ON THE DETECTION OF HATE SPEECH, HATE SPEAKERS AND POLARIZED GROUPS IN ONLINE SOCIAL MEDIA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Dana Warmsley

December 2017

© 2017 Dana Warmsley
ALL RIGHTS RESERVED

ON THE DETECTION OF HATE SPEECH, HATE SPEAKERS AND POLARIZED GROUPS IN ONLINE SOCIAL MEDIA

Dana Warmsley, Ph.D.

Cornell University 2017

The objective of this dissertation is to explore the use of machine learning algorithms in understanding and detecting hate speech, hate speakers and polarized groups in online social media. Beginning with a unique typology for detecting abusive language, we outline the distinctions and similarities of different abusive language subtasks (offensive language, hate speech, cyberbullying and trolling) and how we might benefit from the progress made in each area. Specifically, we suggest that each subtask can be categorized based on whether or not the abusive language being studied 1) is directed at a specific individual, or targets a generalized "Other" and 2) the extent to which the language is explicit versus implicit. We then use knowledge gained from this typology to tackle the "problem of offensive language" in hate speech detection. A key challenge for automated hate speech detection on social media is the separation of hate speech from other instances of offensive language. We present a Logistic Regression classifier, trained on human annotated Twitter data, that makes use of a uniquely derived lexicon of hate terms along with features that have proven successful in the detection of offensive language, hate speech and cyberbullying. Using the tweets classified by the aforementioned hate speech classifier, we extract a set of users for which we collect demographic and psychological attributes, with the goal of understanding how these attributes are related to hate speech use. We first present a binary Random Forest classifier for predicting

whether or not a Twitter user is a hate speaker. We then explore the use of linear and Random Forest regression models as a means of explaining and predicting levels of hate speech use based on user attributes. To the best of my knowledge, this work is the first to present an automated approach for detecting individual hate speakers. Finally, we present a non-negative matrix factorization (NMF) algorithm for identifying polarized groups using tripartite graphs (user-post-tag) gleaned from social media data. This work is heavily inspired by the need for an unsupervised approach that works well in contexts varying in the nature of the controversy, the level of polarization, the number of polarity groups involved, and the presence of neutral entities. I present the first ever analysis of polarization on data from the Tumblr platform, showing improved performance over traditional community detection methods and the state-of-the-art method of NMF on bipartite graphs.

BIOGRAPHICAL SKETCH

Dana Warmsley's journey to the doctorate began at Hunter College (CUNY), where she received bachelors degrees in mathematics and Africana Studies. Her love of mathematics and sociology inspired her to pursue an accelerated masters degree program there, resulting in the completion of a masters degree in Applied Mathematics. While there, she participated in rich scholarly and extracurricular activities as part of the Mellon Mays Undergraduate Fellowship and as member and president of the Africana Puerto Rican/Latino Studies Club. Upon graduation from Hunter College, she spent one year teaching mathematics as an adjunct professor at York College (CUNY) - an experience that only increased her eagerness to pursue a doctoral degree.

Dana began her tenure at Cornell University in 2011 as a member of the Center for Applied Mathematics, under the advisement of Michael Macy and Steven Strogatz. Here, she was able to pursue her love of math and sociology in studying the detection of hate speech and hate speakers. During her time at Cornell, she also took on the positions of Treasurer and Academic Chair of the Black Graduate and Profession Student Association. Summers were often spent developing her teaching skills. The summers of 2013 and 2014 were spent teaching College Algebra to Qatari high school students on Carnegie Mellon University's campus in Doha, Qatar. During the summer of 2015, she acted as an instructor for Cornell University's Engineering Summer Math Institute, exposing Cornell undergraduates to the wonders of applied mathematics and Social Network Analysis. Finally, she participated in an internship with HRL Laboratories during the summer of 2017, where she engaged in research related to identifying polarized groups in online social media.

I dedicate this dissertation to my strongest motivation and my greatest gift,

my beloved son,

Alonzo.

ACKNOWLEDGEMENTS

I would first and foremost like to thank my parents, Diane and Raymond Warmsley. I am often awe-stricken by your unwavering love, support and encouragement. Without you, my journey to the doctorate would not have been possible. A million times, thank you.

I would like to thank my significant other, Alonzo, and my son Alonzo Jr. for their constant encouragement and motivation.

I would like to thank my doctoral committee - Michael Macy, Steven Strogatz, Mark Lewis and Richard Rand. I will forever cherish the time, guidance, advice and learning moments shared with you all. I could not have asked for a more supportive committee.

Cornell's Diversity Programs in Engineering office will always hold a special place in my heart, and I thank Sara Hernandez and Jami P. Joyner for always leaving their doors open for me.

Finally, the chapters of this dissertation were made possible via collaborations with amazing advisors and colleagues, who I thank for engaging in true team work, scholarly advice, and lively exchanges of ideas. Chapter 2 was coauthored with Zeerak Waseem of the University of Sheffield in England, Thomas Davidson of the Sociology department at Cornell University and Ingmar Weber of the Qatar Computing Research Institute. Chapters 3 and 4 were coauthored with Thomas Davidson, Ingmar Weber and my advisor, Michael Macy. Finally, Chapter 5 was coauthored with Jeijun Xu and Tsai-Ching Lu during an internship with HRL Laboratories in Malibu, California.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

With the growing popularity of online social media platforms, the need for automated hate speech and polarization detection methods has been ever-increasing. In recent years, hate speech and other forms of abusive language have plagued online communities, negatively impacting user experience and safety. Social media platforms are being criticized for their inability to combat hate speech in a timely manner, especially by countries that have enacted laws against the use of hate speech. Despite best efforts, the sheer amount of online data to be processed has made it difficult to rise to the challenge.

Identifying polarized groups online has also become of increasing interest. This is especially true in the political realm, where polarized political groups ("echo chambers") have contributed to the spread of misinformation within communities. The ability to reliably classify these groups, especially without having to sift through mass amounts of textual data, has implications not just in politics but in business and finance as well.

While the topics of hate speech and polarization appear to be distinct at first glance, there are indeed threads connecting the two. Hate speech is, in itself, polarizing. The mere use of hate speech has an "othering" effect, especially when directed at racial, sexual, religious and other minorities. Further, as evidenced by the Gamergate controversy, the strategic use of hate speech against opponents has the ability to polarize a cohesive movement. Though Gamergate began as a movement advocating for ethics in journalism, it quickly turned into a hate and harassment campaign against women, forcing many of its early supporters to turn against it.

In this dissertation, I present methods for detecting hate speech, hate speakers and polarized groups in online social media. In chapter 2, a typology for abusive language detection is proposed - one that heavily informs the work presented in the remaining chapters. In chapter 3, I present a hate speech detection classifier that deals with the problem of offensive language. We find that this method is largely successful in distinguishing hate speech from generally offensive language. Chapter 4 includes automated classifier and regression approaches that use demographic, psychological and social network attributes in identifying hate speakers and predicting their levels of hate speech use. Finally, chapter 5 presents a non-negative matrix factorization algorithm for uncovering polarized groups in online social media. Using Tumblr data, I show that this algorithm is adept at identifying polarized groups in varying contexts - from controversies like Gamergate to popular sports tournaments like FIFA.

CHAPTER 2

UNDERSTANDING ABUSE: A TYPOLOGY OF ABUSIVE LANGUAGE DETECTION SUBTASKS

2.1 Introduction

[1] Over the past several years, the desire to detect abusive language online has increased dramatically. With the emergence of the web, internet users have been increasingly exposed to abusive language in all forms. In order to ensure the pleasant experience and safety of users, social media has come under international pressure to tackle problems of abusive language on their sites. Academic and industry researchers alike have taken on the huge task of identifying, understanding and preventing the use of harmful speech online.As the body of research on abusive language detection and analysis grows, there is a need for critical consideration of the relationships between different subtasks that have been grouped under this label.

The terms 'abusive language', 'harmful speech' and 'toxic language' have come to be umbrella terms for several types of abusive language, including profanity, offensive language and hate speech. Identifying one or more of these types of abusive language is a core component in identifying online hate speech, cyberbullying and trolling, but little work has been done to examine the relationship between these aims. Since these subtasks of abusive language detection seek to address specific, yet partially overlapping phenomena, we believe that there is much to gain by studying how these subtasks are related.

[1] This work was co-authored with Zeerak Waseem, Thomas Davidson and Ingmar Weber. It was presented at the Association of Computational Linguistics' Abusive Language Workshop (2017)

The domain of abusive language detection suffers from both definitional and practical issues. There is a general lack of consensus regarding what defines different types of abusive language and what labels to use for them. This is illustrated in two major ways. First, past work uses a variety of different labels to identify similar abusive content. In annotating for cyberbullying events, [80] identifies racist and sexist remarks as a subset of "insults", whereas [60] classifies similar remarks as "hate speech" or "derogatory language." In contrast, the same labels are often used in identifying inherently different types of abusive content. [88] only considers "hate speech" without regard to any potential overlap with bullying or otherwise offensive language, while [23] distinguish hate speech from generally offensive language. Additionally, there is often disagreement regarding what constitutes certain types of abusive language. Some messages considered to be hate speech by [88] are only considered to be derogatory or offensive by [60] and [23]. The overall lack of consensus has resulted in contradictory guidelines for annotating abusive language and precludes the use of annotated data and guidelines across studies.

Further, these issues make it quite difficult to synthesize past literature in abusive language detection. Surely, subtasks related to the identification of online harassment, cyberbullying and trolling can benefit from gains made in the areas of hate speech and offensive language detection. Similarly, annotating for personal attacks as in [92] likely encompasses the identification of cyberbullying, hate speech, offensive language and trolling. Studies aimed at identifying multiple types of abusive language in this manner can benefit from understanding the literature and methods associated with each subtask. While we do not seek to solve the definitional problem present in past work, we do hope to identify some fairly universal characteristics of different forms of abusive language

that will help us to overcome some of the area's inherent problems.

Based on work pertaining to hate speech, cyberbullying, and online abuse we propose a typology that captures central similarities and differences between subtasks and we discuss its implications for data annotation and feature construction. We argue that the differences between subtasks within abusive language can be reduced to two primary factors:

1. *Is the language directed towards a specific individual or entity or is it directed towards a generalized group?*

2. *Is the abusive content explicit or implicit?*

Each of the different subtasks related to abusive language occupies one or more segments of this typology. Our aim is to clarify the similarities and differences between subtasks in abusive language detection to help researchers select appropriate strategies for data annotation and modeling.

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" [79], "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" [23], "Youre one of the ugliest b*tches Ive ever fucking seen" [51].	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" [25], "((((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" [44], "you're intelligence is so breathtaking!!!!!!" [26]
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" [60], "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" [53].	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." [13], "most of them come north and are good at just mowing lawns" [26], "Gas the skypes" [58]

Table 2.1: Typology of abusive language.

2.2 A typology of abusive language

Much of the work on abusive language subtasks can be synthesized in a two-fold typology that considers whether (i) the abuse is directed at a specific target, and (ii) the degree to which it is explicit ([87]).

In terms of targets, abuse can either be directed towards a specific individual or entity, or it can be used towards a generalized *Other* (for example, people with a certain ethnicity or sexual orientation). This is an important sociological distinction as the latter references a whole category of people rather than a specific individual, group, or organization (see [11], [91]). As we discuss below, this entails linguistic distinctions that can be productively used by researchers to identify the different forms of abuse. To better illustrate this, the first row of Table 2.1 shows examples from the literature of directed abuse, where someone is either mentioned by name, tagged by a username, or referenced by a pronoun.[2] Cyberbullying and trolling are instances of directed abuse, aimed at individuals and online communities, respectively. The second row exhibits cases with abusive expressions towards generalized groups such as racial categories and sexual orientations. While previous work has identified instances of hate speech that are both directed and generalized ([13, 88, 23]), it seems [60] come closest to making a distinction between directed and generalized hate.

The other dimension is the extent to which abusive language is explicit or implicit. This is roughly analogous to the distinction in linguistics and semiotics between *denotation*, the literal meaning of a term or symbol, and *connotation*, its sociocultural associations, famously articulated by [5]. Explicit abusive language is that which is unambiguous in its *potential* to be abusive, for exam-

[2] All punctuation is as reported in original papers. We have added all the * symbols.

ple language that contains racial or homophobic slurs. Identifying this type of language should always be done with caution, however, as previous research has indicated a great deal of variation within such language ([83, 23]). The identification of abusive terms alone is not sufficient to identify explicit abuse, since such terms are often used in a colloquial manner or by people who are victims of abuse.

Implicit abusive language is that which does not immediately imply or denote abuse. Here, the true nature is often obscured by the use of ambiguous terms, sarcasm, lack of profanity or hateful terms, and other means, generally making it more difficult to detect by both annotators and machine learning approaches ([26, 20, 49]). Social scientists and activists have recently been paying more attention to implicit, and even unconscious, instances of abuse that have been termed "micro-aggressions" ([76]). As the examples show, such language may nonetheless have extremely abusive connotations. The first column of Table 2.1 shows instances of explicit abuse, where it should be apparent to the reader that the content is abusive. The messages in the second column are implicit and it is harder to determine whether they are abusive without knowing the context. For example, the word "them" in the first two examples in the [generalized and implicit] cell refers to an ethnic group, and the words "skypes" and "Google" are used as euphemisms for slurs about Jews and African-Americans, respectively. Abuse using sarcasm can be even more elusive for detection systems, which don't often benefit from the world experience or situational context necessary to identify sarcasm. As an example, both human annotators and machine learning approaches might find it difficult to pick up on the sarcasm behind the seemingly harmless comment praising someone's intelligence, which was actually a sarcastic response to a beauty pageant contestant's unsatisfactory

answer to a question ([26]).

2.3 Implications for future research

In the following section we outline the implications of this typology, highlighting where the existing literatures indicate how we can understand, measure, and model each subtype of abuse. Specifically, we outline some of the implications that our typology has for creating annotation guidelines and the machine learning models that use them.

2.3.1 Implications for annotation

Directed vs. Generalized Targets

In the task of annotating documents that contain bullying, it appears that there is a common understanding of what cyberbullying entails: an intentionally harmful electronic attack by an individual or group against a victim, usually repetitive in nature ([20]). This consensus allows for a relatively consistent set of annotation guidelines across studies, most of which simply ask annotators to determine if a post contains bullying or harassment ([19, 51, 10]). High inter-annotator agreement on cyberbullying tasks (93%) [20] further indicates a general consensus around the features of cyberbullying ([80]). After bullying has been identified, annotators are typically asked more detailed questions about the extremities of the bullying, the identification of phrases that indicate bullying, and the roles of users as bully/victim ([19, 80, 51]).

We hypothesize that consensus may be due to the directed nature of the phenomenon. Cyberbullying involves a victim whom annotators can identify and relatively easily discern whether statements directed towards the victim should be considered abusive. In contrast, in work on annotating harassment, offensive language and hate speech there appears to be little consensus on definitions and lower inter-annotator agreement ($\kappa \approx 0.60 - 0.80$) are obtained ([68, 84, 77, 9]). Given that these tasks are often broadly defined and the target is often generalized, all else being equal, it is more difficult for annotators to determine whether statements should be considered abusive. Future work in these subtasks should aim to have annotators distinguish between targeted and generalized abuse so that each subtype can be modeled more effectively.

Explicit vs. Implicit Abuse

Annotation (via crowd-sourcing and other methods) tends to be more straightforward when explicit instances of abusive language can be identified and agreed upon ([86]). This is considerably more difficult when implicit abuse is considered ([20, 49, 26]), especially since the connotations of language can be difficult to classify without domain-specific knowledge. While some argue that detailed guidelines can help annotators to make more subtle distinctions ([23]), others find that they do not improve the reliability of non-expert classifications ([68]). In such cases, expert annotators with domain-specific knowledge are preferred as they tend to produce more accurate classifications ([84]).

Ultimately, the type of abuse that researchers are seeking to identify should guide the annotation strategy. In highlighting the major differences between

different abusive language detection subtasks, our typology indicates that different annotation strategies are appropriate depending on the type of abuse. Where subtasks occupy multiple cells in our typology, annotation guidelines should reflect both similarities and differences, and annotators should be capable of (and allowed to) making nuanced distinctions that differentiate between different types of abuse and result in quality annotations.

It should also be noted here that the nature of abusive language can be extremely subjective, and researchers must endeavor to take this into account when using human annotators. [23], for instance, show that annotators tend to code racism as hate speech at a higher rate than sexism. As such, it is important that researchers consider the social biases that may lead people to disregard (or be sensitive to) certain types of abuse.

2.3.2 Implications for modeling

Existing research on abusive language online has used a diverse set of features in training and testing machine learning methods. Moving forward, it is important that researchers clarify which features are most useful for which subtasks and determine where the greatest challenges are presented. We do not attempt to review all of the features used (see [69] for a detailed review), but make suggestions for which features could be most helpful for the different subtasks. For each aspect of the typology, we suggest features that have been shown to be successful predictors in prior work. Many of these features occur in more than one form of abuse. As such, we do not propose that particular features are necessarily unique to each phenomenon, rather that they provide different insights and

should be employed depending on what the researcher is attempting to measure.

Directed abuse

Features that help to identify the target of abuse are crucial to directed abuse detection. Mentions, proper nouns, named entities, and co-reference resolution can all be used in different contexts to identify targets. [9] use a multi-tiered system, first identifying offensive statements, then their severity, and finally the target. Syntactical features have also proven to be successful in identifying abusive language. A number of studies on hate speech use part-of-speech sequences to model the expression of hatred ([83, 37, 23]). This includes determining whether or not the words used are nouns, verbs, pronouns, prepositions, adverbs, conjunctions, participles or articles, and how these words relate to each other. Another syntactical measure - typed dependencies - offers yet a more sophisticated way to capture the relationship between terms ([13]). Overall, there are many tools that researchers can use to model the relationship between abusive language and targets, although many of these require high-quality annotations to use as training data.

Generalized abuse

Generalized abuse online tends to target people belonging to a small set of categories, primarily racial, religious and sexual minorities. Past research has also identified what are considered to be "soft" hate targets, such as people that are called "stupid" or "fat" ([73]). Researchers should consider identifying

forms of abuse unique to each target group addressed, as vocabularies or key words used may depend on the groups targeted. For example, the language used to abuse trans-people and that used against Latin American people are likely to differ, both in the nouns used to denote the target group and the other terms associated with them. In some cases, therefore, a lexical method may be an appropriate strategy. Further research is necessary to determine if there are underlying syntactic structures associated with generalized abusive language.

Explicit abuse

Explicit abuse, whether directed or generalized, is often indicated by specific keywords. As such, dictionary-based approaches may be well-suited to identify this type of abuse ([83, 60]). Still, the presence of particular words should not be the only criteria - even terms that denote abuse may be used in a variety of different ways ([53, 23]). [23], for example, found that homophobic slurs are often used in a joking manner between friends. Additionally, slurs are often reappropriated by the communities they were initially used against. Identifying the negative polarity and sentiment of the text, therefore, may also be helpful as they are likely indicators of explicit abuse that can be leveraged by researchers ([37]).

Implicit abuse

Building a specific lexicon to identify implicit abuse may prove impractical, as in the case of the appropriation of the term "skype" in some forums ([58]). Still, even partial lexicons may be used as seeds to inductively discover other keywords by use of a semi-supervised method proposed by [50]. Additionally,

character n-grams have been shown to be apt for abusive language tasks due to their ability to capture variations of words associated with abuse ([60, 84]). Word embeddings are also promising ways to capture terms associated with abuse ([28, 4]), although they may still be insufficient for cases like 4Chan's connotation of "skype" where a word has both a dominant meaning and a more subversive one. Furthermore, as some of the above examples show, implicit abuse often takes on complex linguistic forms like sarcasm, metonymy, and humor. Without high-quality labeled data to learn these representations, it may be difficult for researchers to come up with models of syntactic structure that can help to identify implicit abuse. To overcome these limitations, researchers may find it prudent to incorporate features beyond just textual analysis, including the characteristics of the individuals involved ([20]) and other extra-textual features.

2.4 Discussion

This typology has a number of implications for future work in the area;

- We want to encourage researchers working on these subtasks to learn from advances in other areas. Researchers working on purportedly distinct subtasks are often working on the same problems in parallel. For example, the field of hate speech detection can be strengthened by interactions with work on cyberbullying, and vice versa, since a large part of both subtasks consists of identifying targeted abuse.
- We aim to highlight the important distinctions within subtasks that have hitherto been ignored. For example, in much hate speech research, di-

verse types of abusive or otherwise offensive language have been lumped together under a single label, forcing models to account for a large amount of within-class variation. We suggest that fine-grained distinctions along these axes allow for the creation of systems that may be more effective at identifying particular types of abuse.

- We call for closer consideration of how annotation guidelines are related to the phenomenon of interest. The type of annotation and even the choice of annotators should be motivated by the nature of the abuse. Further, we welcome discussion of annotation guidelines and the annotation process in published work. Many existing studies only tangentially mention these, sometimes never explaining how the data were annotated.
- We encourage researchers to consider which features are most appropriate for each subtask. Prior work has found a diverse array of features to be useful in understanding and identifying abuse, but we argue that different feature sets will be relevant to different subtasks. Future work should aim to build a more robust understanding of when to use which types of features.
- It is important to emphasize that not all abuse is the same, in terms of its effects, its detection and its consequences. We expect that social media and website operators will be more interested in identifying and dealing with explicit abuse, while activists, campaigners, and journalists may have more incentive to also identify implicit abuse. We also expect that implicit abuse will be more difficult to detect and model, although methodological advances may make such tasks more feasible. Targeted abuse such as cyberbullying may be more likely to be reported by victims and thus acted upon than generalized abuse.

2.5 Conclusion

We have presented a typology that synthesizes the different subtasks in abusive language detection. In particular, we describe the subtasks in terms of whether they aim to identify abuse that targets a specific individual versus a generalized audience and whether that abuse is implicit or explicit. Our aim in creating this typology is to bring together findings in these different areas and to clarify the key aspects of abusive language detection. There are important analytical distinctions that have been largely overlooked in prior work and through acknowledging these and their implications we hope to improve abuse detection systems and our understanding of abusive language. In this vein, we emphasize the practical actions that can be taken by researchers to best approach their abusive language detection subtask of interest.

Rather than attempting to resolve the "definitional quagmire" ([31]) involved in neatly bounding and defining each subtask we encourage researchers to think carefully about the phenomena they want to measure and the appropriate research design. We intend for our typology to be used both at the stage of data collection and annotation and the stage of feature creation and modeling. Finally, we hope that future work will be more transparent in discussing the annotation and modeling strategies used, and will closely examine the similarities and differences between these subtasks through empirical analyses.

CHAPTER 3

HATE SPEECH DETECTION AND THE PROBLEM OF OFFENSIVE LANGUAGE

3.1 Introduction

[1] What constitutes hate speech and when does it differ from offensive language? Much of the difficulty in finding, researching, and countering hate speech is that "hate speech" lacks a universal definition. Still, there is a consensus that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them ([48, 81]). The Merriam-Webster dictionary, for instance, offers the definition "speech that is intended to insult, offend, or intimidate a person because of some trait (as race, religion, sexual orientation, national origin, or disability)" ([24]). Similarly, the International Covenant on Civil and Political Rights (ICCPR) defines hate speech as "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" ([8]).

In the United States, hate speech is protected by the First Amendment - the right to free speech. "Fighting words," which may or may not include hate speech, are one of the only forms of speech to be prohibited by law, as they are likely to incite imminent illegal conduct or violence. Due to this protection, many organizations have had to fight hate speech via extrajudicial means, causing extensive debate in the legal sphere. Colleges and universities across the United States, for example, have enacted various anti-hate speech and anti-

[1] This work was co-authored with Thomas Davidson, Michael Macy and Ingmar Weber. It was presented at ICWSM (2017)

harassment codes to protect their students, many of which have been overturned due to their violation of the First Amendment. In many other countries, including the United Kingdom, Canada, Germany and France, laws have been enacted to prohibit the use of hate speech, which tends to be defined there as speech that targets minority groups in a way that could promote violence or social disorder. People convicted of using hate speech can often face large fines and even imprisonment.

In many cases these laws, by necessity, have been extended to the internet and social media. With the emergence of the web and online platforms that easily allow for self expression, internet users all over the world have the opportunity to quickly produce and access information of all types - including hate content. Online entities, in an effort to make their users feel safe and to comply with local hate speech laws, are tasked with defining and monitoring hate speech and determining a course of action in the event hate speech occurs. Social media websites often define and express their stances against and consequences for hate speech via community policies and guidelines. Facebook, for instance, states in their policy that they "do not permit individuals or groups to attack others based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or medical condition (Facebook 2015). [2] Twitter threatens account deletion or suspension as punishment for violent threats and harassment, and does not allow its users to "promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease" (Twitter 2015). [3] Tumblr also threatens actions against user

[2]Facebook's policy can be found here: `www.facebook.com/communitystandards#hate-speech`.

[3]Twitter's policy can be found here: `support.twitter.com/articles/20175050`.

accounts if users engage in malicious speech, especially when it is meant to instigate violence: "Don't encourage violence or hatred. Don't make violent threats or statements that incite violence, including threatening or promoting terrorism. Especially don?t do so on the basis of things like race, ethnic origin, religion, disability, gender, gender identity, age, veteran status, or sexual orientation." [4]

Guidelines, however, are not enough to keep users from exhibiting hate speech and social media has come under increasing pressure to identify and remove hate speech as fast as possible. As an example, in order to comply with Germany's laws against hate speech and to fight the increase in German hate speech due to an influx of refugees, Twitter, Google, and Facebook agreed to remove any hate speech posted to their sites within 24 hours of its occurrence in late 2015 ([6]). In October 2017, Germany passed a law that would impose huge fines on any social media site with more than two million users that failed to adhere to the law, likely in response to a German study stating that Facebook, Twitter and Youtube each only managed to remove less than 50% of illegal hate speech within the 24 hour deadline ([30]). Even with teams dedicated to identifying and removing hate speech, it proves a daunting task. This speaks to the importance of automated hate speech detection - it is quite difficult to quickly identify and remove hate speech manually given the mass amounts of data to be processed by online websites. The sheer nature of online hate speech makes it difficult to detect, much less eliminate; Online hate speech often spreads much faster than any site can contain, is usually immediately revived after being taken down, and stays around for a long period of time ([34]). Further, (pseudo)anonymity allows users the comfortability of using hate

[4]Tumblr's policy can be found here: `https://www.tumblr.com/policy/en/community`.

speech and expressing possibly unpopular ideas without fear of repercussions. Without automated detection, the web can cause both local law and online policies to be quite useless.

For these reasons, the automated detection of abusive language and hate speech has become of increasing interest in recent years. Indeed, many online entities have begun to employ automated abusive language detection mechanisms. Google created "Perspective" to find toxic comments online based on training data in the form of millions of annotated (very toxic to very healthy) comments ([39]). Yahoo created an abuse-detecting algorithm with 90% accuracy using a combination of machine learning and crowd-sourcing ([14]). Additionally, there has been a surge in the amount of research dedicated to detecting abusive language and hate speech, in particular. In this paper, we hope to contribute to advances in the area of hate speech detection by tackling the problem of offensive language.

3.2 The Problem of Offensive Language

A key challenge for automated hate speech detection on social media is the separation of hate speech from other instances of offensive language. If we conflate hate speech and offensive language then we erroneously consider many people to be hate speakers and fail to differentiate between commonplace offensive language and serious hate speech ([23]). Given the legal and moral implications of hate speech, it is important that we are able to accurately distinguish between the two. Contributing to the issue, as alluded to in Chapter 1, is the existence of a definitional quagmire in abusive language detection. Drawing from guiding

notions of what hate speech looks like and how it has been defined elsewhere, online entities are left to formulate their own definitions of hate speech. Add to these a range of definitions (or simply, differing labels) offered by researchers in various disciplines, and one encounters overwhelming problems associated with identifying hate speech and using past research as a guide in doing so. Data analysts, in particular, must then even further define hate speech by way of their choice of methodology. On a technical level, for example, hate speech can be defined by a list of terms or phrases, or by a set of features found common to hate speech - leaving infinite possibilities for hate speech definitions.

We too, are left to come up with a definition of hate speech that accurately captures the type of behavior we want to detect. Drawing upon earlier definitions, we define hate speech as *language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group*. In extreme cases this may also be language that threatens or incites violence, but limiting our definition only to such cases would exclude a large proportion of hate speech. Importantly, our definition does not include all instances of offensive language because people often use terms that are highly offensive to certain groups but in a qualitatively different manner. Some African Americans often use the term *n*gga*[5] in everyday language online [83], people use terms like *h*e* and *b*tch* when quoting rap lyrics, and teenagers use homophobic slurs like *f*g* as they play video games. Such language is prevalent on social media [82], making this boundary condition crucial for any usable hate speech detection system. Previous work on hate speech detection has identified this problem but many studies still tend to conflate hate speech and offensive language.

[5]Where present, the "*" has been inserted by us.

Bag-of-words approaches tend to have high recall but lead to high rates of false positives since the presence of offensive words can lead to the misclassification of tweets as hate speech ([54, 13]). Focusing on anti-black racism, [54] find that 86% of the time the reason a tweet was categorized as racist was because it contained offensive words. This made the identification of hate speech particularly challenging given the relatively high prevalence of offensive language and "curse words" on social media ([82]). The difference between hate speech and other offensive language is often based upon subtle linguistic distinctions, for example tweets containing the word *n*gger* are more likely to be labeled as hate speech than *n*gga* ([54]). Many can be ambiguous, for example the word *gay* can be used both pejoratively and in other contexts unrelated to hate speech ([82]).

Syntactic features have been leveraged to better identify the targets and intensity of hate speech, for example sentences where a relevant noun and verb occur (e.g. *kill* and *Jews*) ([38]), the POS trigram "DT jewish NN" ([83]), and the syntactic structure I <intensity > <user intent > <hate target >, e.g. "I f*cking hate white people" ([74]).

Other supervised approaches to hate speech classification have unfortunately conflated hate speech with offensive language, making it difficult to ascertain the extent to which they are really identifying hate speech ([13, 89]). Neural language models show promise in the task but existing work has used training data obtained with a similarly broad definition of hate speech ([29]). Non-linguistic features like the gender or ethnicity of the author can help improve hate speech classification, but this information is often unavailable or unreliable on social media ([89]).

In this paper, we used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. We use crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories. Close analysis of the predictions and the errors shows when we can reliably separate hate speech from other offensive language and when this differentiation is more difficult. Our results show that fine-grained labels can help in the task of hate speech detection and highlights some of the key challenges to accurate classification. We conclude that future work must better account for context and the heterogeneity in hate speech usage. We find that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Tweets without explicit hate keywords are also more difficult to classify.

3.3 Data

We begin with a hate speech lexicon containing words and phrases identified as hate speech by *Hatebase.org*. Hatebase.org is dedicated to keeping track of hate speech terminology and its use geographically to "assist government agencies, NGOs, research organizations and other philanthropic individuals and groups use hate speech as a predictor for regional violence." ([61]) Their multilingual lexicon, which we will refer to as the Hatebase lexicon, consists of terms reported as used in hate speech incidents by the global community of Hatebase users. Using these terms as filters for the Twitter API, we collected a sample of tweets presumed to contain hate speech. This resulted in a sample of tweets from 33,458 Twitter users. We then extracted the timelines for each user, result-

ing in a set of 85.4 million tweets in total.

3.4 Tweet Annotation

We used CrowdFlower (CF), a platform that allows researchers to hire CF workers to "collect, clean, and label high-quality large-scale data sets," to have a subset of the 25,360 tweets containing Hatebase terms labeled as either hate speech, offensive language, or language that is not offensive. In order to ensure the quality of the annotations, we provided the following excerpt of instructions to CF's "expert" workers:

Hate speech is defined as "language that is used in reference to certain groups that expresses hatred towards the group or is intended to be derogatory, to humiliate, or to insult the members of the group." Hate speech often targets people based on their race, ethnicity, gender, sexual preference, or religion. This might (but does not necessarily) include calls for action against a targeted group, such as discrimination, deportation, or physical violence.

Offensive speech might use some of the same words we associate with hate speech but do not necessarily constitute hate speech because the words are not used in the same context as "hate speech." Examples include tweets with racial, violent, homophobic or sexist terms that are used jokingly amongst friends, or tweeted as part of song lyrics. Tweets like these should be classified as "offensive," but because they are not used to express hatred, or to denigrate, humiliate, or insult, they do not constitute hate speech.

CF workers were especially urged to consider not just the words contained in the tweet, but also the context of the tweet, as this is key in distinguishing between hate speech and offensive speech. In this vein, they were also provided with example tweets for each category. To illustrate the importance of context, they were asked to consider examples such as the following two tweets, both of which contain the term "f*ggot". While the first user clearly calls for physical violence against gay individuals, the latter uses the term in what appears to be a joking manner.

1. "There really needs to be a Holocaust for f*ggots, y'all taking sh*t too far wearing skirts and leggings, go to Hell."

2. "Oh shush you know I love you f*ggot"

Each tweet was annotated by three CF workers. Having multiple workers annotate a tweet was important due to the subjective nature of hate speech. The benefit of using a lexicon heavily influenced by CF workers is that it allows us to define hate speech based on what human beings believe hate speech to be. The difficulty lies in the fact that one's perception of hate speech is undoubtedly colored by their experiences within the world, in which race, gender, socioeconomic class, sexual orientation and other attributes all come into play. As experience varies from one person to the next, definitions of hate speech are likely to differ. As described in the next section, we find that even when detailed guidelines are provided, it is still quite possible that users' annotations

can differ greatly. Still, by using multiple CF "expert" workers and analyzing inter-annotator agreement, we are able to hone in on terms and tweets that most people can agree is hate speech.

3.4.1 Tweet Annotation Results

Analysis showed that only 1.3% of the 25,360 tweets in the sample were agreed upon to be hate speech unanimously by all three CF workers, while a total of 5% of tweets were agreed upon by at least two of three workers. We found these numbers to be lower than that found in the Twitter-based literature, such as [13], where 11.6% of tweets were flagged as hate speech. We attribute our lower rate of hate speech to two causes. First, we provided the CF workers with a fairly strict definition of hate speech that was meant to weed out offensive language. Second, the imprecision of the Hatebase lexicon resulted in the collection of a large number of false positives in the form of tweets that were either offensive or not offensive at all. While the words contained in the lexicon may have been used in hate speech incidents, many of them (e.g. "bird", "slope" and "sole") are quite innocuous and are more often used in manners that are not offensive.

Most tweets collected were considered to be offensive, with 76% of all tweets having at least 2/3 agreement (53% at 3/3 agreement). And finally, 16.6% of tweets were found not to be offensive at all with at least 2/3 agreement (11.8% at 3/3 agreement). The overall intercoder-agreement score provided by CF was 92%. We used the majority decision for each tweet to assign a label. Some tweets were not assigned labels as there was no majority class. This resulted in a sample of 24,802 annotated tweets. We used this sample of labeled tweets to

further refine the Hatebase lexicon and subsequently construct features used to train a classifier.

3.5 Features

3.5.1 Refining the Hatebase Lexicon

In order to use the Hatebase lexicon as a feature in our classifier, it was vital to improve its precision. To relieve the issue of false positives associated with the Hatebase lexicon, we sought to refine it by removing infrequently used terms, terms that result in high rates of false positives, and terms not related to hate speech at all. We shifted our focus to the use of *n*-grams, which are simply subsequences of tweets that consist of n continuous words in the order they appear in the tweet. We use n-grams instead of the traditional Bag-of-Words approach because the sequential structure of n-grams are better able to relay the context of a sentence. Indeed, the use of n-grams over single words alone has been shown to improve text analysis and classification in previous work ([64]).

Using our annotated subset of 24,802 tweets, the most frequent 1-, 2-, 3-, and 4-grams were found, along with the proportion of time that CF workers classified a tweet containing each n-gram as hate speech (at 2/3 agreement). For each n-gram, we performed a t-test and removed any n-grams not found to be statistically significant with a t-value of $t > 1.96$. The refined lexicon shows improvement over the Hatebase lexicon in terms of precision. As expected, however, it also has a lower recall than that of the Hatebase lexicon, as we are not including nearly as many terms and therefore not capturing every instance of hate speech

possible.

In order to validate this new lexicon of n-grams, up to 50 tweets were collected for each term via Twitter's Streaming API for annotation. As a testament to the rarity of hate speech, even after multiple calls to the API, we were unable to collect the full 50 tweets for some of the terms - the same terms that were found to be most frequent in previous analyses. With the goal of identifying the strength of the connection between the terms in our new lexicon and the use of hate speech, we had the newly collected tweet annotated as "hate speech", "offensive" or "not offensive" using the same CF labeling process described in Section 3.4. We again ranked the n-grams by the percentage of tweets containing each n-gram that were coded as hate speech. After some experimentation, we found that the best approach to further refine our lexicon was by removing n-grams such that 1) > 80% of CF workers found it not offensive, 2) > 80% found it to be offensive, and 3) < 20% found it to be hate speech. By removing the terms often found to be either not offensive or offensive and terms seldom found to be hate speech, we were able to increase precision by decreasing false positives, again at the cost of recall. The final lexicon contains 123 hate speech n-grams.

We chose this final lexicon of 123 hate speech n-grams as the best option for a number of reasons. While the simple use of a Bag-of-Words approach has been found to be inefficient in previous literature ([55]), the frequency of particular unigram (single word) and bigram (two-word) terms in our tweet corpus were overwhelming and needed to be utilized. Further, it allows us to define hate speech as what most people believe it to be. Finally, it performs the best in terms of precision.

Racial and homophobic slurs dominate our final lexicon, with few terms related to women, ethnic background and religion. This is in line with previous research, which suggests a majority of hate speech online targets people belonging to a small set of categories, primarily racial, sexual, and religious minorities ([73]). It also indicates that the CF workers are more sensitive to racial and homophobic slurs than sexist or other terms. It is also interesting to note that the term "f*g" was not included in these lexicons - another indication that certain terms, while inflammatory, are more often used in contexts other than hate speech.

The use of a lexicon-based approach has its limitations. Lexicons pose a number of problems - they can be too inclusive or not inclusive enough, they often do not contain various spellings of terms (which can be used to avoid detection), they must often be updated to be on par with the changing times, and they do not take into account the context or sentiment of the statement ([55]). Lexicon-based approaches work best for offensive language, as no distinction needs to be made between hate speech and offensive language, but can be very useful in detecting hate speech in conjunction with other features when taking a classifier-based approach. As such, we combine our 123-term lexicon with a number of other features, outlined below.

3.5.2 Additional Features

We lowercased each tweet and stemmed it using the Porter stemmer. Stemmers are often used in Natural Language Processing as a means of grouping words together based on their word base. Using these word bases as features for classi-

fication allows the classifier to make associations between words with different inflections. We verified that the stemmer did not remove important information by reducing key terms to the same stem, e.g. *f*gs* and *f*ggots* stem to *f*g* and *f*ggot*.

We created bigram, unigram, and trigram features, each weighted by its term frequency-inverse document frequency value (TF-IDF). TF-IDF, popular in text analysis, is often used as a measure of how important or informative a word is for a document. *Term frequency* counts the number of times a word appears in a document, while *inverse document frequency* is the inverse fraction of the documents containing the term, measuring how unique a word is across documents. TF-IDF varies from 0 to 1, where 0 indicates that a term likely occurs across documents and is therefore not very informative. The n-grams with high TF-IDF scores, therefore, tend to provide the most information about whether or not our tweets contain hate speech. TF-IDF is calculated via the formula $W_{i,j} = tf_{i,j} \times log\frac{N}{df_i}$, where $tf_{i,j}$ is the number of occurrences of n-gram i in tweet j, df_i is the number of documents containing word i and N is the total number of tweets.

To capture information about the syntactic structure we use `NLTK` [7] to construct Penn Part-of-Speech (POS) tag unigrams, bigrams, and trigrams. Though n-grams provide more context than we would obtain with single words, n-grams may still suffer from lack of context in that there may be a great distance between related words ([16]). Identifying the POS tag of each word in the n-gram using the context of the n-gram, which involves marking each word with the part of speech (noun, verb, pronoun, preposition, etc.) it takes on, helps to relieve this issue. Indeed, it has been shown to improve classifier performance

in previous work on abusive language detection tasks such as cyberbullying ([25]).

To capture the quality of each tweet we used modified Flesch-Kincaid Grade Level and Flesch Reading Ease scores ([33]). The Reading Ease formula, which is meant to measure how difficult a text is to read, is calculated as $RE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$, where ASL is the average sentence length, and ASW is the average number of syllables per word. The Grade Level formula is slightly more interpretable in that its output is easily translated into the grade-level at which a reader would need to be in order to understand the text. Grade Level is calculated as $GL = .39(ASL) + 11.8(ASW) - 15.59$. The intuition behind these measures is that longer sentences, and similarly longer words, are likely more difficult to read. These features have been successfully used in previous work on identifying sarcasm ([66]) and sentiment ([41]), and so we use both formulas as features at the tweet level, with the number of sentences fixed at one.

To capture the sentiment of each tweet, we used a sentiment lexicon designed for social media to assign sentiment scores to each tweet ([47]). Sentiment analysis is often used to determine the affective state of a text, often measuring sentiment in terms of the degree of negative and positive sentiment found. While qualitatively different, sentiment analysis is heavily related to the identification of hate speech (one might expect that hate speech is associated with an overall negative connotation) and it has been found to be beneficial in classifying hate speech and offensive language ([83];[28]).

Finally, we include simple binary and count indicators for hashtags, mentions, retweets, and URLs, as well as features for the number of characters, words, and syllables in each tweet. We also attempted to use character n-grams

and a custom word embedding but found that they did not improve our model.

3.6 Model

We first used a logistic regression with L1 regularization to reduce the dimensionality of the data. L1 regularization, also termed LASSO (least absolute shrinkage and selection operator), is often used in machine learning as means of reducing the feature space and preventing the overfitting of the model to the training data. It incorporates a penalty on the parameters of the model that is equal to the sum of the absolute value of the parameters. In doing so, some of the parameters disappear, leaving only features that are most relevant or useful to the task at hand.

We then tested a variety of models that have been used in prior work: logistic regression, naïve bayes, decision trees, random forests, and linear support vector machines. First, we split our data into a training set and a validation set. The training set was split into five equal sized subsamples. In each of five training sessions for a model, a different subsample was held out for use as a test set while the remaining four samples were used to train the model. The validation set, equal to 10% of our original data set, was held out to both prevent overfitting and to later evaluate the models on completely unseen data.

In order to identify the best model, we used a grid-search to iterate over the models and corresponding parameters. We found that the logistic regression and linear SVM tended to perform significantly better than other models. We decided to use a logistic regression with L2 regularization for the final model, where L2 regularization incorporates a penalty equal to the square of the pa-

rameter values in order to further protect against overfitting. This choice of model more readily allows us to examine the predicted probabilities of class membership and has performed well in previous papers ([13];[89]).

We trained the final model using the entire dataset and used it to predict the label for each tweet. We use a one-versus-rest framework where a separate classifier is trained for each class and the class label with the highest predicted probability across all classifiers is assigned to each tweet. All modeling was performed using `scikit-learn` ([62]).

3.7 Results

The best performing model has an overall precision 0.91, recall of 0.90, and F1 score of 0.90. Looking at Figure 3.1, however, we see that almost 40% of hate speech is misclassified: the precision and recall scores for the hate class are 0.44 and 0.61 respectively. Most of the misclassification occurs in the upper triangle of this matrix, suggesting that the model is biased towards classifying tweets as less hateful or offensive than the human coders. Far fewer tweets are classified as more offensive or hateful than their true category; approximately 5% of offensive and 2% of innocuous tweets have been erroneously classified as hate speech. To explore why these tweets have been misclassified we now look more closely at the tweets and their predicted classes. Example tweets for each class, in the form of a confusion matrix, can also be found in Figure 3.2.

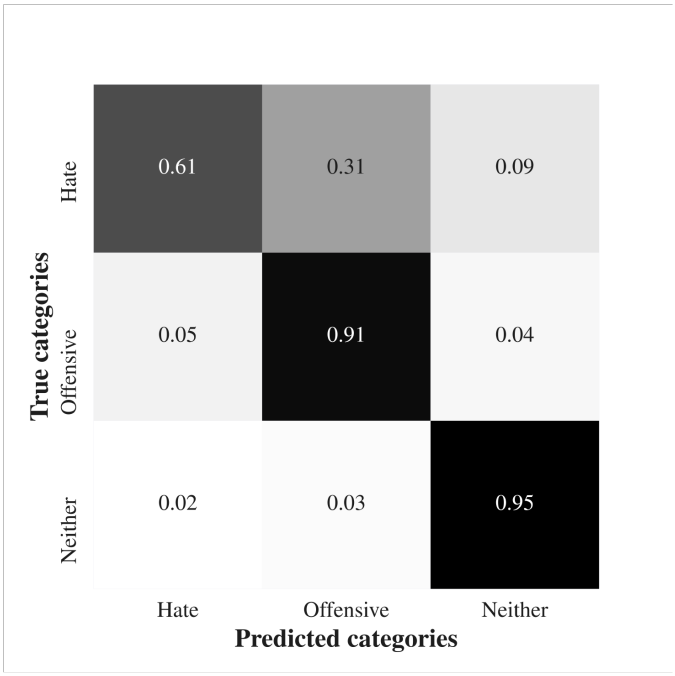
Tweets with the highest predicted probabilities of being hate speech tend to contain multiple racial or homophobic slurs, e.g. *"@JuanYeez shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga"* and *"RT @eBeZa:*

*Stupid f*cking n*gger LeBron. You flipping jungle bunny monkey f*ggot"*. Other tweets tend to be correctly identified as hate when they contained strongly racist or homophobic terms like *n*gger* and *f*ggot*. Figure 3.3 further illustrates our classifier's dependence on strong hate terms. Indeed, the 20 most important features in predicting hate speech are all from our refined lexicon of hate n-grams. Interestingly, we also find cases where people use hate speech to respond to other hate speakers, such as this tweet where someone uses a homophobic slur to criticize someone else's racism: *"@MrMoonfrog @RacistNegro86 f*ck you, stupid ass coward b*tch f*ggot racist piece of sh*t"*.

Turning to hate speech that was incorrectly classified as offensive, we examine tweets based on their predicted probability of being offensive. It appears that hate tweets with the highest predicted probability of being offensive are genuinely less hateful and were perhaps mislabeled as hate speech, for example *When you realize how curiosity is a b*tch #CuriosityKilledMe* may have been erroneously coded as hate speech if people thought that *curiosity* was a person, and *"Why no boycott of racist "redskins"? #Redskins #ChangeTheName"* contains a slur but is actually against racism. It is likely that coders skimmed these tweets too quickly, picking out words or phrases that appeared to be hateful without considering the context. For borderline cases, where the probability of being offensive is marginally higher than hate speech, it appears that the majority are actually hate speech, both directed towards other Twitter users, *"@MDreyfus @NatFascist88 Sh*t your ass your moms p*ssy u Jew b*stard. Ur times coming. Heil Hitler!"* and general hateful statements like *"My advice of the day: If your a tranny...go f*ck your self!"*. These tweets fit our definition of hate speech but were likely misclassified because they do not contain any of the terms most strongly associated with hate speech.

Finally, the hateful tweets incorrectly labeled as not offensive tend not to contain hate or curse words, for example *If some one isn't an Anglo-Saxon Protestant, they have no right to be alive in the US. None at all, they are foreign filth* contains a negative term, *filth* but no slur against a particular group. We also see rarer types of hate speech misclassified; for example, this anti-Chinese statement *Every slant in #LA should be deported. Those scum have no right to be here. Chinatown should be bulldozed*. Overall, the classifier performs well at predicting prevalent forms of hate speech (particularly anti-black racism and homophobia), but is less reliable in detecting types of hate speech that occur infrequently, a problem noted by [60].

Figure 3.1: True Versus Predicted Categories



A key flaw in much previous work is that offensive language is mislabeled as hate speech due to an overly broad definition. Our multi-class framework allows us to minimize these errors; only 5% of our true offensive language was labeled as hate. The tweets correctly labeled as offensive tend to contain curse

Figure 3.2: True Versus Predicted Tweets

JuanJeez shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga RT eBiZa: Stupid f*cking n*gger LeBron. You flipping jungle bunny monkey f*ggot	@SmogBby: These h*es be lyin to all of us n*ggas and My n*gga mister meaner just hope back in the b*tch. @Hova_ ya I know all the slang I'm racist I h8 porch monkies	@mayormcgunn @SenFeinstein White people need those weapons to defend themselves from the subhuman trash your sort unleashes on us.
When you realize how curiosity is a b*tch. An should of let the unknown stay, unknown. #CuriosityKilledMe RT @1BaD_doc: I hate when people get on my bad side cause I honestly don't like to act like a bitch	Why you worried bout that other h*e? Cuz that other h*e aint worried bout another hoe and @KingJames is such a h*e. Wear your shoes ya big b*tch ...you missed the couch shot Ray Allen bailed you out ...p*ssy.	He's a damn good actor. As a gay man it's awesome to see an openly queer actor given the lead role for a major film as #Shoutout to everybody that follow a nicca!!
If some one isn't an Anglo-Saxon Protestant, they have no right to be alive in the US. None at all, they are foreign filth	@MannyDiesel: Def not cowboy lol RT @ArtofFloyd: Terrell Owens was the best Eagle & Cowboy ever" ..dude cried like a bitch on tv, over Romo	cleaned all the bird nests out of the downspouts and ready for more rain, the baby birds are out now. respect the birds.....and bees (:

words and often sexist language, e.g. *"Why you worried bout that other h*e? Cuz that other h*e aint worried bout another h*e"* and *"I knew Kendrick Lamar was onto something when he said "I call a b*tch a b*tch, a h*e a h*e, a woman a woman""*. Many of these tweets contain sexist terms like *b*tch*, *p*ssy*, and *h*e*. Human coders appear to consider racists or homophobic terms to be hateful but consider words that are sexist and derogatory towards women to be only offensive, consistent with prior findings ([89]). Figure 3.4 demonstrates this trend, with more than half of the most important features dedicated to sexist language.

Looking at the offensive tweets misclassified as hate speech we see that many contain multiple slurs, e.g. *@SmogBaby: These h*es be lyin to all of us n*ggas* and *"My n*gga mister meaner just hope back in the b*tch"*. While these tweets contain terms that can be considered racist and sexist it is apparent that many Twitter users use this type of language in their everyday communications. When

they do contain racist language they tend to contain the term *n*gger* rather than *n*gga*, in line with the findings of [54]. We also found a few recurring phrases such as *these h*es ain't loyal* that were actually lyrics from rap songs that users were quoting. Classification of such tweets as hate speech leads us to overestimate the prevalence of the phenomenon. While our model still misclassifies some offensive language as hate speech, we are able to avoid the vast majority of these errors by differentiating between the two.

Finally, turning to the "not offensive" class, we see that tweets with the highest predicted probability of belonging to this class all appear to be innocuous and were included in the sample because they contained terms included in the Hatebase lexicon such as *charlie* and *bird* that are generally not used in a hateful manner (see Figure 3.5). Tweets with overall positive sentiment and higher readability scores are more likely to belong to this class. The tweets in this category that have been misclassified as hate or offensive tend to mention race, sexuality, and other social categories that are targeted by hate speakers. Most appear to be misclassifications caused only by the presence of potentially offensive language, for example *"He's a damn good actor. As a gay man it's awesome to see an openly queer actor given the lead role for a major film"* contains the potentially offensive terms *gay* and *queer* but uses them in a positive sense. This problem has been encountered in previous research [83] and illustrates the importance of taking context into account. We also found a small number of cases where the coders appear to have missed hate speech that was correctly identified by our model, e.g. *"@mayormcgunn @SenFeinstein White people need those weapons to defend themselves from the subhuman trash your sort unleashes on us"*. This finding is consistent with previous work that has found amateur coders to often be unreliable at identifying abusive content ([60];[85]).

Figure 3.3: Most Important Features for the Hate Class

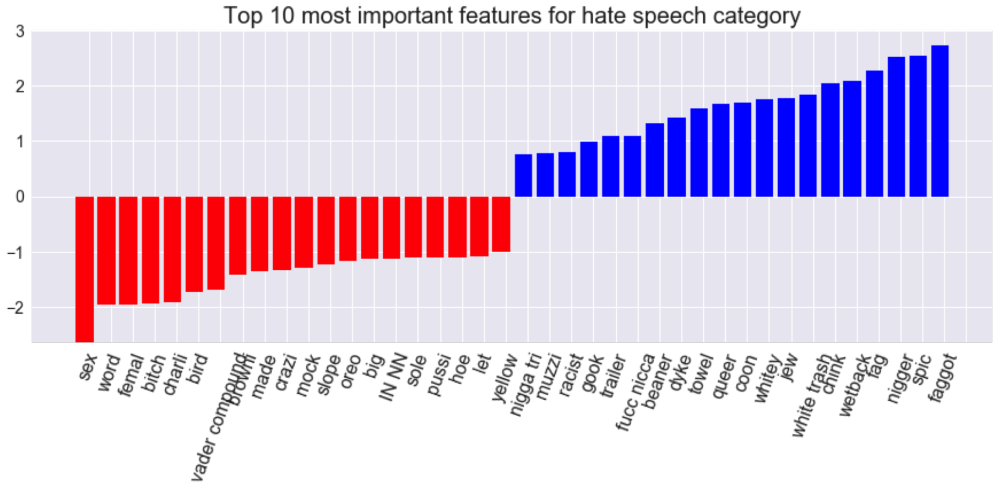
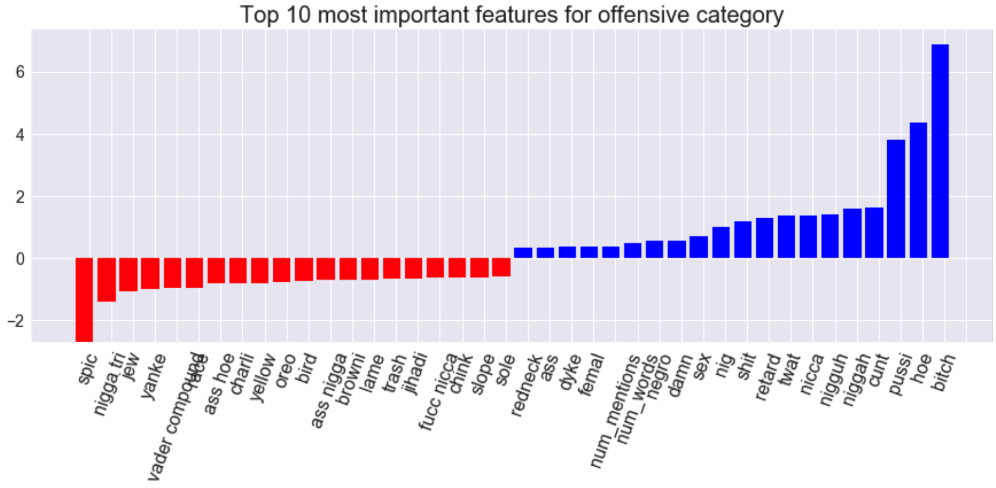


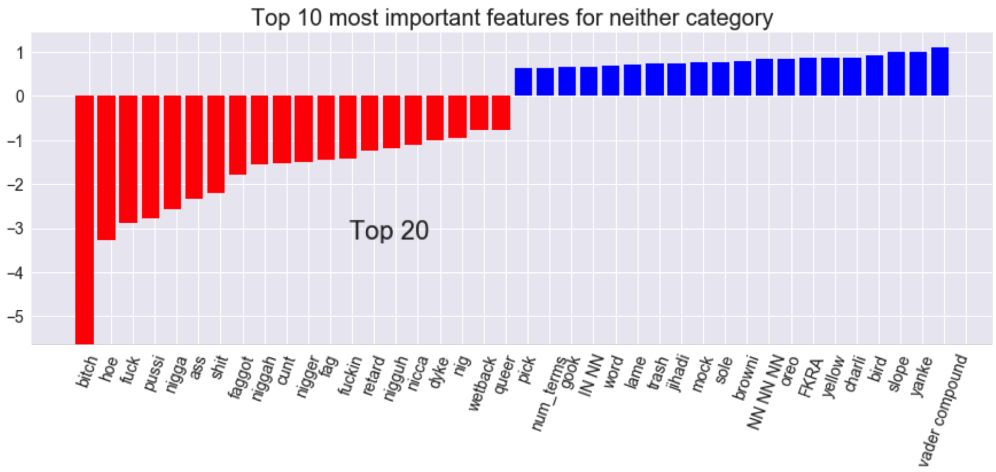
Figure 3.4: Most Important Features for the Offensive Class



3.8 Discussion

We believe the work presented in this chapter has shed light on effective methodologies for detecting hate speech and considerations that must be taken into account when endeavoring to do so. We found that lexical methods are effective ways to identify potentially offensive terms but are inaccurate at identifying hate speech; only a small percentage of tweets flagged by the Hatebase lexicon were considered hate speech by human coders. If a lexicon must be

Figure 3.5: Most Important Features for the Not Offensive Class



used we propose that a smaller lexicon with higher precision is preferable to a larger lexicon with higher recall. [6] While automated classification methods can achieve relatively high accuracy in differentiating between these different classes, close analysis of the results shows that the presence or absence of particular offensive or hateful terms can both help and hinder accurate classification.

Consistent with previous work, we find that certain terms are particularly useful for distinguishing between hate speech and offensive language. While *f*g*, *b*tch*, and *n*gga* are used in both hate speech and offensive language, the terms *f*ggot* and *n*gger* are generally associated with hate speech only. Further, many of the tweets considered most hateful contain multiple racial and homophobic slurs. While this allows us to easily identify some of the more egregious instances of hate speech it also means that we are more likely to misclassify hate speech if it doesn't contain any curse words or offensive terms. To more accurately classify such cases we should find sources of training data that are hateful without necessarily using particular keywords or offensive language.

[6] We have made a more restricted version of the Hatebase lexicon available here: `https://github.com/t-davidson/hate-speech-and-offensive-language`.

Our results also illustrate how hate speech can be used in different ways: it can be sent directly to a person or group of targeted people, it can be espoused to no one in particular (a general *Other*), and it can be used in conversation between people. In line with the typology presented in Chapter 2, future work should distinguish between these different uses, possibly drawing more insight from work in areas such as cyberbullying and trolling. It would also be beneficial to look more closely at the social contexts and conversations in which hate speech occurs. We must also study more closely the people who use hate speech, focusing both on their individual characteristics and motivations and on the social structures they are embedded in. We take this task on in the next chapter.

Finally, hate speech is a difficult phenomenon to define and is not monolithic. Our classifications of hate speech tend to reflect our own subjective biases. We found that people identify racist and homophobic slurs as hateful but tend to see sexist language as merely offensive. While our results show that people perform well at identifying some of the more egregious instances of hate speech, particularly anti-black racism and homophobia, it is important that we are cognizant of the social biases that enter into our algorithms and future work should aim to identify and correct these biases.

3.9 Conclusion

If we conflate hate speech and offensive language then we erroneously consider many people to be hate speakers (errors in the lower triangle of Figure 3.1) and fail to differentiate between commonplace offensive language and serious hate

speech (errors in the upper triangle of Figure 3.1). Given the legal and moral implications of hate speech it is important that we are able to accurately distinguish between the two. In this chapter, we presented a classifier for detecting hate speech that attempts to solve the problem of offensive language. Using crowd-sourced training data that categorized suspected hate speech tweets as "hate speech", "offensive" or "non-offensive," we were able to train a classifier that helped to minimize the amount of offensive tweets that would be incorrectly classified as hate speech using other methods.

Future work includes the creation of a classifier that further prevents the conflation of hate speech and offensive language. While our current classifier does well in correctly classifying offensive language - it only classifies 5% of offensive language as hate speech - it needs improvement in its overall classification of hate speech. In particular, we hope to build a classifier that better recognizes more subtle forms of hate speech that do not necessarily include overtly hateful terms.

CHAPTER 4

HATE SPEAKERS: A HIDDEN POPULATION

4.1 Introduction

[1] While there has been an abundance of research dedicated to detecting abusive language online, the offenders themselves have been largely overlooked. Work that has focused on the parties involved in abusive language incidents have mostly centered around identifying the targets of the abuse, understanding the effects of abuse and mitigating them, especially as related to adolescents and children ([25],[75]). Little work has focused on the perpetrators, with the exception of the cyberbullying literature. This is especially true in the area of hate speech, where individual hate speakers tend to be difficult to identify. Indeed, individual hate speakers that are not tied to larger hate or extremists groups and are not rallying around an indentifiable cause (such as Gamergate) form a sort of hidden population - a population that is not easily accessed by surveys or other traditional means of data collection.

In this chapter, we continue our focus on the abusive language subtask of hate speech, but shift our attention to understanding and identifying individual *hate speakers*. We identified a set of Twitter users whose tweets have been classified using the hate speech classifier presented in the previous chapter. We examine the personal characteristics of these users, some of which exhibit no hate speech and others which exhibit it to varying degrees. Using information from user timelines and machine learning models, we classify users' gender,

[1] This work was co-authored with Thomas Davidson, Michael Macy, Ingmar Weber and Meysam Alizadeh.

race, education level, psychological profiles and location, and assess how hate speech usage is associated with these traits. We test whether this information can be used to predict whether an individual is likely to use hate speech, and the amount of hate speech they are likely to exhibit. In this vein, we hope to illuminate the hidden population of hate speakers by offering new insights into the people who use hate speech online and how we can detect them.

4.2 Related work

There exists a large gap in the abusive language literature as it pertains to the abusers. Past work that has focused on identifying the perpetrators of abusive language has largely focused on two categories: cyberbullies and extremist groups. Unlike with individual hate speakers, these studies are facilitated by the ease with which these groups are identified - largely due to the nature of the abusive language involved.

The repetitive and damaging nature of cyberbullying makes identifying perpetrators one of the top priorities of cyberbullying literature, especially when adolescents are involved ([78], [70], [43]). The directed nature of cyberbullying allows researchers to more easily identify both bully and target, while the repetitive nature often results in a documented history of abusive language, providing context for the researcher that might not be obtained in looking at single instances of abusive text. Extremist groups are easily identifiable for different reason than cyberbullies. These groups tend to be quite visible on social media, often actively recruiting and disseminating hateful information. Since individuals associated with these groups online are usually visible members or

followers, a mere collection of extremist groups' social network data is enough to begin to study these users ([94]).

Though our work may not benefit from methods used to identify cyberbullies and extremist groups, our typology of abusive language suggests it is prudent that we understand the overlapping nature of these studies so that we might benefit from findings and advances in those literatures. This is especially true since members of extremist groups and cyberbullies alike often exhibit the kind of hate speech encountered and classified in the previous chapter. Indeed, the cyberbullying and extremist group literatures point to a number variables that could prove useful in identifying hate speakers.

[67] point to the correlations between cyberbullying and psychological measures of aggression and anxiety, while [2] use psychological and personality measures as a means of explaining the differences between users that follow extreme right groups and users that don't. [18] highlight the importance of incorporating demographic variables, while [45] illustrate the usefulness of combining textual content and social network attributes in detecting cyberbullying instances. [94] take a similar approach in classifying extreme right group online by combining information regarding social network structure with content analysis. [15] performs a hate speaker analysis on Twitter users involved in the Gamergate controversy. Similar to our approach here, they identify tweets containing the #Gamergate hashtag and subsequently use the Hatebase lexicon to filter out abusive and hateful content. In comparing the corresponding hate speakers to a set of random Twitter users, the authors found that Gamergate users tend to have more friends and followers, tweet more, use more hashtags, post tweets with less joy and more hate, and are less likely to have their accounts

suspended.

While these works bear some similarity to the objectives of this paper, there are some key fundamental differences that we believe result in unique contributions to the area. The individuals we study are heterogenous in that they 1) engage in varying types of hate speech (racist, homophic, sexist, etc.), unlike the primarily sexist hate speech associated with Gamergate and 2) are not necessarily organized around extreme group affiliations, movements or rallying causes. These individuals may very well encompass cyberbullies and extremist group members and followers, but will also include hate speakers that don't fall into either category. Further, though [15] make compelling inferences regarding Gamergate tweeters, their hate speech detection method relies too heavily upon a lexicon that we have previously shown is very imprecise in detecting hate speech and offensive language. In this paper, we emphasize the importance of creating of a reliable means of detecting hate speech and speakers in a way that avoids the condemnation of innocent users.

4.3 What Makes a Hate Speaker?

Identifying hate speakers proved to be just as complex as identifying hate speech, if not more. What qualifies a user as a hate speaker? Should their timeline contain a threshold proportion of hate tweets, and how do we choose that threshold? Should they be judged by the number of hate speech targets (race, sex, class, etc.) they reference, or is it enough to just target one? Should they use a certain number of (unique) hate terms in their tweets? Our first attempt to classify hate speakers sought to bypass these very difficult questions by refor-

mulating the problem as a human annotation task. We hoped that categorizing hate speakers, much like categorizing hate tweets, would be a "know it when you see it" task. We hypothesized that human annotators would have high overall agreement on which Twitter users qualified as hate speakers, given their tweets, even if the questions outlined above were not explicitly answered by the authors. In this section, we recount this initial attempt at identifying hate speakers through human annotation and outline possible reasons for why it was largely unsuccessful.

4.3.1 Human Annotation of Hate Speakers: A Cautionary Tale

In an effort not to make arbitrary decisions about what qualifies a user as a hate speaker, we took a crowd-sourcing approach similar to the approach we used for classifying tweets. Running this experiment on Crowdflower, this time using hate speakers instead of tweets, required that we needed a new set of "gold" standard questions. On CF, gold questions test the performance of the CF worker by asking them a question for which the answer is already known. If the worker answers too many gold questions incorrectly, they will be dismissed from the job and their responses deleted. In order to create a set of gold questions, we enlisted the help of three Cornell University undergraduate students. After randomly choosing a set of 250 users from our dataset, we presented each student with a sample of up to 10 tweets from their Twitter timelines. The previously used instructions for labeling a tweet as hate speech were presented, along with the following additional instructions, *"You will read tweets for each user and answer questions about the user of the tweets. First, we want you to classify each tweet based on whether it contains hate speech (as defined above), it is using crude*

and offensive language but does not express hatred, it is against hate speech, or it is inoffensive. It is imperative that you pay attention to the context in which language is used in order to ascertain whether the content qualifies as hate speech. Then, we want you to take all tweets into context in order to answer questions about the user of the tweet." The category "opponent of hate" has been added to this analysis largely in an effort to reduce any false positives associated with hate speakers that often use hate terms to protest them.

Coming up with a set of gold questions proved a difficult task - only two of the 250 sampled users were labeled hate speakers with full 3/3 agreement. An additional 47 were labeled as hate speakers with 2/3 agreement, but we did not feel this agreement was strong enough to justify using them as gold standard questions. It appears that annotation is an easier task for humans when judging tweets than when judging users. We offer some possible hypotheses for such low agreement. First, it is possible that the coders found it more difficult to make a judgment about the "character" of a person, for which there is no direct evidence (via interaction or the like), than the content of a single tweet, for which all of the information is available. Second, hate speaker annotations are triply impacted by the biases of the annotators. We asked that they judge each tweet separately, then make a judgement about the user based on the collection of these tweets. Annotators therefore had to answer three questions in the process: "Is each tweet a hate tweet?", "How hateful is the collection of tweets?" and "How hateful does a collection of tweets need to be in order to classify a user as a hate speaker?" Overall inter-annotator disagreement may have been amplified by disagreement on how each of these judgements should be made.

In a labeling experiment like that described with the students, we tested the

outcome of a CF experiment run without gold questions. We found no additional unanimously labeled users - in fact, no user was labeled a hate speaker with full agreement, and only two with 2/3 agreement. We attribute the major difference in outcomes (students vs. CF workers) to the lack of gold questions. Without these checks in place, there is evidence that even the high quality workers began to answer the questions without thought - possibly without even reading them. This was especially evidenced when looking at the tweets of users that were labeled an "opponent of hate." CF workers had selected this option for a number of offensive and hate speakers that were quite clearly not opponents of hate based on the tweets presented. This serves as a cautionary tale about crowdsourcing - there should always be checks in place (gold questions) to ensure high quality annotations.

As a result of these failed experiments, we ultimately defined hate speakers in ways that would allow us to further study their attributes and potentially identify them via machine learning algorithms. Following one train of thought, we simply classified users as hate speakers if they tweeted at least one hateful tweet on their timeline. We then used these binary classifications as ground truth in training and testing machine learning classifiers created to identify hate speakers in a population. In the second train of thought, we ranked users in terms of the proportion of hate found on their timeline. This not only spared us from making arbitrary decisions regarding how we should identify hate speakers, it also allowed us to study hate speakers in terms of the levels of hate speech they exhibited on Twitter, especially in relation to their demographic and other attributes.

4.4 Methods

Given a set of users and information about their demographic attributes and their timelines, is it possible to determine whether or not these users are hate speakers? To answer this question, we perform a classification task on a subset of the potential hate speakers (found in Chapter 3). We first determined the demographic and psychological attributes of the users, including race, gender, age, grade level and other measures. We then explored a number of regression and classifier-based methods using the proportion of hate found on users' timelines as a target (independent) variable and user attributes as predictor (explanatory) variables.

In the remainder of this section, we first describe methods used to collect attribute data for each of our users. We then present our classifier-based approach to identifying hate speakers online. Finally, we describe our regression-based approach to understanding the effects of our variables on the level of hate speech used.

4.4.1 Data Collection

In what follows, we outline user attributes of interest and the corresponding methodology chosen to collect this data for each user. Along with these descriptions, we present some simple demographic statistics for our set of 33,314 potential hate speakers. We outline the percentage of users for which we were able to collect demographic information, as well as a further break down of users into attribute categories where appropriate. Note that some variables were more eas-

ily obtainable than others. Variables such as proportion of hate, grade level, and psychological variables (negative emotion (negmo), positive emotion (posemo), anger, anxiety and certainty) are based solely on user text and therefore posed less difficulty in terms of data collection. For the attributes dependent upon more than just a user's collection of tweets, we outline the major obstacles encountered for each attribute using the methodologies outlined below.

Proportion of Hate We begin with the aforementioned dataset of 85.4 million tweets identified using the Hatebase lexicon, extracting a corresponding set of 33,314 thousand Twitter users. Each user's set of tweets were sent through our multi-class classifier (outlined in Chapter 3) in order to predict whether or not each tweet in a timeline was in the "hate" class, the "offensive" class, or the "not offensive" class. We then calculated the proportion of hate tweets in each user's timeline.

The frequency distribution of hate proportions across users can be found in Figure 4.1, which resembles a power-law distribution. The majority of our user timelines contain a small proportion of hate, with very few users dedicating their timelines to hateful speech. While 29,000 user timelines exhibit less than 5% hate speech (1,145 of which had no hate speech at all), only 77 timelines exhibit more than 20% hate speech. This again speaks to rarity of hate speech on Twitter first discussed in Chapter 3, both in terms of the overall of existence of hateful tweets on Twitter and the amount of hate found in any given user's timeline.

Location Identifiable locations that were provided by the users in their profile descriptions came in many different forms, including area codes, cities, states,

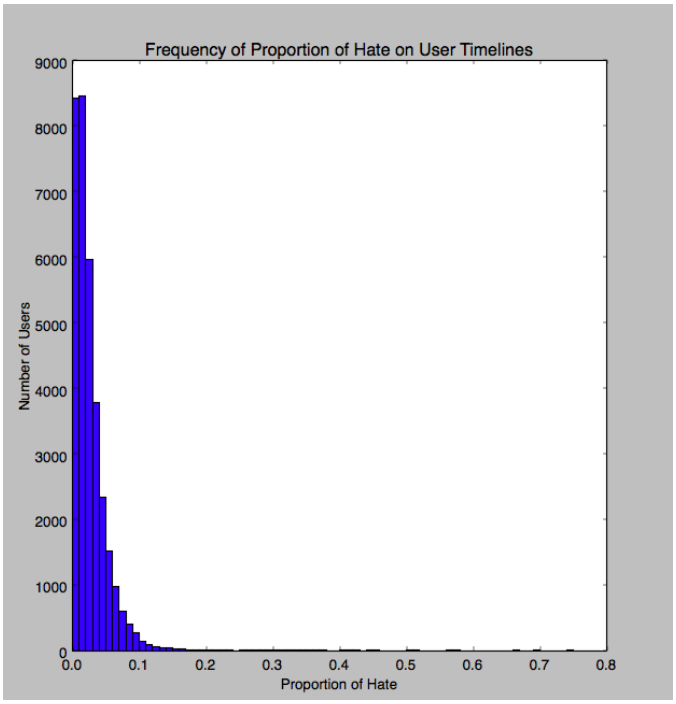


Figure 4.1: Histogram: Frequency of Proportion of Hate

and countries. These locations were compared to 1) a list of United States' area codes, 2) lists of the 50 states in the United States (in abbreviated and full form) and 3) a list of city-state combinations in which each city included was the only one with its name in the country. The third list was included in order to capture instances in which users only provided a city. By cross-referencing these locations with cities that had unique names, we attempted to minimize the error associated with assigning the wrong state to a user. A user that lives in "Springfield," for example, could possibly live in any of the 34 states with a city/town bearing that name.

We further aggregated user locations to the region-level. We annotated each user as living in the north-east, the south, the midwest, or the west. These commonly known regions are based on regions outlined by the United States Census Bureau ([12]).

As expected, geo-located data was quite sparse. Further, many users left the location portion of their profiles empty or entered an unidentifiable location. Via the combination of area codes, state lists and state-city combinations, we were able to identify location at the state-level for 25,141 users (75%). We further aggregated these users by the region of the United States in which they lived, finding that 10,481 (31%) users lived in the south, 6,023 (18%) lived in the midwest, 4,862 (15%) lived in the west, 3,775 (11%) lived in the northeast and the locations of the remaining 8,175 (25%) users were unidentifiable.

Figure 4.2 represents measure a of the hate found in each state. We first found the proportion of hate speakers in each state based on our sample of users. We then found the proportion of people living each state, as compared the the United States population. Percent differences between these values, which range from -3% to 5%, are meant to illustrate the over- or under-representation of hate speakers in each state.

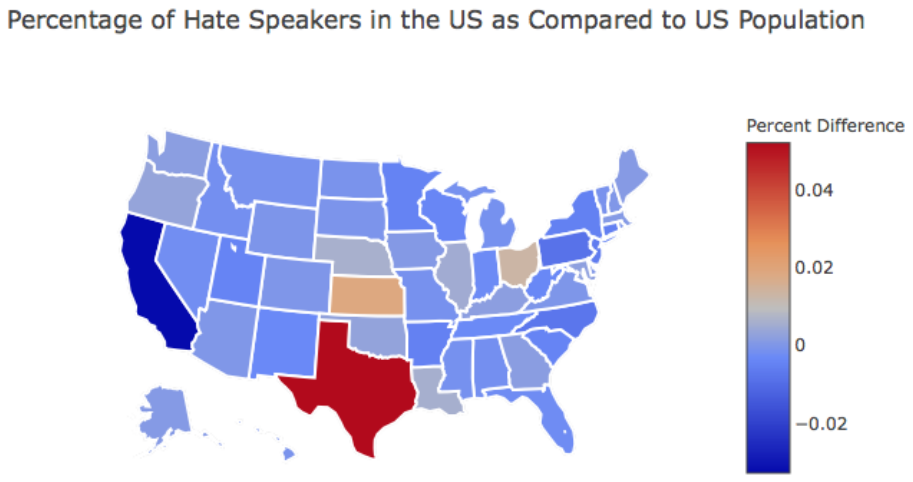


Figure 4.2: Percentage Difference Between Hate Speaker Sample and US Population

Gender We first attempted to determine gender using the FacePlusPlus application, which was only able to return information for approximately 6% of users. We then turned to Crowdflower (CF), where we had all users annotated for race and gender. CF workers were provided with a link to a user's profile page (or profile picture, if their account was deleted or suspended), and asked to determine whether or not the user was "male" or "female." They were also provided with the following excerpt of instructions: *"In this job, you will be given a link to either 1) A Twitter user's profile or 2) A Twitter user's profile picture. Using this link, we would like you to determine the gender and race of the user. In the event that you are provided a link to a user's full profile but are unable to determine race and gender from the profile picture, please use the user's name, profile description, user's pictures and/or most recent tweets to answer the questions."* Further "tips" indicated that the "Other" category should be chosen if the user did not fit into the three racial categories proposed and the "Unable to determine" category should be chosen in instances where 1) the profile represented a brand or company, 2) the profile picture was unclear or 3) the profile picture was not of the actual user *and* it was otherwise impossible to determine race from the profile name, description, tweets or other pictures.

For instances in which a user's profile was no longer available due to deletion or account suspension, we supplemented CF output by cross-referencing gender-name dictionaries with profile information (names and screen names) provided at the time of data collection. Using a ground truth set of 200 annotated users, we found that the name-dictionary method performed quite well with precision and recall scores of 90.7% and an F1-score of 91.1%. Using this combination of Crowdflower workers and gender-name dictionaries, we obtained a gender of "male" or "female" for 27,347 users (82%); 17,110 were male

(51%), 10,237 were female (31%).

Race As aforementioned, race was also determined using Crowdflower. Workers were asked to identify each user as "'white (non-hispanic)", "black (non-hispanic)", "asian (non-hispanic)", "other", or "unable to determine," the latter of which was to be chosen according to the same criteria outlined above. We were only able to determine race for a total of 19,913 (60%) users, with 12,533 (38%) labeled as "white", 6,327 (19%) labeled as "black", 381 (1%) labeled as "Asian", and 672 (2%) labeled as "other."

A major reason that we were not as successful in determining race as we were in determining gender is that the ability to identify one's race in this context usually relies solely upon access to a clear visual of the person. CF workers were often able to recognize the gender of an individual based on their name or screen name, their picture (even if unclear) and/or pronouns used in profile descriptions and most recent tweets. Without a user explicitly stating their race in textual form, it was difficult to overcome problems associated with deleted/suspended accounts and unclear or nonexistent pictures.

Education In order to assess educational attainment, we used the Flesch-Kincaid Grade Level Formula on each user's timeline. Grade level is calculated via the formula .39(ASL) + 11.8(ASW) - 15.59, where ASL refers to the average sentence length (total number of words divided by total number of sentences) and ASW refers to the average syllables per word (total number of syllables divided by total number of words). Each tweet was preprocessed by splitting them into separate sentences where appropriate (as indicated by end-of-sentence punctuation), then gathered into one large "document." Python's

Textstat was used to calculate the number of words, syllables, and sentences ([46]). Grade level was then calculated for each document, or timeline, resulting in grades from 0 - 12. Since we had up to 3,200 tweets for each user, we were able to obtain this attribute for 100% of users.

A brief analysis of the grade level distribution indicates that 27,469 (82%) of users tweeted at a grade level at or below 6th grade. The full grade frequency distribution can be seen in Figure 4.3.

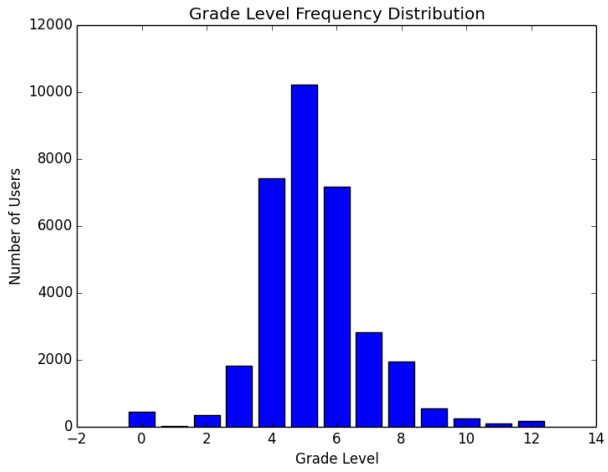


Figure 4.3: Frequency Distribution for Grade Level of Users

Psychological Variables The Linguistics Inquiry and Word Count (LIWC) lexicon ([65]) was used to gauge user affect. LIWC contains five different word lists corresponding to negative emotion (negmo), positive emotion (posemo), anger, anxiety and certainty. For each user, we gathered all of their collected tweets into one "document" and counted the number of words that appeared in each of the five lists. The frequency distribution for each of these scores can be found in Figure 4.4.

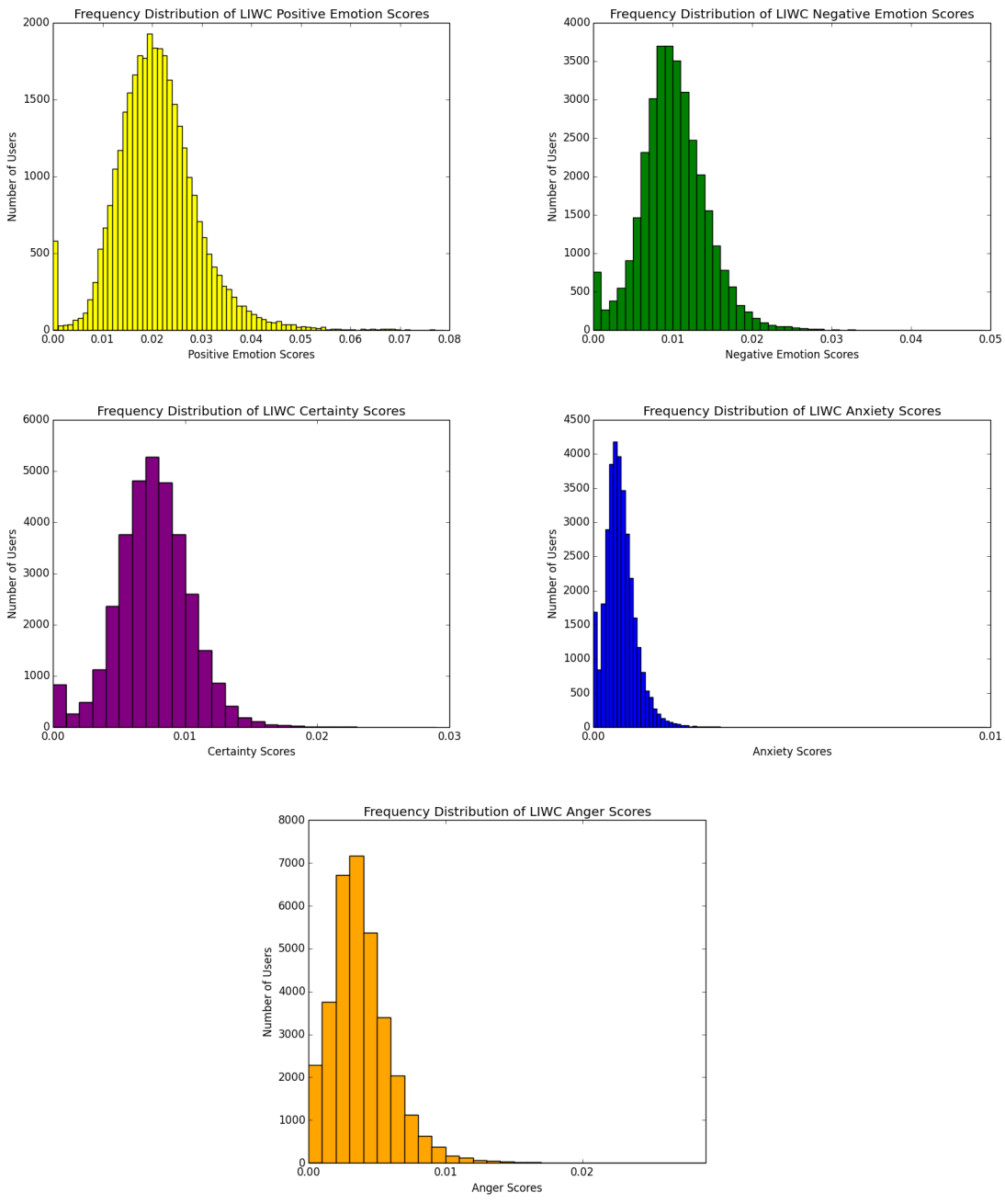


Figure 4.4: Frequency Distributions for each of the LIWC Categories

Extremist Groups Using the methodology presented by [2], we combined information from the U.S. Department of Homeland Security's (DHS) "Domestic Extremism Lexicon" (DHS 2009) with the Southern Poverty Law Center's

(SPLC) database of extremist groups in the U.S., resulting in the identification of 26 left-wing and 45 right-wing extremist groups and organizations. The lists, presented in Table 4.1, contain only the groups with active Twitter handles. We then compared each user's set of friends (accounts followed by the user) to these lists in order to determine the number of extremist groups followed. We found that 7,575 users (22%) were following at least one far right extremist group and 6,554 (20%) were following at least one far left extremist group.

Table 4.1: Extremist Groups

Far Right Groups, Ideology	**Far Left Groups, Ideology**
American Life League, anti-abortion	CrimethInc., Anarchist
Americans United for Life, anti-abortion	Animal Liberation Front, Animal Rights Extremism
National Right to Life Committee, anti-abortion	News and letters Committees, Communist
Federation for American Immigration Reform , anti-immigrant	Workers World Party, Communist
American Family Association, anti-LGBT	Communist Party USA, Communist
American Vision, anti-LGBT	Earth First!, Radical Environmentalist
Bryan Fischer, anti-LGBT	N. American Animal Liberation Press Office, Radical Environmentalist
David Barton, anti-LGBT	DeepGreenResistance, Radical Environmentalist
Dove World Outreach Center, anti-LGBT	Boston Socialism, Socialist
Family Research Council, anti-LGBT	Kshama Sawant, Socialist
Lou Engle, anti-LGBT	Freedom Socialist Party, Socialist
John "Molotov" Mitchell, anti-LGBT	International Action Center, Socialist
Americans for Truth About Homosexuality, anti-LGBT	NYC ISO, Socialist
Illinois Family Institute, anti-LGBT	Peace and Freedom Party, Socialist
Public Advocate of the United States, anti-LGBT	Progressive Labor Party, Socialist
Chalcedon Foundation, anti-LGBT	Party for Socialism and Liberation, Socialist
Faithful Word Baptist Church, anti-LGBT	Radical Women, Socialist
SaveCalifornia.com, anti-LGBT	Socialist Action, Socialist
Tony Perkins,anti-LGBT	Socialist Alternative, Socialist
Traditional Values Coalition, anti-LGBT	Socialist Worker, Socialist
United Families International, anti-LGBT	Socialist Party USA, Socialist
Westboro Baptist Church, anti-LGBT	UMass Boston ISO, Socialist
Faith Freedom International, anti-muslim	Socialist Equality Party, Socialist
Frank Gaffney, anti-muslim	Democratic Socialists of America, Socialist
Robert Spencer, anti-muslim	Liberty Union Party, Socialist
Christian Action Network, anti-muslim	
American Freedom Defense Initiative, anti-muslim	
Tea Party Nation, General Hate	
Tony Alamo Christian Ministries, General Hate	
Michael Hill, Neo-Confederate	
American Nazi Party, Neo-Nazi	
Aryan Brotherhood, Neo-Nazi	
Aryan Nation, Neo-Nazi	
David Irving, Neo-Nazi	
National Socialist Movement, Neo-Nazi	
Chuck Buldwin, Patriot Movement	
Joseph Farah, Patriot Movement	
WND news, Patriot Movement	
The Remnant/The Remnant Press, Radical traditional Catholicism	
American Renaissance, White nationalist	
American Front, White nationalist	
Nationalist Movement, White nationalist	
The Political Cesspool, White nationalist	
VDARE Foundation, White nationalist	

Popularity and Influence Here, we collected the number of friends and follow-

ers of each user. The number of followers a user has is often translated into a measure of that user's popularity. While the number of friends is not an exact proxy for influence, it is a good measure of a user's ability to spread information (in this case, hate) to large audience.

Political Affiliation A naive attempt at imputing the political affiliation of our users included cross-referencing user friends lists with Congress member accounts. This attempt, based on the idea that users follow politicians with whom they align politically, was unsuccessful since none of our users followed Congress members. Later in this paper, we suggest a more refined approach to be used in future work.

Age Age was determined for only 6% of the population using FacePlusPlus. Since this was not enough information to impute the age values of the other users, we excluded this variable from our analyses.

4.4.2 Hate Speaker Binary Classification Task

We first categorized users in our dataset as "hate speakers" if their timelines contained at least one tweet containing hate speech, as per the hate speech classifier presented in Chapter 3. This translated into binarizing the "proportion of hate" variable such that users with a proportion of hate greater than 0 were classified as hate speakers. This variable was used as the target (ground truth) variable in training and testing machine learning algorithms. We then performed One Hot Coding on the categorical variables of race, gender, state and region in order to transform them into numerical variables accepted by sklearn's machine learning algorithms [62]. This method transforms one categorical variable with

n categories into *n* binary variables, arbitrarily dropping one of the binary variables to reduce the possibility of multicollinearity. For each categorical variable, we also added an additional binary variable indicating whether or not information was absent for a user. This was done as an attempt to capture any relationship between hate speaker status and the tendency of users to purposefully hide information about themselves. The latter is evidenced by the discrepancy between information collected on race and gender. We were unable to collect the race of a large proportion of users because those users purposefully avoided posting pictures of themselves online. Users appeared to have less inhibition in exhibiting their gender in less revealing ways. Similarly, many users opted not to enter the location portion of their profile descriptions. Finally, we removed all users with missing data, leaving us with a final set of 27,375 users.

After preprocessing the variables, we split the data into training and testing sets equal to 70% and 30% of the original data, respectively. Both sets were "stratified" in that the proportion of classes in each set was representative of the proportions found in the entire datasets. The training set was used for 5-fold cross validation, in which a machine learning model was trained five separate times on five equally-sized subsamples of the training data. During each training process, the model would train on a different combination of four of the subsamples and be validated on a fifth "hold out" sample. This cross validation procedure allows for the assessment of the generalizability of a model to "out-of-sample" or unseen data. Once a model was trained, it was then applied to our testing set in order to determine how well the trained model performed on completely unseen data.

We used a grid-search approach to identify the best-performing machine

learning model for the task. We first identified Logistic Regression, Support Vector Machines, and Random Forest Classifiers as our models of interest. For each model, we chose sets of parameters that we hypothesized would effect or improve classification outcomes. For example, we separately trained Logistic Regression using L1- and L2- regularization parameters in order to determine which offered superior classification performance. We also tested each model with varying "class weights" as a means of limiting the effect of our extreme class imbalance (788 non-hate speakers, 26,589 hate speakers). In setting the class weight, we ensure that the model applies a greater penalty to the misclassification of the minority class.

Using grid-search, we iterated through each model and its corresponding set of parameters, performing cross validation on every possible combination. We singled out the best model based on both AUC (area under the Receiver Operating Characteristic (ROC) curve) and F1-score values. Once the best model was determined, we fit the model to the entire training set and validated it on the test set. Here, we used AUC, F1-score and Matthew's Correlation Coefficient (MCC) values as measures of a classifier's "out-of-sample" performance. AUC and MCC were chosen over more common measures (such as the *accuracy* measure) because they are not as easily influence by class imbalance. *Accuracy* will often be high for classifiers that perform well on the majority class but poorly on the minority class (effectively classifying most instances as the majority class), while AUC and MCC will measure the classifier's ability to distinguish between the classes. Although the F1-score suffers from similar problems as accuracy, inspection of cross validation results revealed that F1-score was sensitive enough to performance on the minority class that it was worth including in the analyses. The definitions for each of these measures are presented below:

F1-score: The F1-score is a commonly used measure of the overall accuracy of a classifier. The F1-score is the harmonic mean of precision and recall. Their formulas are:

$$prec = \frac{|true\ positives|}{|true\ positives + false\ positives|}$$

$$rec = \frac{|true\ positives|}{|true\ positives + false\ negatives|}$$

$$F1 = \frac{2 * prec * rec}{prec + rec}$$

AUC: The ROC curve is created by plotting the false positive rate against the true positive rate (recall). It is often used in choosing the optimal model based on trade-offs between true and false positives. The Area Under the Curve (AUC) measures the overall performance of the classifier, with a value of 1 indicating that a classifier has perfect performance.

Matthew's Correlation Coefficient: The MCC measure represents the correlation between predicted and true outcomes. The measure ranges from -1 to 1, with 0 indicating performance that is only as good as random classification and 1 indicating perfect classifier performance. Unlike precision and recall, which heavily focus on the positive class, MCC takes true negatives into account when measuring the quality of categorizations. Its formula is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP = true positives, FP = false positives, TN = true negatives and FN = false negatives.

4.4.3 Hate Speaker Regression Task

Using proportion of hate as a means of ranking hate speakers, we created a generalized linear regression model (Ordinary Least Squares (OLS)) and a Random Forest Regression model in order to see if our variables helped to explain and predict the levels of hate speech used. Linear regression (OLS) was carried out using python's statsmodels, due to the in depth results summary the package offers ([71]). Alternatively, Random Forest Regression was performed using scikit-learn ([62]), as this was not an option for statsmodels.

With the exception of the dependent variable (proportion of hate speech), all predictor variables were preprocessed as in the same manner as done for the binary classifier. Hence, we performed regression analyses on the same 27,375 users. As before, we split the dataset into 70% training data and 30% test data and performed a grid-search method, combined with cross validation, to identify the best model. As is traditionally done in regression analyses, we used mean squared error (MSE) and R-squared values to choose models and assess performance.

R-squared: R-squared, otherwise denoted as the "coefficient of determination," represents the percentage of variance in the dependent variable that is explained by the model ($\frac{variation\ explained\ by\ model}{total\ variation}$). It is calculated via the following formula:

$$R^2 = \sum_i (f_i - \bar{y})^2 \sum_i (y_i - \bar{y})^2,$$

where y is the observed data, $\bar{y}$ is the mean of observed data, and f is the vector of predicted values.

MSE: MSE assesses prediction quality by measuring the average of the sum of squared errors. It is calculated as:

$$MSE = (\frac{1}{n} \sum_{i=1}^{n})(\hat{Y}_i - Y_i)^2),$$

where $\hat{Y}$ is a vector of predicted values and Y is the corresponding vector of observed values.

4.5 Results

4.5.1 Hate Speaker Binary Classification Results

Unlike our experience with classifying tweets, we found that the logistic regression and Support Vector Machines (SVM) performed poorly in comparison to the Random Forest Classifier. This makes intuitive sense, since the Random Forest Classifier (RF) is an ensemble model. RF first selects a subset of the given features, or predictor variables. In creating a decision tree, it then determines which feature each node should be split on based on error reduction or information gain, until no more gain can be obtained. By creating multiple decision trees using different subsets of the training data and averaging the results over them, it effectively decreases the influence of noise or outlier results.

The tuned parameters of the best performing Random Forest Classifier are as follows: 1) a maximum tree depth of 3, used to prevent classifier overfitting, 2) a class weight of 15, placed on the minority class of non-hate speakers in order increase penalization on its misclassification, 3) 50 estimators (decision trees), and 4) the "entropy" criterion, which calculates the homogeneity of a sample. If the entropy of a node is equal to 0, the node is homogenous and does not need to be split. Otherwise, the RF algorithm splits the node as long as the maximum tree depth has not been reached.

The top 10 most important features in the creation of decision trees can be found in Figure 4.5. The values of importance of all features sum to 1, and so the results are relative to each other. The most important features include all five of the psychological traits, social network attributes such as number of friends and number of followers, the education level of users, and whether or not the user provided their location (NaN = missing values). The importance of the psychological variables in detecting hate speakers illustrates the importance of including variables gleaned from textual analysis along with other non-textual user attributes. The feature importance values differ from the coefficients offered by a regression analysis, and so we are not able to ascertain exactly how each feature varies with our target variable.

The chosen RF model received an overall precision of .818333, recall of 0.836260, and F1-score of 0.827032. The model has an AUC value of 83.6% and the MCC score is 0.654349, indicating that the classifier does 65% better than random classification would. As expected, the precision, recall and F1-scores for the majority class of hate speakers are all very high at a value of .99 each. The classifier did not perform as well on the minority class, with a precision

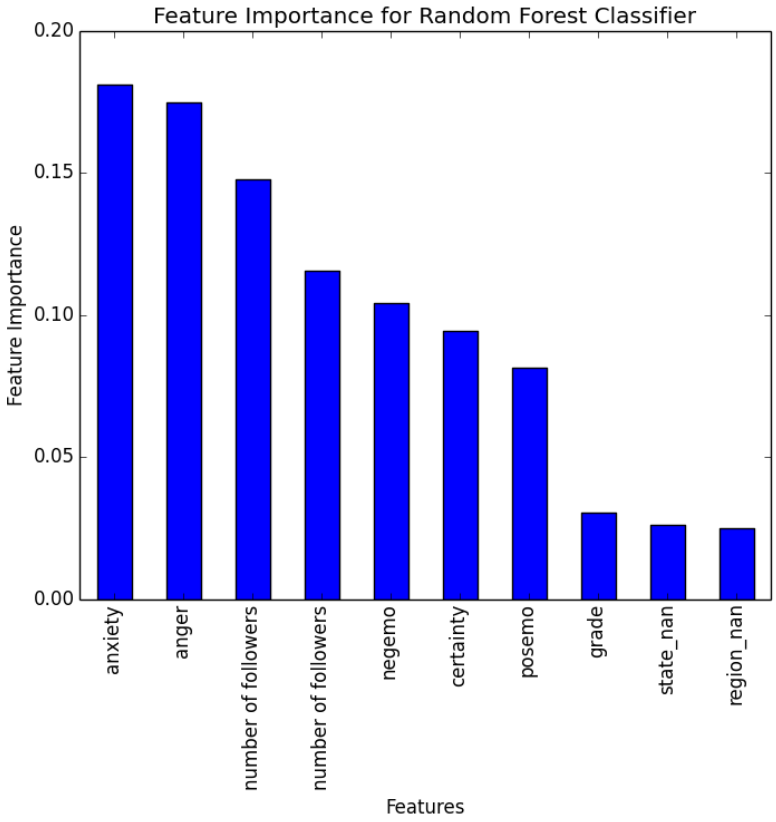


Figure 4.5: Feature Importance Scores for Binary Hate Speech Classification

of .65, a recall of .68 and overall F1-score of .66. Figure 4.6 further illustrates the classifier performance for each class; RF misclassifies approximately 32% of non-hate speakers as hate speakers but only misclassifies 1% of hate speakers as non-hate speakers. Using feature importance information, a deeper look into instances in which true hate speakers were misclassified as non-hate speakers revealed that these hate speakers often exhibited low values of anxiety, anger and negative emotion, while also exhibiting average to high values of positive emotion. Alternatively, non-hate speakers that were misclassified as hate speakers often exhibited higher negative emotions as compared to the sample mean.

Echoing one of the major motivations for creating a hate speech text classifier

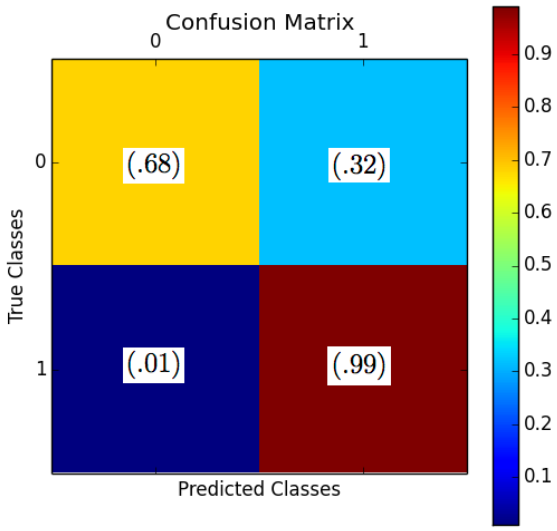


Figure 4.6: Confusion Matrix for the Best Performing Random Forest Classifier

(Chapter 3) that distinguishes between offensive and truly hateful language, an ideal classifier would capture all hate speakers while avoiding labeling innocent users with such a heavy and impactful label. In this vein, we hope to improve the classifier by incorporating additional predictor variables that help distinguish between the two classes. Still, the true success of a classifier is dependent upon the purpose of the model. This classifier is extremely successful in correctly identifying 99% of true hate speakers, and would be useful for those who do not mind dealing with some level of false positives (here, 32%). This might be especially useful, for instance, for online entities that combine machine learning methods with human monitoring in detecting hate speech. By identifying misclassification patterns, humans would be able to hone in on users that exhibit those patterns in order to determine whether or not they are truly hate speakers.

4.5.2 Hate Speaker Regression Results

The Random Forest approach proved a superior method once again, with an R-squared value exceeding that of linear regression by 8% and an MSE of 5.21 - 3% smaller than linear regression. Despite its superior performance over linear regression, RF's R-squared value of 31% meant that our variables were not as descriptive in predicting hate speech proportions as we hypothesized. Additional explanatory variables are necessary to describe the variation found in the data.

The performance enhancing parameters for the regression model differed from that of the binary classification model. Regression parameters included 1) no maximum tree depth, 2) a class weight of 15. placed on the minority class, 3) 1000 estimators (decision trees), and 4) the "MSE" criterion. Here, the decision to split a node was one of error reduction instead of information gain. Figure 4.7 exhibits the top 11 most important features of this regression, most of which mirror those of the binary classifier. They differ in that location variables were replaced by variables related to race and gender. In describing levels of hate speech, then, it appears that demographic attributes race and gender are more predictive than location.

4.6 Discussion

In this chapter, we engaged in tasks related to understanding and identifying individual hate speakers - an area of research that has been largely overlooked in the past. We identified various hate speaker attributes in an effort to understand

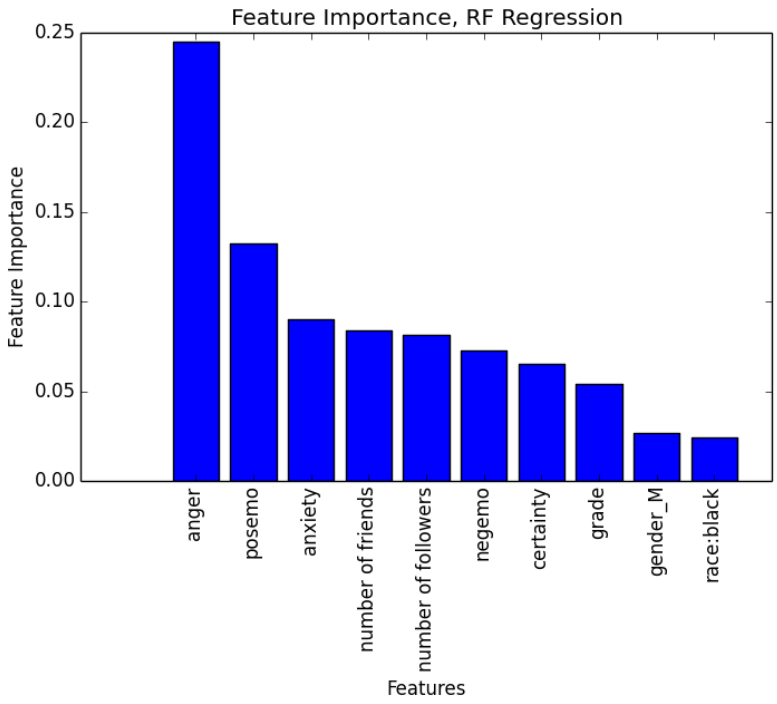


Figure 4.7: Feature Importance Scores for Random Forest Regressor

how they could be used to both classify hate speakers and describe/predict their level of hate speech use. In this vein, we created a binary Random Forest classification model that performed reasonably well in distinguishing between hate and non-hate speakers, with overall classification qualities of greater than 80% on each measure (MCC, AUC, F1-score). We then performed Random Forest Regression, which resulted in a fairly low R-squared value of .32. This low R-squared value indicates that the variables of interest in this paper are more successful in predicting the existence of hate speech than the levels of hate speech exhibited, but that adding additional explanatory variables could improve upon our efforts in both tasks.

This work is not without limitations. First, our analysis is only as good as the classifier (Chapter 3) we used to identify the collection of hate tweets used

to calculate our target variable - proportion of hate speech on each user's timeline. Since our hate speech classifier tends to classify more subtle forms of hate speech as merely offensive, it is quite possible that some of our non-hate speakers have actually exhibited hate speech that was not identified. Still, our classifier performs strongly in identifying the most overt forms of hate speech, and we are confident that we captured the activities of the more extreme hate users. Second, our entire set of users was initially collected using the Hatebase lexicon, which contains terms used in hate speech incidents globally. Future work might benefit from analyses on a truly random set of users. Finally, there is a margin of error associated with the data collection of user attributes. Race, for instance, can be extremely difficult for users to identify - especially when it differs from their own. Despite this, human annotation remains the most effective way to identify demographic attributes and is the ground truth on which many classifiers are built.

Future work aims to improve upon both the Random Forest Classifier and Regressor in an effort to better predict whether or not a speaker exhibits hate speech and the level of hate speech usage. The results of the RF Regressor, in particular, served as a signal that our binary Random Forest Classifier might benefit from the inclusion of additional predictor variables. In this vein, we will attempt to include the excluded variables of age and political affiliation (using the method presented by [72]), as well as identify other variables that would improve performance in both tasks. Supported by the cyberbullying literature, we suspect that including personality traits ("The Big Five") will increase prediction power in both models ([67]). Further, given the importance of the number of friends and followers as features, future classifiers should also take social network information into account. We hypothesize that measures of clustering,

centrality, embededness of hate speakers and relationships between hate speakers will prove vitally important in revealing more information about individual hate speakers, much like they have in the detection of cyberbullying ([45]).

CHAPTER 5

IDENTIFYING POLARIZED GROUPS IN ONLINE SOCIAL MEDIA

5.1 Introduction

[1] Polarization, which we define as the division of a population into two or more groups identified by opposing ideas, values, interests or goals, has become of increasing interest in recent years. With the emergence of the web and popular networking sites like Tumblr, Twitter, Facebook and more, individuals are able to express their opinions to a much larger audience than ever before. Controversial subjects are often widely debated in these public forums, resulting in massive amounts of network and opinion data for researchers to delve into.

With access to such a large amount of social media data, often in real time, a large segment of opinion-related research has been geared to identifying polarized groups ([17], [3]) and understanding how polarization develops ([21],[42]), how to quantify it ([40], [35], [59]), and how to mitigate it ([36]). Polarization research has applications in everything from politics (election prediction, political campaigning, reducing political echo chambers online) to business (product development, consumer response, advertisement). In this paper, we aim to contribute to this area by presenting a nonnegative matrix factorization-based approach to identifying polarity groups that incorporates both content and social network information. Our work using nonnegative matrix factorization (NMF) is heavily inspired by the need for an approach that works well in multiple contexts: ones varying in the nature of the controversy (for example, politics vs.

[1] This work was co-authored with Jiejun Xu and Tsai-Ching Lu during an internship with HRL Laboratories.

sports), the level of polarization, the number of polarity groups involved, and the presence of neutral entities. NMF's ability to uncover latent network properties, coupled with the ease of interpretation of its nonnegative output, makes it an ideal approach for studying these varying types of networks. Further, the approach can be tailored to the amount of information that is available for a given network. Though NMF is unsupervised in its normal execution (a major benefit since annotated data is not always available or obtainable), ground truth data can easily be incorporated into the process.

We build upon previous work in the area by applying NMF to a tripartite graph *user-post-tag*, where a tag serves to annotate post content. Our NMF procedure is outlined as follows: I) Preprocessing: Using our tripartite graph, we construct bipartite graphs *user-post*, *user-tag*, and *post-tag*. We also construct a reblog network R in order to capture user relations. II) Optimization: We then perform NMF on each of the bipartite graphs via a multiplicative update algorithm, separately clustering *users and posts*, *users and tags*, and *posts and tags*. We minimize an objective function containing the cost functions associated with each of the three NMF decompositions, as well as a regularization term that exploits the user information found in the reblog network. Using this objective function, our approach simultaneously and iteratively clusters the set of users, the set of posts and the set of tags, allowing for the clustering of one set to inform the clustering of the others. We find that this method outperforms state-of-the-art approaches to identifying polarized groups online, including the application of NMF to the bipartite graph *user-post*, spectral co-clustering on the bipartite graph *user-post* and popular community detection methods applied to the bipartite graph *user-post*. We confirm these results by applying this method to real-world Tumblr datasets, making our work the first to conduct polarization analyses using the

Tumblr platform.

5.2 Related Work

A number of recent polarization studies have confirmed the efficacy of analyzing social networks in identifying polarity at the user and group levels. An early approach involved using clustering measures, such as modularity, to identify communities within social networks. Approaches like this are limited in the context of polarization because the mere existence of communities does not in itself indicate polarization [40]. A more recent approach has been to collect social media posts referencing a polarizing topic, extract a conversation or interaction network from those posts, and subsequently apply traditional community detection or clustering methods in order to identify polarized groups. Interaction networks (ex: retweet networks) tend to be more useful than social networks (ex: follower networks) because the action of retweeting is a strong indication that the user is interested in the content being shared. [17] found that using a label propagation method on retweet graphs of political content uncovered the network's highly partisan structure, effectively separating Twitter users into right and left-wing clusters. In testing multiple network polarization quantification measures, [35] first used graph partitioning software METIS on retweet graphs to separate networks into polarized groups and later verified that these groups corresponded to the polarized groups they expected.

Though they perform well, community detection approaches have some limitations. First, their performance is often confounded by the existence of neutral users and content ([3]). Neutral users may share content from multiple polar-

ity groups, just as polarized users may share neutral content - actions that are not explicity captured in a retweet network. Further, these methods only take post-related information into account during the stage of data collection by collecting posts containing relevant keywords or hashtags. Given a lack of information beyond social network information, community detection approaches often mislabel users.

An appealing alternative that allows researchers to deal with neutral networks and incorporate post-related information (even without analyzing the actual text of a post) is nonnegative matrix factorization (NMF). NMF has been extensively used in recommendation systems research ([95]; [57]), and to a lesser extent sentiment analysis ([96]) and community detection ([63]). To the best of our knowledge, [3] is the only paper that has applied a NMF-based approach in polarization research. They perform NMF on a source-assertion (user-post) bipartite network to separately cluster sources and assertions into polarity groups, using a social dependency network as a means of regularization. They show that NMF can be more effective in identifying polarized groups than community detection approaches, largely due to NMF's ability to uncover latent relationships in network data.

Building upon this work, we apply an NMF-based approach to a tripartite network that allows us to include even more information about social media users, their relationships, and the content they post. We incorporate the relationships between *posts* and *tags* into the initial *user-post* framework to obtain the tripartite graph *user-post-tag*. Tags, often used to annotate posts, can be a useful source of information in the absence of textual content (for example, a post of a photo, gif or video) or when text analysis in infeasible. Tags are not

only an indicator of post content, but may even express the sentiment ([22]) or point-of-view of the post [90]. As such, understanding how posts are annotated can be beneficial in clustering both posts and users.

5.3 A Nonnegative Matrix Factorization Approach

In this section, we present a nonnegative matrix factorization (NMF) algorithm for identifying polarity groups in social media networks. The traditional NMF problem requires decomposing a matrix (A) into two nonnegative matrices (UV^T), minimizing the error associated with the decomposition. This translates into minimizing an error function similar to the one we choose in this paper - the square of the Euclidean distance between the original matrix and its lower rank approximation $\min_{U,V} \|A - UV^T\|_F^2$, where $\|.\|_F$ is the Frobenius norm. Regularization terms are often added to the objective function to ensure that the results most closely represent the data it approximates. In what follows, we extend this method to a tripartite graph.

5.3.1 NMF on a Tripartite Graph

The identification of polarity groups in social media networks like Tumblr and Twitter is easily characterized as a problem of co-clustering over the tripartite graph *user-post-tag*. Notation related to this problem is located in Table 5.1. We can separate the tripartite graph into three informative bipartite graphs, each with a binary adjacency matrix: user-post (A_{up}), post-tag (A_{pt}) and user-tag (A_{ut}). Matrix $A_{up}(i, j) = 1$ if *user i* shared *post j*, and 0 otherwise. Similarly, $A_{pt}(i, j) = 1$

Table 5.1: Nonnegative Matrix Factorization Algorithm Notation

Notation	Description	Size
A_{up}	user × post matrix	m × n
A_{ut}	user × tag matrix	m × p
A_{pt}	post × tag matrix	n × p
R	reblog network adjacency matrix	m × m
C	co-reblog network adjacency matrix	m × m
T	tag co-occurence (similarity) matrix	p × p
U	user × polarity group	m × k
V	post × polarity group	n × k
W	tag × polarity group	p × k
H_1	association matrix	k × k
H_2	association matrix	k × k
H_3	association matrix	k × k
L_R	Laplacian matrix of R	m × m
L_C	Laplacian matrix of C	m × m
L_T	Laplacian matrix of T	p × p
D_R	Degree matrix of R	m × m
D_C	Degree matrix of C	m × m
D_T	Degree matrix of T	p × p
α	reblog regularization parameter	
β	user similarity regularization parameter	
ρ	tag similarity regularization parameter	

if *post i* is annotated with *tag j* by any user, and $A_{ut}(i, j) = 1$ if *user i* annotated at least one of their posts with *tag j*. The introduction of the latter two matrices into the NMF process is important in clustering both posts and users: two posts using the same tags are likely to be similar in content, two users using the same tags are likely to be sharing similar content. It is clear that performing NMF on each of these adjacency matrices allows us to uncover latent relationships between its rows and columns. Simultaneously performing NMF on each of the bipartite graphs, then, allows us to inform the clustering of one bipartite graph using the intermediate clustering results of another. This gives the following optimization problem:

$$\min_{U,V,W,H_1,H_2} ||A_{up} - UH_1V^T||_F^2 + ||A_{ut} - UH_2W^T||_F^2 + ||A_{pt} - VH_3W^T||_F^2, \text{ s.t. } U^TU = I, V^TV = I, W^TW = I$$

Note that we include orthogonality constraints on U, V, W in order to ensure a better clustering of the rows and columns of our adjacency matrices, resulting in the inclusion of association matrices H_1, H_2 and H_3 ([27]).

5.3.2 Regularization Terms

Regularization terms are added to the optimization problem as a way of ensuring that resulting matrices U, V and W are not simply solutions to the nonnegative matrix factorization problem, but that they reflect our network to the greatest possible extent. In this vein, we introduce the *reblog network (R)*, *co-reblog network (C)* and *tag similarity network (T)*.

Reblog Network (R) Reblog network R captures interactions between users, with $R(i, j) = r$ if *user i* reblogs *user j* exactly r times, 0 otherwise. We assume that users from different polarity groups are not likely to reblog each other extensively, and impose the restriction that *user i* must reblog *user j* more than once ($r > 1$) to avoid capturing instances in which users in different polarity groups reblog each other in disagreeance. Given R, we introduce the regularization term $R(i, j)\|u_i - u_j\|_F^2 = \frac{1}{2}\sum_i \sum_j R(i, j)\|u_i - u_j\|_F^2 = \|U^T L_R U\|_F^2$, where u_i and u_j represent rows in the user cluster (polarity group) matrix U. The term applies a penalty if *user i* reblogs *user j* but they have not been placed in the same polarity group.

Previous work in online polarization acknowledges the importance of analyzing social network structure to uncover polarization patterns. This is because

social networks often exhibit homophily - the tendency of users to connect to and interact with those that are similar to them in terms of values, interests and other characteristics. [1], for instance, found that political blogs more often link to other blogs of the same political orientation. Homophily tends to be even more prevalent when considering actions of endorsement, as in retweet (Twitter) and reblog (Tumblr) networks. [17] found that running a community detection algorithm on Twitters retweet graph performed well in identifying polarized groups because users tend to share or rebroadcast information they agree with. Further, [35] found that graph partitioning on the retweet graph provided more information about the polarized nature of a Twitter network than the follow or content graphs alone. Graphs with links representing actions of endorsement, therefore, can be extremely useful in determining user similarity.

Co-reblog Network (C) Similar to the reblog network, the co-reblog network (C) is also meant to function as a measure of user similarity. The purpose of the co-reblog network is to allow us to uncover latent relationships between users. While two users may not directly interact with each other in the form of *likes*, *follows* or *reblogs*, they may still be similar in terms of the content they share [32]. We assume that if two users are reblogged by a large number of the same users there is an increased likelihood they are sharing similar content, and therefore belong to the same polarity group. Formally, we define the coreblog network C as a symmetric adjacency matrix in which $C(i, j) = c$ if *user i* and *user j* have been reblogged by c of the same users, 0 otherwise. Ultimately, we impose a penalty $C(i, j)\|u_i - u_j\|_F^2 = \|U^T L_C U\|_F^2$ if *user i* and *user j* have been reblogged by the same users, but are not placed in the same polarity group.

Tag Co-occurrence Matrix (T) When little textual content is available for

analysis, it is important to leverage the information that is acessible. In this vein, we exploit not only user similarity in the clustering process, but also tag similarity. Since the co-occurrence of Twitter hashtags has previously been used as a measure of similarity in discovering topics of discussion, we include it here with hopes it will aid NMF in tag clustering. Formally, $T(i, j) = t$ if *tag i* and *tag j* occur on exactly t posts together. Should two tags co-occur quite often but are not of the same polarity group, we impose the penalty $T(i, j)\|w_i - w_j\|_F^2 = \|W^T L_T W\|_F^2$.

5.3.3 NMF using Multiplicative Update Rules

Given the aforementioned regularization terms, we present our complete optimization problem:

$\min_{U,V,W,H_1,H_2} \|A_{up} - UH_1V^T\|_F^2 + \|A_{ut} - UH_2W^T\|_F^2 + \|A_{pt} - VH_3W^T\|_F^2 + \alpha tr(U^T L_R U) + \beta tr(U^T L_C U) + \rho tr(W^T L_T W)$, s.t. $U^T U = I, V^T V = I, W^T W = I$

We solve this optimization problem using a multiplicative update algorithm initially outlined by Lee and Seung [56], dictated by rules later derived by Ding et al. [27]. The algorithm and multiplicative update rules are presented in Table 5.2, with an example derivation offered later in this section.

The multiplicative update rule for U is derived as follows:

The update of U is dependent upon only certain terms of the objective function: $J = \min_{U,V,W,H_1,H_2} \|A_{up} - UH_1V^T\|_F^2 + \|A_{ut} - UH_2W^T\|_F^2 + tr(U^T L_R U) + \beta tr(U^T L_C U)$, such that $U^T U = I$.

A Lagrangian multiplier is included to enforce the constraint $U^T U = I$ and we $\|A\|_F^2 = trace(A^T A)$ to obtain the Lagrangian function L:

$L = tr(A_{up}^T A_{up} - 2V^T A_{up}^T UH_1 + U^T UH_1V^T VH_1^T) + tr(A_{ut}^T A_{ut} - 2W^T A_{ut}^T UH_2 +$

Table 5.2: Nonnegative Matrix Factorization Algorithm

function [U,V,W,H1,H2,H3] = NMF(Aup,Aut,Apt,k,alpha,beta,rho)
Initialize U, V, W, H_1, H_2, H_3 randomly
while not converge:
update U
$U \leftarrow U. * \frac{A_{up}VH_1^T + A_{ut}WH_2^T + \alpha RU + \beta CU}{UU^T A_{up}VH_1^T + UU^T A_{ut}WH_2^T + \alpha D_R U + \beta D_C U + U\lambda_U}$
update H_1
$H_1 \leftarrow H_1. * \frac{U^T A_{up}V}{U^T UH_1^T V^T V}$
update V
$V \leftarrow V. * \frac{A_{up}^T UH_1 + A_{pt}WH_3^T}{VV^T A_{up}^T UH_1 + VV^T A_{pt}WH_3^T}$
update H_2
$H_2 \leftarrow H_2. * \frac{U^T A_{ut}W}{U^T UH_2 W^T W}$
update W
$W \leftarrow W. * \frac{A_{ut}^T UH_2 + A_{pt}^T VH_3 + \rho TW}{WW^T A_{ut}^T UH_2 + WW^T A_{pt}^T VH_3 + \rho D_T W + W\lambda_W}$
update H_3
$H_3 \leftarrow H_3. * \frac{V^T A_{pt}W}{V^T VH_3 W^T W}$

$U^T UH_2 W^T WH_2^T) + \alpha tr(U^T L_R U) + \beta tr(U^T L_C U) + tr(\lambda_U(U^T U - I))$

To find the minimum with respect to U, we take the partial derivative and set $\frac{\partial L}{\partial U} = 0$:

$\frac{\partial L}{\partial U} = -2H_1 V^T A_{up}^T + 2UH_1 V^T VH_1^T - 2H_2 W^T A_{ut}^T + 2UH_2 W^T WH_2^T + 2\alpha L_R U + 2\beta L_C U + 2U\lambda_U$

By KKT Complementarity Conditions,

$(-2H_1 V^T A_{up}^T + 2UH_1 V^T VH_1^T - 2H_2 W^T A_{ut}^T + 2UH_2 W^T WH_2^T + 2\alpha L_R U + 2\beta L_C U + 2U\lambda)_{ij} U_{ij} = 0.$

Finally, we have $\lambda_U = U^T A_{up}VH_1^T - H_1 V^T VH_1^T + U^T A_{ut}WH_2^T - H_2 W^T WH_2^T - \alpha U^T L_R U - \beta U^T L_C U$.

5.4 Data and Methodology

5.4.1 Tumblr

Tumblr is one of the most popular online sites today, ranked the 17th most popular site in the United States, the 46th most popular site globally and the second largest microblogging service available to internet users. The site offers combined aspects of a blogging site and a social network, in many ways distinguishing it from other popular networking sites such as Twitter and Facebook. Each user owns a blog to which they are able to post unlimited text, photos, videos, links, quotes, and audio files. Tumblr offers the supplementary option of adding *tags* to posts, allowing users to annotate their posts with succinct references to its content or intended audience. Via one's blog, users are able to connect to and interact with others by following others' blogs and liking, commenting on and/or reblogging others' posts. Tumblr actions such as *following, liking, commenting* and *reblogging* all result in the formation of social networks. While following patterns indicate general interest in overall blog content, liking and reblogging posts are often considered to be actions of endorsement.

To date, Tumblr data has not been extensively used for research purposes. Network-related and polarization research, in particular, has been dominated by the use of Twitter datasets. We believe Tumblr's vast user base and unique features (unlimited post length, variety of posts allowed) make it an appealing platform to study human behavior, and hope to contribute to research on polarization by analyzing seldomly used data from the Tumblr platform. Further, in leveraging Tumblr's fundamental differences from Twitter, we create an algorithm that works well for both types of data sets. Unlike Twitter, Tumblr allows

its users to post photos, videos, and other non-text content. As such, Tumblr posts are dominated by photos, making it difficult for researchers to use algorithms heavily dependent upon a post's textual content ([93]). An algorithm that is successful in identifying polarization without the examination of textual content can more straightforwardly be applied to a wide variety of datasets, generally requires less human effort, is less computationally intensive, and can easily be updated to incorporate text content if it is available.

5.4.2 Data

We use data collected via the Tumblr Firehose API (100%). We collected Tumblr posts surrounding three separate controversial or polarized topics, listed below. In order to prevent bias, terms used to collect the data were informational or "neutral" in nature and did not favor any polarity group.

2014 FIFA World Cup The World Cup is one of the most prestigious football competitions, occurring every four years. The tournament involves 32 teams globally, though our dataset starts just before the beginning of the semi-finals (7/6 - 7/13). Posts related to FIFA were collected by searching post content and their corresponding tags for the terms "fifa", "fifa 2014", "fifa world cup", "world cup", "world cup 2014" and "wc 2014." The FIFA dataset was used in two different experiments: 1) To identify polarized groups supporting 4 different teams during the semi-finals and 2) To identify polarized groups in support of each of the two teams in the final match.

World Series The World Series is an annual American baseball competition. The tournament involves 2 teams playing for the best of 7 games. We collected

data from 10/21/14 - 10/29/14 using the term "world series."

Gamergate Though Gamergate was intended to be a movement against corrupt gaming journalism, the term came to represent the controversy surrounding use of the tag to conduct a harassment campaign against female gamers. Harassers engaged in hate speech, cyberbullying and doxxing (sharing a user's personal information publicly), mostly directed toward women gamers and other opponents of the movement. In order to capture online discussion about the event and subsequent protests, we collected data from 8/27/2014 - 9/05/2014 using the term "gamergate."

Topics were deliberately chosen so that they would vary in the nature of the controversy and the number of polarized groups involved. Discussion around politicized topics tend to center around two major groups (liberal, conservative). Similarly, individuals are usually "for" or "against" movements/ protests and the event that sparked them. Sports events like FIFA differ in that the number of polarized groups discussing the event will often depend on the number of teams involved in the tournament. Statistics related to each dataset can be found in Table 5.3.

Table 5.3: Polarization Dataset Statistics

	# users	# posts	# tags	# polarity groups
Gamergate	9,799	2,718	1,974	2
World Series (2 teams)	2,353	1,189	1,753	2
FIFA (2 teams)	29,328	6,806	14,796	2
FIFA (4 teams)	981,842	27,105	32,748	4

5.4.3 Methodology

For each dataset described above, we performed 50 runs of the following experiments: 1) NMF on the bipartite *user-post* and tripartite *user-post-tag* graphs, each without regularization, 2) NMF on the bipartite *user-post* and tripartite *user-post-tag* graphs using the reblog network-based regularization term with arbitrarily chosen parameter $\alpha = 0.3$, 3) K-means community detection on bipartite graph *user-post* and 4) Spectral co-clustering on bipartite graph *user-post*. We included spectral co-clustering as a baseline algorithm because it has been successfully applied to many of the same applications (including bipartite document-term clustering) as nonnegative matrix factorization ([22]).

While testing the performance of the NMF algorithm, we varied the set of tags included in the tripartite graph based on the number of times they occurred in the dataset. Intuitively, a tag would need to appear on a number of posts in order to improve the clustering of those posts. We found that we began to obtain better results when we included only the tags that appeared in a dataset at least 10 times. Removing too many tags, however, would worsen results. We hypothesize that removing too many of the lesser-used tags that differentiate polarized groups while retaining more informational tags that are commonly used across polarity groups makes it more difficult for NMF to uncover polarization patterns. The results reported in this paper correspond to tripartite graphs in which each tag included was used on more than 20 occasions.

In order to assess the validity of our model, we had a set of posts from each dataset annotated by humans. For the Gamergate dataset, we asked annotators to categorize each post as "for", "against", or "neutral." For the World Series and FIFA datasets, we asked that they identify which team a post supported. If

the post did not support a team or the team it supported was no longer playing in the tournament, we asked that they categorize those posts as "neutral." We used these annotations to form ground truth clusters to which we compared our results.

In what follows, we outline the four measurements (Accuracy , F1-score, Adjusted Rand Index, area under the ROC curve) used to evaluate and compare classifier performance. Note that the Accuracy, F1-score and Adjusted Rand Index values were found only after calculating the percentage of times a post was classified in each polarity group (out of 50 runs) and assigning it to the polarity group that occurred most frequently.

Accuracy Accuracy gives the percentage of all predictions that were correctly classified. It is calculated as:

$$acc = \frac{|true\ positives|+|true\ negatives|}{total\ predictions}$$

F1-score The F1-score is a commonly used measure of the overall accuracy of a classifier. The F1-score is the harmonic mean of precision and recall. Their formulas are:

$$prec = \frac{|true\ positives|}{|true\ positives+false\ positives|}$$

$$rec = \frac{|true\ positives|}{|true\ positives+false\ negatives|}$$

$$F1 = \frac{2*prec*rec}{prec+rec}$$

Adjusted Rand Index The Adjusted Rand Index (ARI) measures the similarity of two different clusterings of a network. The measure is often used to determine how close the clustering results of a classifier are to the actual clusters. The formula is

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}},$$

where i refers to a cluster in the first clustering, j refers to a cluster in the second clustering, n_{ij} = number of instances clusters i and j share, a_i = number of instances in cluster i and b_j = number of instances in cluster j.

ROC Curves and AUC For each post j in datasets with only two polarity groups, we calculated the percentage of times j was classified in one polarity group ($P1_j$) vs. the other ($P2_j$). We then sorted values $P1_j - P2_j$ in the order of highest likelihood of being in polarity group $P1$ ("for"/team 1) to the highest likelihood of being in $P2$ ("against"/team 2). Going through this list in sorted order and comparing to our ground truth labels, we calculated the true and false positive rates by counting each "for" as a true positive and each "against" as a false positive. For datasets with more than two polarity groups, we took a one-vs-rest approach in letting $P1$ correspond to one polarity group, and $P2$ correspond to the combination of all other polarity groups. This resulted in a ROC curve for each polarity group. We then calculated the Area Under the Curve (AUC) values.

5.5 Results

To assess the baseline performance of NMF performed on a tripartite graph with that of a bipartite graph, we first ran the algorithm without regularization. We present the ROC curves, AUC, Accuracy, ARI, and F1-Score values in section 5.1. For each dataset, we found that our tripartite approach outperforms the bipartite approach on each measure. We then incorporated the reblog network regularization term with a regularization parameter of $\alpha = .3$, arbitrarily chosen. These results are presented in section 5.2. We find that including the reblog network regularization term can increase performance for both the bipartite and tripartite methods, but that this performance boost is not guaranteed.

5.5.1 NMF Results With No Regularization

Gamergate

We used 200 annotated Gamergate posts as ground truth in plotting the ROC curves presented in Figure 4. Of the 200, 60 posts supported Gamergate and 95 were against it. The results, presented in Table 5.4 and Figure 5.1, show that NMF on the tripartite graph outputs better results than the baseline models. In particular, we see a 12% increase in accuracy, a 5% increase in F1-score, and a 21% increase in the ARI value.

Gamergate Results

	Bipartite	Tripartite	K-means	Spectral
Regularization	$\alpha = 0$	$\alpha = 0$		
Accuracy	0.625806451	0.748387096	0.612903225	0.516129032
F1-Score	0.754237288	0.804020100	0.760000000	0.663677130
ARI	0.028070687	0.238096246	0.0	0.02477930
AUC	0.643333333	0.783333333	0.643684210	0.613859649

Table 5.4: Gamergate Data: The tripartite method outperforms the bipartite method and other baselines on accuracy, F1-score, AUC and ARI measures for the Gamergate dataset.

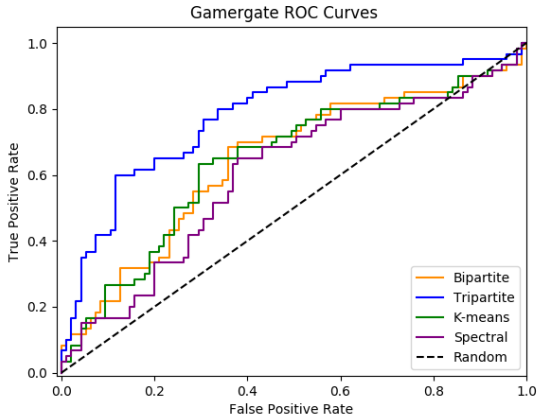


Figure 5.1: Gamergate Data: The tripartite method significantly outperforms the bipartite method and other methods by at least a 14% increase in AUC.

World Series Dataset

We had 478 posts from the World Series dataset annotated based on the team the post supported. Of the 478 posts, 43 supported the 'Kansas City Royals,' 121 supported the 'San Francisco Giants,' and 314 were 'neutral.' The performance measures for each experiment can be found in Table 5.5. We find that applying NMF to a tripartite graph results in approximately a 3% increase in accuracy, 4% increase in F1-score, and 6% increase in the ARI and AUC values. It should be noted that while the K-means algorithm does appear to rival the tripartite method for certain thresholds (Figure 5.2), it does so only by placing most (if not all) of the posts into the same polarity group and therefore does not exhibit

very much overall predictive power in the context of polarization identification.

World Series Results

	Bipartite	Tripartite	K-means	Spectral
Regularization	$\alpha = 0$	$\alpha = 0$		
Accuracy	0.775757576	0.806060606	0.763636364	0.733333333
F1-Score	0.821256038	0.864628822	0.807881773	0.836501901
ARI	0.300564756	0.369964582	0.274550716	0.109881429
AUC	0.872573515	0.937728234	0.835649887	0.537565743

Table 5.5: World Series Data: The tripartite method outperforms the bipartite and other methods on all measures.

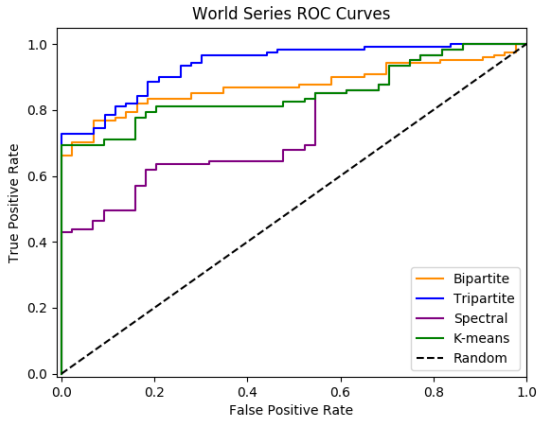


Figure 5.2: World Series Data: ROC Curves indicate that the tripartite method shows enhanced performance with approximately 6% additional area under the curve (AUC).

FIFA (2 Teams) Dataset

Of the 592 FIFA posts annotated, the 344 associated with the final match were used as ground truth labels. Of these, 59 posts supported Germany, 19 supported Argentina, and the remaining 266 posts were neutral. Performance results in Table 5.6 indicate that the tripartite method exhibits an approximately 15% increase in accuracy, an 8% increase in F1-score, and a 37% increase in ARI

value. Further, the ROC curves in Figure 5.3 show a 29% increase in AUC over the bipartite method, and a 20% increase over its closest competitor, spectral co-clustering.

FIFA (2 teams) Results

	Bipartite	Tripartite	K-means	Spectral
Regularization	$\alpha = 0$	$\alpha = 0$		
Accuracy	0.666666667	0.820512821	0.692307692	0.730769231
F1-Score	0.790322581	0.879310345	0.861313869	0.803278689
ARI	-0.012632288	0.370934799	0.0	0.047391903
AUC	0.553077609	0.841213202	0.5941124	0.649420161

Table 5.6: FIFA (2 teams): The tripartite method outperforms the bipartite method and other baselines on all measures.

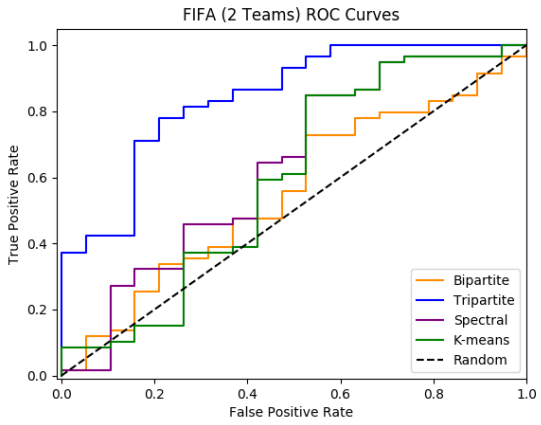


Figure 5.3: FIFA (2 Teams): The tripartite method shows approximately 29% improvement over the bipartite method's AUC.

FIFA (4 Teams) Dataset

For the FIFA (4 teams) dataset, we plotted a ROC curve (Figure 5.4) for each team using 592 annotated posts as ground truth. Of those, 131 supported Germany, 30 supported Argentina, 17 supported Brazil, 20 supported the Netherlands and 393 posts were neutral. In order to match predicted clusters to these ground truth clusters for each of the 50 experiments, we executed the Kuhn-

Munkres matching algorithm to find the cluster matching that would achieve maximum profit ([52]).

We found that our tripartite method outperforms the bipartite method in classifying each of these teams, as determined by the AUC. Additionally, we averaged the F1-scores and ARI scores over the 4 polarity groups for each method. The tripartite method obtained an F1-score of 0.661202185, improving upon the bipartite method's F1-score of 0.606557377 by approximately 5%. Similarly, the tripartite method obtained an ARI score of 0.293969905, about 20% higher than the bipartite method's ARI score of 0.092225707. We do not present results for K-means and spectral clustering, both of which grouped all posts into one polarity group for each of 50 runs. These results can be found in Table 5.7

FIFA (4 Teams): AUC Results

	Germany	Argentina	Brazil	Netherlands
Regularization	$\alpha = 0$	$\alpha = 0$	$\alpha = 0$	$\alpha = 0$
Bipartite AUC	0.635940643	0.801358234	0.817859673	0.865644171
Tripartite AUC	0.759305835	0.869057724	0.850460666	0.905521472

Table 5.7: FIFA (4 Teams): The tripartite method outperforms the bipartite method in clustering the four teams.

5.5.2 NMF Results with Regularization

Regularization terms can be useful in guiding the nonnegative matrix factorization process to a solution that represents real-world data comprehensively.

FIFA (4 teams): ROC Curves

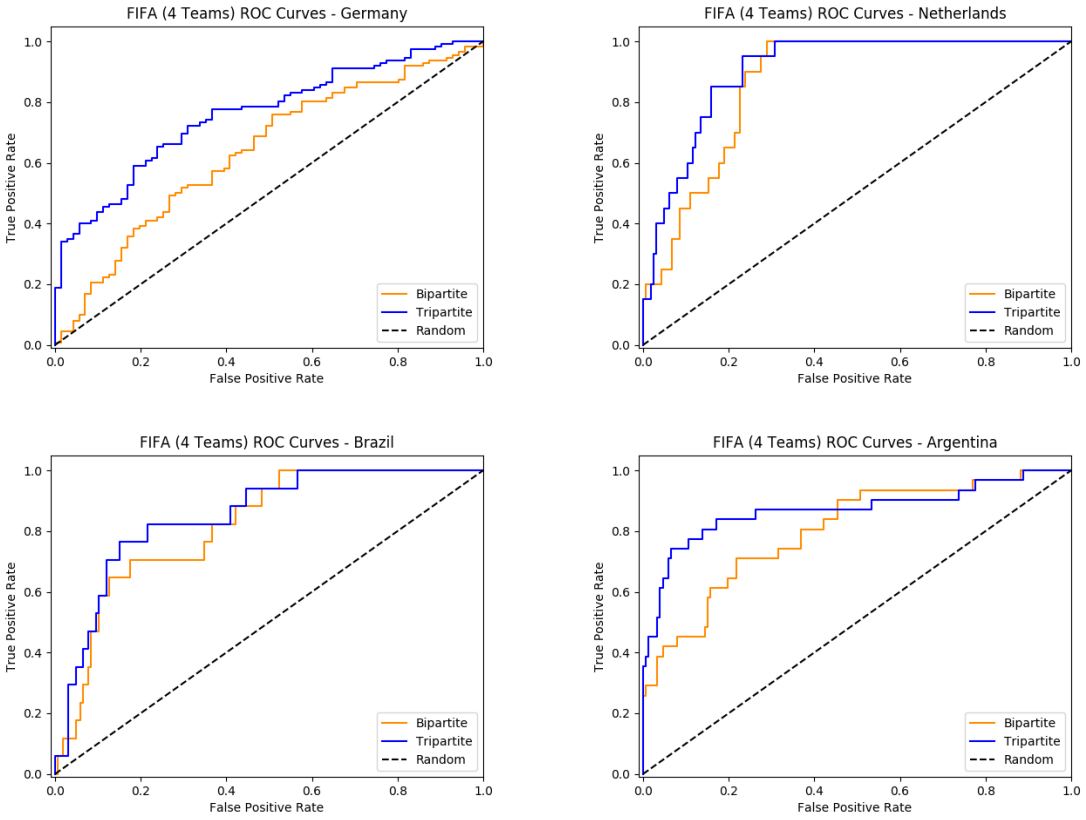


Figure 5.4: FIFA (4 Teams): For each team, the ROC curves indicate that the tripartite method outperforms the bipartite method in terms of AUC.

Reblog Network Regularization

Recall that we incorporated a regularization term that imposes a penalty if there is a reblog relationship between two users but they are not placed in the same polarity group. While we arbitrarily chose a regularization parameter of .3 for this analysis, we acknowledge that different parameter values give different results. Cross-validation would be necessary to determine the appropriate parameter for a given dataset.

In what follows, we compare the results obtained in Section 5.1 to results obtained after integrating regularization into our model. These results, presented

in Table 5.8, exhibit the potential of using the reblog (*endorsement*) network as a means of regularization. We find that reblog regularization improves the overall preformance of NMF on the bipartite graph for all datasets. The results, however, are mixed for the tripartite method. While there is a boost in performance on the World Series and FIFA (4 Teams) datasets, regularization has the opposite effect on the FIFA (2 Teams) dataset. This result is in line with previous research on social dependency regularization [3], and warrants further investigation into when this type of regularization is beneficial.

Regularization Results

World Series	Bipartite	Bipartite	Tripartite	Tripartite
Regularization	$\alpha = 0$	$\alpha = 0.3$	$\alpha = 0$	$\alpha = 0.3$
AUC	0.872573515	0.866423217	0.937728234	0.931193542
F1-Score	0.821256038	0.868852459	0.864628822	0.891774891
ARI	0.294490473	0.369964582	0.337173390	0.466696399
FIFA (2 Teams)				
AUC	0.553077609	0.623550401	0.841213202	0.817127565
F1-Score	0.790322581	0.819672131	0.879310345	0.836363636
ARI	-0.012632288	0.094434278	0.370934799	0.265281156
FIFA (4 Teams)				
F1-Score	0.606557377	0.666666667	0.661202185	0.699453551
ARI	0.092225707	0.153027075	0.293969905	0.326854025

Table 5.8: Regularization improves performance for the bipartite method for all datasets, while it worsens performance for the tripartite method on the FIFA (2 Teams) dataset.

Coreblog and Tag Similarity Regularization

Though we expected the coreblog and tag similarity regularization terms to improve NMF performance, we found that overall performance actually worsened when these terms were included across datasets. We hypothesize that these failed for the same reason community detection performs more poorly on polarized datasets - the existence of neutral users, posts and tags. The beauty of non-negative matrix factorization is that it is better able to identify polarized groups

in the presence of a large amount of neutrality than other methods. Modeling latent similarity into the objective function, therefore, may be counterproductive to NMF's ability to uncover latent properties on its own. The tag similarity metric was further complicated by the tendency of users to co-opt the tags of opposing groups in order to increase post visibility. In these instances, tag regularization would incorrectly impose penalties on opposing tags.

5.6 Discussion

The analysis of polarization on social media networks like Tumblr often requires an approach that works well in many different contexts and for variable amounts of information. In this work, we outlined a nonnegative matrix factorization-based approach that successfully identified polarized groups in two very different contexts, varying both in the nature of the topic (sports vs. movements) and the number of polarity groups involved. We conclude that the inclusion of *tag* and *post-tag* relationship information via a tripartite graph provides for a better clustering outcome than clustering on a *user-post* graph alone. In particular, NMF on a tripartite graph containing tag information exhibits enhanced clustering ability in comparison to three baselines: NMF, K-means community detection and spectral co-clustering on bipartite graphs. We further conclude that regularization terms can be helpful in ensuring that NMF finds the factorization that best fits the real-world dataset being studied. The inclusion of the reblog graph as a representation of endorsement between users improved performance for the bipartite method for all datasets, and for the tripartite method for two out of three datasets tested. We believe these results warrant further investigation into the best approach to incorporating user relation-

ship and user endorsement information into the NMF formulation. It might be useful, for instance, to test the output of the algorithm when a symmetric reblog network is used, possibly with a higher threshold for the minumim number of reblogs between users. Finally, being the the first ever analysis of polarization on the Tumblr network, we show that Tumblr can be an extremely valuable resource for datasets, especially as related to opinion and polarization detection.

Our study is not without limitations. Given the sizes of our datasets, it was infeasible to have every post annotated by humans. The performance measures we present heavily rely on these annotations and would, of course, change with the existence of a fully annotated dataset. Though the posts selected to be annotated were chosen at random and are expected to represent the larger dataset, future analyses on fully annotated datasets would prove beneficial. The inability to fully annotate large social media datasets is a common problem in research, speaking to the need for unsupervised approaches like nonnegative matrix factorization. By showing NMF's versatility (in terms of the information that can be incorporated into the algorithm and the variety of datasets it can be applied to), we hope to aid future research in illuminating polarization patterns in online social media.

BIBLIOGRAPHY

[1] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[2] Meysam Alizadeh, Ingmar Weber, Claudio Cioffi-Revilla, Santo Fortunato, and Michael Macy. Psychological and personality profiles of political extremists. *arXiv preprint arXiv:1704.00119*, 2017.

[3] Md Tanvir Al Amin, Charu Aggarwal, Shuochao Yao, Tarek Abdelzaher, and Lance Kaplan. Unveiling polarization in social networks: A matrix factorization approach. Technical report, IEEE, 2017.

[4] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

[5] Roland Barthes. *Mythologies*. Seuil, 1957.

[6] BBC. *Facebook, google and twitter agree german hate speech deal.*, 2015. `http://www.bbc.com/news/world-europe-35105003`.

[7] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[8] Kevin Boyle. Hate speech–the united states versus the rest of the world. *Me. L. Rev.*, 53:487, 2001.

[9] Uwe Bretschneider and Ralf Peters. Detecting offensive statements towards foreigners in social media. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[10] Uwe Bretschneider, Thomas Whner, and Ralf Peters. Detecting online harassment in social networks. In *ICIS 2014 Proceedings: Conference Theme Track: Building a Better World through IS*, 2014.

[11] Rogers Brubaker. *Ethnicity without groups*. Harvard University Press, 2004.

[12] Census Bureau. *Census Regions and Divisions of the United States*.

[13] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[14] Lulu Chang. *New Yahoo algorithm can spot online abuse in context, not just content*, 2016. `https://finance.yahoo.com/news/yahoo-algorithm-spot-online-abuse-225438302.html`.

[15] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Measuring# gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1285–1290. International World Wide Web Conferences Steering Committee, 2017.

[16] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[17] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. *ICWSM*, 133:89–96, 2011.

[18] Maral Dadvar, Franciska MG de Jong, RJF Ordelman, and RB Trieschnigg. Improved cyberbullying detection using gender information. 2012.

[19] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and machines against bullies: a hybrid approach to detect cyberbullies. In *Conference on Artificial Intelligence*. Springer International Publishing, 2014.

[20] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[21] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, 2013.

[22] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 241–249. Association for Computational Linguistics, 2010.

[23] Thomas Davidson, Dana Warmsley, Micheel Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media*, pages 512–515, Montreal, Canada, 2017.

[24] Merriam-Webster Dictionary. Merriam-webster. *On-line at http://www. mw. com/home. htm*, 2002.

[25] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[26] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02), 2011.

[27] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006.

[28] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30. ACM, 2015.

[29] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *WWW*, pages 29–30, 2015.

[30] M Eddy and M Scott. Delete hate speech or pay up, germany tells social media companies. *The New York Times. Retrieved from https://www. nytimes. com/2017/06/30/business/germany-facebook-google-twitter. html*, 2017.

[31] Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo. Understanding harmful speech online. *Berkman Klein Center Research Publication*, 21, 2016.

[32] Samantha Finn, Eni Mustafaraj, and P Takis Metaxas. The co-retweeted network and its applications for measuring the perceived political polarization. 2014.

[33] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[34] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. UNESCO Publishing, 2015.

[35] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy in social media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 33–42. ACM, 2016.

[36] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 81–90. ACM, 2017.

[37] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[38] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230, 2015.

[39] Google. *What if technology could help improve conversations online?*, 2017. `https://www.perspectiveapi.com/#/`.

[40] Pedro Henrique Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. A measure of polarization on social media networks based on community boundaries. In *ICWSM*, 2013.

[41] Qi Han, Junfei Guo, and Hinrich Schuetze. Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 520–524, 2013.

[42] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

[43] Sameer Hinduja and Justin W Patchin. Cultivating youth resilience to

prevent bullying and cyberbullying victimization. *Child Abuse & Neglect*, 73:51–62, 2017.

[44] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. A longitudinal measurement study of 4chan's politically incorrect forum and its effect on the web. In *Proceedings of the Eleventh International Conference on Web and Social Media*, pages 92–101, Montreal, Canada, 2017.

[45] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6. ACM, 2014.

[46] Matthias Hüning. Textstat simple text analysis tool. *Dutch Linguistics, Free University of Berlin, Berlin*, 2005.

[47] Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, 2014.

[48] James B Jacobs and Kimberly Potter. *Hate crimes: Criminal Law and Identity Politics*. Oxford University Press, 2000.

[49] Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Ins Torres. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124 – 133, 2014.

[50] Gary King, Patrick Lam, and Margaret E Roberts. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 2017.

[51] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 195–204, New York, NY, USA, 2013. ACM.

[52] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.

[53] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against

blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 1621–1622. AAAI Press, 2013.

[54] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

[55] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

[56] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[57] Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284, 2014.

[58] Rijul Magu, Kshitij Joshi, and Jiebo Luo. Detecting the hate code on social media. In *Proceedings of the Eleventh International Conference on Web and Social Media*, pages 608–612, Montreal, Canada, 2017.

[59] AJ Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015.

[60] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *WWW*, pages 145–153, 2016.

[61] Hatebase Organization. *Welcome to the world's largest online repository of structured, multilingual, usage-based hate speech*. `https://www.hatebase.org`.

[62] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[63] Yulong Pei, Nilanjan Chakraborty, and Katia Sycara. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[64] Nick Pendar. Toward spotting the pedophile telling victim from predator in text chats. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 235–241. IEEE, 2007.

[65] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[66] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM, 2015.

[67] Santiago Resett and Manuel Gámez-Guadix. Traditional bullying and cyberbullying: differences in emotional problems, and personality. are cyberbullies more machiavellians? *Journal of Adolescence*, 61:113–116, 2017.

[68] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, 2016.

[69] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics.

[70] Shari Kessel Schneider, Lydia O'donnell, Ann Stueve, and Robert WS Coulter. Cyberbullying, school bullying, and psychological distress: A regional census of high school students. *American journal of public health*, 102(1):171–177, 2012.

[71] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61, 2010.

[72] Yongren Shi, Kai Mast, Ingmar Weber, Agrippa Kellum, and Michael Macy. Cultural fault lines and political polarization. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 213–217. ACM, 2017.

[73] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social

media. In *Proceedings of the Tenth International Conference on Web and Social Media*, pages 687–690, Cologne, Germany, 2016.

[74] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690, 2016.

[75] Barbara A Spears, Carmel Taddeo, and Alan Barnes. Online social marketing approaches to inform cyber/bullying prevention and intervention: What have we learnt? *Reducing Cyberbullying in Schools: International Evidence-Based Best Practices*, page 75, 2017.

[76] Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. Racial microaggressions in everyday life: implications for clinical practice. *American Psychologist*, 62(4):271–286, 2007.

[77] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. The automated detection of racist discourse in dutch social media. *CLIN Journal*, 6:3–20, 12 2016.

[78] Kimberly Twyman, Conway Saylor, Lloyd Adam Taylor, and Cadie Comeaux. Comparing children and adolescents engaged in cyberbullying to matched peers. *Cyberpsychology, behavior, and social networking*, 13(2):195–199, 2010.

[79] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria, 2015.

[80] Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Véronique Hoste, and Walter Daelemans. Guidelines for the fine-grained analysis of cyberbullying. Technical report, LT3, Ghent University, Belgium, 05/2015 2015.

[81] Samuel Walker. *Hate Speech: The History of an American Controversy*. U of Nebraska Press, 1994.

[82] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. Cursing in english on twitter. In *CSCW*, pages 415–425, 2014.

[83] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26. Association for Computational Linguistics, 2012.

[84] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics.

[85] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and CSS*, pages 138–142, November 2016.

[86] Zeerak Waseem. Automatic hate speech detection. Master's thesis, University of Copenhagen, 2016.

[87] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.

[88] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics.

[89] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*, pages 88–93, 2016.

[90] Ingmar Weber, Venkata Rama Kiran Garimella, and Asmelash Teka. Political hashtag trends. In *European Conference on Information Retrieval*, pages 857–860. Springer, 2013.

[91] Andreas Wimmer. *Ethnic boundary making: Institutions, power, networks*. Oxford University Press, 2013.

[92] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[93] Jiejun Xu, Ryan Compton, Tsai-Ching Lu, and David Allen. Rolling through tumblr: characterizing behavioral patterns of the microblogging

platform. In *Proceedings of the 2014 ACM conference on Web science*, pages 13–22. ACM, 2014.

[94] Jiejun Xu, Tsai-Ching Lu, et al. Automated classification of extremist twitter accounts using content-based and network-based features. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 2545–2549. IEEE, 2016.

[95] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 549–553. SIAM, 2006.

[96] Linhong Zhu, Aram Galstyan, James Cheng, and Kristina Lerman. Tripartite graph clustering for dynamic sentiment analysis on social media. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1531–1542. ACM, 2014.