial share of these were made by power users. In the reverse direction, this effect cannot be observed. As Figure 3.4 shows, power users play a more significant role in supporting Trump and Leave respectively.

The #R2U numbers in Table 3.6 confirm this interpretation, and furthermore show that the tweets by power users had additional influence by attracting lots of replies from others.

Compared to the US case, power users in the Brexit case were much less active and showed no indication of one side “hijacking” the other side’s posts. As our notion of power users is given by the account’s activity over its entire lifetime, rather than activity strictly within the adversarial discussion, this low activity profile is not entirely unexpected. Manual sampling

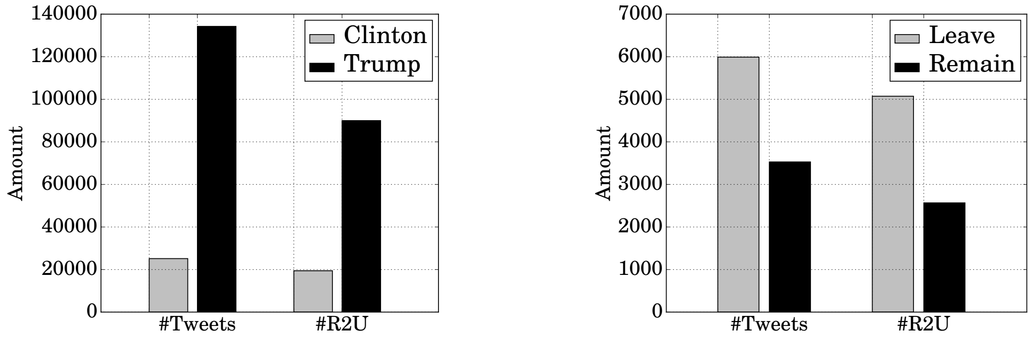
Inclination	pro Clinton		pro Trump	
	#Tweets	#R2U	#Tweets	#R2U
L	2,602	586,335	1,861	549,799
P	25,147	19,439	134,266	89,983
U	338,925	686,541	606,485	297,771

(a) US Election

Inclination	pro Remain		pro Leave	
	#Tweets	#R2U	#Tweets	#R2U
L	1,098	101,193	539	72,190
P	3,529	2,567	5,991	5,072
U	85,455	56,965	77,582	83,310

(b) UK Brexit

Table 3.6: Activity of users and their roles (L = leaders, P = power users, U = regular users).



(a) US Election.

(b) UK Brexit.

Figure 3.4: Activity of power users.

reveals long-lived accounts that were active on earlier or on different political topics, but seldom engaged in replying to one of the Remain or Leave leaders.

3.5.2 Combined View of Topics and Users

To conclude our analyses, we look into affinities between different user roles and the topics of discussion we identified in the previous section, with the goal of further investigating the impact of users in the themes and activity levels of the adversarial discussions.

In the US case, the most expressive topics for the leaders (Clinton and Trump themselves) are pro-candidate topics T0 and T17, encompassing 25% and 22% of tweets made by each respective candidate. In addition to these, topics T15 (Media Bias), T1 (contra-Clinton) and T19 (contra-Trump) also received considerable attention from regular users and together make up 37% of all their tweets.

Interestingly, the biggest differences between power users and regular users are also seen in pro- and contra-Trump topics T16 and T19, with the latter receiving more attention among power users: 8% of their tweets fall into topic T16, compared to 4% of tweets by regular users. In the opposite direction, while T19 is still well represented in tweets by power users, it receives the most attention from regular users and ranks as the most expressive topic for this user group. A similar pattern can be seen in contra-Clinton topics, which receive a slightly smaller share of activity from power users. Thus, activity on pro- and contra-candidate topics suggests that regular users tended to engage in more critical discussions about each party, while power users and leaders were mostly concerned with endorsing or promoting either side.

In the UK case, T0 (Referendum day) is among the strongest topics for all three user categories, accumulating 27% of all tweets made by the Leave campaign leaders, 9% of tweets made by power users and 8% of tweets by regular users. In contrast, less than 1% of the Remain campaign leaders' posts feature in this topic, with most of their activity going into topics T17 (Scotland), T15 (Sadiq Khan's mayoral election) and T19 (pro-Labour party). These are more closely related to the political leaders themselves, as well as the other political events they were involved in, than to topics pertaining to the referendum and its implications. We recall from Section 3.4 that while such factual topics were discussed by both sides of the campaign and by both user groups, no explicit Pro-Remain topic could be identified from the dataset.

Pro-Leave topic T2 saw the largest difference of activity, encompassing 9% of tweets by power users and less than 5% by regular users. As in the US case, such topics expressing support for one side of the campaign tend to be most polarizing, not only in sentiment but in the attention they receive from different user groups.

3.6 Conclusion

In this chapter, we analyzed the Twitter discussions on the 2016 US Election and the UK Brexit as instances of a general model of adversarial discussions on social media. Key insights include our observations on the strength of factual and sentimental (i.e., "post-factual") topics and the notable role and influence of power users. In particular, the US case showed that power users of one side can jump on posts in the opposing side's threads and attract significant follow-up by other users. Such effects were not visible in the UK case.

Future work involves extending our initial findings on the evolution of other adversarial discussions around political events, such as the continued effect of Brexit and upcoming elections around the world. These would allow the investigation of other common and contrasting facets of the discussions, such as the impact of different demographics and public response to the aftermath of political decisions.

4

Traits and Anomalies of Political Discussions on Reddit

Contents

Reddit hosts a number of communities dedicated to local or global politics. Users can contribute to discussions in these communities by submitting relevant news articles and by posting comments in response to these articles or in response to other users' posts. Unlike Twitter and a majority of other social media platforms, posts on Reddit can receive positive and negative feedback from users, which are then used to internally curate a discussion. In this chapter, we describe how we can leverage this explicit community feedback mechanic to distinguish between different types of posts, and propose four conversational archetypes that arise from the presence of these posts throughout a given discussion.

Section 4.3 provides an overview of Reddit and its feedback mechanism, and introduces the notion of *X-posts*, posts which receive significant negative feedback. In Section 4.4, we devise a feature space that further expresses key elements of individual posts and discussions, including the sentiments and topics they convey. These are used in Section 4.5 to characterize and contrast different discussion archetypes that emerge from the occurrence of X-posts. The insights we gained from this analysis are summarized in Sections 4.6 and 4.7.

4.1 Introduction

Motivation. Discussions in online communities, such as Reddit and Twitter, 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 chapter, 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 chapter, 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. 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.

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 et al., 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]).

Recent studies by different groups devised pattern-based characterizations of controversial discussion threads [Coletto et al., 2017, Garimella et al., 2018, Glenski and Weninger, 2017, Zhang et al., 2018b] 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 conversation: users, sentiments, and topics.

Approach and Hypotheses. The unique element in our approach to understanding political discussions in 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, Politics (`reddit.com/r/politics`) and World News (`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. This chapter's salient contributions are:

- We introduce a pattern-based model of different archetypes of online discussions, refining the established notion of controversy into disputes, disruptions, and discrepancies.

- We present the first study of these archetypes by jointly examining 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.

4.2 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 et al., 2017b].

[Liang, 2017] studies the relationship between post scores, participating users, and thread structure in the Q&A sub-reddit, TechSupport. [Zayats and Ostendorf, 2018] tackles 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] monitors the browsing behavior of Reddit users to predict future interactions based on users' voting habits and page-browsing activities.

[Zhang et al., 2018b] studies reply-trees on Facebook in combination with user-user interactions. The authors derive 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 is not considered in this work at all.

[Zhang et al., 2017] develops 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 notes patterns of disagreement chains, particularly in debate-oriented forums such as Political Discussion, but not so in the Politics subreddit.

[Weninger et al., 2013] studies the progression of topics in Reddit threads based on a hierarchical latent model.

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

[Cheng et al., 2015] characterizes antisocial behavior by studying the history of banned accounts in the comment section of three news sites. The resulting features are used to predict whether a user will likely be banned in the future. Subsequent work [Cheng et al., 2017] also investigates 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 leverages the network structure surrounding specific hashtags to quantify the degree of controversy for a given hashtag.

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

4.3 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[1], 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[2].

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.

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

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

4.3.1 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.

4.3.2 Dataset

Our first dataset comes from the Politics subreddit (reddit.com/r/politics), a forum for "current and explicitly political US 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[3]. 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 World News subreddit (reddit.com/r/worldnews), where submission guidelines are similar to those in the Politics subreddit (external links to recent news articles), but specifically excludes US-related news.

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 4.1.

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
World News	24,278	3,727,955	352,055	1,260,515

Table 4.1: Politics and World News subreddit datasets.

4.4 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

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

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.

4.4.1 X-Posts and Normal Posts

In the previous section, 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 World News 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 World News) than normal posts ($M = 1.11, SD = 2.10$ and $M = 1.13, SD = 3.09$)[4], ($p < 0.001$).

We also find that X-posts and normal posts can both be "controversial" with regards to their mentions of controversial issues. To determine this, we compiled a list of phrases related to controversial issues from Wikipedia[5], 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.

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 World News ($p < 0.001$), 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.

4.4.2 Sentiments

As a measure of the sentiments expressed throughout discussions, we evaluate the language used in each post in our datasets using VADER [Hutto and 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

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

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

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 World News 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 World News) 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.

Finally, 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 World News subreddit.

4.4.3 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 into 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 World News 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, **sim($\boldsymbol{p_i}, \boldsymbol{p_j}$)** to be the maximum cosine similarity[6] 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, **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 World News) than posts with low similarity (with similarity scores below the $25th$ percentile) ($M = 0.96, SD = 1.50$ and $M = 0.91, SD = 2.16$).

4.4.4 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 World News.

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 their 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 World News. These users are responsible for 14.7% and 12.9% of all posts (both normal and X-posts) in each respective community.

4.5 Path Patterns

In this section, we turn our focus to the Harmony, Discrepancy, Disruption and Dispute conversational patterns, defined according to how X-posts feature in 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[7] samples.

[6] 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.

[7] iid = identically independently distributed

Pattern	#Paths in Politics	#Paths in World News
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 4.2: Number of path samples for each pattern.

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 4.2 lists the number of sampled paths that fall into each of the path pattern categories.

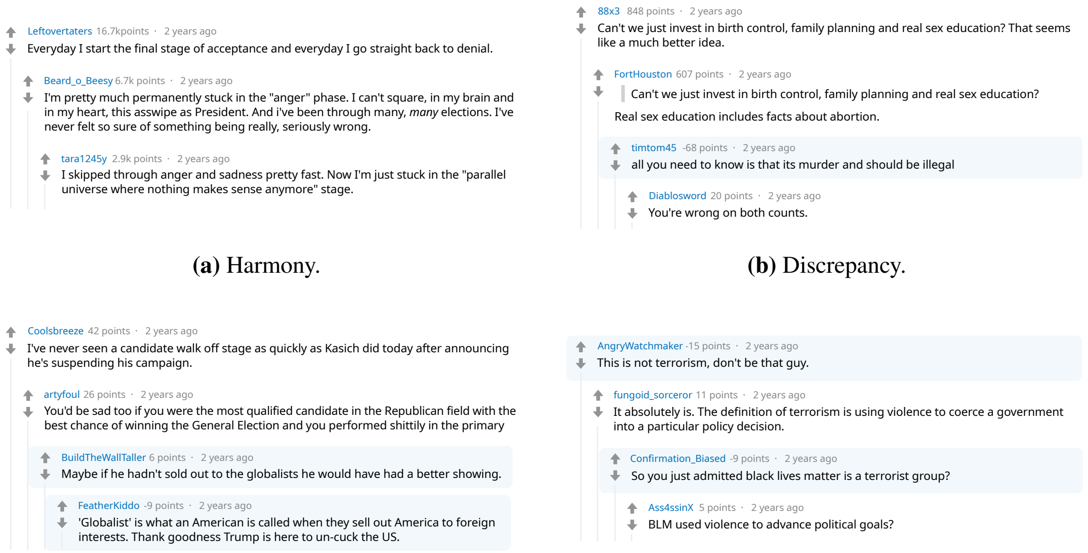
In the following, we express our expectations about each of these patterns as hypotheses and use statistical tests to evaluate how they are expressed in the sentiment, topic, and user dimensions. When examining the role of X-posts in specific path patterns, we employ Student's t-tests to compare them 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 4.3, at the end of this section, shows a summary of our findings.

4.5.1 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 represent agreements, or at least balanced debates, without extreme sentiment polarization. Figure 4.1a 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



Leftovertaters 16.7kpoints · 2 years ago
Everyday I start the final stage of acceptance and everyday I go straight back to denial.

Beard_o_Beesy 6.7k points · 2 years ago
I'm pretty much permanently stuck in the "anger" phase. I can't square, in my brain and in my heart, this asswipe as President. And i've been through many, *many* elections. I've never felt so sure of something being really, seriously wrong.

tara1245y 2.9k points · 2 years ago
I skipped through anger and sadness pretty fast. Now I'm just stuck in the "parallel universe where nothing makes sense anymore" stage.

(a) Harmony.

88x3 848 points · 2 years ago
Can't we just invest in birth control, family planning and real sex education? That seems like a much better idea.

FortHouston 607 points · 2 years ago
> Can't we just invest in birth control, family planning and real sex education?

Real sex education includes facts about abortion.

timtom45 -68 points · 2 years ago
all you need to know is that its murder and should be illegal

Diablosword 20 points · 2 years ago
You're wrong on both counts.

(b) Discrepancy.

Coolsbreeze 42 points · 2 years ago
I've never seen a candidate walk off stage as quickly as Kasich did today after announcing he's suspending his campaign.

artyfoul 26 points · 2 years ago
You'd be sad too if you were the most qualified candidate in the Republican field with the best chance of winning the General Election and you performed shittily in the primary

BuildTheWallTaller 6 points · 2 years ago
Maybe if he hadn't sold out to the globalists he would have had a better showing.

FeatherKiddo -9 points · 2 years ago
'Globalist' is what an American is called when they sell out America to foreign interests. Thank goodness Trump is here to un-cuck the US.

(c) Disruption.

AngryWatchmaker -15 points · 2 years ago
This is not terrorism, don't be that guy.

fungoid_sorceror 11 points · 2 years ago
It absolutely is. The definition of terrorism is using violence to coerce a government into a particular policy decision.

Confirmation_Biased -9 points · 2 years ago
So you just admitted black lives matter is a terrorist group?

Ass4ssinX 5 points · 2 years ago
BLM used violence to advance political goals?

(d) Dispute.

Figure 4.1: Examples of paths following each path pattern (X-posts are highlighted).

$F(4, 110142) = 336.688, p < 0.001$, $\eta^2 = 0.012$ for World News), 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 World News 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 4.1a are examples of posts that would be considered negative and off-topic by our toolset, but which are highly upvoted by the community.

4.5.2 Discrepancy

Discrepancies represent paths where a single post has received a negative or mixed response from the community. Figure 4.1b 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 from 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 World News 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 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 World News. 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.

4.5.3 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 4.1c. 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 World News 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 World News).

Nonetheless, the majority of Disruption paths contain at least one X-post written by an X-user (65% on Politics and 68% on World News), 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.

4.5.4 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 4.1d 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, we 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

Hypothesis	Politics	World News
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 4.3: 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.

Disruptions). This highlights the fact that disputed conversations are often short-lived.

4.6 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 World News 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 World News 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.

4.7 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.

Future work could 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 surrounding controversy.

5

Predicting Controversial Contributions on Reddit

Contents

In the previous chapter, we systematically analyzed how discussions take shape in the presence of X-posts. In this chapter, we turn our focus to X-posts themselves, with the goal of identifying elements in the early stages of a discussions that may lead to the occurrence of X-posts, and how their presence impacts the development of future discussions. Acknowledging the fact that X-posts may embody different characteristics according to context, we extended our analyses to communities dedicated to sports and personal relationships, and we contrast these with the political communities we studied previously.

Section 5.3 revisits the definitions of X-posts and discussion paths given in Chapter 4 and investigates how these are represented in our expanded datasets. In Section 5.4, we propose a new feature space that captures key elements of discussion threads, from their activity levels, to the sentiments being expressed in them, to their main topics and how they evolve throughout the discussion. We then use these features to build logistic regression classifiers that try to predict, based on the initial elements of a discussion, whether it will eventually contain an X-post. The classifiers and their results are presented in depth in Section 5.5, while Section 5.6 discusses their shortcomings and potential extensions.

5.1 Introduction

Motivation. Detecting, analyzing and characterizing sentiments, bias, and controversy in online discussion forums has been a major research topic for years (see, e.g., [Kumar et al., 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., 2018a, Liu et al., 2018], hate-speech [Davidson et al., 2017, Mondal et al., 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.

Approach and Contributions. In this chapter, 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 in Chapter 4.3. 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 controversy 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.

To understand the content signals that lead to an X-post within a discussion, 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.

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.

5.2 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., 2018a] 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 et al., 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 et al., 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 tags 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 interactions 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.

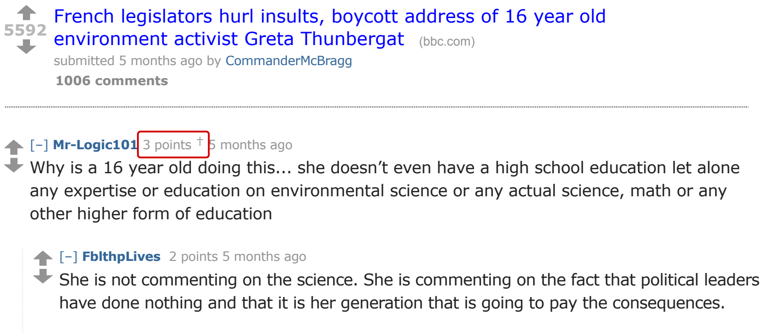
5.3 Data Modeling and Analysis

As explained in Chapter 4.3, 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 comments*. As users reply to these comments, entire discussion *trees* unfold, sometimes comprising a large number of user posts (hundreds or more) and going into considerable depth. 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 5.1 and 5.2 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.



5592

French legislators hurl insults, boycott address of 16 year old environment activist Greta Thunbergat (bbc.com)

submitted 5 months ago by CommanderMcBragg

1006 comments

[–] Mr-Logic101 3 points † 5 months ago

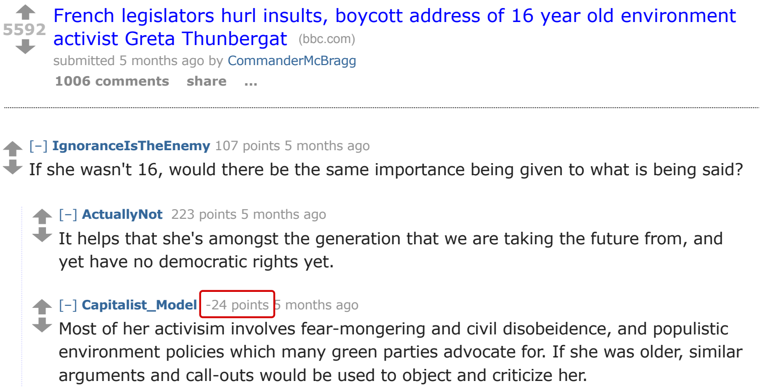
Why is a 16 year old doing this... she doesn’t even have a high school education let alone any expertise or education on environmental science or any actual science, math or any other higher form of education

[–] FblthpLives 2 points 5 months ago

She is not commenting on the science. She is commenting on the fact that political leaders have done nothing and that it is her generation that is going to pay the consequences.

Figure 5.1: Submission and posts from the World News subreddit, with X-post marked by the typographical dagger.

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



French legislators hurl insults, boycott address of 16 year old environment activist Greta Thunbergat (bbc.com)
submitted 5 months ago by CommanderMcBragg
1006 comments share ...

[–] IgnoranceIsTheEnemy 107 points 5 months ago
If she wasn't 16, would there be the same importance being given to what is being said?

[–] ActuallyNot 223 points 5 months ago
It helps that she's amongst the generation that we are taking the future from, and yet have no democratic rights yet.

[–] Capitalist_Model -24 points 5 months ago
Most of her activisim involves fear-mongering and civil disobeidence, and populistic environment policies which many green parties advocate for. If she was older, similar arguments and call-outs would be used to object and criticize her.

Figure 5.2: Submission and posts from the World News subreddit, with X-post indicated by upvote/-downvote difference of -24.

5.3.1 Definitions

We build on the definitions outlined in the previous chapter 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.
- 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**.

5.3.2 Datasets

Our datasets comprise content from four prominent subreddits: Politics (`reddit.com/r/Politics`), World News (`reddit.com/r/WorldNews`), Relationships (`reddit.com/r/Relationships`), and Soccer (`reddit.com/r/Soccer`). On the first two communities, posting guidelines dictate that all submissions must be links to external news articles of

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

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[3] (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 5.1. The distribution of posts that fall under the definition of an X-post in each of the datasets is shown in Table 5.2.

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
World News	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 5.1: Subreddit datasets.

5.3.3 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
- NX: paths that contain only normal posts but are part of a tree containing at least one X-post

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

Source	Year	Controversial	≤ -4 Points	Both
Politics	2016	150,456	95,841	19,574
	2017	23,642	34,917	3,570
Politics	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 5.2: Number of posts that satisfy each criterion for the definition of an X-post.

Source	Year	N	NX	X
Politics	2016	71,898	129,497	117,738
	2017	33,507	37,560	32,375
World News	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 5.3: Number of sampled paths belonging to the N, NX, and X categories.

- 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 frequencies 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 5.4.

While popular entities are frequent across both X and N categories, the ratios do 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

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)
World News	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 5.4: Top 20 entities with highest X/N ratio of occurrence frequencies.

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.

5.4 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 5.5.

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

Notation	Definition
$frac\_pos$, $frac\_neg$, $frac\_neu$	$\frac{\#\{positive, negative, neutral\} posts}{\#posts}$
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 5.5: Feature summary.

the text similarity of two posts p_i and p_j, $\mathbf{sim}(\boldsymbol{p_i}, \boldsymbol{p_j})$ 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 $\mathbf{sim}(\boldsymbol{p_i}, \boldsymbol{p_0})$, as well as the similarity between the original submission and posts in a path, $\mathbf{sim}(\boldsymbol{p_i}, \boldsymbol{s})$.

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.

5.5 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 section 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 5.6. 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 5.7. For each dataset, we present precision (true positives/(true positives+false positives)), recall (true positives/(true

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 5.6: 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 5.7: Prediction results for the X-post class.

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 5.1 and 5.2 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.

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.

5.5.1 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 5.8. 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.

5.5.2 X-post Entities

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

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

Table 5.8: Feature weights.

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 5.4). This result indicates that discussions involving religious topics often evolve in a fairly civilized manner – a good sign

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

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 5.4.

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.

5.5.3 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 5.9.

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.

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.

5.6 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[4] tool and VADER with LIWC[5].

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.

5.7 Conclusion

In this chapter, 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

[4] spacy.io/

[5] liwc.wpengine.com/

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.

6

Comparing Health Forums

Contents

In this chapter, we de-emphasize the aspect of controversy to focus more generally on online discussions and the surrounding context provided by the communities supporting them. For this, we examine discussions in communities dedicated to health topics and patient care. Beyond proving valuable information about medical conditions from a patient's perspective, these communities highlight the importance of online discussions for the exchange of personal experiences and mutual support.

Section 6.3 presents an overview of Health Boards and Patient, the two health-centric communities we examine alongside three Reddit communities dedicated to specific medical conditions. Drawing from our methodology in previous chapters, Section 6.4 characterizes and contrasts the discussions in each of these communities with regard to the intensity of user engagement, the explicit coverage of salient medical entities, and the degree of medical detail expressed by mentions of specific drugs and their dosages. The key findings of our analyses are summarized in Section 6.5.

6.1 Introduction

Motivation. Health discussion forums allow patients and caregivers to seek information and share experiences on medical conditions. They are often a starting point for medical questions by patients interested in checking symptoms and risk factors, and wishing to learn from others who have gone through similar experiences. This information exchange and mutual support is especially relevant for patients with chronic illness and possible complications, and for conditions that involve lifestyle changes. In addition, medical professionals occasionally join the discussion to provide advice, but first-hand accounts of health-issue experiences are valuable for both patients and professionals [Choudhury and De, 2014, Ma et al., 2018].

Contribution. This chapter analyzes and compares three popular health communities: the US-based Health Boards (`healthboards.com`), the UK-based Patient (`patient.info`) and specific health-related subreddits (e.g., `reddit.com/r/diabetes`). The former two are forums exclusively focused on health topics, whereas subreddits are specialized communities within the Reddit social media platform and therefore are typically more diverse in coverage, with personal support being an important component.

Our goal is to contrast the three forums on the principal dimensions of user engagement, salient entities like symptoms or risk factors, and medical detail about specific drugs and their dosages. For instance, a possible hypothesis is that subreddit discussions are more about personal stories whereas the dedicated forums go deeper into medical issues such as specific drugs and their side-effects. The chapter centers on the following research questions:

- **RQ1**: What is the intensity of engagement from users in each community?
- **RQ2**: What are the salient entities, like symptoms, treatments, side-effects and risk factors, reported in the three forums, and are there significant differences between forums in some of these aspects?
- **RQ3**: When discussing treatments, to what extent are specific drugs and drug dosages covered in each community?

Our analysis is based on three representative samples of wide-spread and intensively covered conditions: high blood pressure (hypertension), depression and diabetes. These are chosen as they involve both treatment with medical drugs and concerns about lifestyle issues (both as risk factors and as effects).

6.2 Related Work

There is ample prior work on analyzing and utilizing online content about patients, but the focus is mostly on scientific publications (Pubmed etc.) or clinical records (see, e.g., [Koopman and Zuccon, 2019] and references given there).

Health forums have received less attention; prior work includes examining the role of caregiver support [Hamm et al., 2013], querying and QA for effective search [Nobles et al., 2020, Terolli et al., 2020], and the spread of misinformation [Ghenai and Mejova, 2018, Bal et al., 2020]. The influence of cultural background on how patients express themselves in Talklife, regarding medical vs. lay-user terminology, is investigated by [Pendse et al., 2019].

Specialized health forums, such as TuDiabetes, are studied by [Mamykina et al., 2015, Litchman et al., 2018]. An example of studying specialized communities of cancer patients is the work by [Levonian et al., 2020].

General-purpose social-media platforms also host health discussions. For example, [Dirkson et al., 2019] leverages data from Facebook and Reddit for cancer patients to create language models for user posts. The role of user engagement in discussions about schizophrenia on Twitter was studied by [Ernala et al., 2018]. Another health condition that received attention is depression, for which prior work aimed to detect early indicators of potential self-harm and harm prevention [Yates et al., 2017, Wadden et al., 2021].

6.3 Data Collection

For this comparative study, we collected discussion threads from three kinds of forums:

1. Health Boards, a large community with message boards for over 200 different topics (healthboards.com/boards),

2. Patient, a UK-based forum covering several topics, from specific diseases to general wellness (patient.info/health),

3. Subreddits focused on the topics of blood pressure (hypertension), depression and diabetes: reddit.com/r/BloodPressure and reddit.com/r/Hypertension, reddit.com/r/Depression, reddit.com/r/Diabetes,[1].

We collected all publicly available posts up to 4 April 2020, using a web scraper for the first two forums and Reddit API querying for the four subreddits. Statistics on these datasets are given in Table 6.1. The whole corpus is preprocessed using efficient NLP tools to detect medical entities as described in Section 6.4.2.

Source	Blood Pressure			Depression			Diabetes		
	#Threads	#Posts	#Users	#Threads	#Posts	#Users	#Threads	#Posts	#Users
Health Boards	4,144	26,280	3,545	6,650	46,243	7,992	2,383	12,392	2,548
Patient	910	8,502	66	6,243	72,689	6,849	545	4,440	682
r/Hypertension	482	1,978	504	-	-	-	-	-	-
r/BloodPressure	720	3,033	789	-	-	-	-	-	-
r/Depression	-	-	-	709,116	2,300,273	378,626	-	-	-
r/Diabetes	-	-	-	-	-	-	57,216	622,385	32,358

Table 6.1: Total number of threads, posts and users, and average of posts per thread and users per thread for each dataset.

[1] All health forums last accessed on July 21, 2021.

6.4 Analysis and Comparison of Health Forums

In this section, we examine the key characteristics of discussions in terms of user engagement, salient topics and medical detail.

When comparing observed frequencies between different forums or different conditions, we ensure statistical significance by a Chi-Squared test reporting the chi-squared value and p-value. When comparing the mean values of different observations, we employ a one-way Anova test reporting the F-test statistic and p-value. Each Anova test is followed by a Games-Howell post-hoc test to show differences between pairs of observations.

6.4.1 RQ1: What is the intensity of engagement from users?

As a first measure of engagement, we compare the *lengths of discussion* threads in each community, given by the number of replies per initial post. For all three conditions, threads on Patient are significantly longer than on Health Boards and Reddit ($Hypertension \rightarrow F(2,5533) = 1214.8$, $p < 0.05$; $Depression \rightarrow F(2,60141) = 1020.1$, $p < 0.05$; $Diabetes \rightarrow F(2,60141) = 1020.1444, p < 0.0$). Reddit threads are the shortest on Hypertension and Depression; only the diabetes subreddit has longer threads than Health Boards. This suggests that despite the much larger post volume in Reddit, there is a major point for dedicated health forums where users engage in more intensive exchange of experiences.

A similar observation holds for the frequency of initial posts that receive *no replies* at all. Around 35% of submissions to the health subreddits under consideration received no replies. However, this should not be overinterpreted, as even specialized subreddit communities exhibit high user fluctuation, wide topical diversity and possible digression.

The third measure is the number of *distinct users* who participate in a thread. In this regard, subreddits show the largest numbers, and the Patient forum shows the lowest ($Hypertension \rightarrow F(2,5533) = 1485.4$, $p < 0.05$; $Depression \rightarrow F(2,722006) = 10106.3$, $p < 0.05$; $Dia betes \rightarrow F(2,60141) = 1660.6$, $p < 0.05$). This indicates that Patient users are more likely to repeatedly contribute to a discussion, whereas Reddit users often give merely a single reply.

Key findings. Overall, despite the higher total activity on Reddit, the two dedicated health forums show higher intensity of user engagement. Threads on Health Boards and Patient are longer, more likely to get at least one reply, and users are more likely to participate several times in the same discussion.

6.4.2 RQ2: What are the salient topics of each community?

Entity detection method. To detect mentions of medical entities in each community, we used the method by [Siu et al., 2013], which is an efficient NLP tool for annotating biomed-

Category	**Blood Pressure**	**Diabetes**	**Depression**
Symptoms	Headaches	Weight Change, Fatigue	Anxiety, Insomnia, Fatigue, Suicidal Thoughts
Risk Factors	Smoking, Stress, Salt, Alcohol	Obesity, Cholesterol, Family History	Alcohol, Stress, Anxiety
Complications	Heart Problems, Stroke	Eye Damage, Foot Damage	Anxiety, Panic Disorder, Suicidal Thoughts
Treatments (lifestyle)	Eating, Diet, Exercise	Eating, Exercise	Cognitive Behavioral Therapy, Exercise
Treatments (drugs)	ACE Inhibitors, Beta Blockers, Diuretics	Insulin, Metformin	SSRIs, SNRIs

Table 6.2: Frequent entities and entity categories.

Condition	**Entity Group**	**Health Boards**	**Patient**	**Reddit**	**Statistical Test**
Hypertension	Symptom	0.1325	0.1637	0.0208	$\chi^2(2) = 137.87,\ p < 0.05$
	Risk Factor	0.2245	0.3374	0.0715	$\chi^2(2) = 231.89,\ p < 0.05$
	Drug	0.4262	0.4549	0.0599	$\chi^2(2) = 584.19,\ p < 0.05$
	Lifestyle	0.1477	0.1901	0.0391	$\chi^2(2) = 125.40,\ p < 0.05$
	Complication	0.1663	0.2176	0.0349	$\chi^2(2) = 167.27,\ p < 0.05$
Depression	Symptom	0.2156	0.1714	0.0014	$\chi^2(2) = 99752.77,\ p < 0.05$
	Risk Factor	0.0639	0.0681	0.0032	$\chi^2(2) = 11471.34,\ p < 0.05$
	Drug	0.5072	0.2601	0.0001	$\chi^2(2) = 299327.65,\ p < 0.05$
	Lifestyle	0.0313	0.137	0.0001	$\chi^2(2) = 89432.01,\ p < 0.05$
	Complication	0.1803	0.3167	0.0047	$\chi^2(2) = 78973.72,\ p < 0.05$
Diabetes	Symptom	0.0831	0.1725	0.0022	$\chi^2(2) = 11418.62,\ p < 0.05$
	Risk Factor	0.0827	0.0606	0.0013	$\chi^2(2) = 10149.68,\ p < 0.05$
	Drug	0.4683	0.5248	0.0337	$\chi^2(2) = 16663.06,\ p < 0.05$
	Lifestyle	0.3093	0.3541	0.0154	$\chi^2(2) = 16013.34,\ p < 0.05$
	Complication	0.0869	0.1321	0	$\chi^2(2) = 112837.64,\ p < 0.05$

Table 6.3: Comparing frequencies of entity categories.

ical text and maps mentions to UMLS (Unified Medical Language System). From the top 100 most frequent entities in each community and for each condition, we compile 5 categories of entities: symptoms, risk factors, complications, treatments related to lifestyle, and drug treatments. For this grouping, we used disease-specific pages of the Mayo Clinic Portal (mayoclinic.org/diseases-conditions) as a guideline. Typical entities for each category are shown in Table 6.2, and the relative frequencies of the entity categories in each forum are shown in Table 6.3. To understand how these categories are featured, we drill down into each of the three conditions.

Blood Pressure. Though the condition is often asymptomatic, headaches are a frequently mentioned symptom in all forums. Among risk factors, (high consumption of) salt is frequent in

Reddit and Patient, while anxiety features strongly in Health Boards.

Not unnaturally, a good fraction of users appear to suffer from several chronic diseases, like hypertension and diabetes. Discussions with both conditions co-occurring exhibit notable differences between forums: users on Patient talk more often about nutrition than drugs, while Health Boards users focus more on drugs such as Metformin.

Depression. Across all three communities, depression symptoms like insomnia, fatigue and suicidal thoughts, appear in the most frequent entities. These terms can refer to symptoms, risk factors or complications alike; thus it is hard to differentiate between occurrences referring to post-diagnosis treatment or pre-diagnosis advice seeking.

Treatment plans often involve the use of antidepressants and cognitive behavioral therapy (CBT). The Patient forum contains significantly more mentions of CBT, up to twice as much as both Health Boards and Reddit.

Diabetes. We observed that users on Health Boards often talk about drugs, whereas Patient users have more discussion on lifestyle behavior such as nutrition and exercising.

The same trend shows up when comparing how users discuss symptoms like fatigue, thirst etc. Health Boards shows high co-occurrence frequencies of such symptoms with mentions of drugs, whereas Patient has them more correlated with terms like nutrition or exercise.

Key findings. Health Boards and Patient have a more clinical focus than Reddit, with much stronger coverage of treatment by drugs, across all three conditions. The focus of subreddits is mostly on symptoms (probably before diagnosis), risk factors (for complications) and also lifestyle issues (for prevention as well as treatment). This fits well with the broader themes and more diverse users of Reddit forums in general, whereas the two specialized communities appear to be centered on patients that are already under treatment by doctors. Between Health Boards and Patient, the fractions of drug coverage are similar, except for depression, where Health Boards has significantly higher values (to be revisited under RQ3).

6.4.3 RQ3: How much medical detail is given in each community?

The discussion of RQ2 showed that medical drugs are frequently mentioned, mostly in Health Boards and Patient (and to a much lower degree in the subreddits). We drilled down into which specific drugs or drug families are prevalent for each of the three conditions, and to what extent drug dosages are discussed as well.

For diabetes, unsurprisingly, mentions of Insulin and Metformin are prevalent across all forums ($HealthBoards \rightarrow F(1, 62160) = 240.242,\ p < 0.05$; $Patient \rightarrow F(2, 10920) = 12.251,\ p < 0.05$; $Reddit \rightarrow F(2, 12020) = 10267.376,\ p < 0.05$). For depression, drug mentions are dominated by the SSRI (Selective Serotonin Reuptake Inhibitor) family which includes Zoloft, Prozac, Lexapro and others ($HealthBoards \rightarrow F(1, 59850) = 106.494,\ p < 0.05;\ Patient \rightarrow F(1, 56187) = 559.541,\ p < 0.05;\ Reddit \rightarrow F(2, 5672928) = 2.563,\ p < 0.05$). The family of SNRI drugs (Serotonin–Norepinephrine Reuptake Inhibitor) appears less frequently, perhaps because it is a more recently developed medication. For blood pressure, on the other hand, we see significant differences between the prevalent drugs in Health Boards versus Patient: the former strongly features Beta Blockers (e.g., Metoprolol, Acebutolol) and the latter shows more ACE (Angiotensin-converting-enzyme) Inhibitors (e.g., Zestril, Univasc) ($HealthBoards \rightarrow F(1, 4766) = 240.242,\ p < 0.05;\ Patient \rightarrow F(1, 1090) = 12.251,\ p < 0.05;\ Reddit \rightarrow F(2, 1144330) = 10267.376,\ p < 0.05$).

Additionally, we compared drug dosages across forums. To extract this information from post text, we identified all snippets with a numerical value followed by a dosage unit such as mg, mL, puffs, drops. These are mapped to the drug mention that is closest in proximity.

For blood pressure and diabetes, no substantial differences in drug dosages were found. For depression, however, while the same antidepressants are prevalent, Health Boards and Reddit feature higher dosages. For instance, the most popular drug, Lexapro, is consumed in significantly higher dosages in Health Boards ($\mu = 29.93mg,\ \sigma = 75.74mg$) than in Patient ($\mu = 17.54,\ \sigma = 30.77$) ($t = -2.55,\ p = 0.01$). The same holds for other SSRIs like Zoloft ($HealthBoards \rightarrow \mu = 85.91mg,\ \sigma = 93.21mg;\ Patient \rightarrow \mu = 78.71mg,\ \sigma = 86.33mg$) and Prozac ($HealthBoards \rightarrow \mu = 44.99mg, \sigma = 112.45mg;\ Patient \rightarrow \mu = 34.19mg,\ \sigma = 40.21mg$). Between Health Boards and Reddit, no significant differences were observed.

Key findings. Health Boards and Patient have much higher coverage of drugs than Reddit. Depression and diabetes are largely treated with the same (families of) medications. For blood pressure, however, Health Boards and Patient exhibit two different drug families: Betablockers and ACE Inhibitors, respectively. We believe this is due to differences in regulation and medical practice in the US (Health Boards) versus UK (Patient). Regarding drug dosages, a striking observation is the significantly higher values for antidepressants in Health Boards and Reddit compared to Patient, again likely due to the different geographic foci of the respective forums.

6.5 Conclusion

While there is ample work on analyzing online communities for topics like politics, discussion in health forums have received comparatively little attention. This chapter presents a first step to obtaining insight into the characteristics, benefits and limitations of health communities.

Among our key findings, the most notable observation is that specialized forums like Health Boards and Patient engage more on discussing medical detail like specific drugs and their dosages. In contrast, subreddits with analogous topics appear to be more diverse, with a focus on early-stage advice-seeking and mutual support. Comparing the US-based Health Boards and the UK-based Patient forum on the specific condition of depression, another key observation is that Health Boards features significantly more posts about antidepressant drugs whereas Patient devotes more attention to behavioral therapies.

Future work exploring the detailed demographics of these communities, including user attributes such as age and gender, could reveal more about their users' habits and needs. This information, combined with our initial findings, could guide the development of search and recommendation systems for patients seeking online information and support.

7

Conclusions and Outlook

This dissertation investigates the characteristics of controversial discussions on social media. We propose a general framework that systematically identifies key elements of the discussions, and which can be used to discover patterns and to predict the onset of controversies.

Our initial feature model for adversarial political discussions on Twitter focuses on the posting activity directly involving stakeholders to discover factual and non-factual salient topics present on either side of a campaign, in addition to highlighting the role of power users in the discussion activity. Our findings highlight the mix of topics brought up by stakeholders, with some campaigns focusing much more on sentimental issues than factual ones, and the uneven levels of user activity, indicating that power users were more active in (non-factual) pro-candidate topics.

We then turn to the dynamics of political discussions in another polarized space, made up of Reddit communities dedicated to politics and world news. We introduce the concept of X-posts as posts that have attracted a negative or mixed reaction from the community, and propose four conversational archetypes based on the patterns of occurrences of these posts throughout the discussions. These are characterized via a feature model that captures the nature and intensity of sentiments expressed in individual posts, the textual cohesion between posts in the same discussion thread, and activity signals from users. Among discussions that lack X-posts, we find higher sentiments and lower topic cohesion, while discussions where X-posts are abundant display a greater topical focus and post similarity, higher sentiment variance, and lower sentiments overall.

To understand the flexible nature of X-posts, we generalize beyond political discussions and examine thematically diverse Reddit communities with different notions of what constitutes controversy. Our feature space, describing activity, post sentiments, salient topics, and topical cohesion, is used to build classifiers that can predict the future occurrence of X-posts in a discussion, based on its initial posts. An analysis of classification results further highlights the relationship between the model features, revealing that off-topic content, celebrity mentions, and negative sentiment can all lead to X-posts in the right context.

Finally, we venture outside of controversial discussions and apply our methodological approaches to other forms of long-term, specialized online discussions, in an effort to further our understanding of their dynamics in different contexts. For this, we characterize three different health communities based on a feature space describing user engagement, key medical topics, and medical detail represented by the mention of drug treatments and drug dosages. Our comparison of these forums highlights a difference in user posting behavior and in their topical focus, which may be attributed to community expectations regarding the nature of discussions, as well as the user demographics in each of them.

Several issues pertaining to the development of controversial online discussions still remain open for future research, including:

- **User interaction patterns.** The underlying structure of an online community is determined by the way users interact, and is thus subject to both global and local changes over time. Instances may include subgroups of users who frequently interact, or who interact exclusively under certain conditions (e.g. when a specific topic is being discussed). Finding such groups may provide insights into how different types of discussions develop within a community and how user dynamics relate to the occurrence of X-posts.

- **X-posts and community longevity.** Users may naturally find themselves more or less involved with an online community as their interests, free time, and attention fluctuate. Sudden surges or drops in activity, however, may be a symptom of a significant change in how the user relates to the community. Some users may, for instance, be driven away from communities that become too turbulent, while others may be drawn to them [Chandrasekharan et al., 2017]. By tracking shifts in activity patterns, we may find that user activity is a function of the social feedback they receive from other users (e.g. how many replies their posts receive and whether these are positive or negative), or a function of the dominant topic in a community at a given time.

- **Unified framework for content moderation.** In addition to direct user feedback in the form of voting, flagging, and reporting, administrators and moderators often rely on automated tools to maintain the quality of their online communities [Jhaver et al., 2019]. A tool that can incorporate the knowledge gained from our predictive framework may alert moderators to discussion threads with a greater potential for controversy, allowing these to be monitored and proactively handled as necessary (e.g. by issuing warnings and reminders to the users to remain civil, or to stay on-topic).

LIST OF FIGURES

List of Tables

Bibliography

[Adamic and Glance, 2005] Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, page 36–43, New York, NY, USA. Association for Computing Machinery.

[Addawood et al., 2019] Addawood, A., Badawy, A., Lerman, K., and Ferrara, E. (2019). Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM*, pages 15–25. AAAI Press.

[Aggarwal, 2011] Aggarwal, C. C., editor (2011). *Social Network Data Analytics*. Springer.

[Agichtein et al., 2008] Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In Najork, M., Broder, A. Z., and Chakrabarti, S., editors, *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 183–194. ACM.

[Ahmed et al., 2016] Ahmed, S., Jaidka, K., and Skoric, M. M. (2016). Tweets and votes: A four-country comparison of volumetric and sentiment analysis approaches. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM, Cologne, Germany, May 17-20, 2016*, pages 507–510. AAAI Press.

[Al-garadi et al., 2018] 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 Computing Surveys*, 51(1):16:1–16:37.

[Aragón et al., 2017a] Aragón, P., Gómez, V., , García, D., and Kaltenbrunner, A. (2017a). Generative models of online discussion threads: state of the art and research challenges. *Journal of Internet Services and Applications*, 8(1):15.

[Aragón et al., 2017b] Aragón, P., Gómez, V., and Kaltenbrunner, A. (2017b). To thread or not to thread: The impact of conversation threading on online discussion. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 12–21. AAAI Press.

[Backstrom et al., 2013] Backstrom, L., Kleinberg, J. M., Lee, L., and Danescu-Niculescu-Mizil, C. (2013). Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In Leonardi, S., Panconesi, A., Ferragina, P., and Gionis, A., editors, *Sixth*

ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013, pages 13–22. ACM.

[Bakshy et al., 2011] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the Fourth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 65–74.

[Bal et al., 2020] Bal, R., Sinha, S., Dutta, S., Joshi, R., Ghosh, S., and Dutt, R. (2020). Analysing the extent of misinformation in cancer related tweets. In Choudhury, M. D., Chunara, R., Culotta, A., and Welles, B. F., editors, *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 924–928. AAAI Press.

[Bamakan et al., 2019] Bamakan, S. M. H., Nurgaliev, I., and Qu, Q. (2019). Opinion leader detection: A methodological review. *Expert Systems with Applications*, 115:200–222.

[Barberá et al., 2015] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[Buckels et al., 2014] Buckels, E. E., Trapnell, P. D., and Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102. The Dark Triad of Personality.

[Buntain and Golbeck, 2014] Buntain, C. and Golbeck, J. (2014). Identifying social roles in reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 615–620, New York, NY, USA. ACM.

[Cha et al., 2010] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. *Proceedings of the 4th International AAAI Conference on Web and Social Media, ICWSM*, 4(1):10–17.

[Chandrasekharan et al., 2017] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *ACM Transactions on Computer-Human Interaction*, 1(CSCW).

[Chandrasekharan et al., 2018] Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., and Gilbert, E. (2018). The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

[Chang et al., 2009] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

[Chang and Danescu-Niculescu-Mizil, 2019] Chang, J. P. and Danescu-Niculescu-Mizil, C. (2019). Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 4742–4753. ACL.

[Chatzakou et al., 2017] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017). Measuring #gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 1285–1290, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Chawla et al., 2002] 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, 2017] Chen, M. (2017). Efficient vector representation for documents through corruption. In *5th International Conference on Learning Representations, ICLR*. OpenReview.net.

[Cheng et al., 2016] Cheng, J., Adamic, L. A., Kleinberg, J. M., and Leskovec, J. (2016). Do cascades recur? In Bourdeau, J., Hendler, J., Nkambou, R., Horrocks, I., and Zhao, B. Y., editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 671–681. ACM.

[Cheng et al., 2017] 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 *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW*, pages 1217–1230. ACM.

[Cheng et al., 2015] Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Antisocial behavior in online discussion communities. In Cha, M., Mascolo, C., and Sandvig, C., editors, *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 61–70. AAAI Press.

[Chetty and Alathur, 2018] Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118.

[Choi et al., 2015] Choi, D., Han, J., Chung, T., Ahn, Y., Chun, B., and Kwon, T. T. (2015). Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In Sharma, A., Agrawal, R., and Grossglauser, M., editors, *Proceedings of the 2015 ACM on Conference on Online Social Networks, COSN 2015, Palo Alto, California, USA, November 2-3, 2015*, pages 233–243. ACM.

[Choi et al., 2010] Choi, Y., Jung, Y., and Myaeng, S.-H. (2010). Identifying controversial issues and their sub-topics in news articles. In Chen, H., Chau, M., Li, S.-h., Urs, S., Srinivasa, S., and Wang, G. A., editors, *Intelligence and Security Informatics*, pages 140–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Choudhury and De, 2014] Choudhury, M. D. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the 8th International Conference on Web and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*

[Choudhury et al., 2016] Choudhury, M. D., Jhaver, S., Sugar, B., and Weber, I. (2016). Social media participation in an activist movement for racial equality. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM, Cologne, Germany, May 17-20, 2016.*, pages 92–101.

[Cohen, 1988] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd.

[Coles and West, 2016] Coles, B. A. and West, M. (2016). Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, 60:233–244.

[Coletto et al., 2017] Coletto, M., Garimella, K., Gionis, A., and Lucchese, C. (2017). A motif-based approach for identifying controversy. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 496–499. AAAI Press.

[Conover et al., 2011] Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *Proceedings of the 5th International Conference on Weblogs and Social Media, ICWSM, Barcelona, Catalonia, Spain, July 17-21, 2011.*

[Conover et al., 2012] Conover, M. D., Gonçalves, B., Flammini, A., and Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data science*, 1(1):1–19.

[Das and Lavoie, 2014] Das, S. and Lavoie, A. (2014). The effects of feedback on human behavior in social media: An inverse reinforcement learning model. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS

'14, page 653–660, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Datta and Adar, 2019] Datta, S. and Adar, E. (2019). Extracting inter-community conflicts in reddit. In *Proceedings of the 13nth International Conference on Web and Social Media, ICWSM*, pages 146–157. AAAI Press.

[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM*, pages 512–515. AAAI Press.

[Dimitrov et al., 2021] Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., and Da San Martino, G. (2021). Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

[Dirkson et al., 2019] Dirkson, A., Verberne, S., and Kraaij, W. (2019). Lexical normalization of user-generated medical text. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 11–20, Florence, Italy. Association for Computational Linguistics.

[ElSherief et al., 2018] ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. M. (2018). Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 52–61. AAAI Press.

[Ernala et al., 2018] Ernala, S. K., Labetoulle, T., Bane, F., Birnbaum, M. L., Rizvi, A. F., Kane, J. M., and Choudhury, M. D. (2018). Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses. In *Proceedings of the International 12th AAAI Conference on Web and Social Media Media, ICWSM*, volume 12 of *ICWSM 2018*, pages 62–71, Stanford, California, USA. AAAI Press.

[Fang et al., 2015] Fang, A., Ounis, I., Habel, P., Macdonald, C., and Limsopatham, N. (2015). Topic-centric classification of twitter user's political orientation. In *Proceedings 38th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 791–794. ACM Press.

[Fiesler et al., 2018] Fiesler, C., Jiang, J. A., McCann, J., Frye, K., and Brubaker, J. R. (2018). Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the 12th International Conference on Web and Social Media, ICWSM*, pages 72–81. AAAI Press.

[Flores-Saviaga et al., 2018] Flores-Saviaga, C., Keegan, B. C., and Savage, S. (2018). Mobilizing the trump train: Understanding collective action in a political trolling community. In *Proceedings of the 12th International Conference on Web and Social Media, ICWSM*, pages 82–91. AAAI Press.

[Gao et al., 2014] Gao, H., Mahmud, J., Chen, J., Nichols, J., and Zhou, M. X. (2014). Modeling user attitude toward controversial topics in online social media. In *Proceedings of the 8th International Conference on Web and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.

[Garimella et al., 2016] Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2016). Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 33–42.

[Garimella et al., 2018] Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):3:1–3:27.

[Garimella and Weber, 2017] Garimella, V. R. K. and Weber, I. (2017). A long-term analysis of polarization on twitter. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM*, pages 528–531. AAAI Press.

[Geiger and Ribes, 2010] Geiger, R. S. and Ribes, D. (2010). The work of sustaining order in wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, page 117–126, New York, NY, USA. Association for Computing Machinery.

[Ghenai and Mejova, 2018] Ghenai, A. and Mejova, Y. (2018). Fake cures: User-centric modeling of health misinformation in social media. *PACMHCI*, 2(CSCW):58:1–58:20.

[Gillani et al., 2018] Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., and Roy, D. (2018). Me, my echo chamber, and i: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 823–831, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Gleave et al., 2009] Gleave, E., Welser, H., Lento, T., and Smith, M. (2009). A conceptual and operational definition of 'social role' in online community. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–11. IEEE.

[Glenski and Weninger, 2017] 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 et al., 2011] Gómez, V., Kappen, H. J., and Kaltenbrunner, A. (2011). Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, HT '11, page 181–190, New York, NY, USA. Association for Computing Machinery.

[Gómez et al., 2013] 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.

[Grabowicz et al., 2016] Grabowicz, P. A., Ganguly, N., and Gummadi, K. P. (2016). Distinguishing between topical and non-topical information diffusion mechanisms in social media. In *Proceedings 10th International Conference on Web and Social Media, ICWSM*.

[Grover and Mark, 2019] Grover, T. and Mark, G. (2019). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM*, pages 193–204. AAAI Press.

[Guimarães et al., 2019] Guimarães, A., Balalau, O. D., Terolli, E., and Weikum, G. (2019). Analyzing the traits and anomalies of political discussions on reddit. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM*, pages 205–213. AAAI Press.

[Guimarães et al., 2021] Guimarães, A., Terolli, E., and Weikum, G. (2021). Comparing health forums: User engagement, salient entities, medical detail. In Birnholtz, J. P., Ciolfi, L., Ding, S., Fussell, S. R., Monroy-Hernández, A., Munson, S., Shklovski, I., and Naaman, M., editors, *Companion Publication of the 2021 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2021, Virtual Event, USA, October 23-27, 2021*, pages 57–61. ACM.

[Guimarães et al., 2017] Guimarães, A., Wang, L., and Weikum, G. (2017). Us and them: Adversarial politics on twitter. In Gottumukkala, R., Ning, X., Dong, G., Raghavan, V., Aluru, S., Karypis, G., Miele, L., and Wu, X., editors, *2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 872–877. IEEE Computer Society.

[Guimarães and Weikum, 2021] Guimarães, A. and Weikum, G. (2021). X-posts explained: Analyzing and predicting controversial contributions in thematically diverse reddit forums. In Budak, C., Cha, M., Quercia, D., and Xie, L., editors, *Proceedings of the 15th International Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 163–172. AAAI Press.

[Hamm et al., 2013] Hamm, M. P., Chisholm, A., Shulhan, J., Milne, A., Scott, S. D., Given, L. M., and Hartling, L. (2013). Social media use among patients and caregivers: a scoping review. *BMJ Open*, 3(5).

[Hessel and Lee, 2019] Hessel, J. and Lee, L. (2019). Something's brewing! early prediction of controversy-causing posts from discussion features. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1648–1659. Association for Computational Linguistics.

[Hine et al., 2017] 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 *Proceedings of the 11th International Conference on Web and Social Media, ICWSM*, pages 92–101. AAAI Press.

[Hoffart et al., 2011] 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 *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 782–792. ACL.

[Hutto and Gilbert, 2014] Hutto, C. J. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Web and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.

[Jhaver et al., 2019] Jhaver, S., Birman, I., Gilbert, E., and Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction*, 26(5):31:1–31:35.

[Jiang et al., 2016] Jiang, M., Cui, P., and Faloutsos, C. (2016). Suspicious behavior detection: Current trends and future directions. *IEEE Intell. Syst.*, 31(1):31–39.

[Joseph et al., 2019] 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 *Proceedings of the 13th International Conference on Web and Social Media, ICWSM*, pages 290–301. AAAI Press.

[Jurczyk and Agichtein, 2007] Jurczyk, P. and Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 919–922, New York, NY, USA. Association for Computing Machinery.

[Khalid and Srinivasan, 2020] Khalid, O. and Srinivasan, P. (2020). Style matters! investigating linguistic style in online communities. *Proceedings of the 14th International Conference on Web and Social Media, ICWSM*, 14:360–369.

[Koopman and Zuccon, 2019] Koopman, B. and Zuccon, G. (2019). Wsdm 2019 tutorial on health search (hs2019): A full-day from consumers to clinicians (materials at http:github.comielabhealth-search-tutorial). In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 838–839, New York, NY, USA. Association for Computing Machinery.

[Kumar et al., 2017] Kumar, S., Cheng, J., and Leskovec, J. (2017). Antisocial behavior on the web: Characterization and detection. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 947–950. ACM.

[Kusner et al., 2015] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

[Le et al., 2017] Le, H. T., Shafiq, Z., and Srinivasan, P. (2017). Scalable news slant measurement using twitter. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 584–587. AAAI Press.

[Lerman, 2007] Lerman, K. (2007). Social information processing in social news aggregation. *CoRR*, abs/cs/0703087.

[Leskovec, 2011] Leskovec, J. (2011). Social media analytics: Tracking, modeling and predicting the flow of information through networks. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, page 277–278, New York, NY, USA. Association for Computing Machinery.

[Levonian et al., 2020] Levonian, Z., Erikson, D. R., Luo, W., Narayanan, S., Rubya, S., Vachher, P., Terveen, L., and Yarosh, S. (2020). Bridging qualitative and quantitative methods for user modeling: Tracing cancer patient behavior in an online health community. In Choudhury, M. D., Chunara, R., Culotta, A., and Welles, B. F., editors, *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 405–416. AAAI Press.

[Liang, 2017] 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*, pages 146–154. ACM.

[Litchman et al., 2018] Litchman, M. L., Edelman, L. S., and Donaldson, G. W. (2018). Effect of diabetes online community engagement on health indicators: Cross-sectional study. *JMIR Diabetes*, 3(2):e8.

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

[Liu et al., 2018] 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 *Proceedings of the 12th International Conference on Web and Social Media, ICWSM*, pages 181–190. AAAI Press.

[Ma et al., 2018] Ma, X., Gui, X., Fan, J., Zhao, M., Chen, Y., and Zheng, K. (2018). Professional medical advice at your fingertips: An empirical study of an online "ask the doctor" platform. *PACMHCI*, 2(CSCW):116:1–116:22.

[Ma et al., 2013] Ma, Z., Sun, A., and Cong, G. (2013). On predicting the popularity of newly emerging hashtags in twitter. *JASIST*, 64(7):1399–1410.

[Mamykina et al., 2015] Mamykina, L., Nakikj, D., and Elhadad, N. (2015). Collective sensemaking in online health forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3217–3226, New York, NY, USA. Association for Computing Machinery.

[Matamoros-Fernández, 2017] Matamoros-Fernández, A. (2017). Platformed racism: the mediation and circulation of an australian race-based controversy on twitter, facebook and youtube. *Information, Communication & Society*, 20(6):930–946.

[Matias, 2016] Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1138–1151, New York, NY, USA. Association for Computing Machinery.

[Mejova et al., 2013] Mejova, Y., Srinivasan, P., and Boynton, B. (2013). GOP primary season on twitter: "popular" political sentiment in social media. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 517–526.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

[Mondal et al., 2017] Mondal, M., Silva, L. A., and Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT*, pages 85–94. ACM.

[MonkeyLearn, 2016] MonkeyLearn (2016). Donald trump vs hillary clinton: sentiment analysis on twitter mentions. *Blog Post, blog.monkeylearn.com/donald-trump-vs-hillary- clinton-sentiment-analysis-twitter-mentions/, 20 Oct 2016.*

[Napoles et al., 2017] Napoles, C., Pappu, A., and Tetreault, J. R. (2017). Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM*, pages 628–631. AAAI Press.

[Newell et al., 2016] Newell, E., Jurgens, D., Saleem, H. M., Vala, H., Sassine, J., Armstrong, C., and Ruths, D. (2016). User migration in online social networks: A case study on reddit during a period of community unrest. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM, Cologne, Germany, May 17-20, 2016*, pages 279–288. AAAI Press.

[Nishi et al., 2016] 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.

[Nobles et al., 2020] Nobles, A. L., Leas, E. C., Dredze, M., and Ayers, J. W. (2020). Examining peer-to-peer and patient-provider interactions on a social media community facilitating ask the doctor services. In Choudhury, M. D., Chunara, R., Culotta, A., and Welles, B. F., editors, *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 464–475. AAAI Press.

[Pal and Counts, 2011] Pal, A. and Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*.

[Peddinti et al., 2014] 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, SP*, pages 493–508. IEEE.

[Pendse et al., 2019] Pendse, S. R., Niederhoffer, K., and Sharma, A. (2019). Cross-cultural differences in the use of online mental health support forums. *PACMHCI*, 3(CSCW):67:1–67:29.

[Rizoiu et al., 2018] 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 *Proceedings of the 12th International Conference on Web and Social Media, ICWSM*.

[Samory et al., 2017] Samory, M., Cappelleri, V.-M., and Peserico, E. (2017). Quotes reveal community structure and interaction dynamics. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 322–335, New York, NY, USA. ACM.

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

[Seering et al., 2019] Seering, J., Wang, T., Yoon, J., and Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443.

[Shachaf and Hara, 2010] Shachaf, P. and Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370.

[Siu et al., 2013] Siu, A., Nguyen, D. B., and Weikum, G. (2013). Fast entity recognition in biomedical text. In *Workshop on Data Mining for Healthcare at the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 19 of *SIGKDD 2013*, New York, NY, USA. ACM Press.

[Stromer-Galley et al., 2020] Stromer-Galley, J., Bryant, L., and Bimber, B. (2020). Context and medium matter: Expressing disagreements online and face-to-face in political deliberations. *Journal of Deliberative Democracy*, 11(1).

[Terolli et al., 2020] Terolli, E., Ernst, P., and Weikum, G. (2020). Focused query expansion with entity cores for patient-centric health search. In Pan, J. Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., and Kagal, L., editors, *The Semantic Web*, ICSW 2020, pages 547–564, Cham. Springer International Publishing.

[Tinati et al., 2012] Tinati, R., Carr, L., Hall, W., and Bentwood, J. (2012). Identifying communicator roles in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 1161–1168, New York, NY, USA. Association for Computing Machinery.

[Tumasjan et al., 2010] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010.*

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

[Vosecky et al., 2014] Vosecky, J., Jiang, D., Leung, K. W.-T., Xing, K., and Ng, W. (2014). Integrating social and auxiliary semantics for multifaceted topic modeling in twitter. *ACM Transactions on Internet Technology.*, 14(4):27:1–27:24.

[Vydiswaran et al., 2015] Vydiswaran, V. G. V., Zhai, C., Roth, D., and Pirolli, P. (2015). Overcoming bias to learn about controversial topics. *JASIST*, 66(8):1655–1672.

[Wadden et al., 2021] Wadden, D., August, T., Li, Q., and Althoff, T. (2021). The effect of moderation on online mental health conversations. In Budak, C., Cha, M., Quercia, D., and Xie, L., editors, *Proceedings of the 15th International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 751–763. AAAI Press.

[Wang et al., 2013] 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*, pages 1341–1352. ACM.

[Wang et al., 2012] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings ACL 2012 System Demonstrations*, ACL '12, pages 115–120. Association for Computational Linguistics.

[Welser et al., 2011] Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. (2011). Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA. ACM.

[Welser et al., 2007] Welser, H. T., Gleave, E., Fisher, D., and Smith, M. (2007). Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure*, 8(2).

[Weng et al., 2010] Weng, J., Lim, E., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 261–270.

[Weninger et al., 2013] 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*.

[Wong et al., 2016] Wong, F. M. F., Tan, C., Sen, S., and Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2158–2172.

[Xu et al., 2012] Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, page 656–666, USA. Association for Computational Linguistics.

[Yates et al., 2017] Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

[Yuan et al., 2013] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Magnenat-Thalmann, N. (2013). Who, where, when and what: discover spatio-temporal topics for twitter users. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 605–613.

[Zayats and Ostendorf, 2018] 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 et al., 2017] Zhang, A., Culbertson, B., and Paritosh, P. (2017). Characterizing online discussion using coarse discourse sequences. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM*, volume 11, pages 357–366.

[Zhang et al., 2018a] Zhang, J., Chang, J. P., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., and Thain, N. (2018a). Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1350–1361. ACL.

[Zhang et al., 2018b] Zhang, J., Danescu-Niculescu-Mizil, C., Sauper, C., and Taylor, S. J. (2018b). Characterizing online public discussions through patterns of participant interactions. *Proc. ACM Hum. Comput. Interact.*, 2(CSCW):198:1–198:27.

[Zhao et al., 2015] 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 SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1513–1522, New York, NY, USA. ACM.

[Zhao et al., 2011] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 338–349.

[Zhu et al., 2011] Zhu, H., Kraut, R. E., Wang, Y.-C., and Kittur, A. (2011). Identifying shared leadership in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 3431–3434, New York, NY, USA. Association for Computing Machinery.