The Dissertation Committee for Mohini Tellakat Certifies that this is the approved version of the following Dissertation:

Understanding Intergroup vs. Intragroup Toxicity in Online Communities

Committee:

James W. Pennebaker, Supervisor

Samuel Gosling

Elliot Tucker-Drob

Sarah Abraham

Understanding Intergroup vs. Intragroup Toxicity in Online Communities

by

Mohini Tellakat

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2020

Dedication

I would like to thank my advisor, Jamie Pennebaker for taking a chance on me, a computer scientist, and for all his support and guidance throughout my graduate career. I would also like to thank the members of my committee, Sam Gosling, Elliot Tucker-Drob, and Sarah Abraham for advice and guidance over the years. I could not have completed this dissertation without the help of numerous other graduate students and undergraduate researchers who worked with me at UT and my husband, Alexander Hearn for helping troubleshoot code when things inevitably went wrong.

Abstract

Intergroup vs. Intragroup Toxicity in Online Communities

Mohini Tellakat, Ph.D.

The University of Texas at Austin, 2020

Supervisor: James W. Pennebaker

The digital revolution has changed the world. People around the globe can stay connected with strangers, friends, and family, can learn about in an infinite number of topics, and can be a part of life changing movements and discussions. Anonymity on the internet, however, can lead to negative behaviors such as cyberbullying, trolling, and the formation of groups of highly discriminatory individuals.

Internet-based interactions can lead to toxic (e.g. aggressive, hateful) online environments which can affect people's sense of well-being and mental health in general. Toxic behavior comes in many forms, however, so the goal of this dissertation is to understand the psychological differences between two types of toxicity: Intergroup toxicity and intragroup toxicity. Intergroup toxicity is defined as expressing hatred, disgust, and general ill-will to people who are not within a community's ingroup. Intragroup toxicity is defined as hatred, disgust, aggression, and general ill-will directed to people within a community's ingroup.

Using language analytic techniques on large datasets from Reddit, the current project builds on research done on toxicity in online communities to understand the differences between intergroup and intragroup toxicity. Specifically, the research analyzes psychological differences in terms of linguistic markers of mental health and style (Tausczik & Pennebaker 2010, Martens, Shen, Iosup, & Kuipers 2015). After analyzing the differences in the two kinds of toxicity, the research addresses how participating in a toxic group relates to the way people behave in other online groups they participate in.

For the analyses, 10 subreddits with intergroup toxicity and 10 subreddits with intragroup toxicity were selected based on a network analysis (Datta & Adar 2019) and using Reddit's controversiality measure. The language of the two groups was studied using random sampling and bootstrapping methodologies to compare the means between groups with intragroup toxicity and groups with intragroup toxicity. The full reddit histories of the people used in Study 1 were used to answer the research questions posed in Study 2. Authors were characterized into 4 major subgroups based on individual and group levels of toxicity: Low individual-level toxicity in both groups with intergroup and intragroup toxicity and high-individual level toxicity in both groups with intergroup and intragroup toxicity.

Overall, the research found that there is a personality or trait-based form of toxicity among those who used toxic words at high rates in their home groups. If people used toxic words at high rates in the initial highly toxic 20 subreddits that were collected, they used toxic language in many of the other subreddits that they participated in. Additionally, if people had low individual-level toxicity, their toxic behaviors in groups other than the

initial 20 subreddits tended to be related to which type of initially toxic group they posted in the majority of the time. Specifically, people with high individual-level toxicity behaved as though they had toxic personalities. The same patterns were found among people with low individual-level toxicity who participated in groups with intergroup toxicity. Finally, people with low individual-level toxicity who participated in groups with intragroup toxicity did not use toxic language as much, indicating that people who had low individual-level toxicity could be influenced by the groups that they chose to participate in.

The differences in toxic language use showed that there cannot be a blanket statement or a blanket treatment of toxicity online. Though people with high-individual level toxicity seem to have trait-level toxicity, people who have low-individual level toxicity seem to be influenced by their communities. Being able to stage appropriate interventions for people who do have low-level toxicity becomes important when considering groups with intergroup toxicity promote more hatred than groups with intragroup toxicity do. Recognizing these group members and engaging with them before they spread toxic behaviors in the other groups that they participate in can be key to reducing proliferation of toxic behaviors online. With so many people engaging in important issues in online communities, the use of effective interventions for hate and for discontent become more and more needed. By helping to shed light onto a few aspects of online toxicity, the research from this dissertation can inform the understanding of the behavior of online communities.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Most people around the world spend a large portion of their lives online. According to a survey conducted in 2019, over 5 billion people in the world use mobile phones and over 4.3 billion use the internet (Kemp 2020). With the majority of the world's population connected to the internet and using it for long periods of time, it becomes important to understand the impacts of communicating with people anonymously for hours a day, as well as absorbing more information than they have ever had a chance to before.

With access to multiple sources of information and new routes for communication come consequences that the creators of the internet might not have imagined possible. In a positive light, people around the globe can stay connected with friends, family, global politics at any time of day and night, meet like-minded individuals, and form support groups, for example. However, anonymity on the internet can lead to negative behaviors such as cyberbullying (Lapidot-Lefler & Barak 2012, Elliot 2012), trolling (Cheng, Bernstein, Danescu, & Leskovec 2017, Hopkinson 2013), and the formation of groups of highly discriminatory individuals (Greenwald & Pettigrew 2014, Brewer 1999, Del Vicario et. al. 2016).

So, what makes a positive online interaction and community so different from one that breeds negativity? The current research aims to understand the difference between two major kinds of negative behavior in the online community of Reddit: intergroup toxicity and intragroup toxicity. Intergroup toxicity occurs when people within a group are united

in hatred and toxic behaviors toward an outgroup, whereas intragroup toxicity is characterized by in-fighting within a group.

Reddit, a large online forum, is comprised of millions of subcommunities called subreddits that have both positive and negative reputations. People who have similar interests congregate within a subreddit to have largely text-based conversations related to their interests, whether that interest is professional wrestling or Tetris, or something entirely different.

By analyzing the language from communities that are deemed toxic by redditors and by internet users at large, the research aims to understand whether or not there are any impacts of participating in different kinds of toxic communities and if there are, who displays toxic behavior and where. This way, community managers, social media companies and fellow researchers will be able to develop effective interventions for stopping and addressing toxic behaviors online before they go out of control.

Intergroup Toxicity

Intergroup toxicity, in the context of this research, is defined as expressing hatred, disgust, and general ill-will to people who are not within a community's ingroup. This type of toxicity is typically not directed toward anyone within the ingroup unless they are violating social norms. Some groups that exhibit extreme intergroup toxicity have also been classified as hate groups in the literature and the online support of these groups has been noted by organizations such as the Southern Poverty Law Center (Hankes & Dijk 2019).

How do these hate groups form and why? A lot of past research focuses on the formation of discriminatory views within individuals and groups. One of the theories in

this research is that ingroup favoritism leads toward intergroup discrimination (Greenwald & Pettigrew 2014, Brewer 1999). In this case, people band together because of their similarity to other individuals in the group. They tend to prefer being around people who are like them and thus have a favoritism toward people in their ingroup. Since their preference is for their ingroup, this passively leads to discrimination of people who are not like them, aka an outgroup. Though this does not necessarily mean that these people have an explicit hatred of people who are unlike them, these studies show that people who have ingroup favoritism will choose to interact with people who are more similar to themselves than people who are different from them.

In a study done by Byrne (1961), participants were asked to take a 26-question attitudes survey, and 2 weeks later were asked to evaluate someone who had also taken the attitudes survey. The responses that participants were asked to evaluate were experimentally controlled and were not actual participant responses. They found that liking of the evaluated person was a function of attitude similarity, meaning that participants liked people more if they shared the similar attitudes than if they did not. Additionally, Rokeach and Mezei (1966) did a study that found that perceiving conflicting beliefs triggered more race prejudice than race itself did, which further illustrates the notion that preference for an ingroup or preference for similar values and beliefs can lead to discrimination.

Preferences for ingroups can show up online in the form of online echo chambers. Echo chambers typically form when people select information that confirms their views (i.e. confirmation bias) and/or participate in communities that polarize their opinions (Del Vicario et. al. 2016, Gilbert, Bergstrom & Karahalios 2009, Flaxman Goel & Rao 2016,

Williams et. al. 2015). Though there have been studies that show that social media usage can disseminate ideas across ideologies (Barbera et. al. 2015), many people still choose to partake in groups that foster views and opinions that mirror their own (Van Alstyne & Brynjolfsson 1996). Additionally, there is research that states that the more you participate in social echo chambers, the more negative you become (Del Vicario et. al. 2016). Though the above viewpoints do support the notion that a group can form discriminatory views based on preferences for an ingroup, these behaviors do not seem strong enough to lead to the formation of a hate group.

For a group to be considered a hate group, some research states that shared hatred of another group or entity, instead of shared appreciation for self-similarity is the defining factor (Schaffer & Navarro 2003). People are brought together, not by what makes them similar, but hate of something different from them. In that sense, these people share one similarity, the hate of an outgroup, and because of that hate, they choose to act toward that outgroup in negative ways, and potentially even violence. According to Schaffer and Navarro, haters will gather with each other based on a shared hatred, they will define a group, disparage the target of the hate, taunt the target, attack the target with or without weapons, and then will destroy the target. Though the later parts of this cycle may not always occur, there has been research to show that shared hatred can bring people together in this way (Levin & MacDemott 2013).

Interestingly, though some scholars on hate groups label these groups as toxic (Massanari 2015), the behaviors of these groups toward ingroup members are often encouraging and uplifting. Without the context that a group was a hate group or exhibited

intergroup toxicity, these groups might look like they have a healthy structure and supportive atmosphere. So, where does this perception of toxicity for groups with intergroup toxicity come from? Are people within a group labeled with intergroup toxicity truly toxic?

The current research aims to address this question by comparing a community with intergroup toxicity to a community that exhibits intragroup toxicity, much of which has been seen in areas like politics (Kaarbo & Gruenfeld 1998), violent crime (Hipp, Tita & Boggess 2009), adolescent and educational environments (Felmlee & Faris 2016) , but the comparison has been studied less so in online contexts.

Intragroup Toxicity

Intragroup toxicity, in the context of this research, is defined as hatred, disgust, aggression, and general ill-will to people within a community's ingroup. Typically, it is directed to an individual or small group of individuals within a community and is typically a direct attack. Online, the combination of anonymity and the lack of face to face interaction can lead to even more of these intragroup behaviors (Lapidot-Lefler & Barak 2012). In the current literature on intragroup toxicity, these behaviors, as exhibited online, have been called many things such as toxic disinhibition or flaming, cyberbullying, and trolling.

Toxic disinhibition, or "flaming" is defined as the use of hostile expressions toward another person, especially in an online setting (Lapidot-Lefler & Barak 2012). Toxic disinhibition occurs more in online settings because of the lack of in-person interaction which then allows for people to not censor their thoughts when interacting with others.

Specifically, having the feeling that one is unidentifiable creates this sense of security when expressing negative sentiment toward another person online, which is why most cases of toxic disinhibition are seen in anonymous forums like Reddit. Additionally, in some communities, this "flaming" behavior is a social norm. For example, in some online gaming communities, often, people will play video games with people they do not know, and only communicate via text or through microphones, where a player's only identification is their username (Elliot 2012). Though this behavior is a norm and can sometimes be viewed as humorous to those involved, flaming could result in negative psychological effects for those who are the victims of this behavior as they might in more traditional bullying situations.

Another kind of toxic behavior seen in these communities is trolling. Trolling is defined as intentionally inciting conflict within a conversation or community (Cheng, Bernstein, Danescu, & Leskovec 2017). Trolls (i.e. people who exhibit trolling behavior) tend to come into conversations just to stir up conflict and feed off others' aggression either toward themselves, or toward each other. In many cases, the trolls disguise themselves as people who want to be a part of the in-group, but then purposely exhibit aggressive behaviors to cause discord within a harmonious community (Hopkinson 2013). These trolling behaviors often succeed in dismantling once harmonious communities and can cause people who normally would not be aggressive to act like trolls themselves (Cheng Bernstein, Danescu, & Leskovec 2017), but also can bring members of a community closer together, united against the troll (Hopkinson 2013). In uniting against the trolls, members

of the community feel as though they are taking a stand against the people who are bringing toxicity into what was once a civil discussion or a more civil community in general.

A third form of toxic behavior seen in groups with intragroup toxicity is cyberbullying. Cyberbullying is defined as an act of aggression against another individual in an online setting, typically used to cause harm or distress to an individual and has a repetitive nature (Whittaker & Kowalski 2015). In many cases, cyberbullying has been studied in peer to peer situations (i.e. through texting, and instant messaging), but in other cases, there is a power differential and anonymity associated with cyberbullying behavior that makes it particularly potent (Barlett, Gentile & Chew 2016). In every case, the victim is seen as "weaker" and is picked on by the cyberbully in a repetitive fashion. This behavior differs from both Toxic Disinhibition and Trolling because of the focused and repetitive nature of the aggression.

Cyberbullying, trolling, and flaming can cause people who have been hurt by such behaviors to label a community as toxic. Similarly, people who have been hurt or who have witnessed attacks from hate groups will label people participating in hate groups as toxic even though participation in the two types of groups can look different internally. In one type of group, people are actively bullied, and conflict is stirred, while the other breeds negative thoughts and feelings toward an outgroup and can seem toxic from the perspective of that outgroup. However, though some people hate participating in communities that encourage these behaviors, the communities still live on. What is it, then, that keeps these communities alive? Is the toxicity contained within the community? Is it called out? Do people who exhibit toxic behaviors in these communities go on to be toxic in the other

communities they are a part of? The current research aims to disentangle the effects of being in a community that has intragroup vs. a community that has outgroup hatred to make headway on understanding what motivates toxicity in online communities.

Understanding toxicity in online settings

Though many toxic behaviors can be seen in in-person settings, studying the effect of toxic behaviors in online communities is increasingly important. With many people's social interactions migrating online, and with the development and prevalence of anonymous online forums like Reddit and 4Chan that allow people to access and communicate about a variety of subjects, the impacts of toxic behavior reach farther than a person's real-life social circle (Levin 2002). Since there is more opportunity to interact with random strangers in online settings and to even interact with people you know without seeing their faces, there can be less emotional connection to the people that one interacts with. This anonymity can lead to mobilization toward aggression and violence toward others, organizing against an outgroup, and causing in-person riots, protests, and activities that can physically and emotionally harm others.

However, merely being a member of a group that is deemed toxic either by in-group members, or an outgroup might not necessarily mean that the behavior that a person exhibits within a group leads to any sort of toxic action in the other groups they are a part of. Simply put, being a part of a toxic group might not influence the other aspects of a person's online or real-world life. Maybe the group is just a place where someone goes to vent their frustrations or listen to other like-minded individuals. Or maybe their toxic

behavior was incited by another individual, but outside of the context of that specific group, they are completely fine, upstanding members of society.

The Language of Toxicity

In online forums and communities, text is the primary form of communication. Whether it's replying to someone on social media or having a conversation with multiple strangers in a forum, people use text to communicate ideas and opinions with each other. With recent advances in text analysis and natural language processing, analyzing large data sets comprised of people's texts has become much more attainable. While there have been many developments in all aspects of language analysis (i.e. topic modeling, word embeddings, conversation generation), the piece most relevant to this work is linguistic style.

People's language styles are largely measured in their use of function and emotion words. Function words are words that tend to have little lexical meaning but convey grammatical relationships between words in a sentence. These words fall under the category of pronouns, articles, prepositions, auxiliary verbs, conjunctions, negations, and non-referential adverbs. Studying how a person uses these words in conjunction with things like emotion words and social words can give us insight into a person, or group's frame of mind (Chung & Pennebaker 2007).

In previous work, analyzing linguistic style has been used to determine how the psychology of politicians has changed over time (Jordan, Sterling, Pennebaker & Boyd 2019), how people of different status behave (Kacewicz, Pennebaker, Davis, Jeon, & Graesser 2014), and how people react to and recover from trauma (Cohn, Mehl &

Pennebaker 2004) amongst many other psychological phenomena. Since language is an extremely powerful marker of people's behavior, using this analytical method for data coming from online social environments seemed fitting.

But how can we use language to detect toxic behaviors? Past research has used language as a proxy for toxic behaviors in online communities, specifically in online gaming communities (Martens, Shen, Iosup, & Kuipers 2015). The researchers found that using swearing as an indicator for toxicity captured many toxic interactions, especially when the swearing was directed at another person. Blackburn and Kwak (2014) also found in an analysis of in-game chats from a popular online video game, League of Legends, that matches were deemed toxic if the following behaviors occurred: Assisting the enemy team, inappropriate names, negative attitude, offensive language, and spamming verbal abuse. Though these categories were manually coded, the data was still language data, and markers of these behaviors are like the ones used in the Martens et. al study: swearing, negative emotion, and other forms of offensive language. Based on the above studies and others (Kordis & Smitran 2018), toxic behavior is generally marked by a use of swear words and phrases, slurs, and anger, so any group that is toxic or harbors toxic people should have these indicators.

However, as stated above, intergroup toxicity and intragroup toxicity look different from each other in the ways that people tend to behave toward one another within a group. In groups with intergroup toxicity, members tend to band together and support one another in the hatred of an outgroup, whereas in groups with intragroup toxicity, members tend to

be hateful toward members of their own group. Because of this difference in toxic behaviors, the current research aims to address the following major questions :

1. **To what extent are groups that show intergroup toxicity and/or intragroup toxicity psychologically different from each other?**
2. **How does being a part of a group with intergroup toxicity and/or intragroup toxicity relate to a person's interactions with other groups?**

Toxicity influences people who are perpetrating toxic behaviors, people who are victims of toxicity, and people who are bystanders but still participate in groups known to harbor toxic behavior. Understanding how online toxicity differs in its manifestation and effects on a person's social life online will shine light on different ways to approach the issues of toxic behavior online.

Current Research Goals

RQ1: To what extent are groups that show intergroup toxicity and/or intragroup toxicity psychologically different from each other?

This research question more specifically addresses how these two groups differ psychologically from each other given that they are both toxic. Since the literature on the language of toxicity has measured the construct using swear words and phrases, anger words, and slurs, and both types of groups are deemed to be toxic, the real interest lies in the other psychological differences between the two groups. Base rates for toxic language will be compared as well but will likely not be the only differentiating factor between the two types of toxic groups.

We predict that people who participate in groups that have intergroup toxicity (or externally toxic groups) will be more social, supportive, and more socially well-adjusted than those that have intragroup toxicity (or internally toxic groups) since members of groups that have intergroup toxicity, or hate groups, tend to be united against a cause, whereas those with intragroup toxicity tend to fight amongst themselves.

Having a united, cohesive sense within a group can lead to better mental health outcomes for people in the group (Midtgaard, Rorth, Stelter & Adamsen 2006, Newson, Buhrmester & Whitehouse 2016, Budman et. al 1989) whereas constant in-fighting can lead to victims of abuse having negative mental health outcomes like depression or anxiety (Tsuno et. al. 2009, Jamieson, Valdesolo & Peters 2014). However, participating in groups that have intergroup toxicity can lead to similar outcomes when people from that group interact with others who do not share their opinions.

Understanding the psychological differences between people who participate in groups with intergroup toxicity and intragroup toxicity and how toxicity manifests itself in each type of group can give community managers and other leaders in the online world more specific tools for how to handle these groups and the individuals in these groups who show toxic behaviors since not all toxicity seems to be the same in terms of who it is expressed toward.

RQ2: How does being a part of a group with intergroup toxicity vs. intragroup toxicity relate to a person's interactions with other groups?

After understanding the psychological differences between people who participate in groups with either intergroup or intragroup toxicity, the second research question

addresses how the effects of intergroup vs intragroup toxicity relate to the ways that individual group members interact with people in less toxic or non-toxic groups. Understanding the nature of these effects will allow for dissecting whether toxicity is inherent within a person, whether it is provoked situationally, whether the type of toxicity one is exposed to matters, or some combination of the above.

The phenomenon we are trying to understand is the person – situation interaction with respect to toxicity, and whether this interaction differs between intergroup vs. intragroup toxicity. There has been extensive research on whether or not personality or situations are stronger in influencing people's behavior (Milgram 1963, Mischel 1968, Funder 2009), but now there is more work stating that certain behaviors can emerge based on the interaction of a person's personality and the situation that they either find themselves in (Blass 1991, Mischel & Shoda 1995, Furr 2009) or choose to participate in (Ickes, Snyder & Garcia 1997, Snyder & Gangestad 1982).

In the context of toxicity, if there is an interaction between a person and their situation where their toxic behavior only shows up in one community or only within similar types of communities, then the toxicity in this case would be highly influenced by the situation, but the person might have a personality that chooses to be in toxic communities. However, if a person shows toxic behaviors in a wide variety of groups that they participate in, then the toxicity would be highly influenced by personality, but these people could also choose to show their toxicity in situations where they know they will get the reactions that they want.

The interaction between personality and situation also applies to people who are not usually toxic but might occasionally display toxic behaviors and people who do not display them at all. In these cases, toxicity can emerge from a certain personality type (e.g. someone who chronically wants to "fit in") interacting with another personality type (e.g. someone who is chronically toxic) in a certain situation (e.g. a community where people frequently bully each other). Even if a person does not display toxic behaviors, there might be negative effects of participating in a community with toxicity that can present themselves in interactions with other communities.

Understanding this relationship in the context of toxicity can provide clarity to the question of how much people change when they are around different groups of people, specifically with respect to intragroup toxicity, and also, in the case of people participating in intergroup toxicity, how a person's toxic behavior is either dependent on the groups they participate in or whether they are toxic wherever they go.

People who are toxic in groups with intragroup toxicity are predicted to show toxic behavior in the other groups they participate in as people as trolls and bullies tend to provoke people into behaving in toxic ways (Cheng, Bernstein, Danescu, & Leskovec 2017). In the case of trolls, they are likely to have toxic personalities, whereas the people goaded into performing toxic behaviors are likely to just be toxic in a specific context. Additionally, people who participate in groups with intergroup toxicity, are predicted to be toxic in the other groups they participate in, in certain contexts. If they find themselves interacting with groups they hate or with groups that have different social values, they will likely be a threat and can provoke discord within the group. However, if they are interacting

with groups that are like their home group, then they will likely not be viewed as toxic, since they ascribe to that group's social norms.

In both types of groups, there are likely people who do not display toxic behavior at all, and we predict that these people, though they may be affected by toxicity in other ways (e.g. poorer mental health), will not show signs of toxicity in the other groups that they participate in. Looking at the ways that people who participate in toxic groups engage with communities that are not known for their toxicity can help online community managers determine whether to block certain people from their communities, or to let them be. Understanding the nature of how toxic behaviors spread and how that spread might differ based on what types of toxic groups a person might engage with can help researchers gain a nuanced view on who to focus their efforts on in order to promote healthier online environments.

Chapter 2: Comparing the psychology of groups with intergroup and intragroup toxicity

To answer this first research question, **To what extent are groups that show intergroup toxicity and intragroup toxicity psychologically different from each other?,** ten groups were identified that either exhibit or have been known to exhibit intergroup toxicity, and ten more that either exhibit or have been known to exhibit intragroup toxicity.

Subreddit Determination. To determine which subreddits to use, non-language-based measures were used. Once the subreddits of interest were chosen, their language was analyzed to determine how toxic these groups were, and what language markers differed between the groups given that they were labelled as toxic.

To determine which subreddits exhibit intragroup toxicity, a controversiality score for each subreddit was calculated. Each post has a controversiality score which is either a 1 for controversial or 0 for non-controversial. The way the Reddit algorithm classifies a post as controversial is that it takes a post's upvotes and downvotes into consideration. If a post has a high number of upvotes as well as a high number of downvotes, then, based on a certain threshold, the algorithm classifies the post as controversial with a value of 1, or not controversial with a value of 0. To determine which subreddits were most controversial, the controversiality variable was summed across all posts in a subreddit and then the list of all subreddits was ordered from highest controversiality to lowest.

The summed controversiality score for a subreddit is a good proxy for intragroup toxicity as it reflects the relative controversiality of a subreddit compared to other subreddits. Since the controversial classification is determined by whether a post has many up and down votes, the subreddits with the most controversiality have the most posts where people are taking sides and disagreeing with each other on various issues. For the purposes of this dissertation, only the top 10 most controversial subreddits with at least 10,000 unique authors were considered.

To determine which subreddits to use to study the construct of intergroup toxicity, the literature provides many examples of subreddits that were most associated with hate groups. Additionally, these subreddits that qualified for the intergroup toxicity group also could not be banned before 2015, since the source of the reddit data only started collecting reddit data from January 2015 onward.

According to research done by Datta and Adar (2019), one way to look at intergroup toxicity would be to look at a group's connectedness in their network (i.e. the number of groups a group is connected to), and their conflict intensity, a measure of how often polarizing authors from one group interact with another group. Datta and Adar define a polarizing author as someone who participates in a "social" home and an "anti-social" home, where social means that the author conforms to social norms within a group (as measured by upvotes) and anti-social refers to the author not conforming to social norms within a group (as measured by downvotes). If there is an author who is polarizing, the link between those two groups has a value of 1. The more authors that participate in their social home and an anti-social home, the stronger the link, or conflict intensity.

In their dataset, groups are connected in a directional manner, as in the connection between group A and group B is unidirectional. For example, group A can have polarizing members that create posts in group B, but polarizing members of group B do not have to create posts in group A. The difference in these connections can cause groups to have a different number of indegrees (i.e. the number of groups with an incoming connection to a single group) and outdegrees (i.e. the number of outgoing connections from a single group). Additionally, conflict intensity has directionality as well, meaning that the connection from group A to group B can have a higher conflict intensity score than the connection between group B to group A.

For example, there are 2 subreddits r/MensRights and r/politics. A member of r/MensRights posts in r/politics meaning that there is an outgoing connection from r/MensRights to r/politics, and an incoming connection to r/politics from r/MensRights. This does not necessarily mean that there is an outgoing connection from r/politics to r/MensRights, thus making the connections between the two groups unidirectional. If the post that the member of r/MensRights makes in r/politics is poorly received (i.e. the post gets a lot of downvotes), the post generates conflict, and the conflict intensity of the outgoing connection is given a value of 1. If the post was accepted by the community (i.e. there was no reaction or the post got a lot of upvotes), the conflict intensity of the outgoing connection would be given a value of 0. The outgoing conflict intensity of all of the posts that the r/MensRights member created in politics would be summed to give them an overall conflict intensity score for their outgoing connection between their social home, r/MensRights, and their potentially anti-social home, r/politics. In order to get an average

outgoing conflict intensity score, the conflict intensity scores are averaged for all members of r/MensRights who post in r/politics.

If a group has a high outgoing conflict intensity with a low indegree, the group tends to participate in groups where they are not welcome, but does not have much internal conflict, which is the metric used to determine the subreddits that qualify as having intergroup toxicity. The findings from the Datta and Adar paper were used to gather 10 subreddits that have a low in-degree and have a high outgoing conflict intensity. If the subreddits were banned, they need to have existed after 2015 for us to gather data on them from the reddit data source. The following subreddits were collected based on the above selection criteria:

Table 1. Subreddit Descriptive Data

subreddit	Type of Toxicity	Sum controversy	# of authors	Avg author controversy	outgoing avg conflict intensity	incoming avg conflict intensity
politics	Intragroup	4117000	2113522	1.95	1.87	1.34
nba	Intragroup	2279763	658750	3.46	2.36	0.71
ukpolitics	Intragroup	776146	119775	6.48	2.88	0.99
SquaredCircle	Intragroup	904700	220802	4.10	1.91	0.41
soccer	Intragroup	1875698	592948	3.16	1.84	0.71
nfl	Intragroup	1410240	592559	2.38	2.04	0.42
europe	Intragroup	805465	462589	1.74	3.16	1.63
leagueoflegends	Intragroup	1559787	1273512	1.22	2.19	0.35
canada	Intragroup	801360	349005	2.30	3.37	1.45
conspiracy	Intragroup	770742	404090	1.91	5.24	1.71
UnresolvedMysteries	Intergroup	39004	133408	0.29	20.49	5.49
asktrp	Intergroup	13166	56578	0.23	25.62	0
ShitRedditSays	Intergroup	30895	51697	0.60	22.74	1.33
sjwhate	Intergroup	8405	29910	0.28	22.29	3.85
FULLCOMMUNISM	Intergroup	4348	29447	0.15	42.06	9.13
SargonofAkkad	Intergroup	3209	17265	0.18	32.90	0
european	Intergroup	6200	15218	0.41	20.18	1.41
SocialJusticeInAction	Intergroup	2834	13948	0.20	32.46	0.48
RightwingLGBT	Intergroup	4456	11308	0.39	31.81	0
PanamaPapers	Intergroup	1579	11210	0.14	27.27	0

Note. The subreddits included in the table were the ones used for the dissertation analysis. All subreddits had to have at least 10,000 unique authors and had to be in both the BigQuery and Datta & Adar (2019) datasets. Sum controversy is the sum of all the controversiality scores across all posts in a subreddit. # of authors is the number of unique authors in the subreddit. Avg author controversy is the Sum controversy score divided by the number of unique authors. Indegree is the number of subreddits that interact with 1 specific subreddit. Outgoing avg conflict intensity is the average conflict intensity that a subreddit has when outwardly interacting with other groups (total outgoing conflict intensity / outdegree). Incoming avg conflict intensity is the average conflict intensity that is directed toward a specific subreddit (total incoming conflict intensity / indegree).

Exclusion Criteria. To answer the questions posed in this dissertation, 10 subreddits that have intergroup toxicity and 10 subreddits that have intragroup toxicity as determined using the above criteria were sampled. Then the history of these 20 subreddits were pulled from January 2015 to January 2020, creating the base dataset. Once the base dataset was gathered (20 subreddits and all their text and metadata), a list of all the people who participated in these subreddits during the above timeframe was compiled, and their data, aggregated. In order to make sure that the people studied were definite members of these communities, the mean number of posts per person in each community was calculated and only people who scored at or above the mean # of posts were considered for analysis. People who had less than 100 words were excluded as the language analysis would not be as accurate with a lower number of words.

Participants

Participants in the study were collected from Reddit and more specifically were collected from subreddits that were either highly controversial (intragroup) or could be considered hate groups (intergroup). Reddit is an online forum where hundreds of thousands of people exchange thoughts, information, and opinions on specific topics like politics, sports, video games, religion, and almost any other topic one can think of. Each specific subcommunity, or subreddit, within reddit is centered around a topic, and people who are interested in that topic can create original submissions related to that topic or respond to something that someone has submitted. Original submissions to a subreddit are called Submissions and responses to a submission are called comments. Comments can be nested (i.e. there can be comments made on comments, not just submissions), however,

each comment on their own gets logged in Reddit's system as a post. By looking at these subreddits, one can obtain a large sample of people who participate in communities that have been deemed toxic in different ways by both the public and scholars alike.

Demographics. It is estimated that Reddit's population is not representative of the general population as it skews 67% male, 33% female, and 67% are in the 18-29 age range (Pew 2016). Additionally, 70% are white, non-Hispanic, racially. These estimated population statistics, though not on par for a general population in the US, do give us insight into a population known for harboring toxic people (Marantz 2018). However, one can make assumptions that members of the r/TheRedPill subreddit primarily present as men as the subreddit claims to be devoted to "discussion of sexual strategy in a culture increasingly lacking a positive identity for men", for example. However, these assumptions cannot be made for certain as Reddit is still technically anonymous.

Data Collection and Sampling Methodology

Collection. Data was collected by querying a public Reddit database housed in Google's BigQuery cloud database and was cleaned and processed using a variety of Python scripts. BigQuery is a large database hosted on Google's servers that can be accessed through Google's cloud computing capabilities. Anyone can create database tables within BigQuery that are accessible by the public. To access data located in a BigQuery table, one can query the data using SQL (Structured Query Language) to only extract the subset of the data that they need. From there, the data can be exported into many different formats for additional analyses.

SQL is a programming language that is specifically created to conduct database queries. Typically, using SQL, you can do things like select variables from the database and specify which tables you want to extract the data from, aggregate variables and group them by other variables, create new tables out of old tables and then query that for data amongst other things. One of the reasons using SQL and BigQuery together makes searching for data so quick is that SQL allows for searching a database by an index instead of searching a whole table, and Google's cloud computing capabilities allow for terabytes of data to be searched in a matter of seconds. The Reddit data hosted on BigQuery was made publicly available and ranges from January 2015 to the present day.

The data in BigQuery is organized at the post level (either Comment or Submission) and will typically include metadata variables such as Username, Time-Posted, Upvotes, Downvotes, Post Text (either Comments or Submissions), Parent and Child Posts (which allow you to chain posts, see who responded to who in a conversation). The data can then be aggregated in different ways to answer a variety of research questions.

Sampling. 500 people from each subreddit were randomly sampled to give us a random selection of 10,000 people per sample, 5,000 in each group. When the list of 10,000 people per sample was created, the dataset was searched to just pull the posts and metadata for the selected sample for analysis. Bootstrapping methodologies were used to have more accurate statistical analyses of the data.

Bootstrapping is a resampling method that works by independently sampling the data with replacement from an existing sample dataset with the same sample, n, and performing analyses among these resampled data. By resampling the data multiple times,

one can create precise estimates of the statistics of the sampling distribution of the data, and it allows for assumption of a normal distribution of the data, making any calculations more accurate and much easier to interpret.

Measures

Language. Psychological linguistic markers using dictionary categories from Linguistic Inquiry and Word Count (LIWC) (Pennebaker et. al. 2015) were used to analyze the behaviors of people who participate in toxic communities. LIWC uses an internal dictionary to categorize words into psychologically relevant categories (i.e. linguistic dimensions, psychological constructs, informal language, personal concern, and punctuation). It is also possible to import dictionaries into LIWC in order to understand more specific psychological constructs, so the Hatebase.org dictionary was imported to capture a wider variety of toxic language and add more nuance to the analyses (Tuckwood 2017). The Hatebase.org dictionary is the world's largest online repository of hate-speech, is multilingual, and uses crowdsourcing to build its collection of hate-related words. Since this repository is being updated regularly, using it in conjunction with the standard LIWC dictionary can capture a wider range of hateful speech than LIWC would otherwise.

During operation, LIWC opens a text file, a group of text files, or a spreadsheet containing text and reads in each word (for each unit of measure, like a text file or a line in a CSV file) and looks for a match in one of the categories in the LIWC dictionary. Once all the words in all the texts are read, an output file containing 90 LIWC variables and the percentage of words in each category per text is created. It is then possible to compare the numeric LIWC output of 2 or more groups with each other.

Specifically, to answer research question 1, text from people participating in intergroup toxicity was compared to text from people participating in intragroup toxicity to see whether there were any psychological differences between the two groups.

Metadata. In addition to linguistic differences between the two groups themselves and the types of people within those groups, the number of posts a person has in a subreddit, their net upvotes (upvotes minus downvotes), and controversiality were calculated. These variables were used as behavioral indicators of community acceptance, rejection, or controversiality, which can show what behaviors are condoned within a specific community. The relationships between these variables and the language outcome measures will be looked at to see if there are individual differences in the ways that people participate in these groups.

Analysis

For each random sample of 10,000 people's posts, individual LIWC scores were calculated along with a net number of upvotes, and their total number of controversial posts. Then the LIWC scores between the 2 types of groups (i.e. 5,000 people in the intergroup toxicity group and 5,000 people in the intragroup toxicity group) were averaged as well as net upvotes, and controversial posts. Then the data for known bots were removed which resulted in each sample being a little over 9900 people each (Sample 0 N = 9908, Sample 1 N = 9921, Sample 2 N = 9922, Sample 3 N = 9933, Sample 4 N = 9923, Sample 5 N = 9931, Sample 6 N = 9914, Sample 7 N = 9914, Sample 8 N = 9914, Sample 9 N = 9911).

To assess the differences between groups showing intergroup toxicity and groups showing intragroup toxicity, first overall means (mean values of every author in the intergroup category and the mean values of every author in the intragroup category) of the following LIWC variables were calculated for each sample: social, posemo (positive emotion), negemo (negative emotion), and cogproc (cognitive processing). Then the means of each sample were averaged to produce the bootstrapped mean values for each of these categories. Additionally, to assess differences in toxicity levels between the two groups, the same analyses were conducted using the toxicity dictionary mentioned above. Lastly, to understand how big of a difference in means there were, Cohen's d's were calculated where a positive Cohen's d value indicates that the intergroup value was higher than the intragroup value and a negative Cohen's d value indicates that the intragroup value was higher than the intergroup value. These analyses on each random sample taken were repeated using a bootstrapping methodology.

Social words, emotion words, and cognitive processing words in the LIWC dictionary are all markers of mental health (Tausczik & Pennebaker 2010), so looking at how these categories change along with those related to toxicity such as swear words, and the Hatebase.org dictionary should paint a nuanced picture of the effects of toxicity on things like a person's mental health.

We predict that people who participate in groups with intergroup toxicity should have more social words, higher usage of positive emotion words, and lower usage of negative emotion words and cognitive processing words as compared to people in groups with intragroup toxicity. The reasoning behind the prediction is that people in groups with

intergroup toxicity seem generally supportive of their group members and think positively of their ingroup, even though they have a shared hatred of another group, making their interactions with each other socially healthy. The linguistic trends stated above are typical indicators of good mental health, hence the reason why we think we will see these trends emerge.

People in groups with intragroup toxicity are predicted to have the opposite trend since there is more conflict within the group itself, leading to a lack of cohesion and support network within the group.

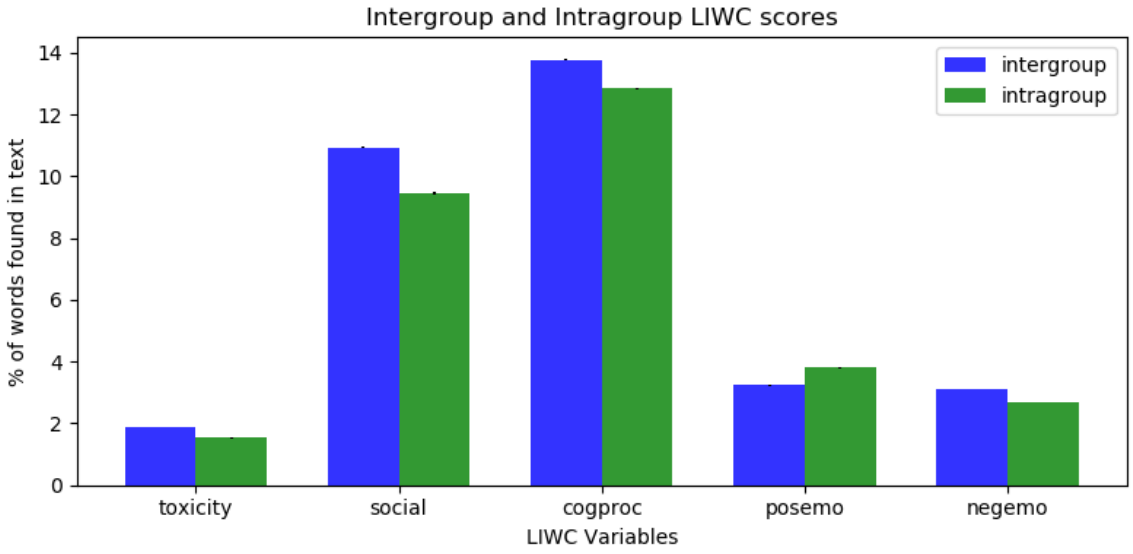
Results

To what extent are groups that show intergroup toxicity and intragroup toxicity psychologically different from each other?

It was predicted that people in groups with intergroup toxicity would outwardly focus their toxic energy and would share similar feelings toward an outgroup and be happy and mentally healthy. Results showed that contrary to the predictions, people who participated in groups with intergroup toxicity used negative emotion words and cognitive processing words at higher rates, positive emotion words at lower rates, but did use social words at a higher rate than people who participate in groups with intragroup toxicity. Additionally, people who participated in groups with intergroup toxicity used toxic words at a higher rate than those who participated in groups with intragroup toxicity (see Figure 1). This means that people who are a part of groups with intergroup toxicity, though they are more social, are more toxic and negative within their home groups than previously predicted. Additionally, the fact that they are using more cognitive processing words than

people in groups with intragroup toxicity can lead us to believe that the mental health of people in these groups are not as stable as we thought. More cognitive processing words indicate that members of the group are working through and are potentially less sure of their ideas. They could be trying to reason why they believe in the mission of the group or are critically thinking about multiple perspectives on an issue.

Figure 1. Bootstrapped means for %-based LIWC categories



Note. Toxicity: (intergroup M=1.88, intragroup M= 1.53, $p<0.05$, d= 0.34, CI for Cohen's d = (0.31, 0.36)), social: (intergroup M=10.92, intragroup M= 9.45, $p<0.05$, d= -0.65, CI for Cohen's d = (0.64, 0.66)), cognitive processing: (intergroup M=13.77, intragroup M= 12.83, $p<0.05$, d=0.39, CI for Cohen's d = (0.38, 0.41)),
positive emotion: (intergroup M=3.25, intragroup M= 3.79, $p<0.05$, d= -0.44, CI for Cohen's d = (-0.47, -0.42)), negative emotion: (intergroup M=3.11, intragroup M= 2.68, $p<0.05$, d=0.38, CI for Cohen's d = (0.37, 0.40))

The number of posts, total controversiality per author, and net upvotes were correlated with the toxicity scores and LIWC variables to better understand the relationship of these linguistic behavioral markers to external behavioral markers (Table 2).

Table 2. Correlations with non-LIWC behavioral variables

	Intergroup-numposts	Intragroup-numposts	Intergroup-controversiality	Intragroup-controversiality	Intergroup-net upvotes	Intragroup-net upvotes
toxicity	0.10***	0.12***	0.11***	0.09***	0.14***	0.02**
social	0.02	-0.03**	0.04***	0.10***	0.03**	0.01
cogproc	-0.07***	-0.17***	0.04***	0.03**	-0.06***	-0.07***
posemo	0.07***	0.18***	-0.004	-0.05***	0.11***	-0.01
negemo	0.12***	0.08***	0.12***	0.11***	0.18***	0.03**

Note. ** = $p<0.05$, *** = $p<0.001$

The most striking findings were that across both groups, a person's number of posts were positively associated with their toxicity score, with a slightly stronger relationship in groups with intragroup toxicity. Number of posts was also positively associated with positive emotion for groups with intragroup toxicity but were positively associated with negative emotion in groups with intergroup toxicity. Though both kinds of toxicity are related to number of posts, the stronger relationship between number of posts and negative emotion in groups with intergroup toxicity suggest that members of this group use negative emotion words at higher rates than members of groups with intragroup toxicity and do so across multiple posts. This means that the toxic behavior is not just concentrated to a few posts, but instead is spread throughout the group. This finding corroborates the above finding that people in groups with intergroup toxicity are toxic and negative within their own groups.

Lastly, net upvotes have stronger positive associations to toxicity and negative emotion words in groups with intergroup toxicity than in groups with intragroup toxicity, though most of these variables have similar trends.

The predictions stated that people who were part of a group with intergroup toxicity would have better mental health outcomes (higher social, higher positive emotion, lower cognitive processing, lower negative emotion) than people who were a part of a group with intragroup toxicity. In the case of the data gathered, however, we see that people from groups with intergroup toxicity use more negative emotion words and higher cognitive processing words, both signs of lower mental health.

A potential reason we see the opposite patterns for negative emotion words than expected is because communities with intergroup toxicity rally around hatred of another group and frequently talk negatively about an outgroup. Seeing opposite trends in cognitive processing could be an artifact of the nature of Reddit. Since Reddit is a place where people come to discuss various topics, the subreddit for a group with intergroup toxicity can attract new members who are searching for a place to belong or to work through thoughts that might be more controversial in other communities. The subreddit essentially acts as a space for the community to cognitively process their thoughts on a subject which they have not been able to do before.

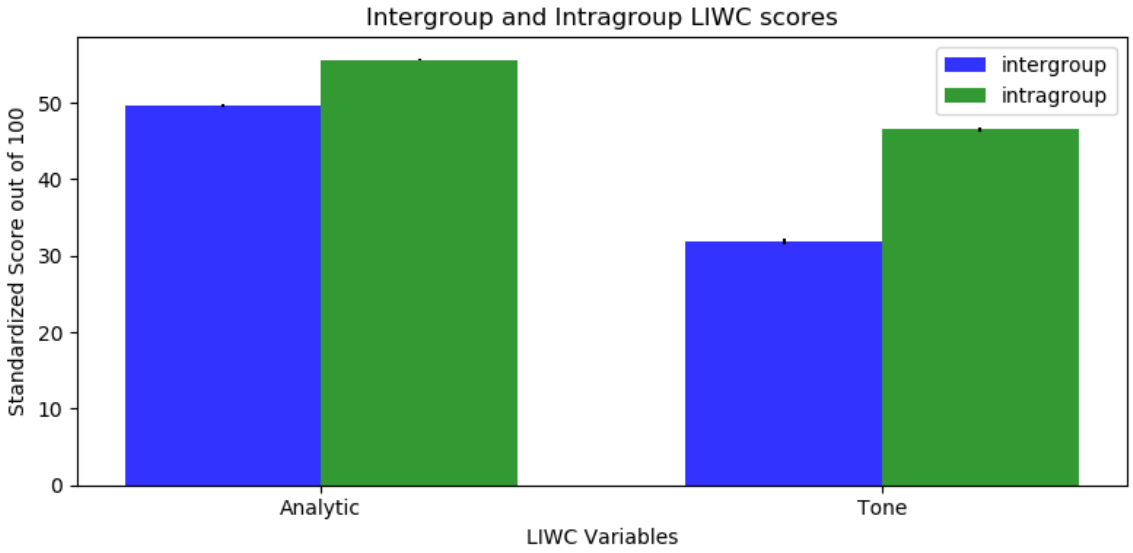
The correlations between toxicity, controversiality, and upvotes bolster the findings we see with negative emotion words being used at higher rates: Toxic behaviors are valued within groups within intergroup toxicity more than they are in groups with intragroup toxicity.

Exploratory Analysis

The LIWC summary variables of Tone and Analytic were also looked at to understand more about the linguistic style of the people participating in groups with

intergroup and intragroup toxicity. The LIWC measure of Analytic thinking is a marker of abstract and hierarchical thinking, on the high end, and narrative or here-and-now style on the other end. It is calculated by the following formula: articles + prepositions -personal pronouns – impersonal pronouns – auxiliary verbs – conjunctions – adverbs – negation (Pennebaker et. al 2014). The LIWC dimension of emotional Tone is Positive Emotion – Negative Emotion (LIWC2015, Pennebaker et. al. 2015). The results show that people participating in groups with intergroup toxicity are lower on the Analytic and Tone scores, meaning that they are more narrative in their writing and they use negative emotion words at higher rates than people who participate in groups with intragroup toxicity (see Figure 2).

Figure 2. Bootstrapped means for summary LIWC variables



Note. Analytic: (intergroup M=49.66, intragroup M= 55.63, $p<0.05$, d= -0.39, CI for Cohen's d = (-0.40, -0.38)), Tone: (intergroup M=31.87, intragroup M= 46.55, $p<0.05$, d= -0.68, CI for Cohen's d = (-0.70, -0.67))

The more narrative style combined with higher cognitive processing words seen in groups with intergroup toxicity can be due to members working through interactions that they have had with outgroup members (either in person or as commentary), and communicating the strong emotions they felt while recounting the events. Recounting the events in a more narrative way while still working through thoughts and emotions is an effective way to rally group members around a cause and can be seen frequently in groups with external toxicity.

Chapter 3: Understanding the effects of toxicity in people's non-toxic communities

In the first study, we examined toxicity at the group level. The primary findings indicated that groups high in intergroup toxicity were generally more negative and toxic, more ruminative (as measured by cognitive processing), and less analytic compared with the toxic intragroup members. In this second study, we examine how being a member of either internally (intragroup) or externally (intergroup) toxic groups impacts their individual members. Specifically, do the most (versus least) toxic members of these already toxic groups spread their toxicity to other groups? How else do high vs. low toxic members from each of these groups differ psychologically?

We predicted that individuals who had high individual-level toxicity in both groups with intergroup and intragroup toxicity would be toxic in the other subreddits that they participated in. We also predicted that individuals with low individual-level toxicity would be more mentally well adjusted (higher positive emotion, lower cognitive processing words, and higher social words) if they came from a group with intergroup toxicity rather than a group with intragroup toxicity.

Data was gathered from all the other subreddits that the people in the above 20 subreddits participate in. This means that from the list of unique authors generated based on the 20 toxic subreddits collected, all those people's entire reddit histories were pulled; not just their posts made in the toxic subreddit they participated in. For example, if someone participated in the r/MensRights subreddit, their entire reddit history from all

other reddit groups would be pulled. For example, such a person might have participated in groups such as r/catsridingroombas, r/cars, and r/hockey. This means, all this person's posts and metadata for r/catsridingroombas, r/cars, and r/hockey would also be pulled in addition to their posts and metadata from r/MensRights.

There were no constraints on the number of other subreddits that the above participants posted in. If participants only chose to post in the toxic subreddit that they were a part of, that behavior provided information about the spread of toxic behavior (i.e. that for those people, their behaviors (toxic or not) were contained to one context).

Participants

The participants for this section of the dissertation are the same people used for the RQ1 analyses. A compiled list of all of the unique authors from the random samples used from the 20 subreddits from RQ1 was compiled, and the data from each person's entire reddit usage was pulled for RQ2 to maintain consistency between the random samples.

Exclusion criteria. Since the research questions address how being a part of a toxic group influences a person's life outside of the toxic group, there is a record of which subreddits and how many posts a person has in subreddits other than the initial 20 subreddits. People whose only posts fall within our initial data set of 20 subreddits are still of interest, but will not be included in these analyses as their potential toxicity has no reaching effects outside of the original dataset of 20 subreddits.

Data Collection and Sampling Methodology

Data collection. The same data collection methods (using BigQuery) used in RQ1 were used to acquire the data for this research question.

Sampling methodology. For each random sample of 500 people from each of the subreddits, the data from the other subreddits they participate in were gathered. This way, there was continuity in the analysis.

Measures

Both the language and metadata measures was the same for the participants' texts and behaviors in groups that are not the target toxic groups.

Analysis Plan

For each sample collected from study 1, all the posts that all the individuals within that sample made on the entirety of Reddit were downloaded and aggregated by person. The research is centered around understanding the effects of participating in a toxic group, so looking at the entire spectrum of group members gave us a sense of what those group-level effects are. We were interested in seeing whether there were personality level effects of toxicity (i.e. a person is toxic across most other subreddits they are a part of), whether there were situational effects of toxicity (i.e. the type of group relates to the ways that people's toxic behavior is expressed), or some combination of the two.

To understand the effects of toxicity outside of the initial subset of 20 subreddits, the data was aggregated by author and subreddit. Then total net upvotes, and total controversiality for each author-subreddit pairing were calculated. Afterward, the text data for each author-subreddit pairing was subjected to LIWC analyses, and each author-

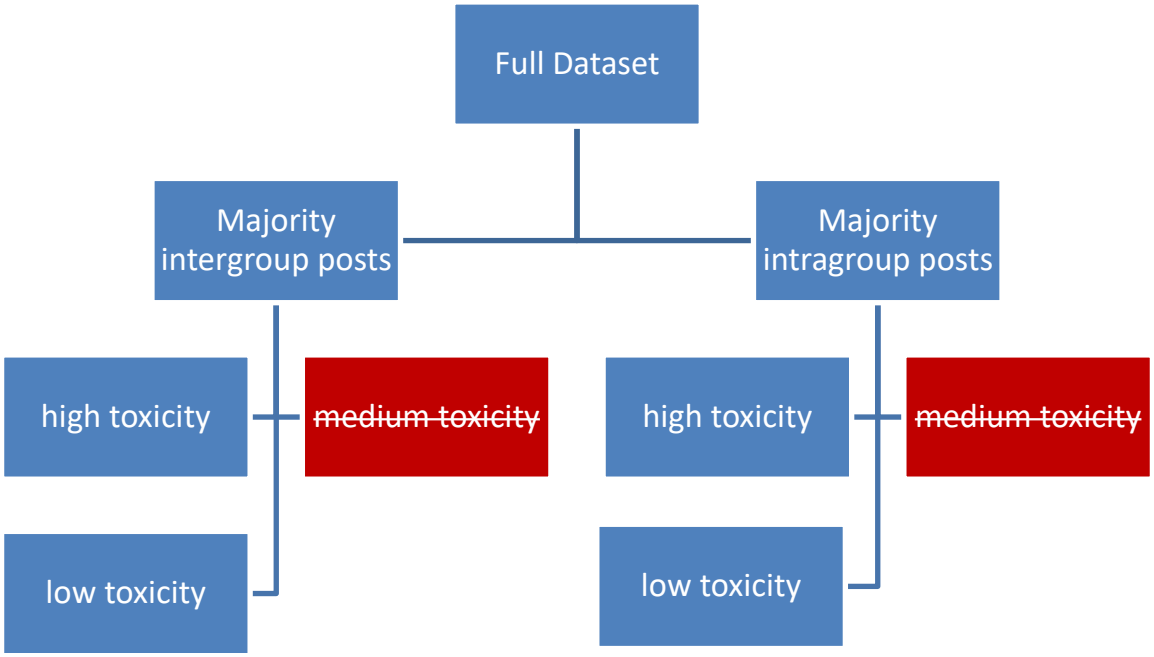
subreddit pairing was either assigned into an ‘intergroup’, ‘intragroup’, or ‘neither’ category based on our initial list of 20 subreddits. If a subreddit was in our initial list of 10 intergroup subreddits, the author-subreddit pairing got the label, “intergroup”. If a subreddit was in our initial list of 10 intragroup subreddits, the author-subreddit pairing got the label “intragroup”. Otherwise, the author-subreddit pairing got the label “neither” as the subreddit was not in the initial list of 20 subreddits.

Additionally, author-subreddit pairs that were in the initial data were categorized into ‘high’, and ‘low’ toxicity based on whether they were in the top or bottom third of toxicity scores.

Afterward, authors were sorted into either majority intergroup, or majority intragroup authors based on the number of their posts that appear in any of the initial 20 subreddits. The number of posts for each author were aggregated by their initial classification of “intergroup” “intragroup” or “neither”. If a person’s intergroup score was higher than their intragroup score, then they were classified as majority intergroup. If their intragroup score was greater than their intergroup score, then they were classified as majority intergroup. If the number of posts in groups with intergroup toxicity and intragroup toxicity were the same, the authors were sorted into an “equal” category but were not included in the major analyses. For the purposes of explanation in the results, reference to “people from groups with intergroup (or intragroup) toxicity” refer to people who were categorized into having the majority of their posts either in groups with group-level intergroup or intragroup toxicity.

After this categorization and joining the toxicity ranking table to the main table, there were 4 major groups for comparison: Intergroup high toxicity, intergroup low toxicity, intragroup high toxicity, and intragroup low toxicity (Figure 3).

Figure 3. Diagram of data classification



Note. People who fell into the middle toxicity bin were excluded since the research questions addressed the more extreme members of groups with intergroup and intragroup toxicity.

Data from the original 20 subreddits were removed after this step to ensure that we would only see the external effects of participating in one (or more) of the initial 20 toxic subreddits.

The 4 groups were compared using t-tests to understand both overall language differences between groups and pairwise language differences. These analyses were conducted on each sample of our data and the values were averaged to get bootstrapped statistical values.

Based on the above information, the following predictions about the effects of intergroup vs. intragroup toxicity on a person's life were made:

1. Highly toxic people in groups that have internal toxicity (i.e. people who are bullies and trolls) or groups that have external toxicity (i.e. people from hate groups) will be toxic in the other groups that they participate in. Trolls and bullies tend to provoke people into behaving in toxic ways, and we believe that they will show similar behaviors across their reddit participation. Additionally, people who participate in hate groups are likely to interact with other groups in a way that is against that group's social norms, making them unlikable when they interact with others.
2. People who have low toxicity (in both types of groups), though they may be affected by group-level toxicity in other ways (e.g. poorer mental health), will not show signs of individual level toxicity in the other groups that they participate in. However, people with low individual-level toxicity who participate in groups with intergroup toxicity will likely have higher positive

mental health indicators (i.e. more social words, less negative emotion and cognitive processing words) than their counterparts who participate in groups with intragroup toxicity. Groups with intergroup toxicity seem to have a sense of cohesion and support that groups with intragroup toxicity do not have, hence the prediction of better mental health.

Prediction 1 essentially posits that people who are highly toxic have a toxic personality and are aggressive and toxic in most groups they participate in. Prediction 2 is centered around the fact that people who are not highly toxic can be influenced by the situations they place themselves in and those situations can dictate whether or not people with low individual-level toxicity show toxic behaviors in the other groups they participate in.

The ways in which different types of toxic communities influence their members lives can help shed light on how to handle the spread of toxicity between and within communities. Understanding potential differences in the spread of toxicity can help researchers and people involved in community maintenance gain a nuanced view on how to promote healthy online environments.

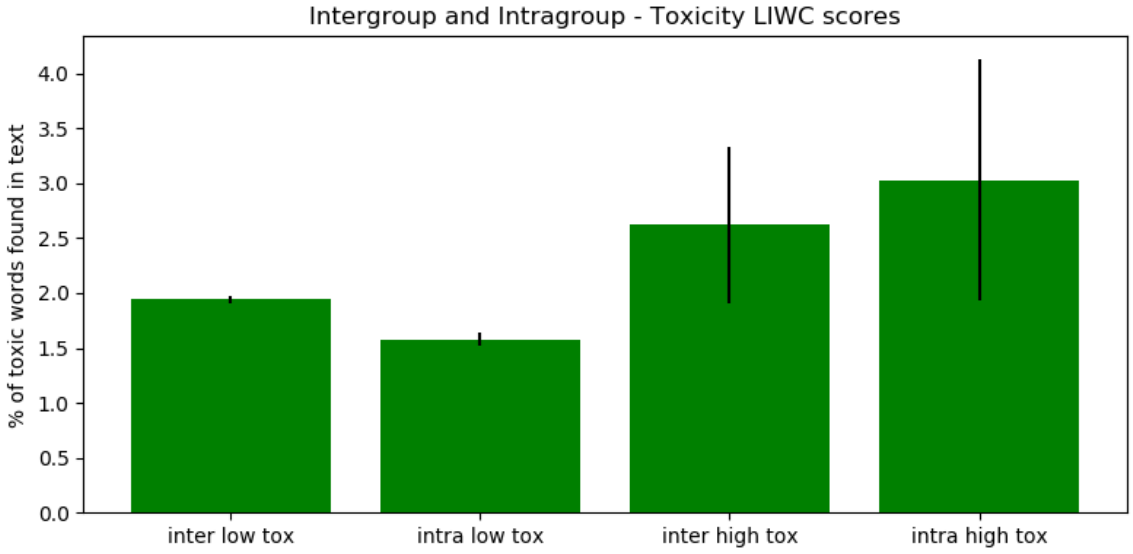
Results

The results showed that in both the case of groups with intergroup toxicity and intragroup toxicity, people who had high individual-level toxicity within the initial 20 subreddits used toxic words at high rates in the other subreddits they participated in (Figure 4). The high individual-level toxicity groups are comparable in the rates in which they use toxic words in other subreddits, but both high individual-level toxicity types are significantly different from people who have low individual-level toxicity from groups

with intragroup toxicity. This means that people with high individual-levels of toxicity behave as though they have a toxic personality. Wherever they go on Reddit, they have similar toxic behaviors and are fundamentally different from people who have low individual-level toxicity and come from groups with intragroup toxicity.

People with low individual-level toxicity and who come from groups with intergroup toxicity are comparable to people with high individual-level toxicity, meaning that their participation in a group with intergroup toxicity causes them to behave like people who have trait toxicity. In this case, there seems to be a strong effect of group-level toxicity on a person's individual level toxicity when they interact with external groups, namely that a person who is a part of a group with intergroup toxicity, even if they were not initially toxic, can become toxic in groups external to the initial 20 toxic subreddits that were analyzed in study 1.

Figure 4. Comparing individual-level toxicity scores between authors with high and low toxicity within groups with intergroup and intragroup toxicity (Not including initial 20 Subreddits)

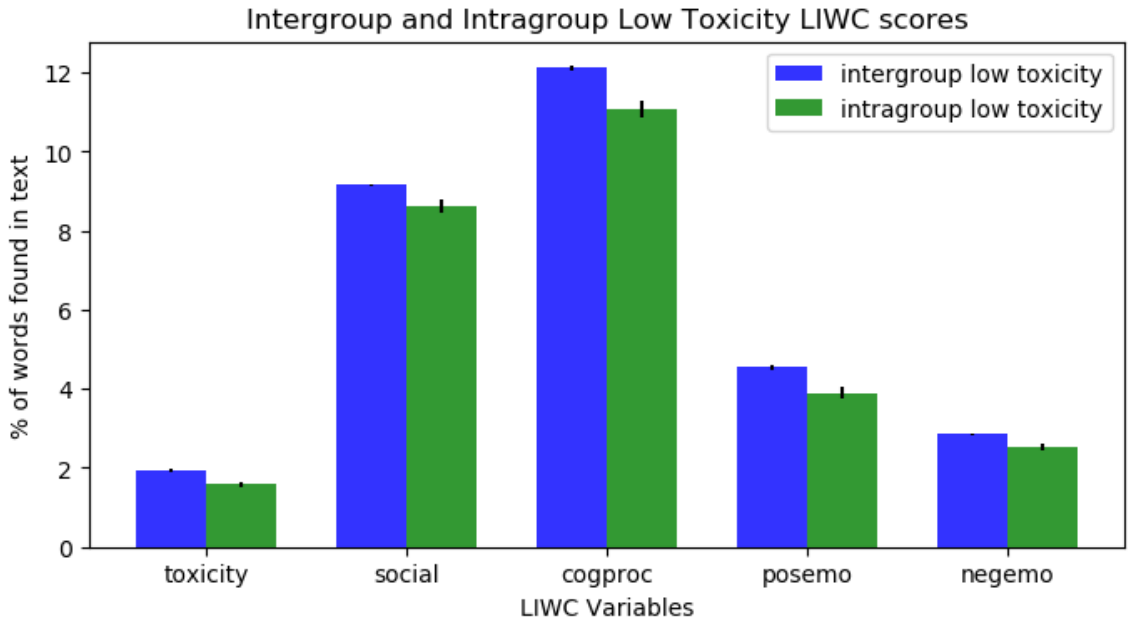


Note. Intergroup low toxicity (M = 1.94), intergroup high toxicity (M=2.62), intragroup low toxicity (M = 1.58), intragroup high toxicity (M=3.03). The x-axis shows 4 groups: Intergroup low toxicity, intragroup low toxicity, intergroup high toxicity, and intragroup high toxicity. The y-axis shows the mean percent of words that fell into the toxicity dictionary category from people's texts within a specific category (e.g. people who had low toxicity scores in a home group that had intergroup toxicity, on average had 1.94% of their words be toxic in the other groups that they participate in). The Intragroup low toxicity was significantly different from the other 3 groups, but Intergroup high toxicity, Intergroup low toxicity, and Intragroup high toxicity were not significantly different from each other.

Additionally, people with low individual-level toxicity from groups with intergroup toxicity were significantly more toxic than people with low individual-level toxicity from groups with intragroup toxicity ($t = 17.19$, $p<0.001$). The difference in percent of toxic words used outside of the initial 20 toxic subreddits further illustrates that people who are from groups with intergroup toxicity but use toxic words less behave more like people with trait toxicity than people who are low on toxicity in general.

The second prediction was that people with low individual-level toxicity from groups with intergroup would be more mentally healthy (higher social, and positive emotion words, lower cognitive processing and negative emotion words) than people with low individual-level toxicity from groups with intragroup toxicity. These trends were predicted because groups with intergroup toxicity were supposed to have healthier atmospheres. The data show, however, that people with low individual-level toxicity from groups with intergroup toxicity had higher toxicity, positive emotion, negative emotion, and cognitive processing than their counterparts who were originally posted in majority groups with intragroup toxicity (Figure 5).

Figure 5. Bootstrapped means for groups with low individual-level toxicity - % based LIWC categories



Note. Low toxicity refers to people who had low toxicity scores in their home groups (i.e. the groups in which they posted the most in our original list of 20 subreddits). Toxicity: (intergroup M=1.94, intragroup M= 1.58, t = 17.19, p<0.001), social: (intergroup M=9.16, intragroup M= 8.62, t = 8.97, p<0.001), cognitive processing: (intergroup M=12.12, intragroup M= 11.09, t = 13.86, p<0.001), positive emotion: (intergroup M=4.54, intragroup M= 3.90, t = 13.40, p<0.001), negative emotion: (intergroup M=2.85, intragroup M= 2.51, t = 11.24, p<0.001)

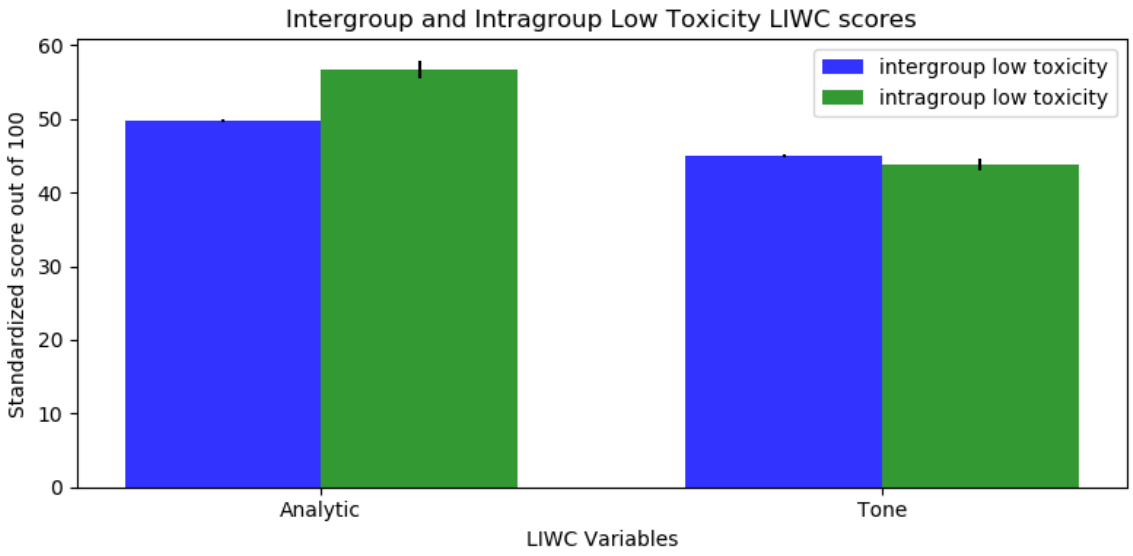
Interestingly, positive emotion words were used at a higher rate for people who came from groups with intergroup toxicity. This finding helps to corroborate the prediction that people who had low individual toxicity that came from groups with intergroup toxicity would have better mental health outcomes (i.e. lower toxicity, higher positive emotion, higher social words, lower cognitive processing, and lower negative emotion words). However, other indicators of mental health (lower cognitive processing words, lower negative emotion words) are not seen in people with low individual-level toxicity that come from groups with intergroup toxicity.

The findings mirror the group-level trends seen for groups with intergroup toxicity suggesting that people from a group with intergroup toxicity generally behave similarly in other groups they participate in, but also use more positive emotion. The higher use of positive emotion and negative emotion in people with low individual-level toxicity from groups with intergroup toxicity can be due to the fact that they seem more emotionally invested in the groups they participate in external to the groups looked at in study 1, and could be attributed to posting in groups where their views are likely to be seen positively.

Exploratory Analyses

The Tone and Analytic variables were analyzed to explore differences among those with low individual-level toxicity in groups with both intergroup and intragroup toxicity. The results show that the Tone was more positive for people with low individual-level toxicity from groups with intergroup toxicity. Also, people with low individual-level toxicity from groups with intergroup toxicity have lower Analytic scores which suggests they use a more narrative style at higher rates than people with low individual-level toxicity from groups with intragroup toxicity (Figure 6). In general, people who originally posted more in groups with intergroup toxicity had lower Analytic (higher narrative) scores than people who originally posted in groups with intragroup toxicity (Table S3).

Figure 6. Bootstrapped means for low individual-level toxicity groups - summary LIWC variables



Note. The language samples used in these analyses exclude any text from the 20 subreddits used in study 1. Analytic: (intergroup M=49.67, intragroup M= 56.76, t = 18.22 p<0.001), Tone: (intergroup M=45.00, intragroup M= 43.76, t=4,34 p<0.001)

The lower Analytic (more narrative) scores for people whose home group had intergroup toxicity are particularly interesting because lower Analytic scores are associated with more personal and emotional language (Pennebaker et. al 2015, Boyd et. al. 2020). More emotion-oriented behavior coming from a group with intergroup toxicity could be related to deeply caring about a cause and defending it in a variety of ways. In groups that have intergroup toxicity, we see members engaging with events and commenting on posts that both support and oppose their cause, either building up those who agree with them by using positive emotional words of encouragement or tearing down those who disagree by using toxic language, and lots of negative affect.

In order to further understand the effects of participating in toxic communities, the percent of external subreddits that had toxic posts was calculated to see how toxic language

usage proliferated in the other subreddits that people participated in (Table 3). People with low individual-level toxicity that come from groups with intragroup toxicity do not seem to spread as much toxic behavior as people with high individual-level toxicity or as much as people with low individual-level toxicity from groups with intergroup toxicity.

Table 3. Percent of subreddits that have toxicity (not including subreddits from RQ1)

Subgroup	Avg # of external subreddits	Mean % toxic
Intergroup_Low_Toxic	32164	59.0%
Intergroup_High_Toxic	727	54.3%
Intragroup_Low_Toxic	111918	39.1%
Intragroup_High_Toxic	1225	58.7%

Note. The average # of subreddits is the avg number of subreddits that were posted in by members of these subgroups. Mean % toxic is calculated by taking the number of subreddits where average toxicity was greater than 0 and dividing it by the total number of subreddits posted in within each subgroup.

Additionally, we see that the number of subreddits that people with high individual-level toxicity participate in are fewer than the number of subreddits that people with low individual-level toxicity participate in. The proliferation of toxic language seems to be similar for people from groups with intergroup toxicity as well as people who are high in toxicity in groups with intragroup toxicity, which can indicate that toxic language has more of a reach if a person participates in a group with intergroup toxicity. These findings also corroborate the earlier findings that there is an effect of situation on people with low individual-level toxicity, and that even people with low-levels of toxicity can be toxic in subreddits external to the initial toxic subreddits looked at in study 1.

Chapter 4: Discussion

The ability to understand the influence of different toxic behaviors in online communities has the potential to contribute to solving the problem of toxic community cultures in online communities. Billions of people around the world participate in and contribute to online forums and many of these people feel excluded from or are harmed by toxic communities regularly. By gaining a greater understanding of the nature of toxicity in online environments, researchers can better develop interventions, teams, and protocols to handle the perpetrators of the behavior in an effective manner.

Differences between groups with intergroup toxicity and intragroup toxicity

We predicted that groups that have intergroup toxicity would be more cohesive and tight knit. We assumed that they would naturally attract people who participate in groups that have a sense of community would have healthier mental health profiles. Initially we predicted that people who participate in groups with intergroup toxicity should use more social words, higher usage of positive emotion words, and lower usage of negative emotion words and cognitive processing words as compared to people in groups with intragroup toxicity. We found, however, that though people who participate in groups with intergroup toxicity use social words at higher rates, they have the opposite trends on all other predicted variables.

One explanation could be that groups that have more externally directed hatred rally around the hate of a specific group or type of person. For example, one person posting in r/sjwhate had the following comments about social justice warriors, specifically a woman giving a TED talk:

> “she's just a smug piece of shit. If i wanted to mention her weight it would be an easy joke.,Is there more context than just this? Sorry, I just think it's bullshit to give a picture and NOTHING else. [...] They feel their dramatic responses are justifying their shitty logic. They have no logical argument, and respond with stupid answers.”

Most comments within groups with intergroup toxicity either look like the one above or are supporting fellow group members in their emotions.

Additionally, the lower score in analytic thinking, indicative of a more narrative thinking style can be attributed to the fact that many people in these subreddits are recounting stories either of things they have experienced or things they have seen related to the topic of the subreddit. Their stories incite reactions of support and discussion within these groups.

The higher cognitive processing score can stem from the fact that people who participate in groups with intergroup toxicity end up ruminating on stories from both sides of the issue they care about. In many cases, as in the quote above, they are incensed by something “idiotic” someone on the opposite side said or did and try to point out the flaws in their arguments or actions. There seems to be a more intellectual involvement with the materials, even though the language used to do so is usually very crass and aggressive.

When correlating the Reddit derived metrics of controversiality, number of posts, and net upvotes to toxicity and LIWC measures of mental health, we find that both controversiality and net upvotes are related to important group dynamics. Reddit’s controversiality metric measures whether a post has both a high number of upvotes and

downvotes. In highly debated subjects, this controversiality score also coincides with more negative emotion words in people's language. Because controversiality is more associated with negative emotion, there is strong evidence that controversiality can be used as a metric to determine types of toxic behavior.

Also, the fact that there is a strong relationship between net-upvotes and toxicity in groups with intergroup toxicity corroborates the earlier prediction that these groups are generally social and united around a cause. Upvotes can be used as a measure of community acceptance, and what the community wants to be known for. In general, the most upvoted posts are some of the first to be seen on the page of a subreddit. In the case of groups with intergroup toxicity, acceptance (or upvotes) are positively associated with toxicity and negative emotion, suggesting that toxic behaviors are valued in the group and are what they want to be known for.

Behaviors of people from toxic groups in other groups on Reddit

As predicted, people who were highly toxic in their home groups had signs of trait toxicity, meaning that they were toxic in the majority of the other groups they participated in, regardless of whether their home group had intergroup toxicity or intragroup toxicity. We predicted this finding because people who are more toxic in both groups with intergroup and intragroup toxicity were expected to meet with conflict in groups that are not their home group, or they would bring conflict with them wherever they went.

When looking at the posts of people who have high individual-level toxicity from groups with intergroup toxicity, they are toxic in 54.3% of subreddits they participate in (Table 3). While not all their language in all their subreddits is toxic, there are many

examples of toxic language that are spread out among the subreddits they participate in. For example, one author posted the following on r/ihatereddit:

> “I hope this sub doesn't get popular because redditors will ruin it like They ruin everything, Pseudointellectual fucks who all take themselves too Damn seriously. All are expert psychologists and condescendingly tell you what they think your problem is just because you post an unpopular opinion.”.

The same person also posted the following on r/TumblrInAction:

> “Welp, I now need to masturbate for a third time today. ,My ex did use crying and emotional tantrums to manipulate me (she admitted later to me that she would do this) and it isn't really uncommon for people to manipulate one another. What the fuck are you fucking implying you cuck!!?? Ur right tho, we do got anger boners here too, I'm sending you my prozac, u need it more than I do, My ex had the same fucking mindset, literally infuriating. I should have known when to chase her and reassure based off of texts that said the opposite. ,Don't be classicist, not everyone can afford to go to^free ^public ^schools ^that ^you ^are ^required ^^by ^^law ^^to ^^go ^^to”.

This person, while they do have some normal posts in subreddits such as r/chess, many of their posts end up being generally aggressive toward the other people they are conversing with. Similar kinds of language can be found in people with high toxicity who participated in groups with intragroup toxicity. For example, an author posted the following on r/SubredditDrama:

> “How is that any different from the retards who spout the "calling me a Nazi only pushes me further to the right" bullshit? ,Imagine a life without a stuffed steak pie. Literally worse than death., […] Don’t cast a black actor for the sake of casting a black actor. Cast an actor based purely on their merit, the best man for the job. Why? Why should a white man (or anybody) who has never discriminated against anybody be discriminated against because of the colour of his skin?, […]The guy’s a total fucking wacko. Again you’re making assumptions. I’m saying the white guy might be *more* qualified but could lose out to hire quotas.”

Though people who are from groups with intragroup toxicity tend to post in many more groups than people whose home group was a group with intergroup toxicity, the relative amount of toxicity being spread by people with high toxicity is very similar (Table 3).

When looking at people who have low individual-level toxicity, the data shows that people from groups with intergroup toxicity had higher scores for toxicity. The fact that even people with low toxicity scores coming from groups with intergroup toxicity had similar trends to people with high toxicity coming from groups with intergroup toxicity suggests that there is a situational effect of being a part of a group with intergroup toxicity on how a person interacts with the other groups they are a part of.

People who have low individual-level toxicity seem to be a lot more influenced by the situations they choose to be in than people with high toxicity (Table 3). In this case, participation in a group with intergroup toxicity seems to make people who might not necessarily be toxic, behave in a toxic way when they participate in other groups, meaning that the group’s toxicity can have residual effects on its members once they interact outside

of that group. It is possible that people who have low individual-levels of toxicity join these groups because they believe in the ideology, and then by being a part of the group, use the group's rhetoric in the other groups they are a part of.

The higher rates of cognitive processing and lower analytic score seen in people coming from groups with intergroup toxicity can be indicative of the group-level patterns seen in study 1: These people are telling stories of either situations they have been in concerning members of an outgroup or reacting to material that is related to their specific outgroups of interest and picking the arguments apart.

The Person-Situation Relationship

The findings from the two studies provide evidence for both trait and state toxic behaviors. Overall, the situations that people choose to participate in (either a group with intergroup or intragroup toxicity) only matter if they are low in toxic behaviors to begin with. If they are already high in toxic behaviors, it does not matter which groups a person participates in because they tend to be toxic wherever they go. In terms of people who are high in toxicity, there seem to be no significant differences between people who are high on intergroup toxicity and high on intragroup toxicity. People in both groups statistically are the same for analytic thinking, cognitive processing, and emotional tone (both positive and negative). They are, however, different in terms of social words, such that people who come from a group with intergroup toxicity use fewer social words than people who have high toxicity scores but come from groups with intragroup toxicity. Since people from groups with intergroup toxicity also participate in fewer groups across Reddit, it seems as though people who participate in groups with intragroup toxicity have tighter social circles

and that there might be more understanding of social constructs within the groups they participate in, so they have less of a need to be explicit about it.

Not only are these groups psychologically similar to each other, but the reach of their toxicity is similar. Both groups with highly toxic people are also toxic in the majority of the other subreddits they participate in. Interestingly, people who are low in toxicity but come from groups with intergroup toxicity have similar toxic reach. In this case, people who have high toxicity in their home groups are psychologically consistent and have similar levels of toxicity to each other and across Reddit, indicating toxicity is a personality trait. However, if a person with low toxicity comes from a group with intergroup toxicity, their toxic spread is similar to those with trait toxicity, whereas if a person with low toxicity comes from a group with intragroup toxicity, their toxic spread is much less. The effect that participating in different groups have on a person's toxicity levels indicate that the situation matters to the spread of a person's toxic behaviors elsewhere.

Future Directions

This research sheds light on the value of understanding how personality and a situation interact when it comes to online toxic behaviors. Looking at two different forms of toxicity and finding that people who participate in these groups are psychologically different can lend itself to further work on understanding group differences in toxicity and the ripple effects that participating in a group has on the rest of a person's life. Though this dissertation focuses on toxic behaviors online, the findings are relevant to other kinds of effects of group participation. For example, how does participating in a support group

affect a person's behaviors in the other groups they participate in? Are there cascading positive effects?

In the future, it would be interesting to explore the following research questions:

1. How does the nature of non-home groups affect toxic behavior?
2. Where do conversations go from normal to toxic?
3. Who are the people who incite toxicity within a conversation? What are their motivations?

It would be interesting to assess the nature of the other groups (i.e., the less toxic groups) that toxic people participate in, which would paint a deeper picture about the online lives of people who spend time on Reddit. Additionally, In-depth conversational analyses can provide a more granular insight into the issue of toxic behavior online and may be more directly related to platforms that deal with live chat or instant messaging capabilities. This way, community moderators or even individuals participating in conversations that turn toxic can have some tools to know how to dissipate toxic behaviors.

The current research also has applications to real-world dilemmas. For example, the findings can help companies (such as gaming companies), as well as other businesses dealing with toxic communities, differentiate between kinds of toxic behaviors and could potentially lead to more effective interventions. For example, if a team of data analysts had access to a community's in-app chats, they would be able to see whether or not any toxic behavior displayed is the result of a conversation a person was a part of or if that person is generally toxic anywhere they go. If a person has trait toxicity, then actions can be to ban the person from using the app.

In cases where toxicity was brought on situationally, community managers could help diffuse the situation. In some situations, companies are only dealing with one kind of community, so if a person does have trait-like toxicity within a company's singular community, banning them should help with the toxicity problem. However, if a company has multiple apps or offerings, decisions to ban people with trait toxicity might not be as clear cut. People with trait-toxicity do have toxic posts in many communities they participate in, but not necessarily all of them. Further steps would need to be taken to assess how disruptive the toxicity is to the other community members, and how far their toxicity reaches. Depending on the severity and reach of the toxicity, companies can choose to ban a user either at a community or a sitewide level.

Understanding the type of community a business, company, or other group has can be useful to figuring out how to diffuse a situation. If the community seems like a hate group, then it is likely that most of the people in the group are toxic. If the company is a company like Reddit and has access to many communities, they can just ban the entire community to stop the congregation of toxic behaviors in a single place. However, this would not stop the spread of toxicity because it is likely that these people will continue to be toxic in the other groups that they participate in. Banning the subreddit and the users could reduce the problem further, and if mistakes have been made, those people could petition to have rights to their account back.

If the community seems like it has infighting, however, looking at individuals' usage of toxic language within that community can shed light onto whether the person should be banned, or a community manager should talk to them. If the person uses toxic

language at higher rates than their peers, they are likely to have trait toxicity and should be banned. However, if they use toxic language at lower rates than their peers, it could be a situationally derived toxic behavior and further actions with that individual could be handled by a community manager.

Shortcomings

There are several shortcomings of the current study. The nature of reddit data is very specific to one kind of interaction that people have in online communities. There are other forms of communications that people have with each other that could contribute to toxicity within a specific online community (i.e. Facebook chats, commenting on YouTube videos, email, instant messaging etc.). By studying only reddit data, we cannot get a full picture of how these people interact with each other across all their online platforms, so they might be missing key components for identifying toxic behavior. One way to mitigate this issue is to run the above analyses on different kinds of data to see if the findings hold up. Though some of these types of data are harder to get a hold of, diversifying the types of data analyzed should help with generalizability of the findings.

Another shortcoming is that we are only using data from subreddits that are typically known for their toxic behavior. Though we believe that these subreddits are the most representative of toxicity online, there could be bias in the subreddits chosen because we could be missing communities that are less well known but are centered around other kinds of toxic subjects. Like the former shortcoming, to try and mitigate issues stemming from this selection bias, diversification of data sources is necessary to be able to see whether any effects found generalize across social platforms.

People who post in toxic groups do post in both kinds of groups. Though most their posts might stem from a certain type of group, people from both kinds of toxic groups post in the other (Table S5). Additionally, people from groups with intragroup toxicity generally post more across Reddit. Further analyses or experiments using people who ONLY post in one group vs. another need to be conducted to continue to clarify the effects that participating in groups with toxic communities have on a person's other communities.

It should also be noted that the selection of subreddits in this project is by no means the full spectrum of toxic communities. There are many other toxic communities out there and some more specific than the ones used in this dissertation. To understand the toxicity on Reddit fully, one would need to find a way to easily identify toxic subreddits and then analyze all of those subreddits' members' behaviors. It is also possible that though we believe we found the top 10 most toxic subreddits for both intergroup and intragroup toxicity, that the other groups that people participate in are also toxic in varying ways, and the current studies do not capture the toxicity of those subreddits.

It is also important to remember that the two studies were observational and relied on correlations and mean differences to assess differences between the two types of groups. This is not an experimental study and as such, no causal claims can be made about the findings.

Ultimately, the study has a very large N, we were only looking at 10 groups of each type and it's not clear that they represent the broader category of intergroup and intragroup, The measures we've used were based on current literature, but more studies need to be conducted to see if there are better measures of these two constructs and how we can

measure them using language analysis. Our measures of toxicity were based off of swear words, anger, and slurs, and though it is a good initial measure of toxicity, swearing is also used in a positive context in everyday speech, so understanding context around the words is a critical next step in clarifying how toxic behaviors are measured.

Conclusion

This dissertation has introduced a new methodology to understand and define toxic behaviors and measure it through language analysis. Defining toxic behavior as the use of anger, slurs, and swear words begins to capture an extremely complex type of behavior that affects millions of people online. Differentiating between toxic behaviors also allows for a greater and more nuanced understanding of where toxic behaviors originate on the internet and how they spread through people's interactions with other individuals and groups.

Understanding how intergroup vs. intragroup toxicity influence people's lives can help diffuse polarizing and toxic conversations in the online world. In an age when broadly different types of groups are politicized from video games to public health to human rights, it is important to understand nuances within these conversations in order to best learn how to turn them around to be more productive, and also when it is necessary to take action.

These differences in toxicity show that we cannot just have a blanket statement or a blanket treatment of toxicity online, and with so many people engaging in important issues in online communities, the use of effective interventions for hate and for discontent become more and more needed. Hopefully, by shedding light onto a few aspects of online toxicity, the research from this dissertation can be used to make online spaces safe and approachable for everyone.

Appendix

Table S1: RQ1 – Bootstrapped Means by subreddit

subreddit	subreddit type	toxicity	Analytic	Tone	posemo	negemo	social	cogproc
asktrp	Intergroup	1.72	33.98	43.62	3.75	2.82	14.11	14.32
european	Intergroup	2.17	55.25	25.51	3.02	3.27	10.20	12.67
FULLCOMMUNISM	Intergroup	1.64	60.37	42.77	3.61	2.71	9.24	12.16
PanamaPapers	Intergroup	1.14	53.83	41.48	2.98	2.25	9.62	14.19
RightwingLGBT	Intergroup	2.12	44.88	36.39	3.46	2.96	11.15	14.24
SargonofAkkad	Intergroup	2.01	52.32	28.44	3.13	3.18	10.50	13.69
ShitRedditSays	Intergroup	2.37	47.34	23.06	3.57	3.92	10.82	13.80
sjwhate	Intergroup	2.69	48.47	24.38	3.23	3.73	11.40	13.49
SocialJusticeInAction	Intergroup	2.26	51.04	23.33	3.03	3.50	11.25	13.74
UnresolvedMysteries	Intergroup	1.17	49.96	27.91	2.66	2.74	10.40	14.80
canada	Intragroup	1.25	57.06	40.61	3.17	2.41	9.36	13.42
conspiracy	Intragroup	1.60	54.25	31.68	2.92	2.72	9.80	13.92
europe	Intragroup	1.30	60.08	36.89	3.03	2.47	8.53	13.28
leagueoflegends	Intragroup	1.34	45.71	65.93	4.87	2.56	9.46	13.12
nba	Intragroup	1.74	56.33	55.46	4.37	2.76	9.96	11.78
nfl	Intragroup	1.87	57.02	51.53	4.44	3.06	9.68	11.67
politics	Intragroup	1.61	54.89	33.50	3.22	2.86	10.23	13.38
soccer	Intragroup	1.54	56.20	58.54	4.55	2.76	9.30	12.04
SquaredCircle	Intragroup	1.77	56.46	53.13	4.15	2.69	9.04	11.77
ukpolitics	Intragroup	1.24	60.63	37.61	3.15	2.53	9.26	14.02

Note. This table shows the breakdown of LIWC scores by subreddit. This breakdown can give contextual insight and differences between scores of the subreddits studied in this dissertation. It also shows the variety of scores found between each of these datasets and can be used for further in-depth analyses into these communities.

Table S2. ANOVA results from RQ2

	ANOVA Results (bootstrapped)	Confidence Intervals
toxicity	408.13	(313.89, 502.37)
Analytic	2824.87	(2132.98, 3516.75)
Tone	106.82	(41.88, 171.77)
social	301.46	(200.79, 402.13)
cogproc	842.22	(561.79, 1122.64)
posemo	392.13	(263.70, 520.56)
negemo	308.07	(207.00, 409.14)

Note. This table shows that the differences between 6 groups (low individual-level toxic people from groups with intergroup toxicity, low individual-level toxic people from groups with intragroup toxicity, high individual-level toxic people from groups with intergroup toxicity, high individual-level toxic people from groups with intragroup toxicity, low individual-level toxic people who posted in both types of groups equally,and high individual-level toxic people who posted in both types of groups equally) are significant. Pairwise relationships are explored in Table S4.

Table S3. Means for 6 subgroups in RQ2

Mean Differences	Toxicity	Analytic	Tone	social	cogproc	posemo	negemo
Intergroup_Low_Tox	1.94	49.67	45.00	9.16	12.12	4.54	2.85
Intergroup_High_Tox	2.62	47.21	45.97	8.94	11.51	4.91	3.48
Intragroup_Low_Tox	1.58	56.76	43.76	8.62	11.09	3.90	2.51
Intragroup_High_Tox	3.03	48.62	42.81	9.93	10.79	5.06	3.82
Equal_Low_Tox	2.06	49.37	44.29	9.36	12.12	4.43	2.94
Equal_High_Tox	2.87	49.30	42.36	9.89	11.45	4.77	3.47

Note. This table shows mean values of toxicity and LIWC scores broken down by individual-level toxicity/group-toxicity pairs.

Table S4. Pairwise comparisons (t-test values and significance)

	Toxicity	**Analytic**	**Tone**	**social**	**cogproc**	**posemo**	**negemo**
low_inter_intra	17.19***	-18.22***	4.34***	8.97***	13.86***	13.40***	11.24***
low_inter_equal	-1.51	0.48	0.77	-1.14	-0.02	0.79	-1.32
low_intra_equal	6.16***	-10.30***	0.55	3.94***	6.63***	3.72**	5.57***
high_inter_intra	-0.99	-0.64	1.32	-2.95**	1.41	-0.46	-0.84
high_inter_equal	-0.88	-1.06	1.47	-3.04**	0.11	0.39	0.07
high_intra_equal	-0.44	0.38	-0.17	-0.12	1.59	-0.69	-0.92
High inter low inter	-3.01**	1.48	-0.64	0.96	1.36	-2.23**	-3.12**
High intra low intra	-4.17***	5.40***	0.50	-5.22***	1.12	-3.92**	-3.81**

Note. *** = p<.001, ** = p<.05. These pairwise comparisons were conducted to better understand the ANOVA results seen in Table S2.

Table S5. Summary metrics for Individual level toxicity groups (ONLY the 20 initial subreddits data)

Subgroup	Avg # of initial subreddits	Mean % toxic	Avg # of toxic posts from initial intergroup subreddits	Avg # of toxic posts from initial intergroup subreddits
Intergroup_Low_Tox	20	100%	249169	35851
Intergroup_High_Tox	8	80.2%	237	53
Intragroup_Low_Tox	20	100%	104016	3625515
Intragroup_High_Tox	14	90.5%	153	5322

Note. This table shows the initial spread of toxicity in the groups that were used to conduct the analyses for study 1. People with high individual-level toxicity seemed to post less than people who had low individual-level toxicity.

References

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological science*, *26*(10), 1531-1542.

Barlett, C. P., Gentile, D. A., & Chew, C. (2016). Predicting cyberbullying from anonymity. *Psychology of Popular Media Culture*, *5*(2), 171.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of social issues*, *55*(3), 429-444.

Budman, S. H., Soldz, S., Demby, A., Feldstein, M., Springer, T., & Davis, M. S. (1989). Cohesion, alliance and outcome in group psychotherapy. *Psychiatry*, *52*(3), 339-350.

Blackburn, J., & Kwak, H. (2014, April). STFU NOOB! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web* (pp. 877-888).

Blass, T. (1991). Understanding behavior in the Milgram obedience experiment: The role of personality, situations, and their interactions. *Journal of personality and social psychology*, *60*(3), 398.

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *The Journal of Abnormal and Social Psychology, 62*, 713–715.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined

through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1-22.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017, February). 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* (pp. 1217-1230).

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. *Social communication*, *1*, 343-359.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, *15*(10), 687-693.

Datta, S., & Adar, E. (2019, July). Extracting inter-community conflicts in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, No. 01, pp. 146-157).

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, *6*, 37825.

Elliott, T. P. (2012). *Flaming and gaming–computer-mediated-communication and toxic disinhibition* (Bachelor's thesis, University of Twente).

Felmlee, D., & Faris, R. (2016). Toxic ties: Networks of friendship, dating, and cyber victimization. *Social psychology quarterly*, *79*(3), 243-262.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, *80*(S1), 298-320.

Funder, D. C. (2009). Persons, behaviors and situations: An agenda for personality psychology in the postwar era. *Journal of Research in Personality, 43*(2), 120-126.

Furr, R. M. (2009). Profile analysis in person–situation integration. *Journal of Research in Personality*, *43*(2), 196-207.

Gilbert, E., Bergstrom, T., & Karahalios, K. (2009, January). Blogs are echo chambers: Blogs are echo chambers. In *2009 42nd Hawaii International Conference on System Sciences* (pp. 1-10). IEEE.

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, *69*(7), 669.

Hankes, K., & Dijk, Z. V. (2019, February 20). Move Slow and Break Everything. Retrieved March 15, 2020, from https://www.splcenter.org/fighting-hate/intelligence-report/2019/move-slow-and-break-everything

Hipp, J. R., Tita, G. E., & Boggess, L. N. (2009). Intergroup and intragroup violence: Is violent crime an expression of group conflict or social disorganization?. *Criminology*, *47*(2), 521-564.

Hopkinson, C. (2013). TROLLING IN ONLINE DISCUSSIONS: FROM PROVOCATION TO COMMUNITY-BUILDING. *Brno studies in English*, *39*(1).

Ickes, W., Snyder, M., & Garcia, S. (1997). Personality influences on the choice of situations. In *Handbook of personality psychology* (pp. 165-195). Academic Press.

Jamieson, J. P., Valdesolo, P., & Peters, B. J. (2014). Sympathy for the devil? The physiological and psychological effects of being an agent (and target) of dissent during intragroup conflict. *Journal of Experimental Social Psychology*, *55*, 221-227.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125-143.

Kaarbo, J., & Gruenfeld, D. (1998). The social psychology of inter-and intragroup conflict in governmental politics. *Mershon International Studies Review*, *42*(2), 226-233.

Kemp, Simon. (2020). Digital 2020: Global Digital Overview - DataReportal – Global Digital Insights." *DataReportal*, DataReportal – Global Digital Insights, 30 Jan. 2020, datareportal.com/reports/digital-2020-global-digital-overview.

Kordiš, L., & Šmitran, M. Multi-label Toxic Language Classification of Wikipedia Comments. *Text Analysis and Retrieval 2018 Course Project Reports*, 60.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, *28*(2), 434-443.

Levin, B. (2002). Cyberhate: A legal and historical analysis of extremists' use of computer networks in America. *American Behavioral Scientist*, *45*(6), 958-988.

Levin, J., & MacDevitt, J. (2013). *Hate crimes: The rising tide of bigotry and bloodshed*. Springer.

Marantz, A. (2018). Reddit and the Struggle to Detoxify the Internet. *The New Yorker*, *12*.

Märtens, M., Shen, S., Iosup, A., & Kuipers, F. (2015, December). Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)* (pp. 1-6). IEEE.

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

Midtgaard, J., Rorth, M., Stelter, R., & Adamsen, L. (2006). The group matters: an explorative study of group cohesion and quality of life in cancer patients participating in physical exercise intervention during treatment. *European Journal of Cancer Care*, *15*(1), 25-33.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of abnormal and social psychology*, *67*(4), 371.

Mischel, W. (1968). *Personality and assessment*. Psychology Press.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, *102*(2), 246.

Newson, M., Buhrmester, M., & Whitehouse, H. (2016). Explaining lifelong loyalty: The role of identity fusion and self-shaping group events. *PloS one*, *11*(8).

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*.

Pew Research Center (2016). Reddit news users more likely to be male, young, and digital in their news preferences. *https://www.journalism.org/2016/02/25/reddit news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/*

Rokeach, M., & Mezei, L. (1966). Race and shared belief as factors in social choice. *Science*, *151*, 167-172.

Schafer, J. R., & Navarro, J. (2003). The seven-stage hate model: The psychopathology of hate groups. *FBI L. Enforcement Bull.*, *72*, 1.

Snyder, M., & Gangestad, S. (1982). Choosing social situations: Two investigations of self-monitoring processes. *Journal of Personality and Social Psychology*, *43*(1), 123.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Tsuno, K., Kawakami, N., Inoue, A., Ishizaki, M., Tabata, M., Tsuchiya, M., ... & Shimazu, A. (2009). Intragroup and intergroup conflict at work, psychological distress, and work engagement in a sample of employees in Japan. *Industrial health*, *47*(6), 640-648.

Tuckwood, C. (2017). Hatebase: Online Database of Hate Speech. *The Sentinal Project. Available at: https://www. hatebase. org*.

Van Alstyne, M., & Brynjolfsson, E. (1996). Electronic Communities: Global Villages or Cyberbalkanization?(Best Theme Paper). *ICIS 1996 Proceedings*, 5.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of school violence*, *14*(1), 11-29.

Williams, H. T., McMurray, J. R., Kurz, T., & Lambert, F. H. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global environmental change*, *32*, 126-138.

Wikipedia (2020). Controversial Reddit communities. (2020, February 24). Retrieved from https://en.wikipedia.org/wiki/Controversial_Reddit_communities