Studying Anti-Social Behaviour on Reddit with Communalytic

Anatoliy Gruzd[1], Philip Mai[2], and Zahra Vahedi[2]

[1]Ryerson University
[2]Affiliation not available

June 16, 2020

Abstract

The chapter presents a new social media research tool for studying subreddits (i.e., groups) on Reddit called Communalytic. It is an easy-to-use, web-based tool that can collect, analyze and visualize publicly available data from Reddit. In addition to collecting data, Communalytic can assess the toxicity of Reddit posts and replies using a machine learning API. The resulting anti-social scores from the toxicity analysis are then added as weights to each tie in a "who replies to whom" communication network, allowing researchers to visually identify and study toxic exchanges happening within a subreddit. The chapter consists of two parts: first, it introduces our methodology and Communalytic's main functionalities. Second, it presents a case study of a public subreddit called r/metacanada. This subreddit, popular among the Canadian alt-right, was selected due to its polarizing nature. The case study demonstrates how Communalytic can support researchers studying toxicity in online communities. Specifically, by having access to this additional layer of information about the nature of the communication ties among group members, we were able to provide a more nuanced description of the group dynamics.

Studying Anti-Social Behaviour on Reddit with Communalytic

Anatoliy Gruzd[1]*, Philip Mai*[2]*, and Zahra Vahedi*[3]
Ryerson University, Toronto, Canada

Abstract

The chapter presents a new social media research tool for studying subreddits (i.e., groups) on Reddit called Communalytic. It is an easy-to-use, web-based tool that can collect, analyze and visualize publicly available data from Reddit. In addition to collecting data, Communalytic can assess the toxicity of Reddit posts and replies using a machine learning API. The resulting anti-social scores from the toxicity analysis are then added as weights to each tie in a "who replies to whom" communication network, allowing researchers to visually identify and study toxic exchanges happening within a subreddit. The chapter consists of two parts: first, it introduces our methodology and Communalytic's main functionalities. Second, it presents a case study of a public subreddit called r/metacanada. This subreddit, popular among the Canadian alt-right, was selected due to its polarizing nature. The case study demonstrates how Communalytic can support researchers studying toxicity in online communities. Specifically, by having access to this additional layer of information about the nature of the communication ties among group members, we were able to provide a more nuanced description of the group dynamics.

Keywords: Reddit, Social Media Analytics, Communalytic, Toxicity Analysis; Social Network Analysis, Politics, Canada.

[1] *Anatoliy Gruzd, Ph.D.* is a Canada Research Chair in Social Media Data Stewardship, Associate Professor at the Ted Rogers School of Management at Ryerson University, and Director of Research at the Social Media Lab. Gruzd is also a Member of the Royal Society of Canada's College of New Scholars, Artists and Scientists; a co-editor of a multidisciplinary journal on Big Data and Society; and a founding co-chair of the International Conference on Social Media and Society.

[2] *Philip Mai, M.A., J.D*. is the Director of Business and Communications and a Senior Researcher at Ryerson University's Social Media Lab at the Ted Rogers School of Management and co-founder of the International Conference on Social Media and Society. In his work, he focuses on tech policy issues, knowledge mobilization, information diffusion, business and research partnerships, and practical application of social media analytics.

[3] *Zahra Vahedi, M.A.* is a Ph.D. candidate in the Department of Psychology at Ryerson University. She is also a research assistant at the Social Media Lab at the Ted Rogers School of Management. Her research interests include the social, psychological, and cognitive effects of information and communication technologies, such as social media platforms and smartphones.

1 Introduction

The main goal of this chapter is to demonstrate a mixed-methods approach and a new research tool to study anti-social behaviour within online communities, and specifically on Reddit. We will refer to anti-social behaviour as any behaviour that may cause, or is likely to cause, harm or distress to one or more persons ("Anti-Social Behaviour, Crime and Policing Act 2014," 2014). Two common forms of anti-social behaviour on social media are trolling and hate speech. *Trolling* is often initiated to disrupt on-topic conversations and provoke other users through deceptive behaviours, accompanied with "inflammatory, extraneous, or off-topic messages" (Buckels, Trapnell, & Paulhus, 2014; Lampe et al., 2014). Trolling may be driven by the perpetrator's own entertainment and online fame without a clear purpose, but it has also been used by state actors to dissuade and incite others online (Howard & Bradshaw, 2017). Alternatively, *hate speech* refers to negative expressions that are typically directed toward a collective of people based on their religious affiliation, ethnicity, race, nationality, sexual orientation, gender identity, disability, or other shared group characteristics (Costello, Hawdon, & Ratliff, 2017; Faris et al., 2016). Although for some online groups, what is described here as 'anti-social' may actually be a communal norm and be practiced by group members to socialize; we are more interested in studying group dynamics where 'anti-social' is not a norm and where such behaviour may negatively affect the overall group cohesion and interactions at the community level and may have psychological and emotional consequences for individuals (Craker & March, 2016; Duggan, 2017; Giménez Gualdo et al., 2015; Hodson et al., 2018; Lindsay et al., 2016). There is also a concern that some forms of anti-social behaviour, such as hate speech, may galvanize xenophobic behaviour offline (Awan, 2014; Awan & Zempi, 2016) and lead to changing social norms at the societal level.

Even though anti-social behaviour on the internet is not a new phenomenon, the number of people who are exposed to it has risen exponentially with the widespread adoption of social media (Anderson, 2018; Runions, Bak, & Shaw, 2017). For example, nearly 60% of Canadian adults have encountered hate speech, racist, or sexist content online at least once a month (Ryerson Leadership Lab, 2019). Given the ubiquity of social media in modern-day society, it is imperative that the extent and types of anti-social behaviour on social media are well-documented and studied. However, one of the major challenges in studying anti-social behaviour online is that not all "anti-social" posts can be easily flagged as offensive or abusive. Are there varying degrees of online anti-social behaviour? How do we distinguish sarcastic texts, 'netspeak' jargon, and other content (e.g., emojis, images, and memes) that may go unnoticed due to their subtlety, but have detrimental effects on groups and individuals? Questions like these require rigorous empirical analysis to better understand how online anti-social behaviour may be changing our society and the practice of public discourse in the 21st century.

To help researchers interested in examining online anti-social behaviour, this chapter introduces *Communalytic*, a new research tool for studying anti-social acts[4] in public groups on *Reddit*, a popular social media site. We will use a case study approach to demonstrate how a researcher can use Communalytic to examine interactions which may lead to toxic exchanges among members of a public group called r/metacanada.

[4] For the remainder of this chapter, we will use the term "act" to refer to a specific manifestation of anti-social behaviour transmitted via social media.

2 Studying Communities on Reddit

While previous research has extensively examined anti-social acts on social media, most of the literature in this area has relied primarily on Twitter data (e.g., Gorrell et al., 2019; Maity, Chakraborty, Goyal, & Mukherjee, 2018; Southern & Harmer, 2019; Theocharis et al., 2016). This over reliance on Twitter data for academic studies is likely due to the public nature of the Twitter platform, and the wide availability of research tools for collecting and analysing data from Twitter. Notably, there are far fewer studies that examine antisocial acts on Reddit (Massanari & Chess, 2018; Massanari, 2017). Communalytic was developed to address a lack of availability of Reddit research tools and to enable internet researchers to study online communities and communication practices on this platform, and more specifically to study how anti-social acts manifest themselves on this social platform. Considering the anonymous nature of online interactions on Reddit and how often it is used for political and polarizing discussions, Reddit is a useful platform for researchers to study and better understand dynamics that drive online anti-social behaviour and their impact on various online communities. This section will provide an overview of Reddit as a social networking platform, as well as a specific subreddit called r/metacanada which we use as a case study in this chapter.

2.1 What is Reddit?

Reddit is a social media platform that was founded by Alexis Ohanian and Steven Huffman in 2005 (Anderson, 2015). Originally billed as “the front page of the internet”, it is comprised of online communities called *subreddits*, where users (also known as *redditors*) can share posts, images, or URLs. These subreddits cover a vast array of topics from history to politics and everything in between, with an estimated 1.8 million subreddits available on Reddit (Redditmetrics, n.d.). Users can “upvote” (i.e., “like”) or “downvote” (i.e., “dislike”) posted content, influencing the rank of that post for both their own main feed, as well as within the subreddit that it belongs to. In other words, more popular content will become more visible, and less popular content will be shuffled down to the bottom of the feed. In addition to upvotes and downvotes, discussion between users are facilitated on the posts in the form of comments and replies. To help supervise the content that is posted in each subreddit, user-appointed moderators – or *mods* – are tasked with regulating each subreddit based on each subreddit’s rules on appropriate content. For example, the subreddit r/politics has a list of eleven rules that users must abide by when posting content, such as “no hateful speech” and “no copy-pasted articles.”

Given the specificity and expansiveness of subreddit topics – as well as the open and public nature of the posted content – Reddit has become a subject of interest for researchers in various fields of study. For example, several recent papers have examined the positive benefits and practicality of Reddit as a platform for informal learning (e.g., Del Valle et al., 2020; Haythornthwaite et al., 2018; Staudt Willet, & Carpenter, 2019). In addition, content analyses of subreddits have revealed that Reddit has facilitated increased public engagement with scientists (e.g., Hara, Abbazio, & Perkins, 2019), and has often been used by researchers to learn more about topics as wide and varied as weight management (Pappa, Cunha, Bicalho, Ribeiro, Silva, Meira, & Beleigoli, 2017), to users’ attitudes toward vaccination (O’Kane, Zhang, Lama, Hu, Jamison, Quinn, & Broniatowski, 2019), to users’ experiences with mental illness (Yoo, Lee, & Ha, 2019). Finally, and more broadly, Reddit has also been touted as a potential platform for participant recruitment (Gutierrez, 2018; Shatz, 2017).

However, despite the positive potentials of Reddit as a social platform – such as the enabling of supportive communities and access to user-generated, niche information – Reddit can facilitate online anti-social behaviour (Ging & Siapera, 2018; Massanari, 2017; Massanari, & Chess, 2018; Topinka, 2018), or what Massanari (2017) calls 'toxic technocultures'. These are defined as "the toxic cultures that are enabled by and propagated through sociotechnical networks such as Reddit, 4chan, Twitter, and online gaming" (p. 333). Such 'toxic technocultures' can be fueled by platform affordances. In the case of Reddit, the ease with which users can create multiple accounts and subreddits, allows for the ability to create and participate in anti-social acts, with little to no repercussions. For example, subreddits that are banned for toxic content – such as r/incels, which promoted mysognistic views and rape – often have similar subreddits where users can flock to (e.g., r/Braincels). In addition, the policies enforced by platform administrators encourage Reddit as a "neutral platform" for discussion. As a result, administrators rarely intervene in disputes, citing their neutrality toward the nature of the content, irrespective of how inappropriate or toxic it may be. Still, users may report behaviour or an entire subreddit community that they deem to violate Reddit's community guidelines on harassment, bullying, and threatening behaviour (Reddit, 2019). In these cases, intervention may result in the banning of users and/or the subreddit community. This course of action is usually reliant on users' self-directed action, such as flagging and reporting toxic and hateful content, including links, comments, and subreddits, on Reddit.

Given the open and permissive structure and policy guidelines of Reddit – coupled with the ease of access to open and public data available on subreddit communities – we chose Reddit as a social media platform to analyse anti-social behaviour in our case study. The following section will describe our selection process of the r/metacanada subreddit used in this case study.

2.2 Case of r/metacanada

To identify a subreddit for the case study, we decided to look for groups that are known to solicit strong reactions from other users, such as those that discuss and espouse nationalistic and extreme right-wing ideologies. To locate potential subreddits, we took the following steps: first, a broad keyword search was conducted on Google Scholar. The following keywords were used: *Reddit, nationalism, altright, right-wing, islamophobia,* and *white nationalism.* This step was conducted in order to help us locate any existing research that may have examined these constructs within Reddit, as well as the specific subreddits of interest. Several studies were located (Nithyanand et al., 2017; Qian et al., 2019; Topinka, 2018), and were reviewed to help us identify what subreddits were generally associated with extreme right-wing sentiment. Finally, the same keyword search was conducted within Reddit in order to identify subreddits relating to our research interest. Only publicly accessible subreddits were searched.

Based on these preliminary searches, the following ten subreddits were identified as potentially being suitable for use as a case study: r/AskThe_Donald, r/Conservative, r/politics, r/ConservativesOnly, r/LeftistWatch, r/POLITIC, r/canada, r/The_Europe, r/askaconservative, r/metacanada. Using Communalytic, we extracted 1 full day's worth of posts for all ten subreddits and analyzed the posts for toxicity (the tool and the process are described in Section 3.2). During the selection process, we examined the following aspects of each dataset: 1) the number of posts and replies extracted in the one-day period, 2) the highest and average toxicity scores for each dataset, and 3) the top ten toxic posts. At this stage, four subreddits (r/ConservativesOnly, r/LeftistWatch, r/The_Europe, r/askaconservative) were eliminated from consideration for low posting activities. We also excluded three subreddits (r/AskThe_Donald, r/politics, r/POLITIC) because their top ten toxicity posts did not include comments that specifically focussed on nationalistic or right-wing topics or issues. For the remaining three subreddits (r/Conservative, r/canada, and r/metacanada), we reviewed a small sample of their posts to gauge their level of toxicty. At the end of the review, we selected r/metacanada for this case study due to the high level of toxic and nationalistic content present in the subreddit.

The r/metacanada subreddit[5] was created on May 6, 2011and is self described as "The only not-retarded Canadian subreddit." We collected data from this subreddit in the two weeks (October 9, 2019 to October 22, 2019) leading up to the Canadian Federal Election which took place on October 21, 2019. At the period of data collection, there were 31.2 thousand subredditors who subscribed to this community. This particular subreddit has ten rules that all members must abide by, including "no doxxing" – or revealing other users' personal information – and "no brigading", which includes the organization of a group of subredditors to attack, harass, and/or downvote another user. Other rules in this subreddit are "use NP for reddit links", where redditors are asked to "replace 'www' in the link with 'np' ", "don't vote/comment in linked threads", "no floodposting/disruptive shitposting", "no racism", "no condoning/threatening illegal activity", "follow rules of reddit", "Mark NSFW posts NSFW", which is an acronym that indicates the posted content is "not safe for work", and "no shitty bots."

3 Method

3.1 Detecting Anti-Social Acts at Scale

Examining anti-social interactions in online communities is a rapidly growing area of inquiry. Recent research has examined a number of different types of anti-social acts, such as hate speech (Southern & Harmer, 2019), impoliteness (Theocharis et al., 2016), rudeness (Su et al., 2018), incivility (Kenski, Coe, & Rains, 2017; Rossini, 2019), offensive comments (Kwon & Gruzd, 2017), and stereotyping (Southern & Harmer, 2019). Considering the volume of available data, we will focus on the automated approaches to detecting anti-social posts in text-based communication on Reddit.

Content-based approaches: Prior literature on detecting anti-social acts at scale has primarily used supervised machine learning that predominantly relies on content-based features to identify relevant posts (Al-Makhadmeh & Tolba, 2019; Kwok & Wang, 2013; Pitsilis, Ramampiaro, & Langseth, 2018; Gorrell et al., 2019; Borkan et al., 2019; Hosseini, Kannan, Zhang, & Poovendran, 2017). For example, Dybala et al. (2010) used support vector machines (SVM) to classify comments posted on unofficial school websites in Japan into those that are potentially harmful and not. The SVM method relied heavily on the use of vulgar words by the purported bullies. Alternatively, Dinakar et al. (2011) developed a binary text classifier to determine whether a message is on a sensitive topic or not. The authors then trained multiclass classifiers to categorize

[5] https://www.reddit.com/r/metacanada/

messages into one of three possible attacks: an attack on 1) sexual minorities (‘sexuality’), 2) race and culture, or 3) one’s intelligence.

Dadvar and colleagues (2013) developed a multi-criteria evaluation system to detect cyberbullying among YouTube commentators. Their system assigns a ‘bulliness’ score to each user based on user information (age and membership duration), content features (post length, presence of profane words, profanity and bullying sensitive topics, the use of first and second person pronouns, non-standard spelling), and activity features (number of uploads, subscribed channels, posts). The researchers further improved the performance of their expert system by adding supervised machine learning using a Naïve Bayes classifier (Dadvar et al., 2014).

One of the most promising works is by Nahar and colleagues (2014), which involves a fuzzy SVM approach to cyberbullying detection designed to handle noisy, imbalanced, and streaming text from social media. The advantage of their approach is that it only requires a small training set which can then be expanded based on unlabelled streaming data. The feature set included keywords, the number of swear words, presence of pronouns, the degree of users’ emotions, the number of capitalized letters that may indicate shouting, additional metadata, and users’ age and gender. The evaluation, based on three different datasets from *Myspace*, *Kongregate*, and *Slashdot*, demonstrated the superior performance of the proposed approach over more traditional, fully supervised approaches.

Although showing a lot of potential, solutions based on content-based classifiers, as described above, suffer from several shortcomings. They require training and as such tend to be domain and context dependent, making them less effective in environments where bullies or trolls may – and often do – use slang, image-based messaging, or other subversion techniques to attack others. Another limitation of such approaches is that content-based only techniques focus on individual messages and are, therefore, not well equipped to detect a coordinated campaign by a set of users (or a set of accounts managed by a single entity) to disrupt an online group or discussion.

Graph-based approaches: To address some of the limitations of the content-based approaches, we can turn to graph-based approaches which focus on user *accounts* (instead of posts) and connections between them. From a graph perspective, online participants can be considered as nodes, and interactions between them as edges. A graph-based approach has the benefit of not relying on the content of messages and therefore removes the need to train a text classifier to support different languages, communities, and platforms. Another advantage is that such approaches are capable of identifying clusters of related accounts based on certain network properties (e.g., densely-connected accounts). This, therefore, allows for the detection of coordinated anti-social acts. Existing graph-based approaches rely on the detection of anomalies in the network structure. They can generally be divided into three broad categories: feature-based methods, community-based method or relational learning (Aggarwal, 2013).

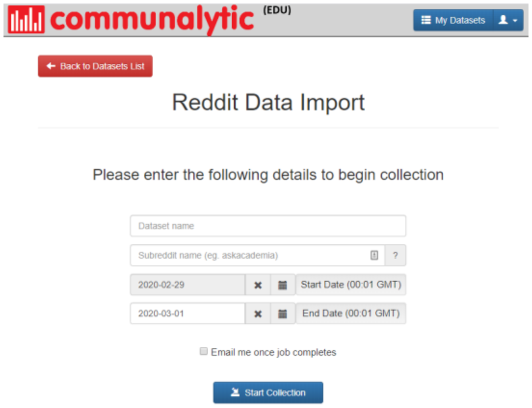
First, ***feature-based*** methods “transform the graph anomaly detection problem to the well-known and understood outlier detection problem” (Akoglu et al., 2014). Features may include node-level measures such as various node centralities, dyadic measures such as the number of common neighbours, or group-level measures such as density, reciprocity and modularity (Gruzd & Tsyganova, 2015). An example is a technique called OddBall (Akoglu et al., 2010) which uses graph-based measures, such as the number of neighbours and the number of triangles, for each ego network (that is a node/ego, all of its neighbours and connections among the neighbours) in order to detect those ego networks that deviate from the majority. The second type are ***community-based*** methods. They usually rely on partition or community detection techniques that are able to identify densely connected groups of nodes. Usually these would be the nodes that bridge different communities. For example, gSkeletonClu algorithm (Huang et al., 2013) finds outlier nodes as a by-product of the graph clustering algorithm. In FocusCO – another implementation of a community-based approach (Perozzi et al., 2014) – the algorithm also requires clusters to include nodes that have similar attributes. Nodes that are placed in the same cluster but differ from the other nodes in some attribute values, are labelled as outliers. The third approach is based on ***relational learning*** methods. This is a binary classification approach that classifies graph objects such as nodes and edges, while considering their inter-dependencies. For example, if one node is labelled as a 'troll', then this would increase the chance that the node connected to it is also a 'troll'; in other words, nodes connected to each other will likely have the same class label. Thus, in addition to node attributes, relational learning algorithms exploit class labels and attributes of node neighbours. Algorithms in this category often rely on an inference procedure to classify unlabelled nodes iteratively (Macskassy & Provost, 2007).

In this work, we propose to combine both a content-based and a graph-based approach. We start by discovering a communication network that represents w*ho interacts with whom* in an online group. Next, we apply a content-based machine learning classifier to determine whether an interaction between any two nodes in the communication network can be viewed as an anti-social act. Specifically, we rely on Perspective API, a machine learning classifier developed by Google that can recognize different types of anti-social acts such as toxicity, identity attack, insult, and threat (Chakrabarty, 2020). While the earlier versions of the Perspective API has been criticized for assigning high toxicity scores for non-toxic posts mentioning one's identity such as posts with LGBTQ+ related words (Hosseini et al., 2017; Jain et al., 2018), the most recent iteration of the Perspective API (the current being version 6) has shown a high level of accuracy (~80%) in offensive language detection (Jigsaw, 2019; Pavlopoulos et al., 2019), and has been used in a number of recent empirical studies (Delisle et al., 2019; Hopp & Vargo, 2019; Mittos et al., 2019; Obadimu et al., 2019).

Using Perspective API, we score each interaction between two users on a scale from 0 to 1 based on the likelihood of that text-based interaction exhibiting an anti-social act. We then assign these scores as individual weights to each edge in the network. Edges with higher scores closer or equal to one are more likely to denote 'anti-social' exchanges between users. The following section describes our approach in more detail including data collection from subreddits, the exporting of datasets in various formats, and analysis, as implemented in an online research system called Communalytic. Once communication networks are discovered and anti-social scores are assigned to edges, we use Gephi, a popular social network analysis tool (Bastian et al., 2009), to examine anti-social patterns in the network.

3.2 Introduction to Communalytic

Communalytic is a web-based research tool that can collect and analyze publicly available data from Reddit[6]. It is designed to study patterns of anti-social behaviour and can display the results of analyses visually in a variety of ways. Currently, the main data source for Communalytic is Reddit, specifically the subreddits within Reddit. Subreddits are the online forums that comprise Reddit and are denoted with an "r/" before the subreddit title. They are often dedicated to a specific topic, which users can subscribe to, post, and comment exclusively to that subreddit. Using Communalytic, researchers can collect publicly available submissions, comments, and replies posted within a subreddit. When importing data from Reddit, users are asked to specify the subreddit that they wish to collect data from and indicate the length of data collection (see Figure 1).



Reddit Data Import

Please enter the following details to begin collection

Dataset name

Subreddit name (eg. askacademia) ?

2020-02-29 Start Date (00:01 GMT)

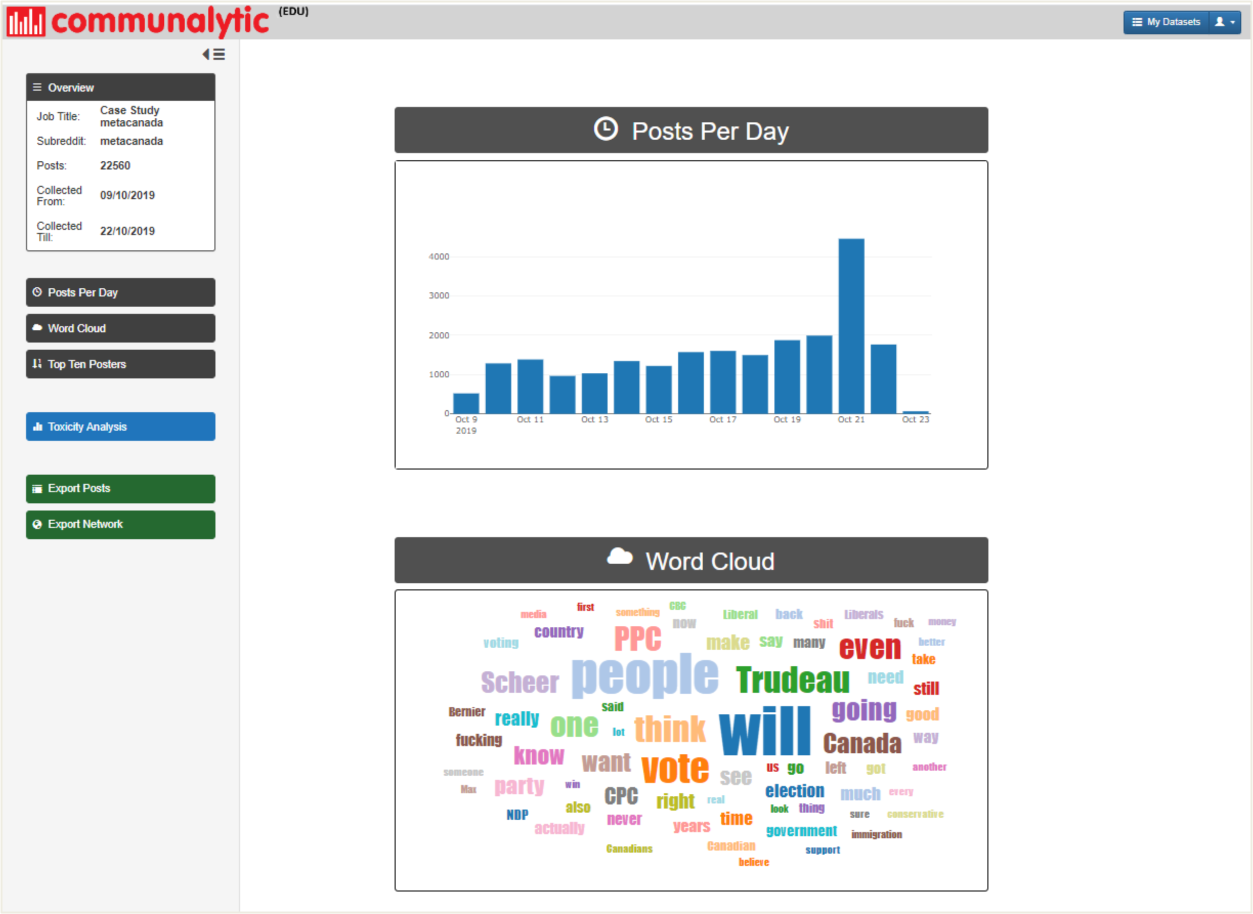
2020-03-01 End Date (00:01 GMT)

□ Email me once job completes

Figure 1. Data collection form in Communalytic

Once the data is collected, users can perform a variety of tasks with the dataset. Communalytic provides an overview of the dataset, including the subreddit name, the number of posts extracted, and the time period of extraction (see Figure 2). In the Dataset Overview screen, users can view visual representations of the number of posts per day, a word cloud depicting the most frequent words used in the dataset, as well as the top ten posters of a selected subreddit. In addition, this view enables users to export both posts and communication network data to their own computer for further analysis. Posts are exported as a CSV file, which also contains metadata about the collected posts, such as the author's username, date published, the content of the post, and the number of upvotes, as provided by Reddit's Public API. The network data file is exported as a GraphML file and can then be imported to other software – such as Gephi – for social network analysis. A snapshot of the network is also automatically generated by Communalytic, providing users with a static preview image of the dataset.

[6] Communalytic is developed by the Social Media Lab at Ryerson University, and is available for use at https://communalytic.com.



Job Title: Case Study metacanada

Subreddit: metacanada

Posts: 22560

Collected From: 09/10/2019

Collected Till: 22/10/2019

Figure 2. Dataset overview screen in Communalytic

After data collection has been completed, users can run a ‘toxicity analysis’ on the dataset. During this stage, with the help of Perspective API, Communalytic generates seven types of anti-social scores (between 0 to 1) for each post in the dataset, with scores closer to one indicating higher levels of toxicity. Communalytic uses the following scores as provided by Perspective: toxicity, severe toxicity, insult, identity attack, profanity, threat, and attack on commenter. Table 2 includes definitions and corresponding sample posts for each category.

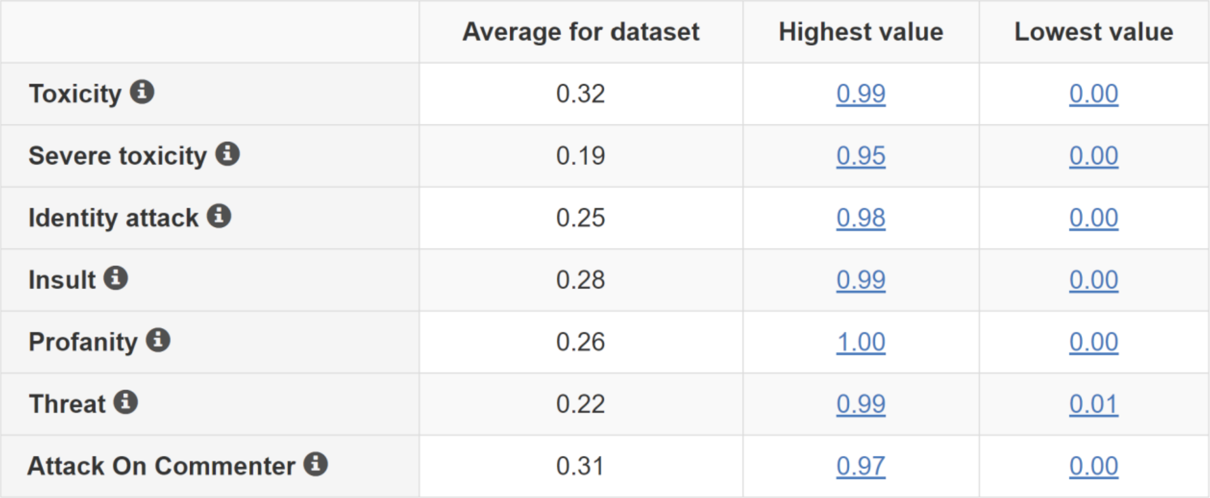
Table 2. Seven categories of anti-social acts from Perspective API available for analysis in Communalytic, their definitions and examples

	Definition[7]	Sample post
Toxicity	Rude, disrespectful, or unreasonable post	*“This is one of stupidest things I've read but fuck me I laughed at the second line”*
Severe toxicity	Very hateful, aggressive, disrespectful post. This score is less sensitive to posts that include positive uses of curse words	*“Fuck off pathetic loser, no one cares about your worthless opinion”*
Insult	Insulting, inflammatory, or negative post toward an individual or a group	*“How fucking stupid is [Name]? That is pretty fucking stupid. What's next - a deep fake*

[7] As defined by Perspective API https://github.com/conversationai/perspectiveapi/blob/master/2-api/models.md

		having him say racist things as a "social experiment"?
Identity attack	Negative post attacking someone because of their identity (including race, gender, sexual orientation, ideology, religion, nationality, etc.)	*"You people are a bunch of fags. And I voted for [Political Party Name]"*
Profanity	Post with swear words or other obscene language	*"Why vote for the [Political Party Name] when you know they won't win shit."*
Threat	Post with an intention to inflict pain, injury, or violence against an individual or group	*"Shoot all yellow vests! We have to kill all Nazis!"*
Attack on commenter	Post directly attacking another user	*"You are a disgusting human and less valuable than any one of Canada's millions of hard-working blue-collar people."*

Using the toxicity analysis option in Communalytic, users are able to determine the overall level of 'anti-social' in the dataset by examining the average scores, distributions of scores for all posts in the datasets, as well as by reviewing the top 10 posts that received the highest and lowest scores (see Figure 3a, 3b). The scores for individual posts and replies are also downloadable as a CSV file.



	Average for dataset	Highest value	Lowest value
Toxicity ℹ	0.32	0.99	0.00
Severe toxicity ℹ	0.19	0.95	0.00
Identity attack ℹ	0.25	0.98	0.00
Insult ℹ	0.28	0.99	0.00
Profanity ℹ	0.26	1.00	0.00
Threat ℹ	0.22	0.99	0.01
Attack On Commenter ℹ	0.31	0.97	0.00

Figure 3a. Toxicity analysis summary table for r/metacanada subreddit.

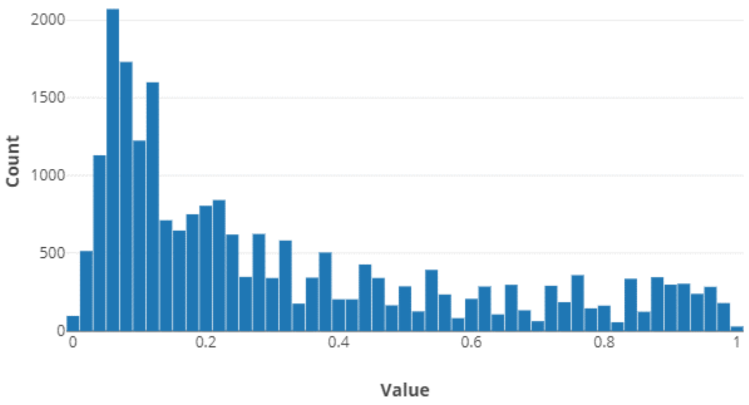


Figure 3b. The distribution of toxicity scores for r/metacanada subreddit, showing that the toxicity score for most posts is under 0.3. (Y axis: post count; X axis: toxicity score)

One particularly useful feature of Communalytic is that within the exported network file, one can access edge-level weights corresponding to each of the seven Perspective scores of anti-social acts available in Communalytic. For example, based on the data in Table 1, a user *n6* sent a highly toxic (toxicity=0.86) and insulting (insult=0.85) post to user *n2125* (edge *e2*). And since each reply is recorded as a single edge, some users will have multiple edges between them. For instance, user *n8* has two edges connecting her to user *n428* (edges *e8* and *e9*). These additional metadata fields embedded in the network file allow researchers to visualize and examine different communication layers based on different types of anti-social acts. This will be demonstrated in Section 4.2.

Table 1. Edge-level attributes for each of the seven categories of anti-social acts (highlighted values are referenced in text above).

Id	Source	Target	Toxicity	Severe toxicity	Insult	Identity attack	Profanity	Threat	Attack on commenter
e1	n5	n1117	0.19	0.10	0.18	0.10	0.10	0.25	0.03
e2	n6	n2125	0.86	0.69	0.85	0.81	0.75	0.36	0.80
e3	n7	n1435	0.24	0.12	0.24	0.25	0.16	0.47	0.12
e4	n7	n877	0.90	0.58	0.66	0.22	0.96	0.21	0.59
e5	n7	n877	0.72	0.43	0.65	0.26	0.80	0.24	0.02
e6	n7	n2038	0.17	0.05	0.13	0.17	0.07	0.16	0.27
e7	n8	n722	0.24	0.15	0.25	0.39	0.15	0.23	0.37
e8	n8	n428	0.61	0.44	0.67	0.80	0.54	0.29	0.76
e9	n8	n428	0.39	0.26	0.34	0.61	0.18	0.20	0.49
e10	n8	n635	0.50	0.37	0.51	0.63	0.44	0.26	0.30

4 Results

4.1 Toxicity analysis

In total, there were 22,560 posts, including 1,717 submissions (posts that start a new thread), and 20,843 replies. Table 3 shows how many posts are automatically classified as one or more of the seven Perspective's anti-social scores available in Communalytic. For example, 5.3% to 15.0% of posts can be characterized as *toxic*, whereas only 0.2% to 6.2% of posts are characterized as *severe toxic*. These ranges vary depending on the threshold used. The table shows the counts based on the three different thresholds: 0.7, 0.8, and 0.9. Considering the polarizing nature of this group, we expected a larger portion of posts to be toxic, but only a small fraction of them really are. It is likely that the level of toxicity was limited due to the active moderation by eight moderators in this subreddit. Future work will need to compare these levels to other subreddits to establish the baseline.

Table 3. Number and percentage of toxic posts

	Number and Percentage of Posts with the Scores …					
Threshold	>=0.7		>=0.8		>=0.9	
Toxicity	3376	15.0%	2287	10.1%	1198	5.3%
Severe toxicity	1401	6.2%	497	2.2%	54	0.2%
Insult	2658	11.8%	1515	6.7%	709	3.1%
Profanity	3358	14.9%	2671	11.8%	1595	7.1%
Identity attack	1114	4.9%	538	2.4%	99	0.4%
Threat	386	1.7%	241	1.1%	52	0.2%
Attack on commenter	3100	13.7%	1902	8.4%	1251	5.5%

To determine a cut-off value for the Perspective scores, we recommend testing different thresholds to identify a suitable level based on the research questions and the focus of a given subreddit. This is because by lowering the threshold to 0.7 or lower, the system will more likely catch most anti-social acts, but at the same time, it will increase the likelihood of labelling a post as 'anti-social' when it is not; thus, introducing false positive results. On the other hand, by setting the threshold to 0.9 or higher, we will reduce the chance of false positives but will be at risk of missing some anti-social posts that scored below 0.9. In general, if your project aims to identify more severe cases of toxicity, and explicit cases of anti-social, then setting the threshold to 0.9 may be appropriate. But if your project is seeking to examine all possible anti-social acts, then you may consider casting a wider net by lowering the threshold to 0.7. To evaluate the accuracy of the Perspective scores, we recommend recruiting human coders who would review and score a smaller, random sample of the collected posts to compare their scores with the ones assigned by Perspective API. This way, you would be able to establish and report accuracy, precision and recall measures for how well Perspective detects anti-social acts in your specific dataset. The calculation of these evaluation metrics is outside the scope of this chapter, but well covered in other texts (see, for example, Dhaoui et al., 2017). In this case study, we use the threshold of 0.8.

While Perspective calculates several different scores, some of them are interrelated. For example, based on correlation analyses (see Table 4), the following four scores are highly correlated with each other (Pearson correlation > 0.9): *toxicity*, *severe toxicity*, *insult*, and *profanity*. This suggests that depending on one's research questions, it might be enough to examine one of the above-mentioned scores. For the purpose of this chapter, out of the four highly correlated scores, we will examine the *toxicity* score.

We also note that the *threat* score is the most 'conservative' metric because it flags a smaller proportion of posts as anti-social, between 0.2% to 1.7% of posts, depending on the set threshold. Considering the limited scope of the *threat* score relative to the other scores, we exclude it from further analysis. In sum, for the remainder of this chapter, we will examine the types of interactions and resulting communication networks based on the following anti-social scores: *toxicity*, *identity attack* and *attack on commenter* scores.

Table 4. Pearson correlation analysis among Perspective API scores

	Toxicity	Severe toxicity	Insult	Profanity	Identity attack	Threat	Attack on commenter
Toxicity	1	0.948	0.962	0.96	0.688	0.475	0.069
Severe toxicity	0.948	1	0.908	0.942	0.668	0.517	0.019
Insult	0.962	0.908	1	0.917	0.728	0.457	0.106
Profanity	0.96	0.942	0.917	1	0.578	0.402	-0.015
Identity attack	0.688	0.668	0.728	0.578	1	0.503	0.030
Threat	0.475	0.517	0.457	0.402	0.503	1	-0.017
Attack on commenter	0.069	0.019	0.106	-0.015	0.030	-0.017	1

Note: All correlation values are significant at the 0.01 level (2-tailed)

4.2 Social network analysis

While the review of toxicity scores offers a general sense of how toxic a particular group is, this analysis alone does not highlight which users tend to instigate such behaviour, which users are on the receiving end, and whether there is a specific pattern to the spread of anti-social behaviour. For example, is the anti-social behaviour a norm in the group as whole or are there users who are more likely to engage in such behaviour than others. Furthermore, are there signs of coordination among users (a behaviour known as *brigading*) to target others in the group? To help us answer these questions, we use Communatlytic to discover and export a communication network representing who replies to whom in the group. As noted earlier, the uniqueness of this network data is that edges have additional attributes assigned to them in the form of Perspective API scores.

To fully understand the inner-dynamics of r/metacanada, we turned to Gephi, an open-source software. Communatlytic exports network files in the GraphML format, which is supported by a wide variety of programs for social network analysis (SNA), including Gephi. Previous studies of various online groups suggest that by examining communication network structures, we may be able to predict the level and quality of group participation, and even the group's longevity (Chua et al., 2007; Gruzd & Haythornthwaite, 2013).

Excluding isolates – that is, posters who have not received any replies – the resulting network consisted of 2454 nodes and 14579 edges (see Figure 4). In this case, each node represents a redditor, and each edge represents a reply to an original post or another reply. The size of each node corresponds to the number of other users they replied to or received a reply from (also known as *total degree centrality*). Similar to other online groups (Yang et al., 2018), the r/metacanada network exhibits a *core-periphery* structure; that is, there is an active group of users in the core of the network who post and reply to each other, with less active group members found at the periphery of this network. Another metric that can be used to describe this network is *modularity*. It is a network-level measure that ranges from 0 to 1 where values closer to 0 suggest a highly connected network (Gruzd et al., 2016). By applying the label propagation algorithm (Raghavan et al., 2007), we calculated the value of *modularity* as 0.264. Because this value is closer to 0, it indicates that most conversations were primarily among the same group of users. We also note that the overall reciprocity of the network is 0.38, meaning that 38% of all ties among users are reciprocal (that is, they received a reply); this value is consistent with other online discussion groups sharing similar interests (i.e., politics and identity) (Del Valle et al., 2020; Sun, 2019).

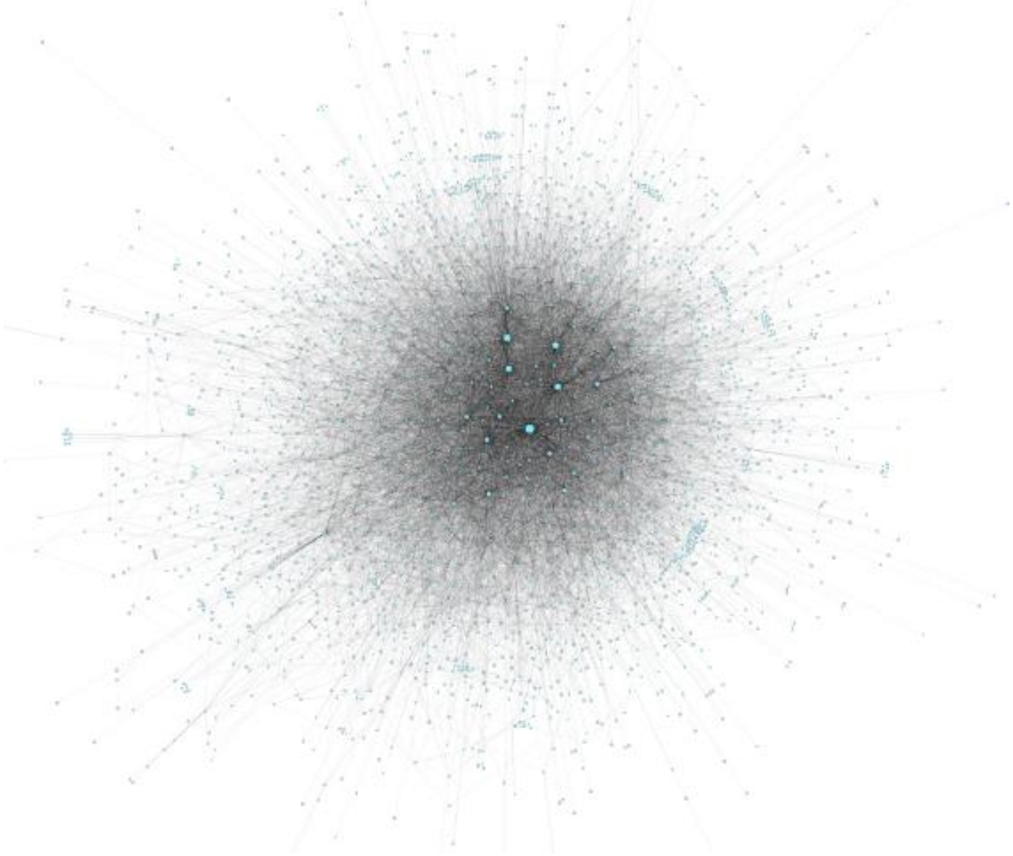
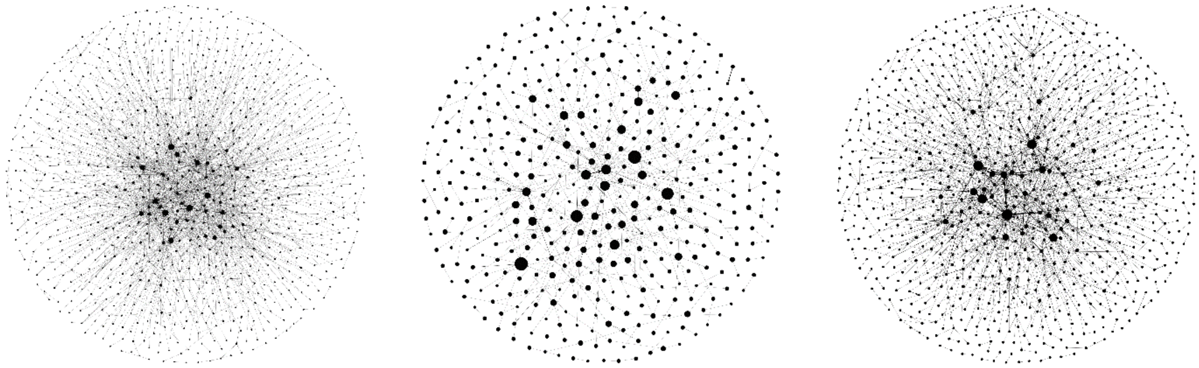


Figure 4. Network visualization of r/metacanada communication network.
(Node size = total degree centrality)

In order to identify the most toxic users and their interaction patterns, we used the Filter Tool in Gephi to show only edges with values higher than or equal to the selected threshold of 0.8 in accordance with the three scores we are examining: *toxicity*, *identity attack* and *attack on commenter* scores. Figure 5 shows the resulting network visualizations after the filter was applied. In each network visualization, the node size represents the number of other nodes the user replied to (out-degree centrality). This way, larger nodes represent users who tend to attack others. Table 5 lists the network-level properties and metrics for each of these sub-networks.



(a) Toxicity Sub-networks (b) Identity Attack Sub-networks (c) Attack on Commenter Sub-networks

Figure 5. Network visualization of anti-social interactions in r/metacanada with a threshold of 0.8 (Node size = out-degree centrality)

Table 5. Sub-Network level properties and metrics for the three selected scores of anti-social acts

	Toxicity	Identity Attack	Attack on Commenter
Number of Nodes	987	435	922
Number of Edges	1979	485	1504
Density	0.002	0.003	0.002
Clusters	35	49	53
Reciprocity	0.142	0.067	0.234
Modularity (based on label propagation community)	0.134	0.712	0.598

Even though all three networks in Figure 5 display a similar core-periphery structure, there are some structural differences between them. For example, the modularity scores for the Identity Attack and Attack on Commenter networks are much higher than for the Toxicity network. This suggests that interactions classified as Identity Attack and Attack on Commenter tend to stay within closely-connected groups of users (higher values of modularity), likely due to the directed nature of such attacks.

Based on the network visualizations, we can see how some users tend to be the primary spreaders of anti-social acts in this group (users depicted as larger size nodes). Furthermore, based on relatively low values of reciprocity, users who are attacked do not tend to reply in kind. The only exception is the Attack on Commenter network which has the highest reciprocity value of 0.234 relatively to the other two networks. This means that about 23.4% of interactions that were classified as “Attack on Commenter” are reciprocal, as opposed to only about 6.7% for the Identity Attack type interactions.

To explore this pattern of interactions further, for each of the three scores, a researcher can use the Data Laboratory option in Gephi to locate and examine users with the greatest number of incoming edges (in-degree centrality) and those with the greatest number of outgoing edges (out-degree centrality). Users with the highest in-degree values would be recipients of anti-social content from the greatest number of nodes. And users with the highest out-degree values are those who post anti-social replies to the greatest number of nodes in the network. While outside the scope of this chapter, future research may include employing a more qualitative and content-driven approach in examining the anti-social acts of these key users within the network in more detail. Coupled with the network-level data analysis and visualization, a qualitative approach would provide a more nuanced understanding of anti-social behaviour present within an online community, and help add richness to this line of research.

5 Conclusions

The rising tide of online anti-social behaviour has elevated public concern and skepticism over the perceived benefits and promise of social media in society (Bauman & Baldasare, 2015). A darker side of social media has emerged and remains evident today, with various countries, governing bodies, and citizens grappling with the impending normalization of aggressive behaviour, hostility, and negative discourse in online spaces. This realization has led to an influx of research examining the patterns of ‘anti-social’ in online communities, as well as the development of the necessary tools required for systematic investigations.

One such tool is Communalytic. It provides researchers with an accessible and easy to use approach for analyzing public groups on Reddit. Its ability to export data for social network analysis, along with anti-social scores, makes it a useful tool for researchers to examine and analyze anti-social behaviour both at the group and user levels. These functions are supported by the use of Google’s Perspective API, which uses a machine learning classification system to score the content across various categories of anti-social behaviour and in six different languages (English, French, German, Italian, Portuguese and Spanish). Furthermore, the ability to export network-level data allows for an additional analysis of online exchanges using metrics from social network analysis.

This chapter serves as an introductory guide to the functionalities and features of Communalytic, as well as a case study on how to use this tool. Specifically, the tool was used to examine online toxicity within the r/metacanada subreddit during a 10-day time-period preceding the 2019 Canadian Federal election. Based on a threshold of 0.8, approximately ten percent of the posts collected during this 10-day period were categorized as toxic. The network-level data was also exported and examined using Gephi software, to visualize the overall network, as well as those depicting general toxicity, identity attacks, and attacks on commentators.

To conclude, Communalytic is a tool that offers researchers the ability to study anti-social interactions on Reddit at scale. Given the ubiquity of social media in modern-day society, the significance of research examining anti-social behaviour on sites like Reddit is imperative. It can also support moderators of these communities in their efforts to sustain a healthy and welcoming environment for its members, and to evaluate their effectiveness.

References

Aggarwal, C. C. (2013). *Outlier Analysis*. Springer New York. http://link.springer.com/10.1007/978-1-4614-6396-2

Akoglu, L., McGlohon, M., & Faloutsos, C. (2010). oddball: Spotting Anomalies in Weighted Graphs. In M. *J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), Advances in Knowledge Discovery and Data Mining* (pp. 410–421). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13672-6_40

Akoglu, L., Tong, H., & Koutra, D. (2014). Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, 29(3), 626–688. https://doi.org/10.1007/s10618-014-0365-y

Al-Makhadmeh, Z., & Tolba, A. (2019). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*. https://doi.org/10.1007/s00607-019-00745-0

Anderson, Katie Elson. (2015). Ask Me Anything: What is Reddit?. *Library Hi Tech News 32*(5). Retrieved from doi:10.7282/T3D220BR.

Anderson, M. (2018). *A Majority of Teens Have Experienced Some Form of Cyberbullying*. Pew Research Center. Retrieved from http://www.pewinternet.org/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/

Anti-social Behaviour, Crime and Policing Act 2014. (2014). Retrieved October 23, 2019, from http://www.legislation.gov.uk/ukpga/2014/12/contents

Awan, I. (2014). Islamophobia and Twitter: A Typology of Online Hate Against Muslims on Social Media. *Policy & Internet*, *6*(2), 133–150.

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, *27*, 1–8.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Third International AAAI Conference on Weblogs and Social Media.*

Bauman, S., & Baldasare, A. (2015). Cyber Aggression Among College Students: Demographic Differences, Predictors of Distress, and the Role of the University. *Journal of College Student Development*, *56*(4), 317–330.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. ArXiv:1903.04561 [Cs, Stat]. http://arxiv.org/abs/1903.04561

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102.

Chakrabarty, N. (2020). A Machine Learning Approach to Comment Toxicity Classification. In *Computational Intelligence in Pattern Recognition* (pp. 183–193). Springer.

Chua, Z., Goh, K.-Y., Kankanhalli, A., & Phang, C.-W. (2007). Investigating participation in online policy discussion forums over time: Does network structure matter? *ICIS* 2007 Proceedings, 117.

Costello, M., Hawdon, J., & Ratliff, T. N. (2017). Confronting Online Extremism: The Effect of Self-Help, Collective Efficacy, and Guardianship on Being a Target for Hate Speech. *Social Science Computer Review*, *35*(5), 587–605.

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, *102*(Complete), 79–84.

Cyberbullying. In *Proceedings of the International Conference on Weblog and Social Media* (Social Media Web Workshop).

Dadvar, M., Trieschnigg, D., & Jong, F. de. (2013). Expert knowledge for automatic detection of bullies in social networks. 57–64. http://bnaic2013.tudelft.nl/proceedings/papers/paper_79.pdf

Dadvar, M., Trieschnigg, D., & Jong, F. de. (2014). Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In *M. Sokolova & P. van Beek (Eds.), Advances in Artificial Intelligence* (pp. 275–281). Springer International Publishing. https://doi.org/10.1007/978-3-319-06483-3_25

Dadvar, M., Trieschnigg, R. B., & de Jong, F. M. G. (2013). Expert knowledge for automatic

Del Valle, M. E., Gruzd, A., Kumar, P., & Gilbert, S. (2020). Learning in the Wild: Understanding Networked Ties in Reddit. In *Mobility, Data and Learner Agency in Networked Learning* (pp. 51-68). Springer, Cham.

Del Valle, M. E., Gruzd, A., Kumar, P., & Gilbert, S. (2020). Learning in the Wild: Understanding Networked Ties in Reddit. In *N. B. Dohn, P. Jandrić, T. Ryberg, & M. de Laat (Eds.), Mobility, Data and Learner Agency in Networked Learning* (pp. 51–68). Springer International Publishing. https://doi.org/10.1007/978-3-030-36911-8_4

Delisle, L., Kalaitzis, A., Majewski, K., de Berker, A., Marin, M., & Cornebise, J. (2019). A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter. *ArXiv Preprint ArXiv*:1902.03093.

detection in social networks. In *Databases Theory and Applications, LNCS'12*, 160–171.

detection of bullies in social networks. In *25th Benelux Conference on Artificial Intelligence, BNAIC 2013* (pp. 57-64). Delft: Delft University of Technology.

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, 34(6), 480–488. https://doi.org/10.1108/JCM-03-2017-2141

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the Detection of Textual

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the Detection of Textual Cyberbullying. *The Social Mobile Web*, Papers from the 2011 ICWSM Workshop. https://www.researchgate.net/publication/221297959_Modeling_the_Detection_of_Textual_Cyberbullying

Duggan, M. (2017, July 11). Online Harassment 2017. Retrieved November 14, 2017, from http://www.pewinternet.org/2017/07/11/online-harassment-2017/

Dybala, M. P. P., Masui, T. M. F., Rzepka, R., & Araki, K. (2010). Machine Learning and Affect Analysis against Cyber-Bullying. Proceedings of *the Linguistic and Cognitive Approaches to Dialog Agents Symposium.*

Faris, R., Ashar, A., Gasser, U., & Joo, D. (2016). *Understanding Harmful Speech Online* (SSRN Scholarly Paper No. ID 2882824). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2882824

Giménez Gualdo, A. M., Hunter, S. C., Durkin, K., Arnaiz, P., & Maquilón, J. J. (2015). The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. *Computers & Education*, *82*(Complete), 228–235.

Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies, 18*(4), 515–524. https://doi.org/10.1080/14680777.2018.1447345

Gorrell, G., Bakir, M. E., Greenwood, M. A., Roberts, I., & Bontcheva, K. (2019). Race and Religion in Online Abuse towards UK Politicians: Working Paper. *ArXiv:1910.00920 [Cs]*. http://arxiv.org/abs/1910.00920

Gruzd, A., & Haythornthwaite, C. (2013). Enabling Community Through Social Media. *Journal of Medical Internet Research*, 15(10), e248. https://doi.org/10.2196/jmir.2796

Gruzd, A., & Tsyganova, K. (2015). Information Wars and Online Activism During the 2013/2014 Crisis in Ukraine: Examining the Social Structures of Pro- and Anti-Maidan Groups. *Policy & Internet*, 7(2), 121–158. https://doi.org/10.1002/poi3.91

Gruzd, A., Paulin, D., & Haythornthwaite, C. (2016). Analyzing Social Media And Learning Through Content and Social Network Analysis: A Faceted Methodological Approach. *Journal of Learning Analytics*, 3(3), 46–71. https://doi.org/10.18608/jla.2016.33.4

Gutierrez, J. A. W. (2018). Students evaluate music theory courses: A reddit community survey.
Hara, N., Abbazio, J., & Perkins, K. (2019). An emerging form of public engagement with

Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve Del Valle, M., & Paulin, D. (2018). Learning in the Wild: Coding for Learning and Practice on Reddit. *Learning, Media and Technology, 43*(3), 219–235. https://doi.org/10.1080/17439884.2018.1498356

Hodson, J., Gosse, C., Veletsianos, G., & Houlden, S. (2018). I get by with a little help from my friends: The ecological model and support for women scholars experiencing online harassment. *First Monday*, *23*(8).

Hopp, T., & Vargo, C. J. (2019). Social Capital as an Inhibitor of Online Political Incivility: An Analysis of Behavioral Patterns Among Politically Active Facebook Users. *International Journal of Communication*, 13, 21.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. *ArXiv:1702.08138 [Cs]*. http://arxiv.org/abs/1702.08138

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. *ArXiv*:1702.08138 [Cs]. http://arxiv.org/abs/1702.08138

Howard, P., & Bradshaw, P. (2017). Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. *Oxford Internet Institute*, *2017.12*. Retrieved from https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6

Huang, J., Sun, H., Song, Q., Deng, H., & Han, J. (2013). Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network. *IEEE Transactions on Knowledge and Data Engineering*, 25(8), 1876–1889. https://doi.org/10.1109/TKDE.2012.100

Jain, E., Brown, S., Chen, J., Neaton, E., Baidas, M., Dong, Z., Gu, H., & Artan, N. S. (2018). Adversarial Text Generation for Google's Perspective API. *2018 International Conference on Computational Science and Computational Intelligence* (CSCI), 1136–1141. https://doi.org/10.1109/CSCI46756.2018.00220

Jigsaw. (2019, December 2). *Increasing Transparency in Perspective's Machine Learning Models*. Medium. https://medium.com/the-false-positive/increasing-transparency-in-machine-learning-models-311ee08ca58a

Jigsaw. (2019, December 2). Increasing Transparency in Perspective's Machine Learning Models. Medium. https://medium.com/the-false-positive/increasing-transparency-in-machine-learning-models-311ee08ca58a

Kenski, K., Coe, K., & Rains, S. A. (2017). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, 1-20.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137.

Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. *AAAI*.

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research*, *27*(4), 991–1010. https://doi.org/10.1108/IntR-02-2017-0072

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research*, 27(4), 991–1010. https://doi.org/10.1108/IntR-02-2017-0072

Lampe, C., Zube, P., Lee, J., Park, C. H., & Johnston, E. (2014). Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, *31*(2), 317–326.

Lindsay, M., Booth, J. M., Messing, J. T., & Thaller, J. (2016). Experiences of Online Harassment Among Emerging Adults: Emotional Reactions and the Mediating Role of Fear. *Journal of Interpersonal Violence*, *31*(19), 3174–3195.

Macskassy, S. A., & Provost, F. (2007). A Brief Survey of Machine Learning Methods for Classification in Networked Data and an Application to Suspicion Scoring. In *E. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, & A. X. Zheng (Eds.), Statistical Network Analysis: Models, Issues, and New Directions* (pp. 172–175). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73133-7_13

Maity, S. K., Chakraborty, A., Goyal, P., & Mukherjee, A. (2018). Opinion Conflicts: An Effective Route to Detect Incivility in Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–27. https://doi.org/10.1145/3274386

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Massanari, A. L., & Chess, S. (2018). Attack of the 50-foot social justice warrior: The discursive construction of SJW memes as the monstrous feminine. *Feminist Media Studies*, *18*(4), 525–542. https://doi.org/10.1080/14680777.2018.1447333

Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2019). “And We Will Fight For Our Race!’’" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan.” *ArXiv Preprint*:1901.09735.

Nahar, V., Unankard, S., Li, X., & Pang, C. (2014). Semisupervised learning for cyberbullying

Nithyanand, R., Schaffner, B., & Gill, P. (2017). Online Political Discourse in the Trump Era. *ArXiv:1711.05303 [Cs]*. Retrieved from http://arxiv.org/abs/1711.05303

O’Kane, N., Zhang, C., Lama, Y., Hu, D., Jamison, A., Quinn, S. C., & Broniatowski, D. A. (2019). Characterizing Trends in Human Papillomavirus Vaccine Discourse on Reddit (2007-2015): An Observational Study. *JMIR Public Health and Surveillance*, *5*(1). https://doi.org/10.2196/12480

Obadimu, A., Mead, E., Hussain, M. N., & Agarwal, N. (2019). Identifying Toxicity Within YouTube Video Comment. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 214–223.

Pappa, G. L., Cunha, T. O., Bicalho, P. V., Ribeiro, A., Silva, A. P. C., Jr, W. M., & Beleigoli, A. M. R. (2017). Factors Associated with Weight Change in Online Weight Management Communities: A Case Study in the LoseIt Reddit Community. *Journal of Medical Internet Research*, *19*(1), e17. https://doi.org/10.2196/jmir.5816

Pavlopoulos, J., Thain, N., Dixon, L., & Androutsopoulos, I. (2019). ConvAI at SemEval-2019 Task 6: Offensive Language Identification and Categorization with Perspective and BERT. Proceedings of the *13th International Workshop on Semantic Evaluation*, 571–576.

Perozzi, B., Akoglu, L., Iglesias Sánchez, P., & Müller, E. (2014). Focused Clustering and Outlier Detection in Large Attributed Graphs. Proceedings of *the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1346–1355. https://doi.org/10.1145/2623330.2623682

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 48(12), 4730–4742. https://doi.org/10.1007/s10489-018-1242-y

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A Benchmark Dataset for Learning to Intervene in Online Hate Speech. *ArXiv:1909.04251 [Cs]*. Retrieved from http://arxiv.org/abs/1909.04251

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*(3), 036106.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3), 036106.

Reddit. (2018a). The conversation starts on Reddit. Retrieved November 2019 from https://about.reddit.com/

Reddit. (2019). Account and Community Restrictions. Retrieved from https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-threaten-harass-or-bully

Redditmetrics. (n.d.). New subreddits by month—Reddit history. Retrieved February 29, 2020, from https://redditmetrics.com/history/month

Rossini, P. (2019). Toxic for whom? Examining the targets of uncivil and intolerant discourse in online political talk. P. Moy & D. Matheson (eds). Voices: Exploring the Shifting Contours of Communication, 221-242. New York: Peter Lang.

Runions, K. C., Bak, M., & Shaw, T. (2017). Disentangling functions of online aggression: The Cyber-Aggression Typology Questionnaire (CATQ): Cyber-Aggression Typology Questionnaire. *Aggressive Behavior*, *43*(1), 74–84.

Ryerson Leadership Lab. (2019). Rebuilding the Public Square. *Public Report.* Ryerson University, Toronto. Canada. Available at https://www.ryersonleadlab.com/rebuilding-the-public-square
science: Ask Me Anything (AMA) sessions on Reddit r/science. *PLOS ONE*, *14*(5), e0216789. https://doi.org/10.1371/journal.pone.0216789

Shatz, I. (2017). Fast, Free, and Targeted: Reddit as a Source for Recruiting Participants Online. *Social Science Computer Review*, *35*(4), 537–549. https://doi.org/10.1177/0894439316650163

Southern, R., & Harmer, E. (2019). Twitter, Incivility and "Everyday" Gendered Othering: An Analysis of Tweets Sent to UK Members of Parliament. *Social Science Computer Review*, 089443931986551. https://doi.org/10.1177/0894439319865519

Staudt Willet, K. B., & Carpenter, J. (2019, March). Educators on the Front Page of the Internet: Education-Related Subreddits as Learning Spaces. In Society for Information Technology & Teacher Education International Conference (pp. 2787-2795). Association for the Advancement of Computing in Education (AACE).

Su, L. Y.-F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society*, *20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205

Sun, Y. (2019). How conversational ties are formed in an online community: A social network analysis of a tweet chat group. Information, *Communication & Society*: 1–18. https://doi.org/10.1080/1369118X.2019.1581242

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates: Incivility in Interactions With Candidates on Twitter. *Journal of Communication*, *66*(6), 1007–1031. https://doi.org/10.1111/jcom.12259

Topinka, R. J. (2018). Politically incorrect participatory media: Racist nationalism on r/ImGoingToHellForThis. *New Media & Society*, *20*(5), 2050–2069. https://doi.org/10.1177/1461444817712516

Yang, J., Zhang, M., Shen, K. N., Ju, X., & Guo, X. (2018). Structural correlation between communities and core-periphery structures in social networks: Evidence from Twitter data. *Expert Systems with Applications*, 111, 91–99. https://doi.org/10.1016/j.eswa.2017.12.042

Yoo, M., Lee, S., & Ha, T. (2019). Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. *Information Processing & Management*, *56*(4), 1565–1575. https://doi.org/10.1016/j.ipm.2018.10.001