Computational Models of Identity Presentation in Language

Michael Miller Yoder

CMU-LTI-21-006

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Carolyn Penstein Rosé, Chair
Yulia Tsvetkov
Geoff Kaufman
David Jurgens

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy In Language and Information Technologies

Computational Models of Identity Presentation in Language

Michael Miller Yoder

September 30, 2021

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Carolyn Penstein Rosé, Chair
Yulia Tsvetkov
Geoff Kaufman
David Jurgens

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Keywords: Computational modeling of identity, identity and language, computational sociolinguistics, computational social science, natural language processing, Tumblr, fanfiction

Abstract

Researchers in computer science and natural language processing have built models of how language use varies according to the latent, stable identities of language users. However, such a conception of identity is ill-equipped to investigate dynamic, contextual expressions of identity in online communities. This thesis draws on theories from sociolinguistics, linguistic anthropology, and the social sciences that view identity not as fixed and predetermined, but constructed in language and interaction. We pair these theories with techniques from machine learning, statistics, and natural language processing in a new framework for computational investigations of identity presentation in online communities. This framework identifies the role identity presentation plays in online contexts, as well as its relationship to social interaction and outcomes within and outside particular online communities. We demonstrate this framework on datasets of linguistic and social interaction from two online contexts known for identity talk, especially regarding gender and sexuality. The first context is Tumblr, a social media and micro-blogging site, where we examine direct identity presentation of users in textual self-descriptions and communities in network data. The second context is fanfiction, narratives that transform and expand on original media, where we investigate the indirect presentation of character identity in narrative.

With data from Tumblr, we first examine associations between self-presented identity labels and content sharing through the network of users. We extract both identity categories and specific labels presented by Tumblr users in free-text bio boxes and consider whether similarities or differences in self-presentation affect the propagation of content. To test this hypothesis, we use self-presentation features in a machine learning task predicting whether a user will share content from another user. We find that identity features provide an informative signal often overlooked in previous work on content propagation. Interpreting the learned feature weights in a linear model, we find that alignment on different "levels" of identity self-presentation (broad categories or exact label matches) have differing effects on content propagation in a social network. Interactions between labels that indicate shared experience or values, such as conceptualizations of gender, are particularly informative. Though we cannot directly observe the construction of social solidarity or alignment that comes from self-presentation in language, in this way we are able to use computational tools to discover its effects.

Identity alignment plays a role in how content propagates in Tumblr, but communities are also salient in this social media site. From communities emergent from network connections among users, we investigate the effect of community alignment on content propagation. We find a non-random association, yet this effect is quite small compared with the influence of features of the content. The most informative content features relate to communities, however, which points to the importance of community-based identity organized around content rather than direct network connections among users on Tumblr.

We then use this framework to examine how implicit identity positioning in nar-

rative is associated with social change within and beyond online communities. For this we use corpora of fanfiction, stories written by fans that expand or change original narratives from TV shows, comics, books and movies. To extract which characters are present in stories and which text is used to portray them, we first introduce and evaluate a processing pipeline that adapts natural language processing tools for entity coreference and quote attribution to the fanfiction domain of varied, informal narrative.

Using this pipeline, we investigate a suite of computational approaches that use word vectors to represent characters and relationships between them in narrative. In particular, we investigate the ability of such approaches to identify when fanfiction writers change the depiction of a relationship between characters from its portrayal in the original, source narrative as romantic or not. A qualitative analysis reveals that this predictive model picks up on emotionally intense language used around relationships that have been changed from original media. We also construct visualizations of the learned representations for characters and analyze the extent to which contrasts in these representations reflect contrasts in the positioning of characters between original and derived works. However, some difficulty in separating out social phenomena from more surface-level features in language such as genre persists.

Fanfiction has been considered a "queer space" for its considerable representation of same-gender relationships. The community of online fanfiction readers and writers has grown at the same time of significant offline action by the LGBTQ social movement for rights. We investigate how the representation of LGBTQ characters in number and in textual portrayal varies along with changes in the offline social movement. We find that representations of characters in same-gender marriages in fanfiction increases after 2015 US marriage equality. We also find correlations between the representation of LGBTQ characters in fanfiction with trends in mainstream news of LGBTQ issues from 2010-2020. Cultural events such as Pride drive this effect more than legal and law-focused news, suggesting the nature of the relationship between fanfiction and the LGBTQ social movement.

In these computational projects, we operationalize identity as a dynamic, contextual presentation in a new framework for computational analysis of identity in online communities. The findings from these projects demonstrate how this framework can be used to identify the values around identity that are central in online communities– and how these identity presentations relate to social interactions online and offline. These include which combinations of identity labels relate to the flow of content on social media, and which presentations of LGBTQ characters in online narrative relate to the LGBTQ social movement offline. Our goal is for this framework, theory, and example case studies to be useful examples to others investigating the effects of identity in online communities at a large scale.

Acknowledgments

There are so many people who shared their time, teaching, mentorship, and collaboration with me to make this dissertation a reality. I thank my advisor, Carolyn, for always pushing me to think more deeply about data, methods, and implications. Her ability to find a way forward no matter how research enfolds has been a great example for me. I also thank my committee, who always had the right advice and the right questions to move the work forward. Their insights across disciplines–both theory and methods–guided and enriched the work. Friends and colleagues at Carnegie Mellon University and the University of Pittsburgh supported me in every project. I thank them for their time, ideas, work, and companionship amidst the stress of paper deadlines. I could not have done this dissertation without the loving support of my family, and especially of my spouse and partner in life, Lydia. Finally, this dissertation would not be possible without the work and interaction of social media users and fanfiction writers, the data for projects in this thesis. I thank them for their creativity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Identity is often assumed to be some sort of essence, and this is reflected in the way people speak about identity. Individuals *have* an ethnicity, someone *is* a man, I *am* the author of this text. Aspects of identity seem to appear naturally given and can color our entire perceptions of people. One can hardly speak of another person in many languages without choosing a gendered pronoun. Identity labels operate as a convenient metonymy for grouping people, as seen in abstracted phrases such as: "*Scientists* agree...", "The *Hispanic* vote", or "Go, fight, win, *Lady Tigers*!". These labels not only refer to *who*, but subtly imply *what* the people referred to are: which aspects of their identity are most relevant in grouping them together and which are subsequently backgrounded.

On closer inspection, and especially with historical review, we find that this identity essence is much more squishy and unstable that previously assumed. First is the issue of philosophical nominalism: can a categorization of people exist without a name for them? Who was a *gay* person in the ancient world? Can there be *Americans* before the United States of America was formed? Can there be *hipsters*, *punks*, or *jocks* before those names were given to people who expressed certain ways of looking, being, and interacting? If people with green hair were given a name other than "people with green hair", would they become some sort of a group, a more salient part of someone's identity, and perhaps even be noticed more often?

Then of course, there are the boundary cases to think about. Why is Barack Obama *black* when his mother was perceived as *white*? Is being *Jewish* an ethnicity or religion? If sex is defined by certain configurations of chromosones and body parts, what about people whose body parts and chromosones do not easily fit these categories? Situations matter as well: what if someone is recognized as *Asian* when walking around in the US and *Punjabi* when walking in India? What if someone "is" Latinx but passes as white in most daily interaction in the US? Perhaps identity is what people mark in surveys, such as a census. But what if someone marks different categories on different surveys? What if the categories do not fit the person's own ideas of who they are? What if the categories on the census itself change? We take these cases not as exceptions to a rule, but indicative of the the very way in which identity operates in language.

What comes out of all this speculation is a conception of identity as much more contingent than previously assumed. Identity is contingent on language, especially language situated in particular social contexts. And curiously, identity may not be some sort of personal essence that

exists prior to interaction, but emerges after an individual is named *as* something, or assumed to be something, or claims to be something, in an interaction. This seems rather turned around from the intuitive sense of identity we started with: someone's identity comfortably lodged in their own person, preceding any social interaction.

Many researchers who study language and identity in fields such as sociolinguistics, discourse analysis, linguistic anthropology, and gender studies argue for this "turned-around" view in which identity emerges in social and linguistic interaction. This conception is frequently termed the *social construction* of identity, where identity is not viewed as something natural or biological, but rather grouped and named by society in particular social and historical contexts, often to serve the interests of groups with power. In this thesis, we adopt this view of identity as a starting point for computational analyses of linguistic data.

This is not the usual thing to do in computer science. Computational work associating behavior with identity usually considers identity to be exactly what we started with: a latent, personal, even psychological essence that produces variation in language or other behaviors (Wang and Kosinski, 2018; Huang et al., 2020). This may not be directly stated but is often an assumption realized in computational data and models. I argue that such a view is limiting, often does not fit the observed behavioral data of social interaction, and can even lead to the construction of tools that reinforce problematic assumptions.

Conversations and interactions in online communities shape notions of what identities are possible, what associations are made with identities, and how individuals' perceive their own identities. Some of these online spaces contribute to growing freedom of expression around identity, such as those that for years have supported LGBTQ youth (Pullen and Cooper, 2010). Too often, movements to restrict identity expression have gained traction online; hate speech and extremist organizing in new media have often preceded racist, sexist, and homophobic violence. The default theorization of identity in computer science as stable, individual, and "natural" is a poor fit to studying how associations with identities are being shaped and transformed in these emergent online contexts.

In turning the tables and acknowledging the role of language in *constructing* identity, we can build computational models that explore how identity depends on context and is *presented*, not just reflected, in interaction. This change in mindset shifts how we extract features of identity in projects within this thesis. We seek to capture how people make choices to *present* their own identities and those of others in context, not inferring some notion of "true" identity for people. These features include self-presented identity labels and hashtags on social media, as well as tags and text descriptions of characters in online fanfiction.

But the goal of this thesis is not to study identity presentation in isolation. We are interested in how identity presentation in online communities relates to social interaction–within those communities or externally. The relationships that we discover between identity presentation and social interaction allow insight into how identity presentation is valued and has an impact in these online communities. More broadly, discovering these values and patterns between identity presentation and social interaction illustrates the ways in which online communities are shaping and changing associations made with identity. For example, evidence for fanfiction's relationship to the LGBTQ social movement suggests that portrayals of LGBTQ characters have the potential to reference or reinforce calls for rights happening offline. On Tumblr, the unique combinations of identity labels found to be associated with content propagation suggest new forms of social

solidarity–though we find evidence that these communities are primarily around content rather than direct user-user connections.

The main contribution of this thesis, however, is not findings about the nature of particular online communities. Rather, we offer a framework for discovering how identity is working in online communities: which aspects of identity are presented and valued in a community, and how the presentation of identity relates to social interactions within and beyond a community. This framework for identifying features of identity presentation and relating them to social interaction in large datasets from online communities is more fully described in the next section.

1.1 Framework of analysis

The projects in this thesis follow the same framework of analysis, which we offer as a main contribution of this thesis. This framework can be viewed as a set of tools for researchers studying the relationships between identity presentation and social interaction in large datasets from online communities. The tools include a theorization of identity to guide feature extraction and research questions, as well as recommendations for selecting outcome variables to correlate those identity features with. Projects presented in this thesis can be viewed as demonstrations of this framework to find the values placed on identity presentation in specific online contexts, as well as how that identity presentation relates to social outcomes within and outside the community. Such a framework allows researchers to design experiments that identify what types of identity presentations are relevant, and thus valued and reinforced, in particular online communities. It also allows researchers to estimate the relationship between users presenting forms of identity and others interacting with them in online platforms.

This framework requires two main variables: features of identity presentation and measures of social interactions or outcomes. Both are discussed in the remainder of this section.

Theory-driven identity feature extraction First, we offer a theoretical background from the humanities and social sciences to draw on: the social construction of identity. Adopting this perspective shifts focus from identifying the "true" identities of users to looking for how users themselves *present* their identity or those of others. Quantitative work requires operationalizing concepts such as identity as clearly defined variables that can be measured in large datasets. In machine learning terms, this step is referred to as *feature extraction*. The social construction of identity is especially relevant to how features of identity are operationalized within this framework.

The features of identity we use in this thesis are not inferred about people, which assumes a static, "natural" view on identity. Instead, we look for how choices in how people themselves present identities in online communities. Looking for identity presentation rather than inferred, latent, identity is more consistent with theories of the social construction of identity from the humanities and social sciences. Not defining rigid specifications for identity *a priori* is more inclusive to those marginalized by society who do not generally fit such categorizations, such as transgender people. Crucially, focusing on the presentation of identity in a bottom-up fashion also allows for agency in how the participants in online communities present their own identities or those of others. We view identity as an impression achieved through a set of behaviors in

context, which reflect choices made by online community participants. Users on Tumblr, for example, make choices about what identity labels to put in their blog's description, or whether or not to put identity labels there. These choices reflect which dimensions of identity are relevant to a user in this particular context. We argue that identities that are presented by users rather than inferred by researchers are more likely to be relevant to social interaction on the platform.

Our goal is to extract features that capture how people use language to position the identities of themselves or others and as described in the next section, explore how this presentation relates to social interaction and outcomes. In modeling the associations between identity presentation and social interactions, this framework allows inspection of the values a community places on forms of identity presentation. For example, in Chapter 4, we learn weights on combinations of identity labels between users that indicate a higher likelihood of sharing content between those users. This provides evidence for the relevance of those particular forms of identity presentation on the platform.

Capturing the presentation of identity comes with its own challenges when applied to large datasets. Instead of a controlled set of possible values for identity categories (*male* and *female* for gender, for example), we are interested in a much larger range of possibilities driven by how people are presenting themselves and others in online contexts. This thesis demonstrates a number of approaches for handling this, including bottom-up grouping of labels in self-descriptions on Tumblr (Chapter 4), learning word embeddings (Chapter 7), and estimating topic models (Chapter 8) over text used to portray characters of different genders and sexualities in narrative. We also contribute an example of using NLP methodology and tools such as coreference resolution and quote attribution to extract the identity portrayal of characters from free-form text.

Relating identity presentation to social interaction The framework of analysis in this thesis incorporates measures of social interaction to understand the values placed on identity presentation and how identity relates to the work of communities. We choose social behaviors or outcomes that are measurable and express value in an online community or that are exemplary about a community that draw researchers to study it. These measurable outcomes, which can be regressed onto or predicted, provide "hooks" into large datasets. They also ground investigations into identity by using a measure with social significance. For example, content propagation in Chapters 4 and 5 is one of the major outputs of interactions on social media and directly affects what people see and what shapes them on the platform. Similarly in Chapter 7, one of the most dramatic departures from original work made by fanfiction is the changing of character relationships, particularly from the obsessive focus of traditional media on heterosexual romantic relationships to a queer-friendly space with majority gay male relationships. This informed the choice of the social outcome we wished to measure in Chapter 7: fanfiction authors changing relationships from canon stories.

These outcomes also relate to users' choices: we aim to isolate key choice points that reflect values of the community and how identity plays a role in those choices. In work on Tumblr in Chapters 4 and 5, we examine the choice to share content, which can be seen as currency that enables the flow of content through this network. In work on fanfiction (Chapters 7 and 8), we look at how an external movement related to LGBTQ inclusion gives a larger context for LGBTQ character representation in fanfiction.

This thesis demonstrates a variety of quantitative and computational approaches for relating identity presentation to social outcomes, including the applied machine learning in Chapter 4 and statistical methodology used in Chapter 8.

1.2 Definition of identity

From sociocultural linguists Bucholtz and Hall (2005; 2010), we adopt the broad conception of identity as *the social positioning of self and other*. The idea of identity as *positioning* is a useful metaphor in that the functional role of identity labeling is akin to drawing a map. Certain people are located in one spot in mental or conceptual space, others are located in a different place due to differences in certain personal or group characteristics. Note that such positioning necessarily foregrounds certain aspects of identity similarity and erases other aspects, just as what coordinate system or perspective a map uses affects distances between points.

Bucholtz and Hall's definition of identity is not just about a single person, the self. Rather, it includes the positioning of *self and other*. This is not by accident: they argue for a conception of identity as inherently relational and social. A conception of self cannot exist alone, but is necessarily tied to the ideas of who is like and unlike oneself. This is most visible in social interaction, where one's expressions as a particular kind of person can create similarity or difference with other participants–present or not–in the discourse (the positioning of the *other*). Can one "be" or talk as a *teacher* without positioning who they're talking to as *students*? Can a person be *white* without making others *black*, *brown*, or otherwise racially marked? We take the position that they cannot: that identity, even if expressed just as a positioning of self, implicitly or explicitly positions others as similar or different from that self. This is apparent in our first investigation of identity labeling on social media, where we consider the impact of similarities and differences in self-labeling between pairs of users (Chapter 4). The relational component of identity is also clear in our focus on associations with other-centered identities constructed in narrative (Chapter 7)

To analyze this concept of identity further, it may be useful to discern 3 levels of identity labels, also from Bucholtz and Hall (2005), which we use to distinguish identity categories expressed in communities on social media (Chapter 4).

1. **Macro-level demographic categories.** These are the dimensions of identity probably most often considered prototypical "identity", such as *age, gender, nationality, ethnicity, occupation, class*, and *sexual orientation*. These undoubtedly play a role in the interactions we study in this thesis and are the generalizable categories which social scientists most often try to correlate with behavior. However, to assume identity is a sum or collection of these categories ignores the particularities of lived experiences at the intersections of multiple categories (Crenshaw, 1989; Herbelot et al., 2012). Furthermore, these broad categories are not always equally relevant to participants in linguistic and social interaction (Coupland, 2007). These "natural-seeming" categories themselves were constructed with particular needs in history to group large numbers of people together for capitalist and nationalist goals such as market analysis or apportioning social goods with census-taking.
2. **Local, ethnographically specific cultural positions.** These categories of identity are less general than demographic categories, but are still relatively durable groupings of people

that are relevant to participants in interaction. These categories are often ethnographic in the sense that they hold specific meaning within language users' own frames of reference or local communities. A classic example of these categories from sociolinguistics is Eckert (2000)'s work in a Detroit area high school looking at language among *jocks* and *burnouts*. Though these ethnographically specific categories correlated with class and race, they often eclipsed these demographic categories in influencing specific interactions. The importance of which *fandom* a Tumblr user belongs to is an example of this level in this thesis (Chapter 4).

3. **Temporary, interactionally specific participant roles.** On the opposite end of the spectrum from demographic categories, these are the rather transient roles such as *joke-teller* in a group of friends, *discipliner* when a baby-sitter scolds a child, or *presenter* of a thesis defense. They arise in specific interactions, though can play a more durable role if participants are repeatedly in particular situations that give rise to them. These interactional roles are also inflected by the cultural positions and demographic categories of people believed to inhabit these roles more often. For example, being asked to say a prayer before a large group meal in American Christian households may be associated with being an older, male person. If the person asked to say a prayer doesn't fit that positionality, it may be seen as striking or even lead to questions, surprise, or laughter. This thesis largely considers more enduring, general categories of identity than interactional roles. However, it is important to note that these roles play a significant role in behavioral and linguistic interaction.

We consider categories and labels at each of these levels to still be *identity*. Separating these levels is useful conceptually, but it is important to note that each can affect behaviors and language choices in interaction, and that these effects can vary (Coupland, 2007). Within a single conversation between coworkers working on a project, for example, issues of race and gender may only tangentially arise, whereas positions within the company (local, culturally specific identity) may affect the speech. Over lunch, performing interactional roles such a story-teller may affect language choices and conversational behavior the most.

It can be difficult to disentangle these levels and their effects. In online stories, a narrative may cast characters in particular genders, but what they say may largely be due to the roles they are playing in a particular scene, or even because of their cultural positioning as a member of a particular group of characters in the original media series. However, what roles characters play are most likely not distributed evenly among characters fitting larger demographic categories such as gender. We model this layering of roles (operationalized as topics) with sexuality in our experiments investigating how the presentation of LGBTQ characters in online narrative varies with changes in the offline LGBTQ social movement (Chapter 8).

Finally, a note on terminology used for identity in this thesis. I have tried to hew as closely as possible to terminology used by those who have certain identities themselves. I also tend toward terms that activists in social movements use, as is particularly relevant in Chapter 8 with LGBTQ social movements. I have tried to be inclusive with terms, such as using "Latinx" instead of the more gender-exclusive and patriarchal "Latino". However, following Stulberg (2018), I also want to be as descriptive and accurate as possible and not include groups that are not actually included in the research. For example, there are significant online movements from asexual and intersex groups, and such groups are not included in the term "LGBTQ". In Chapter 4, we do capture self-

descriptions with asexual terms such as "ace", and so I use the more expansive term "LGBTQ+". But asexual and intersex movements are not highly visible in the social movements whose events and trends I measure from 2010-2020 in Chapter 8, so for accuracy I do not include them in my use of "LGBTQ". No choice of terminology is perfect, however, and I fully acknowledge the limitations of such choices.

1.3 The author's identity

I adopt the view from anthropology and other social sciences that it is important to recognize a researcher's own social positionality. This identity influences research questions chosen, experiments performed, and the analysis completed. So, here are some identity labels that may be relevant: I am white, relatively young (though not as young as when I started this Ph.D.), American (Midwestern), liberal, straight, cisgender, male, and well-educated. (This document, of course, is an attempt to claim to be even more educated with the label of *Ph.D.*)

1.4 Structure of this thesis

We present four computational analyses of identity presentation in language in this thesis (Chapters 4, 5, 7, and 8). In Chapter 6, we present a pipeline for extracting text relevant to characters in fanfiction, which is used in Chapters 7 and 8. Each of these analyses relate features of identity presentation, usually presented by participants in online communities themselves, with measures of social interaction.

See Table 1.1 for an overview of the identity features and social interactions associated in each project. Using data from Tumblr, Chapters 4 and 5 use the social outcome of content propagation and estimate the effects of identity presentation alignment, or community alignment, on that outcome. In Chapter 7, we use the social outcome of whether or not the relationship presented in a fanfiction story matches the original canon relationship with respect to being romantic. We then estimate associations with how the author portrayed the characters in the relationship (the presented identity) in the story text. In Chapter 8, we use a measure of social interaction external to the online community, events and topic shifts in the LGBTQ social movement. We test for associations between the LGBTQ social movement and factors representing the portrayal of LGBTQ characters in fanfiction.

The analyses vary across several axes, which may be helpful for conceptually organizing this thesis:

- **Data.** One source of data for these projects is social media data from Tumblr, including text self-descriptions, posts (both content and hashtags), and network behavior such as following and sharing others' content. With this type of data, we consider how social media users use these affordances to position the identity of themselves and other users in particular ways. See Section 3.1 for details on the social media data from Tumblr that we use. Another source of data is online narrative, specifically fanfiction, written by amateur writers. With this type of data, we consider how writers make associations with identities

through portrayals of the characters they write. See Section 3.2 for details on the fanfiction data that we use.

- **Identity presentation features.** Language features of identity presentation used include identity labels that people use on social media to describe themselves, text used to portray characters in narrative, and hashtags used to label and make commentary on social media content. Network features include social media users following each other and sharing each others' content. Finally, we use the structured metadata of relationship types, characters, and free-form tags that fanfiction authors annotate stories with on the platform we draw data from, Archive of Our Own. All of these are indicators of user choice in presenting the identities of themselves or others, which we use as independent variables in regressions or as features in machine learning classifiers.
- **Social interaction features.** We relate identity presentation to measures of social interaction as outcome measures in online contexts. These include content propagation in social media, the transformation of relationships between characters in narrative, and events in a social movement offline. Though we sometimes label these as "outcomes", this just refers to their status as dependent variables in regressions or target variables to be predicted in machine learning classifiers. All findings and relationships between variables are correlational, from observational data, in quasi-experimental designs.
- **Identity focus.** Central to our conception of identity is that one's own identity is socially constructed in relation to positioning the identity of others. However, our analyses feature self-positioning and other-positioning to varying degrees. They also vary according to which "other", positioned in text and interaction, is represented in our computational models. This ranges from other specific individuals in social media to associations made with broad identity types in narrative.

A table comparing analyses according to these dimensions is presented below:

Chapter	**Data**	**Identity features**	**Social interaction**	**Identity focus**
4	social media	text identity labels, profile images content hashtags	content propagation	self, other users
5	social media	network community, content hashtags	content propagation	self, other users
7	fanfiction	text portrayal of characters	character relationship transformations	characters, relationships
8	fanfiction	text portrayal of characters story metadata	LGBTQ movement events, topics	characters, identity types

Table 1.1: Comparison of experiments.

Chapter 2

Related Work

2.1 Critical approaches to the study of identity

This thesis draws on critical approaches to theorizing identity from the humanities and social sciences. By "critical", I mean theories that do not take identity as a given notion, which often defaults to a "common-sense" or folk approach to identity as individual, stable, and psychological. Instead, critical approaches reflect and critique how notions of identity are not natural, but arise in particular social, cultural, and historical contexts (Barry, 2002). These socially constructed notions of identity often play a role in reinforcing existing power structures or challenging them. Crucially, these approaches draw attention to how the work of research itself conceptualizes identity. Research plays a role in reinforcing or challenging particular notions of identity, with possible consequences for those who are labeled with particular identities. I argue that if computational research is done "uncritically", without stating how it is theorizing concepts such as identity and how those conceptual frameworks arose within particular contexts and amidst particular power structures, it risks reinforcing whatever dominant notion of identity is held in a society regardless of how that notion may support existing social inequalities. For example, uncritical computational work in user attribute inference that predicts dimensions of identity such as gender from user data can reinforce harmful stereotypes about who men and women "are", and reinforce the notion that gender is simply binary.

In this section I describe the specific areas I draw on in the humanities and social sciences for these critical theories of identity. This thesis aims to apply these theories in computational research.

2.1.1 Sociocultural linguistics

The main theoretical approach this thesis draws on is from researchers studying the relationships between language and society, an area Bucholtz and Hall (2005) describe as "sociocultural linguistics". This encompasses areas of sociolinguistics, linguistic anthropology, and discourse analysis, which often draw upon shared theoretical resources in developing conceptual frameworks for how social structures, such as identity, are both reflected and constructed in language. Classic sociolinguistic work relates broad social variables such as social class with linguistic

variables such as phonetic variation. For example, Labov (1972) finds that the voicing of post-vocalic *r* among retail workers in New York patterns neatly by social class. Some work in such "variationist" sociolinguistics does consider the social and historical context of particular identity variables (Labov, 1963; Le Page and Tabouret-Keller, 1985). However, approaches that complicate and critique notions of fixed identity more often come from "interactional" sociolinguistics.

Interactional work in sociolinguistics considers identity as emergent in discourse, dynamic throughout an interaction, and expressed in relation to other perceived identities (Bucholtz and Hall, 2005; Eckert, 2000; Coupland, 2007). The concept of identities formed through interaction in language has a history in socially-oriented linguistics through the study of *speech acts* (Austin, 1962). The ability of language to create and change identities is most clearly illustrated by performative speech acts, such as pronouncing the name of a ship or declaring two people to be married. This work has had influence in gender and sexuality studies (Butler, 1990), but increasingly within sociolinguistics as well. Eckert (2012) describe work using such a perspective as a "third wave" within sociolinguistics. This is in contrast to the first wave of variationist work that often viewed identity as static, given variables. The second wave looked for the impact of locally specific identities, instead of simply broad demographic categories, but did not fully acknowledge the role of language in constructing those identities, as work in this third wave does. Work in sociocultural linguistics has also studied how forms of language and identities are often co-constructed to serve those with power in society (Barrett, 2014; Rosa and Flores, 2017; Charity Hudley et al., 2020). For example, stigmatized ways of speaking are often cast onto racially marked bodies under colonialism and white supremacy. See Section 10.2.3 for comments on extending the work in this thesis toward similar social justice aims.

Two concepts from sociocultural linguistics are particularly relevant to this thesis. The first is *indexicality*, the notion that language use "points to" the social context in which it is often used. The second concept is that of *language ideologies*, beliefs about language and language users, including what language should be used for, what is "good" language and who speaks properly or improperly.

Language use plays a role in the associations made with particular identity groupings. The ability of language to reference certain social contexts is referred to as *indexicality*, a concept originating from the work of nineteenth-century semiotician CS Peirce (Atkin, 2013). For example, the satirical tagline of a Reddit group organized to make fun of and antagonize Tumblr content, r/TumblrInAction, at one point sarcastically read "Seen a horribly oppressed transethnic otherkin blog their plight? Wept at how terrible it is for the suffering of multiple systems to go unheard every day? Been unable to even live with the thought of the identities of someone's headmates being cisdenied? Then you've come to the right place!" Particular terms here, often reformulated to not make sense, are meant to refer to, to *index*, the language often found on Tumblr. This language is particularly associated with the social identity of "tumblrinas", idealized prototypical young, female Tumblr users who care deeply about social justice issues (LaViolette, 2017). Note that indexical meanings are not stable, but can change and be objects of contention from groups with different interests (Silverstein, 2003; Yoder and Johnstone, 2018). In Chapter 8, we wish to measure and track associations made with how straight and queer characters are voiced and portrayed in fanfiction. Following Ochs (1992)'s model of indirect indexicality between language features and identity, we relate individual words used to describe these characters first to topics learned in an unsupervised manner, and then to identity labels given by fanfiction

authors.

Another relevant concept developed in sociocultural linguistics is that of *language ideologies*, representations that arrange and mediate language and people in a social world (Schieffelin et al., 1998). These beliefs about language include what language should be used for, what is "good" language and who speaks properly or improperly. The online communities studied in this thesis, Tumblr and fanfiction, are known to have high LGBTQ representation and are known as spaces that are more accepting of marginalized identities. Language is often used in both of these spaces with a focus on emotionality, or from a common expression on Tumblr, "the feels". This ideology of language to express and develop emotion fits what cultural theorist Raymond Williams identifies as "structures of feeling", which often include art, that form "alternative hegemonies" as options outside of dominant narratives about gender and sexuality (Philips, 1998). In Chapter 7, we found emotionality as a marker of fanfiction authors changing relationships from the original source in mainstream media. In Chapter 8, we explore how language is used in the community of fanfiction of writers at a large scale to describe and voice characters of different sexualities.

Online settings are a natural fit to this interactional perspective; without a physical presence, individuality can only be constructed in discourse. This thesis studies how such interactions online give meaning and shape notions of identity. This perspective of identity as socially constructed and dynamic shaped choices made in the design of experiments using computational modeling, for example the focus on self-presented identity in Chapter 4 and the examination of identity portrayal of characters signalling relationship changes (Chapter 7).

2.1.2 Sociological perspectives

Work in sociocultural linguistics often draws on theories developed in disciplines that focus more specifically on social structures. These areas include sociology, anthropology, cultural studies, philosophy, gender and sexuality studies, and area studies. Much of this work converges around the *social construction of identity*, the idea that identity is formed in particular social and historical contexts and that these contexts ascribe salience or difference on biological or physical features, rather than particular differences being "natural" or "innate" (Allen, 2010). This perspective destabilizes notions of categories such as race and gender and shifts focus to the history and power relations that lead societies to "make people up", that is make up distinctions that matter (Hacking, 1986). Examples of this include the invention of "scientific" racism in post-Enlightenment Europe, which helped justify colonialism, and the move from classifying sexual acts to making those acts parts of identities of sexual deviance in nineteenth-century Europe (Foucault, 1990). Though social constructionism posits that dimensions of identity are not as tied to relevant physical differences as we might think, this does not make them less "real", as anyone who has experienced differential treatment based on race, gender, sexuality, or ability could confirm. At a large scale, particular social relations in history have given rise to marking differences between groups as relevant, such as "racial projects" that produced white supremacy and an emphasis on skin color to justify exploitative systems such as colonialism and slavery (Omi and Winant, 2014). At a smaller scale, the social construction of identity emphasizes how interpersonal interaction gives rise to our sense of what identities we have and others have, what associations and implications are wrapped up with those identities. A classic example

of this from philosopher Louis Althusser is that a criminal "comes into being" when a police officer yells, "Hey you!". Language and other interactions, physical or semiotic, can produce and construct identities. Goffman (1959)'s classic metaphor of the self as a performance on a stage draws attention to how individuals *produce* their own identities, to a degree that can be changed or understood within societal norms. This thesis attempts to use quantitative and computational techniques to study how this is happening in online spaces.

Though notions of identity designed to support Western, patriarchal, capitalist hegemony separate out notions such as race, gender, sexuality, and class, people's lived experiences cannot cleanly be separated into the effects of such categories. There are particular oppressions and privileges at the intersections of social identities, a concept develeoped in critical race theory by Crenshaw (1989) as *intersectionality*. This thesis attempts to integrate this perspective by looking at how identity is presented in online spaces with a bottom-up approach of what identity labels are being used (Chapter 4), though we do separately consider the effects of some identity dimensions.

The social constructionist view of identity is not without critique. Though it began as an attempt to focus on the inequitable social power structures that are enabled by racism, sexism, homophobia, classism, and ableism, the notion of a fluid, presented identity can be unhelpful and unrealistic to those who experience oppression based on their identities. Africana Studies scholars Saucier and Wood (2016) critique the social construction of identity and particularly Omi and Winant (2014)'s racial formation theory as unable to explain and confront the anti-black racial violence that emerges out of the quite stable structures of white supremacy and anti-black racism (see Charity Hudley et al. (2020) for a discussion of this in light of sociocultural linguistics). Others note that a focus on "identity" and identity performance as personal actions and interactions avoids confronting larger, systemic social structures and inequalities. This individualization is a common tactic in neoliberal capitalism, distracting from widespread, systemic oppressions by casting them as personal issues and responsibilities (Fisher, 2009).

Gender studies, trans and queer theory The online communities from which data was gathered in this thesis, Tumblr and fanfiction, are known for large LGBTQ communities, and thus we look particularly to theories of identity in queer theory and gender studies. Butler (1990) describes gender as the "repeated stylization of the body in culturally intelligible ways", drawing attention to the social construction of many of the notions assumed to be "innate" about gender, and even the social construction of sex when medical professionals pronounce a baby to be a boy or girl based on criteria that can be ambiguous and uneven. This builds on Foucault (1990)'s classic account of how hegemonic power is reinforced by increasing regulation and the creation of sexual norms and deviations. Working within literary theory, Halperin (1990) and Sedgwick (1990) develop these theories more specifically for gay and lesbian experiences with an eye toward language. This thesis adopts this theoretical view, but does explicitly focus on the embodied, physical, and material implications of sexual otherness and marginalization, especially for transgender people (Stryker, 2004). See Chapter 10) for a discussion of how the work in this thesis builds toward such work with more direct impact on online users.

Applications in online spaces The notion of "performed" identity is often used as a framework to describe self-presentation (Goffman, 1959; Butler, 1990), and this framework has been frequently applied in social media analysis. Bullingham and Vasconcelos (2013) applied Goffman's performance metaphor to personas on Second Life and found that users often present similar online selves (even in completely constructed avatars) to their offline selves—but with selective edits. Working with Tumblr blog descriptions, Oakley (2016) found that LGBTQ Tumblr users use labeling practices that challenge dominant binary conceptions of gender and sexuality, and which may enable a "performed" self to more closely match their perceived "true" self.

2.1.3 Social identity theory

The social construction of identity also relates to social identity theory in social psychology, which stresses how social groupings shape individual identities and how the nature of these social groupings have a different influence on interaction than do collections of individual identities (Tajfel, 1974). This background is especially drawn on in considering how identity interacts with community in Chapter 5.

2.2 Computational approaches to the study of identity

This thesis uses computational tools to investigate the presentation of identity in language in large datasets, and thus a review of relevant computational work is necessary. Much of this work is not "critical", in the sense that it takes categories and values of identity as given, stable, individual, and absolute. This fits well within a data science methodology in which someone's identity is simply a value in a column in a table of data. I argue that this sort of work often reinforces existing identity-based inequalities and can be harmful to those who do not fit cleanly into the neat values of identity in a spreadsheet, such as transgender people or people from smaller ethnic groups, who already face considerable marginalization. Though participants in such work are not usually intending to be malicious, it should be recognized that such computational work is a natural progression from the roots of many of the quantitative social sciences in which quantifying identity has been a central technique for aims of nationalist, racist, queerphobic classification and control (Birhane and guest, 2021; Hanna et al., 2020). This history demands a critical stance toward the methods, purposes, and implications of computational work on identity. Today, techniques such as user attribute inference are primarily used for personalized, targeted advertising, which is a major funding source of large international tech companies. Such an aim fits within a neoliberal capitalist agenda that views individuals (constructs subjects) along one primary dimension: potential consumers. This value is at the expense of other possible values that are not about making money. User privacy, bias and fairness, or challenging racism, queerphobia, patriarchy and other power structures that reinforce historical inequalities are all values that are generally sidelined for capitalism aims.

This dissertation also posits that by not taking the social construction of identity seriously, our models as computer scientists do not fit the reality of identity categories being shaped, reinforced, and challenged in dynamic online contexts. For example, Nguyen et al. (2014) find that the degree to which someone presents gender associations affects tweet text more strongly than

which gender they present. However, there is some work that does allow for more complexity in modeling identity, and some that, like this thesis, explicitly draws on theories from the humanities and social sciences to inform its methodology. And of course, there is no firm line between these trends of work. I review both types of computational work that involves identity in this section, organized by specific area of study.

2.2.1 User attribute inference

User attribute inference, sometimes known as demographic inference or author profiling, is computational work that predicts attributes about people from digital trace data. This data includes voices, images, or text such as tweets (Chen et al., 2015). Predicted attributes are usually broad demographic traits such as gender, age, race, and ethnicity. Though such work is designed for practical applications such as targeted advertising, digital forensics, or population surveying, it must be noted what such a set-up assumes about identity. For one, it assumes that behavior recorded in data is simply a function of people's identity. There is always uncertainty in machine learning and applied statistics, but for this prediction to work at all behavior must be different based on gender, race, or sexual orientation, or at least correlate with such differences.

Drawing on conceptions of identity from the humanities and social sciences, this thesis takes issue with such work on a number of fronts. First, it ignores that different factors of identity are more or less salient in different situations (Coupland, 2007). For example, gender and race certainly matter in educational contexts, but likely more relevant to what each person says in the classroom is their role: teachers, students, or administrators. Computational researchers often classified attributes such as age and gender with a popular blog corpus (Schler et al., 2006), but what was a greater discriminator of the text produced in these blogs was occupation, often the topic of discussion. Second, such prediction ignores agency and clashes with the social constructionist position that identity is constructed in interaction. People know associations with social groups in society (such as how men and women are "supposed to" talk), and consciously or unconsciously, present these in varying degrees depending on the situation and how much they want to (Nguyen et al., 2014, 2016). This is particularly relevant online, where there is more agency over self-presentation than in in-person interactions, for example in choosing a profile picture to represent oneself visually. These online contexts thus skew toward what people intentionally choose to "give" versus what they unintentionally "give off" about themselves (Goffman, 1959). Third, such work reinforces the idea that whatever values it chooses for identity are stable, mutually exclusive, and the only ones possible. Trans-exclusive work in computer vision that predicts gender from images (Keyes, 2018), even though gender is not a physical feature, is an example of the problems that such categorizations have.

User attribute inference from text. There is a body of work from NLP researchers predicting attributes of people from behavior on social media. Twitter is a popular platform for this work (Volkova et al., 2015). Other than predicting attributes of those who write tweets, Jurgens et al. (2017) predict attributes only from those who write *to* users. Reddy and Knight (2016) develop methods for obfuscating gender in social media text, taking into account the agency of authors.

Other related areas in NLP include authorship attribution and text forensics in which the goal

is to identify specific authors of text. Style transfer also often considers attributes of users who produce text as aspects of a demographic style, which again assumes that writing is a function of user attributes (Kang et al., 2019). There is also interest in learning how people with different demographic attributes talk or write in developing persona-based dialogue agents (Li et al., 2019).

One of the first widely-used datasets for user attribute inference of online text was the Blog Authorship Corpus (Schler et al., 2006). Informative features from a linear model predicting binary gender found associations with males and "tech" words such as "linux" and females with more chat speak and pronouns, though these features are confounded with other sources of variation such as occupation. Most models were able to predict gender with 70-90 percent accuracy, and these rates are similar to those from Twitter data (Nguyen et al., 2013; Burger et al., 2011) Nguyen et al. (2014) find that there is much variation in how much users emphasize gender and thus how accurate their classifiers are, which complicates this task and points to the limitations of assuming gender affects the language of all people in the same way in all situations.

More recent work predicting user attributes from text focuses on multilingual corpora of tweets (Rangel et al., 2017) or on language generation (Kang et al., 2019).

User attribute inference beyond text. Age, gender, race, or even sexual orientation are commonly predicted from facial images, often erasing and mislabeling anyone who does not fit whichever values are chosen as possible for these dimensions. Keyes (2018) illustratees that binary gender classifiers erases non-binary people and often misgenders transgender people. Moreover, this work problematically reinforces the assumption that such features of identity are so "natural" that something like sexual orientation can be read off of someone's face (Wang and Kosinski, 2018), which also has a long racist and queerphobic history in the pseudo-science of phrenology (Browne, 2015; Hanna et al., 2020; Birhane and guest, 2021).

Much work has also used features from multiple modes to predict demographic attributes of images. Fang et al. (2015) use text features along with profile images and post photos to predict attributes on Google Plus. Others use network information, relying on the property of *homophily*, that users with similar attributes are more likely to have links in a social graph (Gong et al., 2014; Yang et al., 2017; Kim et al., 2017c).

Signals of self-presentation from different modes work in concert; Chapter 4 uses both text and profile image data to predict reblogging behavior. An important difference in our work is that we focus on self-presented labels and information instead of inferring labels as in user attribute inference. There is much variation in how user attribute inference assigns gold labels in their datasets to be predicted, such as from names, user-provided labels (structured or unstructured, required or not), from human annotation or from another modality such as photos if the task is predicting from text. User attribute inference ignores the possible differences in self-presentation across modalities and assumes that they all predict the one true label for an individual, which again theorizes identity as stable, psychological and internal instead of dynamic, social, and constructed with agency in interaction.

2.2.2 Computational social science

Beyond researchers who focus exclusively on language, identity often relates to work in computational social science. Such work often incorporates the impacts of multimodal signals, network structure, and community membership in constructing representations of individual or group identities.

This work commonly focuses on the impact of platform design on factors of self-presentation and community identity (LaViolette and Hogan, 2019). We measure both the influence of text and image self-presentation on reblogging in Tumblr (Chapter 4) and from communities inferred from network structure (Chapter 5).

Though predicting identity labels in user attribute inference is more common as a task in related computational work (see previous section), there have been a few computational studies that predict the effects of online self-presentation. Wang and Jurgens (2018) examined how the presentation of gender on Reddit, StackExchange, and Wikipedia affects reactions of support. Bareket-Bojmel et al. (2016) examined how self-enhancing and self-derogating posting strategies affected responses to users with different goals on Facebook. Vedres and Vasarhelyi (2019) find that behaving in ways that are typical of male or female users on GitHub has more of a relationship with success than the "actual" gender of users. In Chapter 4, we use content propagation as a measure of how particular forms of self-presentation come across in the context of Tumblr.

The well-known property of homophily suggests that users who share attributes are more likely to be connected within social networks. Gong et al. (2014), for example, find that inferring user attributes improves link prediction performance in Google+. However, identifying what relationships between identity labels are meaningful in context is not trivial. Bucholtz and Hall (2005), for example, identify authenticity and legitimacy as axes other than sameness/difference on which speakers position themselves. Piergallini et al. (2014) discuss challenges with modeling the complex alliance systems in online street gang discourse. Chapter 4 builds a framework for identifying combinations of identity labels that hold social currency in their association with content propagation, while Chapter 5 tests for the influence of community alignment.

Though this thesis fits more closely with work that is analyzing social behavior, issues of identity are closely linked with research on ethics, bias, and fairness in machine learning and AI as well. Such work has considered how NLP operationalizes gender (Larson, 2017; Strengers et al., 2020) and race (Field et al., 2021) while considering how important conceptions of these categories are to the language technologies that affect users. Similarly, Cao and Daumé III (2020) challenge folk notions of binary gender that are operationalized into trans-exclusionary coreference systems. They draw on Archive of Our Own for data using non-binary pronouns, the same site of data that we study in fanfiction projects (Chapters 7 and 8).

Online communities Research that focuses on how online communities function often overlaps with computational social science (CSS), since data of social interaction for CSS is usually drawn from online settings. The data in this thesis are drawn from recordings of interaction in online communities, so such work is very relevant. boyd (2007)'s conception of online spaces as "networked publics" and Renninger (2015)'s development of this idea in counter-hegemonic "networked counterpublics" is particularly relevant in Chapter 5, where we explore the nature of community on Tumblr through quantitative tests of what influence content propagation. Such

work is also relevant in investigating fanfiction's relationship with the broader LGBTQ social movement, which challenges dominant queerphobic narratives (Chapter 8).

Critical data studies and science and technology studies Work from science and technology studies usually adopts a critical perspective that challenges how identity is theorized in machine learning and data science, and what the implications are of "flattening" social constructs like identity into "given" realities in datasets (Gitelman, 2013; D'ignazio and Klein, 2020). Such work also situates data science within the broader history of post-Enlightenment classification of peoples for control (Bowker and Star, 2000). Work in science and technology studies has built a critical framework for arguments for challenging the treatment of identity and other social structures in computer science (Hanna et al., 2020).

2.2.3 Digital humanities

Theorizing about identity in the digital humanities is much more likely to adopt a critical approach since it is more closely tied the humanities and English departments, where such a conception of identity is dominant. This includes computational work on literary texts that considers when machine learning classifiers fail as crucial to understanding the instability of such broad categories as race (So et al., 2019). Drucker (2011)'s reframing of data as "capta" to stress the role of the researcher in recording or capturing data about people is one example, and Bode (2020) reinforces the importance of critical, political, and ethical issues in computational literary investigation.

We attempt to adopt this perspective in this thesis's work on fanfiction (Chapters 7 and 8), considering how identities are presented dynamically and associations are made in discourse with unsupervised topic models, not predetermined by researchers.

2.2.4 Computational sociolinguistics

In its aims to understand social dynamics in online communities and how identity is expressed in language in these communities, this thesis aligns with computational sociolinguistics, an area of research that uses computational tools to study social meaning in language (Nguyen et al., 2016). Though work from computer scientists in this area often fits more closely with uncritical user attribute influence, there has been work that givens more attention to the construction of social context, dynamism instead of immediately assuming latent, personal attributes.

This includes Bamman et al. (2014a) who challenge notions of binary gender in social media data and Nguyen et al. (2014)'s recognition that attributes such as gender have large variance in the amount of expression on Twitter.

See Nguyen et al. (2016) for a survey of computational linguistic work that engages with social data, including research on language and social identity as well as language and social interaction. They conclude that a central challenge in such work is the dynamic relationship between social and linguistic variables, evident in speakers' linguistic agency in presenting identity. This thesis works to address this challenge by incorporating the dynamic presentation of identity in language into computational models.

2.3 Conclusion

The primary audience of this thesis is researchers in NLP and computational social science, but it is intentionally interdisciplinary. Many areas of research are relevant to the study of identity and language in online contexts, including sociocultural linguistics, computational social science, digital humanities, and others reviewed above. These areas are not always in conversation, however, and often rely solely on theories and methods from their respective disciplines. There is a particular bifurcation between those who adopt the social construction of identity as dynamic, contextual, and emergent in interaction, who generally use qualitative methods, and those who use computational approaches and generally treat identity uncritically as natural, individual, and stably preceding interaction. We attempt to bridge these communities by pairing the social construction of identity from the humanities and social sciences with computational methodology familiar to NLP and computational social science. In this way, we hope to move the field beyond user attribute inference to discover how forms of identity in language are shifting and forming online, and how this identity presentation relates to social interactions and outcomes.

Chapter 3

Data

This section provides an overview of the contexts in which data was collected for this thesis. These include Tumblr, a social media site; and Archive of Our Own, a fanfiction repository. We also review related work relevant to these sources.

Specific details on the datasets used in projects can be found in their respective chapters.

3.1 Social media: Tumblr

3.1.1 Background

With its design and user culture, Tumblr provides an ideal platform for studying the relationship between identity and social interaction. As of December 2018, Tumblr had 167 billion posts over 450 million blogs.[1] The site is advertised as a place to "express yourself", "be yourself", and "connect with your people". [2] Self-presentation is important on Tumblr, as there are no centralized communities to officially belong to or required fields for listing personal information. Thus, the only information other users will know about a user from their blog is what is consciously expressed. Finally, Tumblr users are known for being interested in social justice issues which often intersect with identity; issues of gender and sexuality, for example, are prominent themes in content on Tumblr (Oakley, 2016).

One of the main forms of interaction on Tumblr is reblogging posts (analogous to retweeting on Twitter), which we choose to relate identity presentation to in Chapter 4. When a user chooses to reblog content, it appears as a post on one of their blogs. The user can also expand on the reblogged content by adding responses or using their own set of hashtags on the resulting post. This content is then disseminated to the followers of that blog on users' individualized "dashboards" (feeds of current activity).

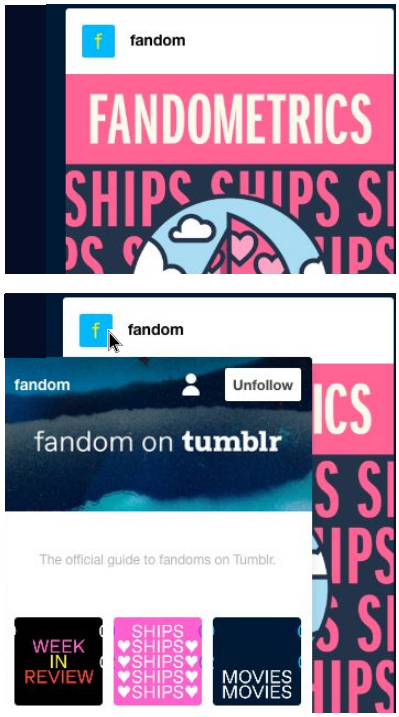

Figure 3.1: Example of information displayed about a Tumblr blog when mousing over the profile image on a post.

3.1.2 Self-presentation on Tumblr

Tumblr users engage in self-presentation through a variety of means in connection with their personal blogs. Most of these self-presentation traces are either visible or accessible when other users make decisions to reblog posts from that user.

Affordances for self-presentation on Tumblr blogs include:

- **Blog descriptions**. Users often use the free-text blog description field to report short spans of identity information such as age, gender, sexual orientation, interests, etc. (Oakley, 2016). This blog description appears on the blog as well as with a mouseover of a user's profile image next to posts on users' dashboards (Figure 3.1).
- **Profile images**. Each blog has either a default profile image or an image chosen by the owner of the blog. This profile image appears on the blog and also with posts appearing on other users' dashboards.
- **Blog names**. Each blog has a name which is often descriptive of the content or user behind the blog. This blog name appears in the URL of the blog, as well as with a mouseover over the profile image in the dashboard.
- **Posts and reblogs**. The most recently posted material is displayed on the top of the blog. This content would also play a role in the characterization of a user to other users viewing

[1] https://www.tumblr.com/about

[2] `https://www.tumblr.com/about`, accessed 9 September 2021.

that blog. Users know that content they reblog is posted on their blog and becomes part of their own self-presentation, which is one of the motivating factors for the influence of self-presentation on sharing content.

- **Other blog aesthetics**. Tumblr users can choose from a myriad of "themes" to customize the appearance of their blogs. Aesthetic choices include background images, wallpaper, fonts, borders around posts, and even mouse changes and default music for blog visitors.

Users have the opportunity to make decisions about sharing content either when viewing another user's blog or when viewing content on their own dashboard. Information about the user is readily available when viewing a blog, and some information is available with a mouseover of the profile image next to a post on the dashboard (Figure 3.1).

3.1.3 Prior research on Tumblr

Related social media sites Twitter and Facebook have been the focus of much more research than Tumblr. But Tumblr has still been the focus of a wide variety of research (Attu and Terras, 2017). This research can roughly be divided into a) social science research, often focusing on identities expressed on the platform, and b) computational research that has focused in a practical way on network analysis, recommendation systems, and search tools for identifying relevant or problematic content. Our research examines the effects of identity self-presentation on interaction on Tumblr, and in doing so uses computational tools for social science investigations.

In their survey of Tumblr research, Attu and Terras (2017) find that sexuality and other identity issues are some of the most commonly studied subjects on Tumblr. They find that many of these studies use qualitative methods to examine how identity is expressed through Tumblr content. Fink and Miller (2014), for example, use auto-ethnographic dialogue to relate how trans and queer Tumblr users created an artistic space that challenged dominant straight cisgender norms. In their interview- and content-based research on Tumblr users posting NSFW selfies, Tiidenberg (2014) similarly finds that Tumblr users create a space for body and sex positivity outside what is deemed "sexy" by mainstream society. Haimson and Hayes (2017) focus on gender transition blogs and find that transgender users used words indicating negative affect and fewer words related to family after the divisive 2016 US presidential election. With the goal of informing sex educators and clinicians, Zeglin and Mitchell (2014) choose Tumblr to research public understandings of sexuality to contrast with proposed theoretical models of sexuality. They find an emphasis on sexual identity issues on Tumblr over other aspects of sexuality, such as intimacy.

In computational research, Chang et al. (2014) aim for a statistical overview of the site in comparison with other social media platforms and find that multimedia content is more prevalent on Tumblr. In a similar large-scale statistical analysis, Xu et al. (2014a) find that more than half of Tumblr posts have no tags.

In the space of recommendation systems, Xu and Lu (2015) infer user interests over topics (Louvain clusters of tags) using both homophily from a reblog network and tag co-occurrence, while Kozareva and Yamada (2016) use a collective matrix completion method for post recommendation to Tumblr users.

Some researchers have addressed issues specific to Tumblr's young demographic. Xu et al. (2014b) build information extraction methods for early detection of planned civil unrest events

among activists on the site. Milner (2013) use methods of critical discourse analysis to explore voicing in populist memes during the 2011 Occupy Wall Street movement. Tumblr's problematic pro-anorexia community has also been a focus. Choudhury (2015) characterize "pro-anorexia" against "pro-recovery" communities and find lexicon markers that provide accuracy in identifying pro-anorexic content, while Chancellor et al. (2017) describe changes to hashtags used by the pro-anorexia community to avoid censorship.

As for computational work that specifically addresses identity in Tumblr, Grbovic et al. (2015) build classifiers for user (binary) gender using names matched to a baby name database as gold labels. They construct user profiles from blog description and title unigram and bigram features, as well as from tag use and liking, following and reblogging behavior, to predict these labels.

Our work attempts to bring themes of identity and self-presentation that are well-studied in qualitative Tumblr work into a computational analysis of the effects of self-presentation on content propagation. Specifically, we take the common view in qualitative identity research that focuses on the discursive *construction* of identity. With this lens, we focus on the effects of self-presented identity labels in a large corpus instead of labels inferred by an automated system.

3.2 Narrative: Fanfiction on Archive of Our Own

This section describes the second source of data used in this thesis: fanfiction drawn from Archive of Our Own[3]. This data includes the story narratives posted on Archive of Our Own, as well as the rich metadata and tags that authors provide about these stories.

3.2.1 Fanfiction terminology

Fanfiction has an extensive vocabulary of specific jargon. In an attempt to make this work accessible, I have avoided much of this terminology. However, some terms' precision and efficiency make them useful to include. Learning the terminology of the fanfiction community is also useful to understand community culture and values, visible through which terms have developed specific meanings and which concepts are important enough to have developed particular terms to express them. Here is a list of fanfiction terminology that may be found in this thesis:

1. **canon**: the original media series on which fanfiction is based.
2. **fandom**: fan communities organized around particular media series.
3. **fic** or **fics**: fanfiction or fanfiction stories.

3.2.2 Prior research on fanfiction

Fanfiction has been extensively studied with qualitative methods, generally under a cultural studies lens as in the work of media scholar Henry Jenkins, who frames fanfiction as "participatory

[3] `http://archiveofourown.org/`

culture" (Jenkins, 1992). Currently, there is expanding interest in quantitative exploration of fanfiction in part due to the enormous volume of stories accessible online. Over 7.8 million stories are hosted on the fanfiction website Archive of Our Own alone, as of July 2021.

Though specific divergences from canon in fanfiction may seem like trivial alterations to the source material, trends in fanfiction can also reflect broader sociological and cultural shifts. For example, Milli and Bamman (2016) used computational techniques to examine which characters were emphasized in fanfiction compared to canon. They found an emphasis on female characters in fanfiction congruent with an understanding of fanfiction as a female space (Lothian et al., 2007). Many qualitative studies speculate on the reasons why fans might choose to shift the original canon work. One analysis through a literary lens presents fanfiction as a means for authors to practice technical skills like characterization (Kaplan, 2006). With a more sociological outlook, Goodman (2007) frames fanfiction as a space for fans to craft their own ideal of the canon while examining, critiquing, or even outright defying the original work. One important way this divergence occurs is dramatically higher representation of LGBTQIA+ characters and pairings in fanfiction than in mainstream media, leading researchers to label fanfiction a "queer space" (Lothian et al., 2007; Fazekas, 2014).

3.2.3 NLP work on fiction and fanfiction

Many computational tools and studies have focused on fiction and characterization (Iyyer et al., 2016; Rahimtoroghi et al., 2017; Rashkin et al., 2018). Our pipeline for extracting text relevant to characters in fanfiction (Chapter 6) builds on and extends prior NLP work. In particular, Bamman et al. (2014b) developed BookNLP, an pipeline for character identification, coreference, and feature extraction in novels. We adapt this pipeline to work more specifically with fanfiction for our purposes in Chapter 6.

Other NLP work on characters in fiction includes that of Kim et al. (2017b), who model emotion in character relationships, and Kim and Klinger (2019), who annotate character relationships in a set of 19 fanfiction stories. In contrast, we focus on capturing how relationships are presented differently from canon, a question that lends itself to a much larger corpus of fanfiction.

Directly relevant to our project in Chapter 5, Fast et al. (2016) measure gender stereotypes in the online fiction community of Wattpad. They find that gender stereotypes such as violent, sexual men and domestic, submissive women are largely reproduced.

Data from fanfiction has been used in NLP research for a variety of tasks, including authorship attribution (Kestemont et al., 2018), action prediction (Vilares and Gómez-Rodríguez, 2019), fine-grained entity typing (Chu et al., 2020), and tracing the sources of derivative texts (Shen et al., 2018).

Part I

Identity, Community, and Interaction on Tumblr

Chapter 4

Self-presentation and Interaction on Social Media

4.1 Abstract

Research on content propagation in social media has largely focused on predictive features from the content of posts and the network structure of users. However, social media platforms are also spaces where users present their identities in particular ways. How do the ways users present themselves affect how content they produce is propagated? In this chapter, we address this question with an empirical study of interaction and self-presentation data from Tumblr. We use a pairwise learning-to-rank framework to predict whether a given user will reblog (share) another user's post from features comparing self-presented textual and visual identity information. We find evidence that alignment in identity presentation is associated with content propagation, as these features increase performance over a baseline of content features. Interpreting learned feature weights on self-presented text identity labels, we find that users who present labels that match or indicate shared interests and values are generally more likely to propagate each other's content.

4.2 Introduction

Social media is a space where users not only share content, but also construct identities and position themselves in relation to others. However, it is difficult to directly measure how presentations of identity come across to others at a large scale. In this chapter, we develop a framework to identify such self-presentation effects on a primary form of interaction on social networks: content propagation, sometimes known as information propagation.

Prior work quantifying patterns of social media content propagation has not addressed a relationship to self-presentation. Most existing studies rely primarily on content and network features (Naveed et al., 2011; Zhang et al., 2014, 2016). However, content propagates through a social network from individual decisions to share others' posts. At this local level, content propagation is an interaction between users. Users with similar attributes are known to have stronger network connections (the property of homophily). In this way, self-presented identity attributes may af-

fect content sharing as a form of network connection. This chapter investigates this connection for two reasons: in order to broaden understanding of factors associated with content propagation, and to provide an experimental paradigm in which reactions to self-presentation practices can be investigated.

We use data from Tumblr, a blogging and social media platform. Sharing content is one of the primary modes of interaction on Tumblr; more than 90% of posts are "reblogs" of other posts (Xu et al., 2014a). Identity construction is also an important part of the participatory culture on Tumblr. Users on Tumblr each have a personal blog, an individualized artifact that reflects a user's identity (Hogan, 2010). Tumblr's multimedia content, personalized blog layouts, and affordances for users to maintain multiple blogs without being tied to a real name have created a unique environment for identity expression without many of the social pressures found on other social media sites (Devito et al., 2018). Talk about identity issues such as gender, sexuality, and ethnicity—as well as their intersection with media, culture, and fandom—is common on Tumblr (Fink and Miller, 2014). This makes the identity positioning of users who create content especially relevant on the platform. For these reasons, we expect users' self-presentation of identity to play a role in how content is propagated on Tumblr. We investigate this role through two research questions:

> **RQ 1:** Is there evidence of an association between identity alignment and content propagation?

To explore the possibility of this effect, we construct a classification task predicting whether a user will reblog another user's posts given how both users express their identity in profile images and text blog descriptions. We find that incorporating identity-related features improves classification performance when used in addition to post content features, and that profile images and text blog descriptions provide complementary effects. This shows evidence that identity presentation is associated with content propagation.

However, this result does not specify the nature of this effect: features signifying similarities, differences, or some other interaction between text or image features could be especially predictive. We therefore explore a further research question:

> **RQ 2:** What is the nature of the association of identity alignment with content propagation?

We investigate this question with logistic regression models trained on identity comparison features between text blog descriptions. Tumblr users often use these free-text blog description fields to provide identity labels such as 'girl', 'canadian', or 'intj' (Oakley, 2016). We find that providing identity labels that indicate shared values or experiences generally increases the likelihood of users reblogging.

4.3 Theoretical motivation, research context, and prior work

We conceive of online self-presentation as the construction of an artifact that reflects identity in the particular space of Tumblr (Hogan, 2010). This artifact, a Tumblr blog, contains both linguistic and visual elements of self-presentation: written blog descriptions, profile images, and multimedia posts. Much of this self-presentation consists of symbols that point to culturally specific

understandings of types of people. For example, being a female fan can be expressed in a blog description explicitly with labels (*i'm a super fangirl*) or more implicitly (♥♥*OBSESSED*♥♥ *with exo*) (Johnstone, 2010; Gee, 2011).

See Chapter 2 for a discussion of this dissertation's theoretical background on identity. See Chapter 3 for background on self-presentation and related research on Tumblr. We review relevant related work on content propagation in the following section.

4.3.1 Content propagation

Content propagation and virality on social media are often predicted from features of individual post content or patterns in how that content spreads through a network (Zhang et al., 2016; Vosoughi et al., 2018). Naveed et al. (2011), for example, predict whether a tweet will be retweeted based on text features from the tweet. Zhang et al. (2014) studied how the network feature of reciprocity is associated with reblogging on Sina Weibo. On Tumblr, where content can take diverse forms (text, photos, videos, audio, etc.), work more often focuses on network-based features. To predict whether a Tumblr post will become viral, Xie et al. (2017) centered their analysis on "early adopters" (the first users to share a post) in combination with content-specific features. Others have examined network features in relation to reblog cascades. Chang et al. (2014) examined Tumblr reblog cascades based on structure of the follower network, while Alrajebah et al. (2017) examine both structural and temporal aspects of reblog cascades on a set of popular posts on Tumblr.

Our work extends this research by additionally viewing content propagation as a social interaction in which the identities of the users involved may play a role.

4.3.2 Self-presentation on Tumblr

See Chapter 3 for background on the different self-presentation affordances in Tumblr, and note that we use text blog descriptions for prediction and interpretation, and profile images for prediction. Though this information is readily available with a mouseover of a post or view of a blog, the analysis in this chapter does not assume that users necessarily check this self-presentation information before choosing to reblog. We certainly do not assume that this information has more of an effect in reblog choices than the content of posts. Rather we test whether the presence of those features has a measurable statistical effect in relation to content propagation when we control for content features.

For this analysis, we use two prominent affordances for identity presentation: profile images and text blog descriptions. In a sample of 1 million blogs that made at least 10 reblogs from June through November 2018, 61.2% had filled in blog descriptions. We sampled profile images for 810,800 blogs from this set of 1 million (the remainder could not be accessed). Of these 810,800, 69.6% provided a non-default profile image.

Note that unlike the conventions in many other social media sites such as LinkedIn, Facebook, Twitter, and Instagram, many of these profile images on Tumblr do not contain images of the user. In a sample of 1,000 profile images from blogs that made at least 10 reblogs from June to November 2018, another colleague and I found that only 29.3% supplied a profile image of a

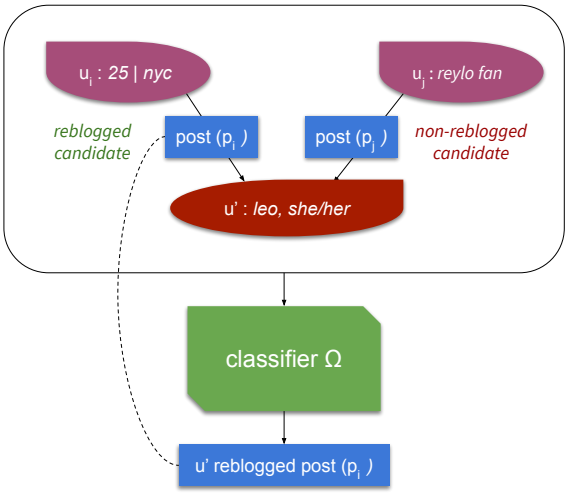

Figure 4.1: Illustration of the pairwise learning-to-rank formulation. From the perspective of user $u^{'}$, we want the classifier Ω to decide whether to reblog post p_i or p_j from different users $u_{i,j}$ that user $u^{'}$ is following. Fabricated examples of text blog descriptions are included in the diagram, though profile images are also used in practice.

person who is likely the user, whereas 42.5% supplied another type of image. This makes conventional gender, age and ethnicity image detection systems not as effective; we use continuous, dense, neural representations of profile images instead (see Section 4.5).

4.4 Experiments

We design an experimental paradigm to evaluate whether similarities and differences in the identity presentation of two users are associated with patterns in their decisions to propagate each others' content (**RQ 1**). This framework allows us to identify patterns in how self-presentation comes across in interaction, specifically associations with content propagation. We train logistic regression, SVM, and neural network models on features that represent identity alignment between users (matches, mismatches, and other interactions) and look for changes in accuracy over a baseline of post content features. A significant increase in accuracy suggests that comparisons and contrasts in self-presentation are relevant for predicting content propagation between users.

We treat the choice of reblogging as a selection problem, where a user is exposed to a wide variety of content and chooses to propagate some posts rather than others (see Figure 4.1). For example, if issues related to sexual orientation are important to a user and they present themselves in those terms, we expect that all else being equal, they would choose to reblog a post from a user who also signals sharing that value compared with another user they follow who does not share that identity framing. We use a pairwise learning-to-rank paradigm to embody such a comparison.

4.4.1 Learning-to-rank formulation

The learning-to-rank method we use in our experiments is a variant of the RankSVM algorithm (Joachims, 2002). The RankSVM algorithm enables the use of traditional classifiers, like support vector machines (SVM), to make pairwise comparisons by considering items in a comparison feature space. Given a set of pairwise post comparisons P, their corresponding ranking labels R, and a classifier $\Omega(X, Y)$ that can be optimized given feature vectors X and corresponding labels Y, RankSVM performs the following transformation:

1. For every pairwise post comparison $\langle p_i, p_j \rangle \in P$:
   (a) Map p_i and p_j into a common feature space $\mathcal{F}$ using feature function $\Phi(p) : p \to \mathcal{F}$
   (b) Create a feature vector representing the comparison by calculating the difference between the feature vectors for p_i and p_j. The resulting comparison feature vector $c_{ij} \in C$ is now in a pairwise comparison space $\mathcal{F}'_C$.

$$c_{ij} = \Phi(p_j) - \Phi(p_i)$$

2. Train a classifier $\Omega(C, R)$.

The trained $\Omega(C, R)$ takes a feature vector in the pairwise comparison space and produces a ranking between the pair of posts: rank 0 for the post the user reblogged, rank 1 for for the post they did not. While SVM with a linear kernel was traditionally used in RankSVM, the algorithm can be extended to other classifiers.

Reblog classification could be formed as a simple prediction task over a sample of posts that a user did or did not reblog. Howerver, users on average reblog fewer than 1% out of all posts from blogs they follow (Chang et al., 2014), which leads to heavily skewed data. The pairwise learning-to-rank formulation addresses this issue by directly representing the reblogged/non-reblogged post comparisons, allowing us to rebalance the skewed dataset in a meaningful way. This is similar to previous work that handles highly skewed distributions in information cascade size prediction by constructing balanced binary classification tasks (Cheng et al., 2014; Krishnan et al., 2016).

4.4.2 Experimental dataset

We use Tumblr posts and text blog descriptions from a data dump ranging from 1 June 2018 to 30 November 2018[1]. Profile images for users were obtained from the Tumblr API.

To construct our dataset of paired reblogged and non-reblogged posts, we first sample a set of 1,000 blogs[2], U, which have reblogged at least 10 posts as a minimum level of activity. For these 1,000 users, we find all users they have followed, F. We collect reblogged posts made by blogs in U from blogs in F after the associated user began following the blog in F. In the case of reblog cascades with content that is reblogged multiple times, we take identity features from the

[1]This period is before the adult content ban was announced in December 2018, so any changes due to this ban are not reflected in our data.

[2]We also refer to blogs occasionally as *users*. While a Tumblr user may have multiple blogs associated with their account, for simplifying purposes, we consider users to be on the blog level, as the identity labels in blog descriptions we examined apply to the user and not the blog.

	Prediction dataset	Interpretation dataset
Identity features	text + image	text
# Users	14,177	34,801
# Reblog prediction instances	228,424	712,670

Table 4.1: Dataset statistics. The prediction dataset contains users who provide both text blog descriptions and profile images. The interpretation dataset contains users who provide a blog description but not necessarily a profile image.

most recent reblogger of the content. These features are from the immediate user who the user follows, whose self-presentation information is more readily apparent than that of the original poster.

For each reblog in this set, we sample candidate non-reblogged posts to act as a comparison in our learning-to-rank framework. Since the details of Tumblr's dashboard ranking algorithm are not public, we assume recent posts from followers likely appear on a user's dashboard. We restrict both reblogged and non-reblogged candidate posts to only be from blogs the user follows. To increase the likelihood that paired non-reblogged posts were seen by the user, we select those that were posted within 30 minutes of each paired reblogged post. Since the user reblogged a post within 30 minutes, there is a greater chance they were active and saw these other, non-reblogged posts from blogs they follow. We sample up to five non-reblogged posts, from unique blogs, for every reblog.

From this initial data collection, we extract two datasets (Table 8.1). For prediction we use identity features from profile images and blog descriptions, and so we filter to users who provide both. For interpretation we only use text blog descriptions, and so we incorporate users who provide blog descriptions but not necessarily profile images into a slightly larger dataset.

Ranking labels are generated by randomly shuffling the order of the posts within each comparison so that the reblogged post appeared as p_i 50% of the time and p_j 50% of the time. However, we want feature weights in the model to be consistent and interpretable (i.e. positive weights indicate a higher likelihood of reblogging). So in practice, we always treat the reblogged post as p_i but flip the sign on the label and features when it should be considered p_j. Each dataset is randomly split into a training/test split of 90% and 10%.

4.4.3 Model hyperparameters

Logistic regression classifiers are trained with L2 regularization; constants are selected using grid search from 10^{-4} to 10^{4} on a base 10 log scale on 10-fold cross-validation on the training set.

SVM models were trained with linear kernels due to the traditional use of linear SVM with RankSVM and the large size of our training set. L2 regularization constants were chosen using grid search from .01 to 100 on a base 10 log scale.

For our neural network, we use a multi-layer perceptron (MLP) over the same feature set as the logistic regression and SVM models. The MLP consists of three hidden layers of size 100,

50, and 32 with ReLU activation in each layer. We train this model with L2 regularization with a constant $C = 10^{-4}$. We used the Adam optimizer (Kingma and Ba, 2015) with $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$. The MLP was trained with early stopping, where 10% of the training data was randomly set aside as a validation set.

4.5 Feature extraction

We describe here the feature function $\Phi(p)$ that is applied to candidate posts p_i and p_j and their associated users in each pairwise comparison. Identity features are extracted from both text blog descriptions and profile images.

4.5.1 Text blog descriptions

We extract identity information from blog descriptions at two levels: (1) what specific identity labels, such as 'trans man' and 'british', are given, and (2) which broader identity categories, such as *gender* or *age*, those labels indicate. Our intuition is that providing similar categories of identity, even if labels are different, may orient users to the platform in similar ways. We use a bootstrapping approach to find labels that indicate identity and to group them into categories for automatic annotation.

Similarly to profile images, neural embeddings could be learned for blog descriptions and various similarity metrics computed for comparison between them. However, for interpreting which kinds of similarities and differences in self-presentation are associated with content propagation, we choose to extract one-hot features.

Bootstrapping delimiters As a convention, Tumblr users often provide identity labels separated by delimiters (such as commas or pipes) in blog descriptions (Oakley, 2016). For example, note the pipes as delimiters in the fabricated blog description "22yo — she/they — too many fandoms" and the periods as delimiters in "andre . nyc . manga". To identify such delimiters, we started with a small list of identity labels manually identified from blog descriptions. We then searched for these labels in a separate set of blog descriptions and found characters in between these terms as potential delimiters. Manually reviewing this list of potential delimiters, we kept those that could function as separators between labels in a list, primarily punctuation and emojis. This resulted in a list of 95 delimiters.

Bootstrapping identity labels To find additional identity labels, we extracted short texts (maximum 25 characters) in between any of these delimiters on the larger set of blog descriptions. Long spans often indicated quotes or other unrelated material. We extracted and ranked identity label candidate *n*-grams in these short segments by frequency, discarding stopwords and other terms that were not indicative of identity.

Choosing identity categories We want categories that are:

1. Popular, and thus relevant on Tumblr.

2. Largely about the user, not the content. Our goal is to identify how users position themselves, not the main topics of a blog.
3. A relatively limited set of possible values so they can be accurately identified outside of a training set.

Guided by Bucholtz and Hall (2005), we manually group labels into categories that encompassed broad demographics as well as labels more specific to Tumblr (such as fandoms and interests). Some of these popular labels are creative and not well-known outside of Tumblr, such as 'phans' for fans of Phil Lester, a YouTube personality, and 'stans' for obsessive fans. Others are more specific to Tumblr but intersect with larger demographic categories, such as 'cishets' for cisgender heterosexuals. Our final list of 11 identity categories, with example labels, is shown in Table 4.2.

Identity Category	**Label Examples**
age	21, seventeen
ethnicity/nationality	latina, haitian
fandoms	shipping, crossovers, star wars, lotr
gender	woman, husband, mtf, nonbinary
interests	photography, running, makeup
location	australia, london, socal
personality type	intp, slytherin
pronouns	she/her, they
relationship status	married, single
sexual orientation	bi, lesbian, aro-ace
zodiac	virgo, capricorn

Table 4.2: Our identity categories with examples of labels.

Annotating identity categories A colleague and I manually annotated for the presence of these categories in a random sample of 1200 blog descriptions, from which we pulled 100 samples for a development set and 100 for a test set. From our manual annotations, we built regular expressions to automate annotation. On the development set, we iteratively added or refined regular expression patterns for each category. Sometimes it was unclear which category a label indicated, such as 'LGBT' indicating gender, sexual orientation, or both. In these cases we added the pattern to all categories that may be indicated (both gender and sexual orientation for 'LGBT').

For a subset of identity categories, we compared our regular expression annotation approach with Naive Bayes and SVM models. We trained on unigrams and character 1-4grams in the blog descriptions to predict which categories are present. We found that the regular expressions performed better on the test set (over 80% F1 average compared with under 50%), likely due to the small amount of available training data. Percentages of users in our interpretation data training set who present each category are shown in Figure 4.2.

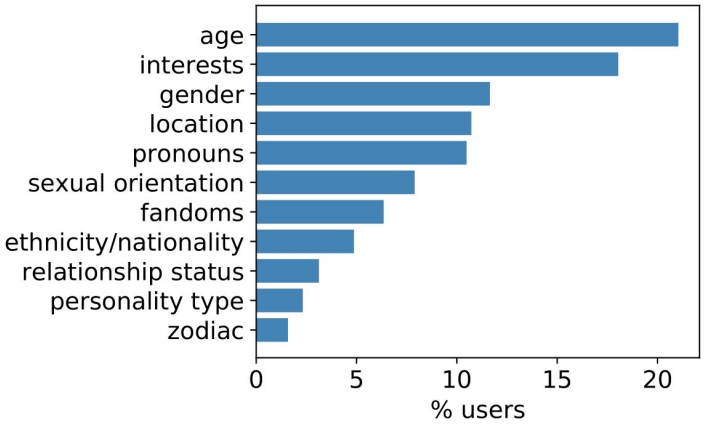

Figure 4.2: Proportion of users in our interpretation dataset who present each identity category.

Category alignment features

Let u' be the user making the comparison and u_* be the user associated with the candidate post p_* when applying $\Phi(p_*)$. Features included are listed below.

- **Category Match** (c): A binary variable indicating if u' and u_* both provide identity category c.
- **Category Mismatch** (c): A binary variable indicating if only one of u' or u_* provides identity category c.
- **Directional Category Mismatch** (c, u', u_*): Directional version of **Category Mismatch** (c) indicating if u' provided identity category c but not u_*, or vice versa.

Label alignment features

- **Label Match** (c): A count variable indicating the number of labels used by both u' and u_* in category c.
- **Label Mismatch** (c): A count variable indicating the number of labels that are unique to u' + the number of labels that are unique to u_* for category c.
- **Label Interaction** (l', l_*, c): A count variable of how many times u' used label l' and u_* used l_* for category c. Label interactions across categories were not considered since this would lead to a very large input dimensionality.

All text features are normalized to have unit variance over the training set after generating comparison feature set C.

4.5.2 Profile images

To extract features from profile images, we used the 1,000-dimension layer before the softmax layer from ResNet-152 (He et al., 2016), a popular computer vision benchmark that was pre-trained on 1,000 image categories from ILSVRC-2012 (Russakovsky et al., 2015). Since we are

	LR	SVM	MLP
Content	64.65	64.60	66.28
Content + text	77.66*	77.52*	80.86*
Content + image	81.99*	81.81*	89.97*
Content + text + image	**87.72***	**87.56***	**92.81***

Table 4.3: Learning-to-rank accuracy with text blog description and profile image identity alignment features. Text identity features include category and label alignment. $* \ p < 0.05$ compared to the content features (McNemar's test). A random baseline would achieve 50% accuracy.

interested in alignment between followers and users they follow, we use three different comparisons between profile image embeddings: **Cosine Similarity**, **Euclidean Similarity**, and **Vector Difference** (element-wise subtraction between vectors).

4.5.3 Content features

Post content features likely capture much of the signal in content propagation. In our experiments, we look for any additional signal provided by self-presentation alignment features above these baseline features. Note that we do not use actual post content of text, images, or videos. The wide variety of formats that post contents take make this inclusion non-trivial, and so we use features that are uniform across all post types.

- **Post Tags**: Post hashtags, similarly used by Naveed et al. (2011) to represent post content in predicting content propagation. All post tags are lowercased and only tags that are used by more than one user are considered in the tag vocabulary. Post tag features are binary variables indicating whether the post contains tag in the vocabulary.
- **Note Count**: A count variable indicating the number of notes (previous likes, reblogs, and comments) attached to the post. This is included to control for popularity.
- **Post Type**: A categorical variable indicating the type of the post. Posts can be of type text, photo, quote, video, audio, chat, link, or answer. Xie et al. (2017) also use the type of post as their content features in content propagation prediction.

4.6 Prediction model results

Results on the reblog prediction task for logistic regression, SVM, and MLP models are shown in Table 4.3. We find significant performance increases with the addition of identity alignment features from both blog descriptions and profile images, evidence for an association between identity presentation and content propagation on Tumblr.

Profile image alignment provides a stronger signal for reblog prediction than alignment in the text blog descriptions. This could be due to the visual focus of Tumblr (Xu et al., 2014a), the greater expressiveness of the feature vectors for image alignment, or both. It is also apparent that visual and text self-presentation are complementary, as performance improves with the

Category	Category Features	Label Features
Content features only	62.69	62.69
+ age	63.25*	66.29*
+ ethnicity/nationality	62.78	63.97*
+ fandoms	62.80	63.49*
+ gender	62.80	64.51*
+ interests	63.35*	65.82*
+ location	63.10*	65.03*
+ personality type	62.69	63.00*
+ pronouns	63.05*	63.89*
+ relationship status	62.82	63.17*
+ sexual orientation	63.10*	63.59*
+ zodiac	62.98*	63.07*
+ *all*	64.72*	74.30*

Table 4.4: Learning-to-rank reblog prediction accuracy using logistic regression on text blog description features for interpretation. **Category/Label Features** refers to content + category or content + label features. Each row refers to a separate model trained only on the features for that identity category. $* \ p < 0.05$ compared to a baseline of only content features.

combination of both signals.

The MLP is the best-performing model on reblog prediction. This may be explained by its ability to exploit non-linear combinations of content features with identity information (e.g. when users list a fandom and a post contains related hashtags).

4.7 Interpretation

Overall, we find that alignment in self-presented identity labels is associated with content propagation on Tumblr. However, this tells us little about the nature of this effect. For example, are users who present similar information more likely to reblog each other's content? Are users who present dissimilar images and labels less likely to reblog each other's content? Or do more unexpected interactions play an important role?

To investigate the nature of this effect (**RQ 2**), we focus on the text blog description features. Our image representations are dense, continuous vectors that are more difficult to interpret. Although there have been efforts to improve the interpretability of visual inference models (Selvaraju et al., 2017; Kim et al., 2017a), existing techniques are difficult to apply to our learning-to-rank setup, and we leave this to future work. We train logistic regression models separately across categories on a larger dataset including users who provide blog descriptions but not necessarily profile images. See Table 8.1 for details on this dataset and Table 4.4 for results from these models.

In general, we find that users who give matching labels and categories are more likely to

reblog each other's posts. We also find that users who give labels that indicate shared values around issues such as conceptions of gender and sexuality, and shared interests such as around visual content, are more likely to reblog each other's content. From the effects we see on content propagation, identity presentations on Tumblr seem to come across to establish solidarity and common ground. Specific findings are discussed below.

4.7.1 Category vs. label alignment effects

The effects of category features are inconsistent across categories (Table 4.4). Five out of the 11 categories do not show significant improvement over the content baseline, while 6 categories do improve: interests, age, pronouns, location, sexual orientation, and zodiac. For some categories, this is likely due to the presence of the category itself indicating an alignment of shared values or interests. Providing pronouns can indicate shared conceptualizations of gender, providing any location an interest in local (often visual) content, and any zodiac label indicates an interest in astrology.

For other categories, labels are skewed such that providing the category acts as a proxy for providing popular labels that indicate shared values or experiences. For example, 54% of users in our training set who provide an age present an age from 18 to 22, and 30% of users who provide interests list visual interests such as 'art', 'draw', or 'photos'. For sexual orientation, only about 10% of users provide labels that indicate being straight, so providing any sexual orientation likely indicates identifying as LGBTQ+.

Label alignment features significantly improve performance over content features (Table 4.4), suggesting that matches and mismatches of identity labels are associated with content propagation. Though category alignment features for the ethnicity/nationality, fandoms, gender, and relationship status categories did not significantly improve over the post content baseline, the use of label alignment features in these categories do lead to significant improvements. For these categories, we hypothesize that distinctions between specific label alignments are necessary to indicate shared values or experiences, rather than simply framing one's participation in a kind of interaction by listing any value in the category. For example, presenting any common gender label, such as 'male' or 'female', does not express an ideological position, whereas giving pronouns can express an ideology on gender issues.

4.7.2 Category alignment interpretation

For all categories except for pronouns and ethnicity/nationality, models trained with only baseline and category alignment features learned positive weights on the **Category Match** feature (Figure 4.3). This indicates that users are more likely to reblog content from other users who present the same category. Listing one's sexual orientation or interests—regardless of the labels used in these categories—signals that these categories are important to users' self-presentation.

The model trained on pronouns placed a negative weight on category match: both users giving pronouns was associated with a slight decrease in the likelihood of reblogging (odds ratio = 0.855). However, if the user deciding whether to reblog presents pronouns while the user they follow does not, the follower is less likely to reblog this post (OR = 0.781). There have been calls from transgender activists for cisgender people to share pronouns to normalize the practice

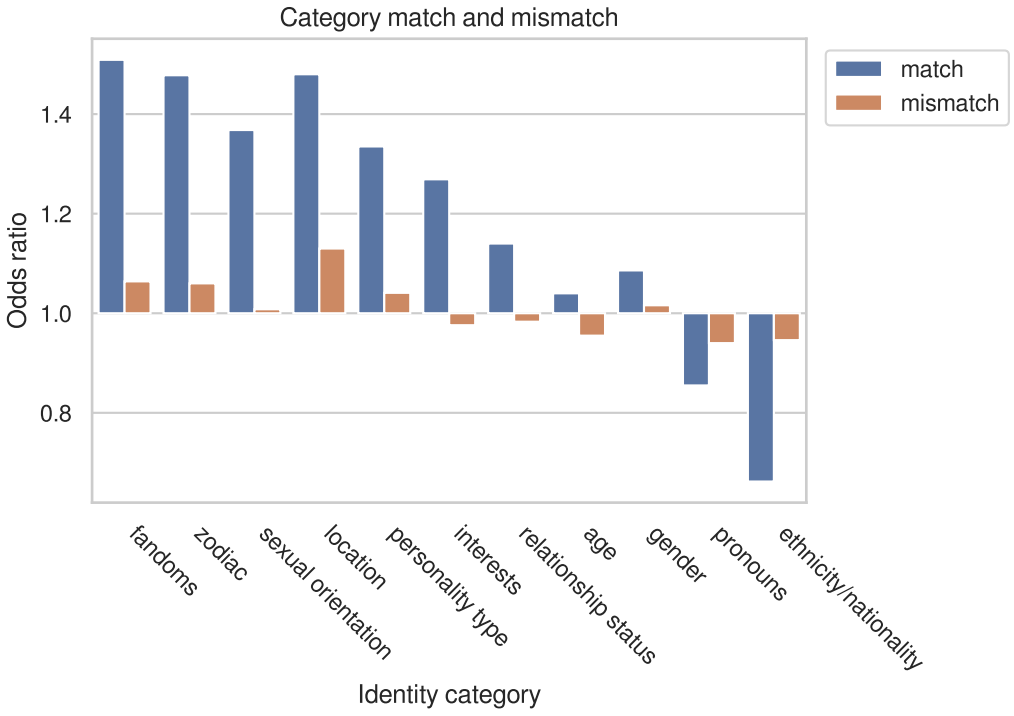

Figure 4.3: Odds ratios for the **Category Match** and **Category Mismatch** features from logistic regression models trained separately across identity categories. Categories are sorted by the difference between match and mismatch.

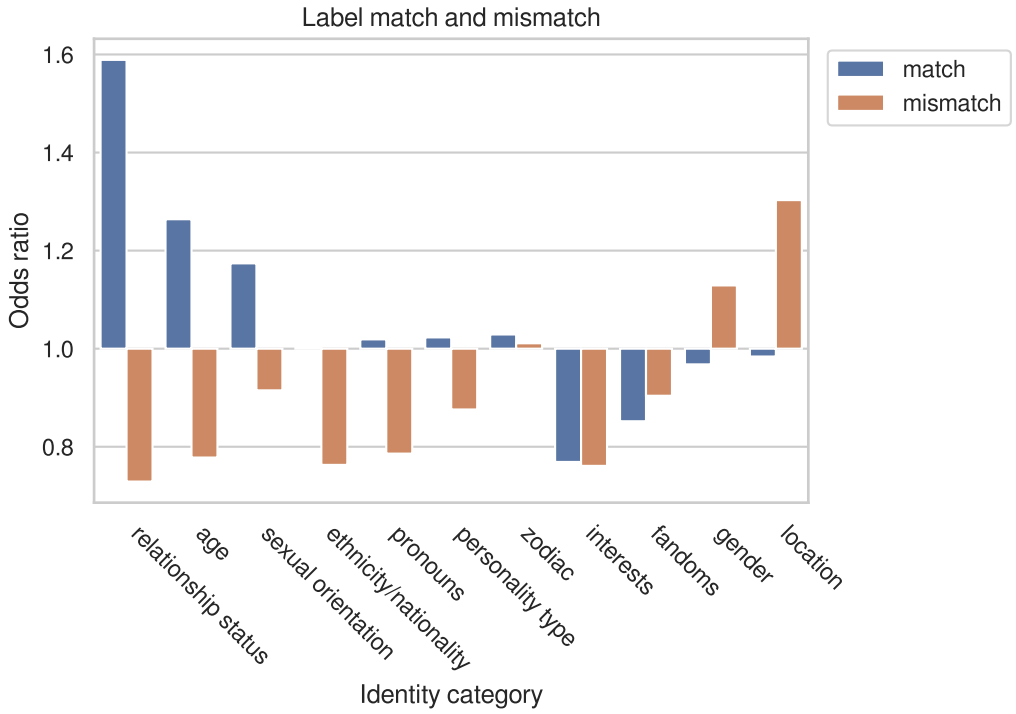

Figure 4.4: Odds ratios for the **Label Match** and **Label Mismatch** features from logistic regression models trained separately across identity categories.

of not assuming gender.[3] Not giving pronouns may signal that a user does not share a view of gender that encourages listing pronouns, and we see an association with content propagation.

Similar trends are observed for ethnicity/nationality, where followers who present any ethnic or national label are less likely to reblog users who do not present such labels (OR = 0.609), a stronger association than the negative weight learned on reblogging between users who both present the category (OR = 0.662).

For the zodiac category, the model placed most positive weight on the directional category alignment mismatch in which the user choosing to reblog presented zodiac but the user providing the post did not (OR = 2.395). This suggests that users who list zodiac are more open to propagating content from blogs that do not list zodiac.

In almost all of these categories that showed significant improvement, negative weight was placed on the directional mismatch in which the user making a decision to reblog does not provide a category, while the user they follow does. If a user does not place value in presenting an identity category, they appear less likely to propagate content from a user who does.

4.7.3 Label alignment interpretation

Models using only baseline and label alignment features learned positive weight on the **Label Match** feature for most categories (Figure 4.4). Most models learned negative weight on the **La-**

[3] `https://www.glaad.org/transgender/allies`, accessed 19 February 2020.

bel Mismatch feature. This provides evidence that matches in identity labels generally increase, and mismatches decrease, the likelihood of reblogging.

However, weights learned on specific **Label Interaction** features were often higher in absolute value, and thus more informative, than matches or mismatches. Looking more closely at these interactions, users who present labels that indicate similar interests in content often are more likely to reblog each other. For fandoms, one user listing 'star wars' and the other listing 'reylo', a popular character pairing in Star Wars, increased the likelihood of reblogging. For the interests category, users who presented 'anime' were more likely to reblog those who presented 'design'. Users who presented 'gaming' as an interest were more likely to reblog those who presented 'manga'. The model placed negative weight on some specific interactions that indicated differences in tone, such as users who listed 'memes' as interests reblogging those who list 'history'.

Other times specific interactions likely indicative of shared experiences were more informative. For example, in the age category, users presenting similar ages (e.g. users who present an age of 20 following users who present 21), were more likely to reblog each other's content. Positive weight was learned on interaction features of trans-identified users reblogging users who give no gender terms (OR = 1.23) and those who give other trans labels (OR = 1.06). The same holds for users giving 'non-binary' as a label: they were slightly more likely to reblog those who give no gender label (OR = 1.13). This may point to a preference for content from users who do not specify terms that are explicitly on the gender binary, which forms the bulk of the extracted gender labels. Similarly, those listing 'queer' were less likely to reblog others who present no term for sexual orientation and who may be more often straight without giving a label (OR = 0.85). Users presenting 'straight' were more likely to reblog others presenting 'straight' (OR = 1.25). Explicitly cis-identified users were slightly more likely to reblog content from other cis-identified blogs (OR = 1.02). Note that this feature only applied to users who explicitly identified as cisgender, not users who simply did not give any transgender labels. Identifying as 'cis', a relatively new and rarer term for gender, likely shows a knowledge of the discourse around transgender issues.

4.8 Limitations and future work

In this study, we only observe user behavior (posting profile images, writing blog descriptions, and sharing posts) and use statistical models to understand this behavior at scale. User interviews or surveys may provide more insight on motivations for these behaviors. For example, what do users themselves say they are trying to signal by using certain categories or labels in blog descriptions?

We use a regression-based analysis on records of naturally occurring interaction to find associations between identity alignment and content propagation. While this approach allows us to make correlational conclusions, we cannot draw any conclusions regarding causality. Though we sample scenarios where it is likely that users will see both posts in our comparisons, Tumblr's own ranking of posts in users' dashboards has an effect that is difficult to measure. Other factors such as what users are known as authorities on certain issues may also have an effect on reblogging behavior. A larger sample of users may provide a more comprehensive picture of the

relationship between identity and content propagation on Tumblr. Narrowing in on particular communities within Tumblr could reveal community-specific responses to self-presentation.

Since our datasets contained up to five data points for each reblogged post, each time paired with a different non-reblogged post, there is a concern that the model could somehow be memorizing reblogged posts. We split our dataset randomly, and so this would especially be a concern if a reblog appears in both training and test sets paired with different non-reblogged posts. However, we checked the results on a smaller sample without duplicate reblogged posts and the results and feature weights were very similar.

In future work, profile image alignment could be interpreted based on the 1,000 ILSVRC-2012 categories (Russakovsky et al., 2015). We could also explore image features relevant to Tumblr profile images, such as drawings and cartoons versus human faces, or facial expressions and emotions (Baltrusaitis et al., 2018). Complementary effects from visual and textual self-presentation could then be investigated. Identity alignment features could also be extracted from blog names.

Nonlinear feature combinations used by the MLP could be investigated. For example, do certain post hashtags increase the likelihood of reblogging in combination with profile images or blog description features? Tumblr has a particular emphasis on identity, so it remains to be seen what effects hold in other social media contexts.

4.9 Ethics and privacy

Tumblr profile images and blog descriptions contain sensitive personal information, so care was taken to protect users' privacy and remove the possibility of identifying any blogs in this study or amplifying any content from these blogs (Fiesler and Proferes, 2018). We only included public Tumblr blogs accessible without a Tumblr log-in. All examples of blog descriptions, except the staff Fandom account in Figure 3.1, have been fabricated so as to not easily trace back to any individual blog (Bruckman, 2002). Though researchers did view blog descriptions and profile images, these descriptions were not matched with blog names or URLs, which were also never used in analysis.

Regular expression patterns for identity label annotation were constructed from aggregate blog description n-grams, and are not shared. Only labels occurring in more than one blog description were extracted to protect any users who used unique labels. Similarly, hashtags were only considered in aggregate, and we removed any hashtags used by only one blog for feature extraction.

We did not construct classifiers that predict user identity attributes from text or images. However, our classifiers did include self-presentation features in predicting content propagation. This approach could possibly be used for targeted marketing, though this is not the intended purpose of this work, and no classifiers, data, or feature extractors have been made publicly available.

4.10 Conclusion

To explore the effect of identity-based features on content propagation on Tumblr, we constructed a machine learning task predicting which posts users propagate among posts they would have likely seen. We found that features from profile images and text descriptions were informative for predicting content propagation between users. Visual and textual self-presentation information provided complementary signals. Investigating the nature of this effect for text features, we found that users who presented similar identity categories and labels were generally more likely to reblog each other's content. Specific interactions between labels were also an informative part of this signal; users who presented labels that indicated shared interests in content or shared values around gender and sexuality, for example, were more likely to reblog each other.

These results suggest that homophily may support content propagation, though we caution that the pattern we have identified is only correlational. This raises the question of what the most effective balance between homophily and diversity would be to support content propagation while encouraging communication across subcommunities.

Chapter 5

Identity Representations and Community in Social Media

5.1 Introduction

Identity is inherently social, a shorthand for one's socially recognized position in a family, a workplace, a community, a society, or other grouping (Bucholtz and Hall, 2005; Gee, 2011). Thus notions of identity are not just about attributes of individuals by themselves, or how attributes between pairs of users align (operationalized in Chapter 4), but also about membership in groups and communities. Theories of identity from the social sciences, such as social identity theory from social psychology, stress how social groupings shape individual identities and how the nature of these social groupings have a different influence on interaction than do collections of individual identities (Tajfel, 1974; Seering et al., 2018). For example, being a fan of the TV show *Voltron*, whose fans are known to be fiercely loyal and even antagonistic, may be highly relevant to interacting about fandom controversies on Tumblr. Identities that are more at the individual level, such as being a young person or being a college graduate, likely has a different effect on interaction as these are not cohesive communities with strong, emotional associations.

From Chapter 4, we establish that features of identity alignment between users are informative in predicting content propagation (reblogging) on Tumblr. But this is simply at the level of pairs of users and does not incorporate the particular possible influence of *social* identities that emerge with membership in communities.

In this chapter, we set up statistical analyses to test for the influence of community alignment on reblogging behavior in Tumblr. We examine the influence of communities that are based on interaction between users, specifically those that emerge from follow links between users. With a logistic regression model on a dataset from a broader sample of users than in Chapter 4, we test the influence of community alignment, along with features of the content of posts, on reblogging behavior.

We find a significant positive effect of community alignment on reblogging, but this effect is small compared with that of features related to the content itself, such as hashtags placed on posts. This finding aligns with qualitative work on LGBTQ+ Tumblr users that suggests users' experience of "community" on Tumblr is often focused more on content than sustained interper-

sonal interaction (Byron et al., 2019). Identity and community are still relevant to reblogging practices on Tumblr: hashtags that are the strongest positive or negative influence on reblogging in our dataset are often related to identity and indicative of communities interested in similar content. However, on this platform we find that identity and community seem largely mediated through interests in content, and less prominently through direct interpersonal interaction.

5.2 Communities on Tumblr

The notion of communities is a driving force on Tumblr, even though there are no pages or groups clearly defined as communities (as subreddits on Reddit, for example). The importance of communities is evident even in promotional material from Tumblr staff, which portrays the site as "where you interests connect you with *your people*"[1] (emphasis added). This Tumblr "About" page further specifies that Tumblr's users are in "millions of communities" that users have access to.

Researchers often organize academic work on Tumblr around communities as well, such as fandom communities (Hillman et al., 2014), NSFW communities (Tiidenberg, 2016), asexual communities (Renninger, 2015), and transgender communities (Haimson et al., 2021). In particular, Tumblr is known for communities organized against mainstream or dominant expectations around topics like gender and sexuality expression (Renninger, 2015; Oakley, 2016).

There is no generally accepted notion of community or strict boundaries of interaction on Tumblr. Communities on Tumblr can be conceptualized as groups of users, groups of blogs, groups of content, or a mixture of these. Tumblr staff do curate a list of top fandom communities for reporting on interesting changes in popularity, maintained in a blog post[2] This list is based on manual curation of content hashtags by content matter experts on fandom.

Tumblr staff is focused on organizing and presenting content for presentation to users who may be interested in it. In this project, we encode and test the effects of a different view on community, centered on interaction between users themselves. The next section expands on theoretical motivation for this notion of community, as well as conceptual groundwork for the distinctions we investigate.

5.3 Theoretical motivation

Renninger (2015) characterizes Tumblr as a "networked counterpublic". This term combines boyd's (2010) notion of a "networked public" for describing online communities with cultural critic Michael Warner's (2002) idea of "counterpublics" working against dominant, hegemonic narratives. Such a term applies at the level of site-wide analysis, but how does Tumblr as a "networked counterpublics" operate at the micro-level of interaction between users? We examine influences on content sharing, or reblogging, one of the most frequent and influential interactions

[1] `tumblr.com/about`, accessed 21 May 2021

[2] See `https://fandom.tumblr.com/post/189334873364/2019-tumblr-communities` for a 2019 example.

on Tumblr. How much is reblogging driven by community behavior based on bonds between users, and how much by signals and properties of the content itself?

Such distinctions between interpersonal attachment to groups and interest- or identity-based attachments have a history of investigation by social psychologists and education researchers. In social psychology, Prentice et al. (1994) formalize differences between common-bond and common-identity motivations for engagement. Researchers have then applied this framework to online community settings such as IRC chat (Sassenberg, 2002), Usenet newsgroups (Arguello et al., 2006), WeChat (Pan et al., 2017), and in experimental settings with crowdworkers (Tausczik et al., 2014). However, differences in consequences between these types of attachment are difficult to determine (Yang et al., 2017; Pan et al., 2017). In this work, we do not attempt to operationalize this distinction directly, as our Tumblr features do not cleanly separate between the two notions. However, our operationalization of community as emergent from follow links is closer to a bond-based form of attachment, while post hashtags are closer to an interest- or identity-based relationship.

From semiotic and cultural studies traditions, Tardini and Cantoni (2005) identify a similar distinction between communities defined by interaction (*syntagmatic* communities) and those defined by simply sharing something in common (*paradigmatic* communities). The latter notion is closer to Anderson (2006)'s idea of the nation-state as an "imagined" community, or to "communities of interest", which are more loosely and informally connected by shared interests rather than shared practices as in communities of practice (Jones and Preece, 2006). Communities on Tumblr defined only by content or hashtags align more closely with Tardini and Cantoni (2005)'s interest-based paradigmatic communities, while communities emergent in follow graphs are closer to syntagmatic communities based on actual interaction between users.

5.4 Data

To test the influence of community alignment, we again examine reblogging behavior with a dataset set up in a similar way to that used in Chapter 4. To operationalize community alignment, we use the Louvain community detection algorithm on a directed follow graph among users present in this dataset.

In contrast to the more limited sample used in Chapter 4, we construct a new dataset with a broader sample of users. This new dataset contains 110,922 reblog "opportunities", instances in which one user chooses to reblog a post from one of the users they follow instead of a post from another of their followers, posted within 30 minutes of the reblogged post (and thus more likely to have been seen by the user while they were active). The biggest difference from the dataset used in Chapter 4, however, is a much larger sample of users: 294,253 unique users instead of the 34,801. By expanding the dataset, we aim to capture trends that are more robust and generalize more readily outside of the sample. We also just sampled one negative sample per reblog instead of up to 5 for the previous dataset to maximize the number of unique users in the dataset.

The sampling process was as follows. All data is from a data dump of Tumblr blog information from June 1, 2018, through November 30, 2018 (6 months). We restrict blogs to those which have added content to blog descriptions and which post or reblog a minimum of once per week on average, 26 times in the 6 month period. This a higher threshold for activity than that

of 10 reblogs used for the dataset in Chapter 4. Among these users, we then search for follow relationships and filter to reblogged posts during the period of time that users follow each other. For each reblogged post, we look for negative examples, posts that occur from other users a user follows from 30 minutes before the reblog to 30 minutes after. We filter to just those instances which we can pair with a negative example.

Network structure We build a directed, weighted network among users in our dataset from follow behavior. Reblog behavior could also be considered when building a network between Tumblr users, but we exclude reblog links since we later use this network information to predict reblogging behavior. We build a directed follow graph with edges pointing in the direction of influence (assuming the blog that is followed influences the blog that follows it, since that user will now see that blog's post on their feed). If Blog A follows Blog B, an edge is thus drawn from Blog B pointing to Blog A. Note that edges can be bidirectional between nodes if both blogs follow each other. This network is used for community identification.

5.5 Methods

In Chapter 4, we establish evidence for an association between self-expressed identity alignment and reblogging behavior by testing for increases in performance with the addition of identity alignment features across a set of machine learning classifiers. Here we wish to narrow in on what exactly drives this effect, especially to see if community alignment is a factor. Our focus is not about what features contribute to good reblog prediction in a potential machine learning classifier. In fact, the features used in Chapter 4 do not significantly improve performance on reblog prediction in this new dataset, which has a much larger and sparser sample of users. For the aims in this chapter, we adopt a smaller feature set and statistical hypothesis testing framework more common in social science research. With a logistic regression model, we assess the correlational effect of community alignment along with features that represent post content and virality.

5.5.1 Feature extraction

Community alignment As a measure of community alignment, we extract whether users are members of the same community as detected in the follow graph. We identify communities using the popular Louvain algorithm (Blondel et al., 2008) over a directed follow graph. We do not include reblogs in this graph since we wish to compare to our approach, which is supervised by reblog behavior.

More information on the most popular communities identified, along with the top TF-IDF terms from blog descriptions of users in the communities, can be found in Table 5.1.

Hashtag topic representation In order to identify significant factors in statistical analysis, we wish to reduce the large dimensionality of post hashtag features, which are important content features. There is a culture among Tumblr users of using multi-word hashtags, including entire

Community	Selected TF-IDF blog description terms	Number of users
Art, aesthetic	art, love	111,205
Straight NSFW	nsfw, love, like, only	43,234
Gay NSFW	nsfw, gay, men, just, please, love	32,602
Fandom	love, masterlist	13,910
Black, social media	ig, black, instagram, love, sc, twitter, snapchat	13,182
Eating disorders	gw, lbs, kg	3,698
Taylor Swift	taylor, tour, swift, rep, swiftie, followed	2,999
The Sims	wcif, sims, cc, friendly, simblr	804

Table 5.1: Communities above 500 users detected with the Louvain algorithm on a directed, weighted Tumblr follow graph. Communities made up of primarily non-English-speaking users are not included.

Topic name	Selected TF-IDF blog description terms
Gay NSFW	anonymous, daddy, boy, lgbt, happy, man
Straight NSFW	me, sexy, nude, sex, hot, porn
Fashion, memes	fashion, personal, funny, memes, gayman
Fandom	reblog, fanart, lol, voltron, pokemon, omg
Anime, art	my art, anime, lmao, wow, supernatural, pink
Comics	gif, marvel, love, 1k, aes, comics, mcu
Cute, animals	nsfw, video, cute, animals, fave, aesthetic
Music, food, fandom	text, mine, bts, bnha, music, food, photo
Illustrations, vintage	art, queue, fav, illustration, vintage, cats
Photos, nature	q photography nature culo words

Table 5.2: LDA topics from post hashtags. Topic names are manually given based on top terms.

sentences, since Tumblr allows spaces within hashtags. This makes the space of hashtags very sparse.

To reduce the number of possible hashtags, we select the most frequent hashtags over a threshold. We wish to select a threshold that maximizes the coverage of instances in our dataset (so that many instances have at least one hashtag feature) while also minimizing the number of hashtags selected (the sparsity). Empirically, we find that a threshold of 10 strikes a balance between these goals.

Filtering tags by this threshold first, we then further reduce the dimensionality of post hashtag features with Latent Dirichlet Allocation (LDA; Blei et al., 2003). Top terms in the 10 LDA topics are shown in Table 5.2.

5.5.2 Regression analysis

To estimate the effect of follow graph community alignment on content propagation, we use a logistic regression analysis predicting reblogging in the learning-to-rank framework of Chapter

4. We include a feature that encodes community match between follower and followee, as well as features representing post content. These include the post note count (number of likes, comments, and previous reblogs), presence of the LDA hashtag topics, and the post type (answer, audio, chat, photo, text, or quote). All features are normalized to zero mean and unit variance.

5.5.3 Propensity score matching

To measure the effect size of follow graph community alignment on reblogging beyond that of the post-specific variables, we use propensity score matching. Propensity score matching is an approach that more closely approximates causal inference from observational data, specifically estimating the propensity for certain data points to have received the treatment since this assignment is not random in observational data. In this case, we consider the treatment to be community alignment between the user choosing to reblog and the one who is followed. However, since this is not randomly assigned in our observational data, this community alignment could correlate with post features, such as those indicating viral content. We thus wish to compare sets of posts that differ with respect to community alignment but have the same or similar likelihood of being reblogged based on the post (control) features.

To do so, we first remove a test set of instances in which one followee differs from the community of the follower, and one is the same. This test set will be where we draw instances for comparing the influence of community alignment.

Using all other instances as a training set, we learn a logistic regression model predicting reblogging from post (control) features. Using this model on our test set, we select posts for which predicted probabilities for the positive class are 0.4-0.6, close to the threshold of 0.5 and thus are roughly equally likely to be reblogged based on the control features.

We then compare the reblog rate between posts that match the community of the follower and those that do not within these selected instances from the test set.

5.6 Results

Features found to be significant in the logistic regression analysis are listed in Table 5.3

Even with much information on users and content, virality is hard to predict (Martin et al., 2016). Yet we recognize some statistically significant patterns in which features are informative in reblogging behavior on Tumblr. Unsurprisingly, the post note count, how much activity the post has garnered so far, is the most informative feature positively associated with reblogging. The community match feature is given positive weight in reblogging.

Hashtag topics are generally significant factors in the analysis. The presence of NSFW (gay or straight) tags is positively associated with reblogging, while other topics received negative weight or were non-significant. Note, however, that some hashtags indicating erotic or pornographic content are mixed into other topics as well, illustrating how such content often played a role in expression across communities in Tumblr before the 2018 adult content ban (Haimson et al., 2021; Sybert, 2021). Hashtag topics indicating the popular “aesthetic” or “vintage” content, as well as the more general appeal of “cute” and “animals”, are negatively associated with

Feature	**Coefficient**	p**-value**
Post note count	0.494	<0.001
Gay NSFW hashtag topic	0.295	<0.001
Follower-followee community match	0.143	<0.001
Text post type	0.068	<0.001
Quote post type	0.064	<0.001
Video post type	0.064	<0.001
Chat post type	0.019	0.003
Straight NSFW hashtag topic	0.013	0.020
Fandom hashtag topic	-0.031	<0.001
Anime, art hashtag topic	-0.046	<0.001
Photos, nature hashtag topic	-0.050	<0.001
Comics hashtag topic	-0.062	<0.001
Photo post type	-0.077	<0.001
Music, food, fandom hashtag topic	-0.091	<0.001
Cute, animals hashtag topic	-0.100	<0.001
Illustrations, vintage hashtag topic	-0.106	<0.001
Answer post type	-0.139	<0.001
Audio post type	0.009	0.146
Link post type	-0.009	0.174
Fashion, memes hashtag topic	-0.010	0.099

Table 5.3: Logistic regression coefficients in reblog prediction. Factors significant at p <0.05 are shown above the horizontal separator.

reblogging. Post types are also often significant. Text, quote, and video posts have similar positive weight, whereas photos and especially the answer post type are negatively associated with reblogging.

The effect size of the community alignment feature, controlling for post-related features with propensity score matching, is $\phi = 0.15$ from a χ^2 test. This is slighly larger than what is considered a small effect of 0.1. Overall, we find that there is a non-random effect of community alignment on reblogging behavior on Tumblr, but that the size of this effect is fairly small. Much more informative are features of the post content.

5.7 Discussion

Our finding that community matters in content propagation, but that this effect is small, is consistent with Byron et al. (2019)'s interview study of LGBTQ+ Tumblr users. This study found that users' experiences of "community" are not as simple or uniformly experienced as finding others with similar marginalized identities and interacting. Direct interaction with other users was short and fairly limited. Many interviewees used Tumblr as a place to build a repository of content they liked rather than a place to directly interact with other users. Finding and reblogging content was still central to identity exploration and an empowerment that one's interests were shared, but the content itself played a crucial role in this process. In this way, identity and community are not experienced primarily through interpersonal bonds, but through interaction with content. Our findings align with the importance of community and identity *through content* in this space.

This challenges a possible notion of "community" on Tumblr as relying just on bonds between users. Unlike other social media platforms, Tumblr is not a place where people commonly connect with people they know offline (Devito et al., 2018), so communities are more likely to tend toward identity- and interest-based attachment than bond-based attachment. We see this reflected in factors associated with content propagation.

It is also evident that identity and community play a role in the hashtag topics–but again, mediated through interests and content. Some of the hashtag topics with the strongest positive or negative association with reblogging are ones that have more salient associations with social identites: gay NSFW content, fandoms, and the "vintage" aesthetic on Tumblr, which is associated with a large, known, female-based community on Tumblr (see Table 5.1). This is in contrast to topics with likely less identity associations, like "photos, nature", "music, food, fandom", and "fashion, memes", which were not significant or whose coefficients are closer to zero.

5.8 Conclusion

In this project, we examine the effect of community alignment on content propagation in Tumblr. We find a significant effect, but that this effect is rather small compared to the influence of post content features. This points to the importance of interests and content in interaction on Tumblr. The features of post content most informative for reblog prediction were relevant to identity and community, however, suggesting that identity and community play a role in content propagation in this space, but it is primarily through content, not as much through direct interaction.

More broadly, this finding points to the importance of being open to how identity is defined in a space. In the case of Tumblr, identity seems inextricably wrapped up with interests/information, whereas this may not be the case in other online platforms where users may be more likely to know each other offline, for example. Paying attention to social value in this space (in this case, reblogging) and studying emergent communities from a follow graph allowed us to discover this importance instead of imposing *a priori* notions of identity or community. The entangling of information with identity and community on Tumblr is emblematic of the importance of considering how identity and community play roles in how information is presented or received. One need only look to the politicization of Covid-19 information in the United States to see the importance of this entanglement.

In terms of this thesis, this entanglement between interests questions the value of our particular framework for measuring reactions to identity in isolation, which was part of our goal outlined in Chapter 4. There is still value in the paradigm of measuring the influence of various factors on predicting an important social behavior on a platform, such as reblogging on Tumblr, to illustrate the values that are important on a site. In this case the values we found were the centrality of interests and content to communities on Tumblr.

Part II

Characters, Relationships and Identity in Fanfiction

Chapter 6

FanfictionNLP: A Text Processing Pipeline for Fanfiction

6.1 Abstract

Fanfiction presents an opportunity as a data source for research in NLP, education, and social science. However, answering specific research questions with this data is difficult, since fanfiction contains more diverse writing styles than formal fiction. We present a text processing pipeline for fanfiction, with a focus on identifying text associated with characters. The pipeline includes modules for character identification and coreference, as well as the attribution of quotes and narration to those characters. Additionally, the pipeline contains a novel approach to character coreference that uses knowledge from quote attribution to resolve pronouns within quotes. For each module, we evaluate the effectiveness of various approaches on 10 annotated fanfiction stories. This pipeline outperforms tools developed for formal fiction on the tasks of character coreference and quote attribution.

6.2 Introduction

A growing number of natural language processing tools and approaches have been developed for fiction (Agarwal et al., 2013a; Bamman et al., 2014b; Iyyer et al., 2016; Sims et al., 2019). These tools generally focus on published literary works, such as collections of novels. We present an NLP pipeline for processing fanfiction, amateur writing from fans of TV shows, movies, books, games, and comics.

Fanfiction writers creatively change and expand on plots, settings, and characters from original media, an example of "participatory culture" (Jenkins, 1992; Tosenberger, 2008). The community of fanfiction readers and writers, now largely online, has been studied for its mentorship and support for writers (Evans et al., 2017) and for the broad representation of LGBTQ+ characters and relationships in fan-written stories (Lothian et al., 2007; Dym et al., 2019). See Chapter 3 for more background on fanfiction. Fanfiction presents an opportunity as a data source for research in a variety of fields, from those studying learning in online communities to social science analysis of how community norms develop in an LGBTQ-friendly environment. For NLP

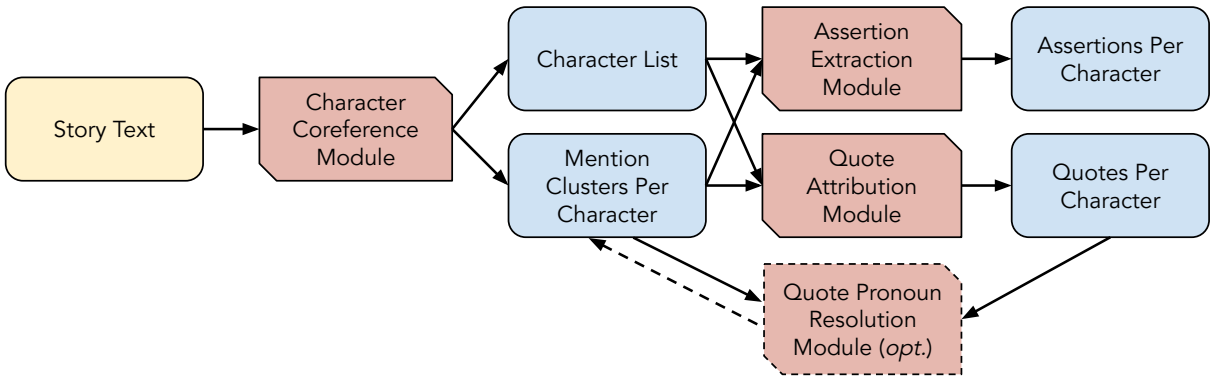

Figure 6.1: Fanfiction NLP pipeline overview. From the text of a fanfiction story, the pipeline assigns character mentions to character clusters (character coreference). It then attributes assertions and quotes to each character, optionally using the quote attribution output to improve coreference resolution within quotes (see Section 6.3.3).

researchers, fanfiction provides a large source of literary text with metadata, and has already been used in applications such as authorship attribution (Kestemont et al., 2018) and character relationship classification (Kim and Klinger, 2019).

There is an vast amount of fanfiction in online archives. As of March 2021, over 7 million stories were hosted on just one fanfiction website, Archive of Our Own, and there exist other online archives of similar or even larger sizes (Yin et al., 2017). We present a pipeline that enables structured insight into this vast amount of text by identifying sets of characters in fanfiction stories and attributing narration and quotes to these characters.

Knowing who the characters are and what they do and say is essential for understanding story structure (Bruce, 1981; Wall, 1984). Such processing is also useful for researchers in the humanities and social sciences investigating identification with characters and the representation of characters of diverse genders, sexualities, and ethnicities (Green et al., 2004; Kasunic and Kaufman, 2018; Felski, 2020). The presented pipeline, which extracts text related to characters in fanfiction, can assist researchers building NLP tools for literary domains, as well those analyzing characterization in fields such as digital humanities. For example, the pipeline could be used to explore how characters are voiced and described differently when cast in queer versus straight relationships.

The presented pipeline contains three main modules: character coreference resolution, quote attribution, and extraction of "assertions", narration that relates to particular characters. We incorporate new and existing methods into the pipeline that perform well on an annotated set of 10 fanfiction stories. This includes a novel method using quote attribution information to resolve first- and second-person pronouns within quotes.

Fanfiction is written by amateur writers of all ages and education levels worldwide, so it contains much more variety in style and genre than formal fiction. It is not immediately clear that techniques for coreference resolution or quote attribution that perform well on news data or formal fiction will be effective in the informal domain of fanfiction. We demonstrate that this pipeline outperforms existing tools designed for formal fiction on the tasks of character

coreference resolution and quote attribution (Bamman et al., 2014b).

In this chapter, we present a fanfiction processing pipeline that outperforms prior work designed for formal fiction. The pipeline includes novel interleaving of coreference and quote attribution to improve the resolution of first- and second-person pronouns within quotes in narrative text. We also introduce an evaluation dataset of 10 fanfiction stories with annotations for character coreference, as well as for quote detection and attribution. We are not aware of any text processing system for fanfiction specifically, though BookNLP (Bamman et al., 2014b) is commonly used as an NLP system for formal fiction. We evaluate our pipeline's approaches to character coreference resolution and quote attribution against BookNLP, as well as against other task-specific approaches, on an annotated evaluation dataset of fanfiction.

6.3 Fanfiction processing pipeline

We introduce a publicly available pipeline for processing fanfiction.[1] This pipeline is a command-line tool developed in Python. From the text of a fanfiction story, the pipeline extracts a list of characters, each mention of a character, as well as what each character does and says (Figure 6.1). More specifically, the pipeline first performs character coreference resolution, extracting character mentions and attributing them to character clusters with a single standardized character name (Section 6.3.1). After coreference, the pipeline outputs quotes uttered by each character using a sieve-based approach from Muzny et al. (2017) (Section 6.3.2). These quote attribution results are optionally used to aid the resolution of first- and second-person pronouns within quotes to improve coreference output (Section 6.3.3). In parallel with quote attribution, the pipeline extracts "assertions", topically coherent segments of text that mention a character (Section 6.3.4).

6.3.1 Character coreference module

The story text is first passed through the coreference resolution module, which extracts mentions of characters and attributes them to character clusters. These mentions include alternative forms of names, pronouns, and anaphoric references such as "the bartender". Each cluster is then given a single standardized character name.

Coreference resolution We use SpanBERT-base (Joshi et al., 2020), a neural method with state-of-the-art performance on formal text, for coreference resolution. This model uses SpanBERT-base embeddings to create mention representations and employs Lee et al. (2017)'s approach to calculate the coreferent pairs. SpanBERT-base is originally trained on OntoNotes (Pradhan et al., 2012). We further fine-tune SpanBERT-base on LitBank (Bamman et al., 2020), a dataset with coreference annotations for works of literature in English, a domain more similar to fanfiction. The model takes the raw story text as input, identifies spans of text that mention characters, and outputs clusters of mentions that refer to the same character.

[1]The pipeline is available at `https://github.com/michaelmilleryoder/fanfiction-nlp`.

Character standardization We then assign representative character names for each coreference cluster. These names are simply the most frequent capitalized name variant, excluding pronouns and address terms, such as *sir*. If there are no capitalized terms in the cluster or if there are only pronouns and address terms, the most frequent mention is chosen as the name.

Post-processing SpanBERT-base resolves all entity mentions. In order to focus solely on characters, we post-process the cluster outputs. We remove plural pronouns (*we*, *they*, *us*, *our*, etc.) and noun phrases, demonstrative pronouns (*that*, *this*), as well as *it* mentions. We also remove clusters whose standardized representative names are not named entities and have head words that are not descendants of *person* in WordNet (Miller, 1995). Thus clusters with standardized names such as "the father" are kept (since they are descendants of *person* in WordNet), yet clusters with names such as "his workshop" are removed.

For each character cluster, a standardized name and list of the mentions remaining after post-processing is produced, along with pointers to the position of each mention in the text. This coreference information is then used as input to quote attribution and assertion extraction modules.

6.3.2 Quote attribution module

To extract quotes, we simply extract any spans between quotation marks, a common approach in literary texts (O'Keefe et al., 2012). For the wide variety of fanfiction, we recognize a broader set of quotation marks than are recognized in BookNLP's approach for formal fiction.

The pipeline attributes quotes to characters with the deterministic approach of Muzny et al. (2017), which uses sieves such as looking for character mentions that are the head words of known speech verbs. We use a standalone re-implementation of this approach by Sims and Bamman (2020) that allows using the pipeline's character coreference as input. Muzny et al. (2017)'s approach assigns quotes to character mentions and then to character clusters. We simply assign quotes to the names of these selected character clusters.

6.3.3 Quote pronoun resolution module

Recent advances in coreference resolution, such as the SpanBERT-base system incorporated in the pipeline, leverage contextualized word embeddings to compute mention representations and to cluster these mentions from pairwise or higher-order comparisons. They also concatenate features such as the distance between the compared mentions to their representations. However, these approaches to not capture the change in point of view caused by quotes within narratives, so they suffer when resolving first- and second-person pronouns within quotes. To alleviate this issue, we introduce an optional step in the pipeline that uses the output from quote attribution to inform the resolution of first- and second-person pronouns within quotes.

In prior work, Almeida et al. (2014) proposed a joint model for entity-level quotation attribution and coreference resolution, exploiting correlations between the two tasks. However, in this work, we propose an interleaved setup that is modular and allows the user of the pipeline to use independent off-the-shelf pre-trained models of their choice for both coreference resolution and quote attribution.

Fandom	Primary media type(s)
Marvel	Comics, movies
Supernatural	TV show
Harry Potter	Books, movies
DCU	Comics, movies
Sherlock Holmes	Books, TV show
Teen Wolf	TV show
Star Wars	Movies
Doctor Who	TV show
The Lord of the Rings	Books, movies
Dragon Age	Video game

Table 6.1: The most popular 10 fandoms on Archive of Our Own by number of works, as of September 2018. We annotate 1 story from each fandom to form our test set.

More specifically, once the quote attribution module predicts the position of each quote (q_i) and its associated speaker (s_i), the first-person pronouns within the quote (e.g. *I*, *my*, *mine*, *me*) are resolved to the speaker of that quote, s_i. For second-person pronouns (e.g. *you*, *your*, *yours*), we assume that they point to the addressee of the quote (a_i), which is resolved to be the speaker of the nearest quote before the current quote ($a_i = s_{i-j}$ such that $s_{i-j} \neq s_i$). We only consider the previous 5 quotes to find a_i.

Since there are no sieves for quote attribution that consider pronouns within quotes, the improved coreference within quotes from this optional step does not affect quote attribution. Thus, this "cycle" of character coreference, then quote attribution, then improved character coreference, need only be run once. However, the improved coreference resolution could impact which assertions are associated with characters.

6.3.4 Assertion extraction module

After coreference, the pipeline also extracts what we describe as "assertions", topically coherent segments of text that mention a character. The motivation for this is to identify longer spans of exposition and narrative that relate to characters for building embedding representations for these characters. Parsing these assertions would also facilitate the extraction of descriptive features such as verbs for which characters are subjects and adjectives used to describe characters.

To identify such spans of texts that relate to characters, we first segment the text with a topic segmentation approach called TextTiling (Hearst, 1997). We then assign segments (with quotes removed) to characters if they contain at least one mention of the character within the span. If multiple characters are mentioned, the span is included in extracted assertions for each of the characters.

Evaluation Dataset	
# stories	10
# words	22,283
# character mentions	2,808
# quotes	876

Table 6.2: Fanfiction evaluation dataset statistics

6.4 Fanfiction evaluation dataset

To evaluate our pipeline, we annotate a dataset of 10 publicly available fanfiction stories for all mentions of characters and quotes attributed to these characters, which is similar in size to the test set used in LitBank (Bamman et al., 2020). We select these stories from Archive of Our Own[2], a large fanfiction archive that is maintained and operated by a fan-centered non-profit organization, the Organization for Transformative Works (Fiesler et al., 2016). To capture a representative range of fanfiction, we choose one story from each of the 10 most popular *fandoms* on Archive of Our Own when we collected data in 2018 (Table 6.1). *Fandoms* are fan communities organized around a particular original media source. For each fandom, we randomly sampled a story in English that has fewer than 5000 words and does not contain explicit sexual or violent content.

A colleague and I annotated the 10 stories for each of the tasks of character coreference and quote attribution. All annotators were graduate students working in NLP. Statistics on this evaluation dataset and the annotations can be found in Table 6.2.

The stories in this set illustrate the expanded set of challenges and variety in fanfiction. In one story, all of the characters meet clones of themselves as male if they are female, or female if they are male. This is a variation on the practice of "genderswapping" characters in fanfiction (McClellan, 2014). Coreference systems can struggle to keep up with characters with the same name but different genders. Another story in our test set is a genre of fanfiction called "songfic", which intersperses song lyrics into the narrative. These song lyrics often contain pronouns such as *I* and *you* that do not refer to any character.

For quote attribution, challenges in the test set include a variation of quotation marks, sometimes used inconsistently. There is also great variation in the number of indirect quotes without clear quotatives such as "she said". This can be a source of ambiguity in published fiction as well, but we find a large variety of styles in fanfiction. One fanfiction story in our evaluation dataset, for example, contains many implicit quotes in conversations among three or more characters, which can be difficult for quote attribution.

Annotation details and inter-annotator agreement for this evaluation dataset are described below. An overview of inter-annotator agreement is provided in Table 6.3.

[2] `http://archiveofourown.org/`

6.4.1 Character coreference annotation

To annotate character mentions in our evaluation dataset, annotators were instructed to identify and group all mentions of singular characters, including pronouns, generic phrases that refer to characters such as "the boy", and address terms. Possessive pronouns were also annotated, with nested mentions for phrases such as `<char1><char2>his</char2> sister</char1>`. Determiners and prepositional phrases attached to nouns were annotated, since they can specify characters and contribute to characterization. For an example, `<char1>an old friend of <char2>my</char2> parents</char1>`. Note that "parents" is not annotated in this example since it does not refer to a singular character. Appositives were annotated, while relative clauses ("the woman who sat on the left") and phrases after copulas ("he was a terrible lawyer") were not annotated, as we found them to act more as descriptions of characters than mentions.

After extracting character mentions, annotators grouped these mentions into character clusters that refer to the same character in the story. Note that since we focus on characters, we do not annotate other non-person entities usually included in coreference annotations. Full annotation guidelines are available online[3].

To create a unified set of gold annotations, we resolved disagreements between annotators in a second round of annotation. The final test set of 10 annotated stories contains 2,808 annotated character mentions.

In Table 6.3, we first provide inter-annotator agreement on extracting the same spans of text as character mentions by comparing BIO labeling at the token level. Tokens that begin a mention are labeled B, tokens that are inside or end a mention are labeled I, and all other tokens are labeled O.

Which mentions are identified affects the agreement of attributing those mentions to characters. For this reason, we provide two attribution agreement scores. First, we calculate agreement on mentions annotated by either annotator, with a NULL character annotation if any annotator did not annotate a mention (Attribution (all) in Table 6.3). We also calculate agreement only for character mentions annotated by both annotators (Attribution (agreed) in Table 6.3). Character attribution was labeled as matching if there was significant overlap between primary character names chosen for each cluster by annotators; there were no disagreements on this.

For all these categories, inter-annotator agreement was 0.84 Cohen's κ or above, "near perfect", for character coreference (Table 6.3).

6.4.2 Quote attribution annotation

A colleague and I annotated all quotes that were said aloud or written by a singular character, and attributed them to a list of characters determined from the character coreference annotations. Annotation was designed to focus on characters' voices as displayed in the stories. Thus characters' thoughts were not annotated as quotes, nor were imagined or hypothetical utterances. We also chose not to annotate indirectly reported quotes, such as "the friend said I was very strange"

[3] `https://github.com/michaelmilleryoder/fanfiction-nlp/annotation_guidelines.md`

	Character Coreference	Quote Attribution
Extraction (BIO)	0.95	0.97
Attribution (all)	0.84	0.89
Attribution (agreed)	0.95	0.98

Table 6.3: Inter-annotator agreement (Cohen's κ) between two annotators for each task, averaged across 10 fics. Extraction (BIO) is agreement on extracting the same spans of text (not attributing them to characters) with token-level BIO annotation. Attribution (all) refers to attribution of spans to characters where missed spans receive a NULL character attribution. Attribution (agreed) refers to attribution of spans that both annotators marked.

since this could be influenced more by the character or narrator reporting the quote than the original character who spoke it. However, we did annotate direct quotes that are reported by other characters.

Inter-annotator agreement on quote attribution was 0.89 Cohen's κ on the set of all quotes annotated by any annotator (see Table 6.3). Attribution agreement on the set of quote spans identified by both annotators was very high, 0.98 κ. Token-level BIO agreement for marking spans as quotes was 0.97 κ. The final test set of 10 stories contains 876 annotated quotes.

6.5 Pipeline evaluation

We evaluate the pipeline against BookNLP, as well as other state-of-the-art approaches for coreference resolution and quote attribution.

6.5.1 Character coreference evaluation

We evaluate the performance of the character coreference module on our 10 annotated fanfiction stories using the CoNLL metric (Pradhan et al., 2012; the average of MUC, B^3, and CEAFE) and LEA metric (Moosavi and Strube, 2016).

We compare our approach against different state-of-the-art approaches used for coreference resolution in the past. Along with BookNLP's approach, we consider the Stanford CoreNLP deterministic coreference model (CoreNLP (dcoref); Raghunathan et al., 2010; Recasens et al., 2013; Lee et al., 2011) and the CoreNLP statistical model (CoreNLP (coref); Clark and Manning, 2015) as traditional baselines. As a neural baseline, we evaluate the more recently proposed BERT-base model (Joshi et al., 2019), which replaces the original GloVe embeddings (Pennington et al., 2014) with BERT (Devlin et al., 2019) in Lee et al. (2017)'s coreference resolution approach.

Micro-averaged results across the 10 annotated stories are shown in Table 6.4. The FanfictionNLP approach is SpanBERT-base fine-tuned on LitBank, with the post-hoc removal of non-person and plural mentions and clusters (as described in Section 6.3.1). Note that these results are without the quote pronoun resolution module described in Section 6.3.3. Traditional

	CoNLL (Avg.)			LEA
	P	R	F1	F1
BookNLP	67.7	27.4	38.5	28.7
CoreNLP (dcoref)	26.9	49.5	29.6	21.9
CoreNLP (coref)	39.8	47.0	40.5	36.7
BERT-base O	45.8	53.2	49.2	50.9
BERT-base OL	55.0	62.3	58.4	63.1
SpanBERT-base OL	60.3	**71.1**	64.8	69.4
FanfictionNLP	**72.6**	70.1	**71.4**	**73.5**

Table 6.4: Character coreference performance on CoNLL and LEA metrics. **O**: Model is trained on OntoNotes. **L**: Model is also fine-tuned on LitBank corpus. **FanfictionNLP** is the SpanBERT-base OL model with post-hoc removal of non-person entities. Note that none of the approaches had access to our fanfiction data. These results are without the quote pronoun resolultion module described in Section 6.3.3.

approaches like BookNLP and CoreNLP (dcoref, coref) perform significantly worse than the neural models, especially on recall. Neural models that are further fine-tuned on LitBank (OL) outperform the ones that are only trained on OntoNotes (O). This suggests that further training the model on literary text data does indeed improve its performance on fanfiction narrative. Furthermore, the SpanBERT-base approaches outperform their BERT-base counterparts with an absolute improvement of 4-5 CoNLL F1 percentage points and 6 LEA F1 percentage points. Post-hoc removal of non-person and plural entities improves CoNLL precision on characters by more than 12 percentage points over SpanBERT-base OL.

6.5.2 Quote attribution evaluation

Using our expanded set of quotation marks, we reach 96% recall and 95% precision of extracted quote spans, micro-averaged over the 10 test stories, compared with 25% recall and 55% precision for BookNLP.

For attributing these extracted quotes to characters, we report average F1, precision, and recall under different coreference inputs (Table 6.5). To determine correct quote attributions, the canonical name for the character cluster attributed by systems to each quote is compared with the gold attribution name for that quote. A match is assigned if a) an assigned name has only one word, which matches any word in the gold cluster name (such as *Tony* and *Tony Stark*), or b) if more than half of the words in the name match between the two character names, excluding titles such as *Ms.* and *Dr.* Name-matching is manually checked to ensure no system is penalized for selecting the wrong name within a correct character cluster. Any quote that a system fails to extract is considered a mis-attribution (an attribution to a NULL character).

As baselines, we consider BookNLP and the approach of He et al. (2013), who train a RankSVM model supervised on annotations from the novel *Pride and Prejudice*.

	System coreference			*Gold coreference*			*Gold quote extraction*		
	P	R	F1	P	R	F1	P	R	F1
BookNLP	54.6	25.4	34.7	66.8	38.9	49.2	65.0	49.7	56.3
He et al. (2013)	54.0	53.3	53.6	56.5	55.7	56.1	56.7	56.0	56.3
Muzny et al. (2017) **(FanfictionNLP)**	**68.7**	**67.0**	**67.8**	**73.5**	**75.4**	**74.4**	**77.5**	**77.5**	**77.5**

Table 6.5: Quote attribution evaluation scores. Scores are reported using the respective system's coreference (*system coreference*), with gold character coreference supplied (*gold coreference*) and with gold character and gold quote spans supplied (*gold quote extraction*). Attribution is calculated by a character name match to the gold cluster name. If a quote span is not extracted by a system, it is counted as a mis-attribution. Micro-averages across the 10-story test set are reported. We include Muzny et al. (2017)'s approach in the FanfictionNLP pipeline.

The quality of character coreference affects quote attribution. If an entire character is not identified, there is no chance for the system to attribute a quote to that character. If a system attributes a quote to the nearest character mention and that mention is not attributed to the correct character cluster, the quote attribution will likely be incorrect. For this reason, we evaluate quote attribution with different coreference settings. *System coreference* in Table 6.5 refers to quote attribution performance when using the respective system's coreference. That is, BookNLP's coreference was evaluated with BookNLP's quote attribution and FanfictionNLP's coreference with FanfictionNLP's quote attribution. We test He et al. (2013)'s approach with the same coreference input as FanfictionNLP. Evaluations are also reported with gold character coreference, as well as with gold character coreference and with gold quote extractions, to measure attribution without the effects of differences in quote extraction accuracy.

The deterministic approach of Muzny et al. (2017), incorporated in the pipeline, outperforms both BookNLP and He et al. (2013)'s RankSVM classifier in this informal narrative domain.

6.5.3 Quote pronoun resolution module evaluation

We test our approach for resolving pronouns within quotes (Section 6.3.3) on character coreference on the fanfiction evaluation set. We show results using gold quote attribution as an upper bound of the prospective improvement, and using quote attributions predicted by Muzny et al. (2017)'s approach adopted in the fanfiction pipeline. As shown in Table 6.6, post-hoc resolution of first-person (*I*) and second-person (*you*) pronouns with perfect quote annotation information (Gold QuA) substantially improves the overall performance of coreference resolution (by 1.7 CoNLL and 3.7 LEA F1 scores).

Similarly, coreference resolution using information from a state-of-the-art quote attribution system (Muzny et al., 2017) also results in statistically significant, although smaller, improvements across both metrics (by 0.3 CoNLL and 1.0 LEA) on the 10 fanfiction stories. These results suggest that our approach is able to leverage the quote attribution outputs (speaker information) to resolve the first and second-person pronouns within quotations. It does so by assuming that the text within a quote is from the point of view of the speaker of the quote, as attributed by the

	CoNLL			LEA
	P	R	F1	F1
FanfictionNLP	72.6	70.1	71.4	73.5
+ I (Muzny QuA)	72.9	70.2	71.6	74.4
+ I + You (Muzny QuA)	73.1	70.2	**71.7**	**74.5**
+ I (Gold QuA)	73.9	71.2	72.5	76.0
+ I + You (Gold QuA)	74.6	71.6	**73.1**	**77.2**

Table 6.6: Quote pronoun resolution evaluation scores. Coreference resolution scores on the 10 fanfiction evaluation stories are reported. Improvements gained from changing the attribution of *I* and *you* within quotes are shown, with both the Muzny et al. (2017) quotation attribution system used in the FanfictionNLP pipeline, as well as the upper bound of improvement with gold quote annotation predictions.

Quote	**Speaker** (Muzny QuA / *Gold QuA*)	**Addressee** (Muzny QuA / *Gold QuA*)	**FanFictionNLP**	**FanFictionNLP + I + You** (Muzny QuA / *Gold QuA*)
"Alright , give me [**your**] phone . These questions are lame ."	Caitlin / *Caitlin*	Cisco / *Cisco*	**your** = Caitlin	**your** = [Cisco / *Cisco*]
"Would [**you**] rather give up showering for a month or the Internet for a month ?"	Caitlin / *Caitlin*	Cisco / *Cisco*	**you** = Caitlin	**you** = [Cisco / *Cisco*]
"[**You**] know what , do n't reply to that one , [**I**] do n't want to know ."	Cisco / *Caitlin*	Caitlin / *Cisco*	**I** = Cisco **You** = Cisco	**I** = [Cisco / *Caitlin*] **You** = [Caitlyn / *Cisco*]

Table 6.7: Coreference Resolution of first- and second-person pronouns in three consecutive quotes from one of the fanfiction stories in our dataset. Results show the impact of the Quote Attribution predictions on the performance of the algorithm described in Section 6.3.3.

quote attribution system.

Table 6.7 shows qualitative results on three consecutive quotes from one of the stories in our fanfiction dataset. For the first two quotations, FanfictionNLP incorrectly resolves *your/you* to the character *Caitlin*. However, FanfictionNLP + I + You correctly maps the mentions to *Cisco*. In the third example, we find that FanfictionNLP + I + You (Muzny QuA) does not perform correct resolution as the speaker output by the quote attribution module is incorrect. This shows the dependence of this algorithm on quality quote attribution predictions.

6.5.4 Assertion extraction qualitative evaluation

There is no counterpart to the pipeline's assertion extraction in BookNLP or other systems. Qualitatively, the spans identified by TextTiling include text that relates to characterization beyond simply selecting sentences that mention characters, and with more precision than selecting whole paragraphs that mention characters.

For example, our approach captured sentences that described how characters were interpreting their environment. In one fanfiction story in our test set, a character "could see stars and planets, constellations and black holes. Everything was distant, yet reachable." Such sentences do not contain character mentions, but certainly contribute to character development and contain useful associations made with characters.

These assertions also capture narration that mentions interactions between characters, but which may not mention any one character individually. In another fanfiction story in which two wizards are dueling, extracted assertions for each character includes, "Their wands out, pointed at each other, each shaking with rage." These associations are important to characterization, but fall outside sentences that contain individual character mentions.

6.6 Ethics

Though most online fanfiction is publicly available, researchers must consider how users themselves view the reach of their content (Fiesler and Proferes, 2018). Anonymity and privacy are core values of fanfiction communities; this is especially important since many participants identify as LGBTQ+ (Fiesler et al., 2016; Dym et al., 2019). We informed Archive of Our Own, with our contact information, when scraping fanfiction and modified fanfiction examples given in this paper for privacy. We urge researchers who may use the fanfiction pipeline we present to consider how their work engages with fanfiction readers and writers, and to honor the creativity and privacy of the community and individuals behind this "data".

6.7 Conclusion

We present a text processing pipeline for the domain of fanfiction, stories that are written by fans and inspired by original media. Large archives of fanfiction are available online and present opportunities for researchers interested in community writing practices, narrative structure, fan culture, and online communities. The presented text processing pipeline allows researchers to extract and cluster mentions of characters from fanfiction stories, along with what each character does (assertions) and says (quotes).

We assemble state-of-the-art NLP approaches for each module of this processing pipeline and evaluate them on an annotated test set, outperforming a pipeline developed for formal fiction on character coreference and quote attribution. We also present improvements in character coreference with a post-processing step that uses information from quote attribution to resolve first- and second-person pronouns within quotes. Our hope is that this pipeline will be a step toward enabling structured analysis of the text of fanfiction stories, which contain more variety than published, formal fiction. The pipeline could also be applied to other formal or informal narratives outside of fanfiction, though we have not evaluated it in other domains.

Chapter 7

Portrayal of Characters and Relationships in Fanfiction

7.1 Introduction

Computational literary studies often use NLP techniques in tasks such as genre prediction (Underwood, 2016), character type identification (Bamman et al., 2014b), and character relationship classification (Agarwal et al., 2013b; Chaturvedi et al., 2016; Iyyer et al., 2016). This work typically focuses on identifying and classifying features of characters and plot within individual texts or corpora. Yet fiction is not created in isolation. The influence of source texts on derivative texts is especially visible in folk tales, mythology, and re-makings of classic stories, but tropes and character types are repeated or modified throughout narrative fiction.

Social relations depicted in narrative writing can both reflect and communicate social values of a society (Booth, 1961). How authors transform prior narratives in their depiction of relations between characters can also carry social meaning (Fairclough, 1992). For example, consider Nancy Springer's *Sherlock Holmes* remake, which focuses on a sister of Sherlock who outshines her brother in detective mastery (Springer, 2007). A comparison to the original work reveals how the story projects different, perhaps more contemporary, gender roles. Automated methods could track such changes to trace influence and reveal larger-scale trends in how authors change or align with source narratives.

In this work, we take first steps toward modeling social shifts from source to derivative narratives. We focus on relationships between characters, which are central to the development of stories (Bruce, 1981) and are commonly studied with computational methods (Agarwal et al., 2013a; Iyyer et al., 2016; Chen et al., 2019). We present a new task classifying character relationships in fanfiction as similar or different from a source narrative. *Fanfiction*, informal narrative writing based on existing narratives, is an ideal space for this study since stories explicitly indicate their source. Character relationship changes in fanfiction are also relevant as an expression of shifts in social attitudes around identity. For example, fanfiction has dramatically higher representation of same-gender romantic relationships than *canon*, the mainstream narratives which fanfiction is based on (Lothian et al., 2007).

To extract text related to characters and relationships, we use the pipeline presented in Chap-

ter 6. We then evaluate a variety of unsupervised approaches for building embeddings for character relationships over this extracted text. Using these embeddings to predict relationship changes, we find that a) focusing on contexts that are meaningful for characterization within stories and b) measuring differences between a text and its source in constructed embedding spaces, are informative for predicting relationship shifts in derivative texts. A qualitative analysis identifies emotionally intense language indicating changes in character relationships from the original narrative.

7.2 Computational literary studies and intertextuality

Computational literary studies is an area that applies NLP and other computational approaches to answer literary questions (Bamman et al., 2019). Prediction tasks in this area typically concern character and plot within individual stories or corpora, for example finding latent character types (Bamman et al., 2014b) or extracting character attributes in novels in order to track how relationships between characters change over time within a story (Iyyer et al., 2016; Chaturvedi et al., 2016). We extend this work by focusing on changes *across* texts, setting up a task that predicts how derivative works change the portrayal of character relationships from a source.

Considering the relations between texts, referred to as *intertextuality*, is a common analytical framework adopted in literary studies (Kristeva, 1986; Holquist, 2003; Allen, 2011). Discourse analyst Norman Fairclough (1992) argues that intertextual processes of transforming past texts contributes to social change. This is evident in fanfiction, where relationships are reinterpreted in a community that centers marginalized gender and sexual identities. Learning how authors transform aspects of stories allows us to understand how narrative contributes to an ongoing societal discourse around identity and sexuality (Hall, 2005).

Prior work in natural language processing that relates to intertextuality has largely focused on textual references, tracking the reuse of phrases, quotes, or wording. For example, Niculae et al. (2015) trace quoting behavior of political media coverage and Leskovec et al. (2009) track the adoption of short phrases to model news cycles. Other work has focused on tracing the origins of specific pieces of texts, such as in Wikipedia and fanfiction (Shen et al., 2018) or in biblical texts (Lee, 2007). Our work differs in that we investigate changes between texts at a more abstract level than surface text: the portrayal of character relationships.

7.3 Fanfiction and NLP

We choose fanfiction as a collection of narrative texts with explicit links to the original narratives which they are based on, e.g. movies, TV shows, books, or comics. Online fanfiction presents an opportunity for researchers in NLP and digital humanities to study a community built around creative recontextualization on a large scale (Jenkins, 1992; Tosenberger, 2008; Milli and Bamman, 2016).

Computational work considering character identity in fanfiction includes that of Milli and Bamman (2016), who found that fanfiction writers are more likely to emphasize female and secondary characters. Studying a platform that includes fanfiction along with original online sto-

ries, Fast et al. (2016) find that portrayals of gendered characters generally align with mainstream stereotypes. Most similar to our work, Kim and Klinger (2019) manually annotate types of emotional relationships between characters in 19 fanfiction stories. In contrast, we focus on capturing how relationships are presented differently from canon. For more background on fanfiction, see Chapter 3.

7.4 Modeling intertextual relationship changes in fanfiction

Though fanfiction writers change settings, plots, styles and other narrative elements from original media, the community often focuses on romantic relationships between characters, "ships"[1] in fandom jargon. Tagging these relationships is also the most common way of indicating that a character's sexual orientation has been changed; labels for character sexuality are much less frequent. A fanfiction reader with the same shared canon background as the writer would understand when a relationship has been changed from its portrayal in canon; our goal is for a computational model to extract these differences automatically for analysis. Toward this end, we set up a prediction task predicting whether stories present character relationships similarly or differently than in canon.

7.4.1 Data

For our analysis, we selected J.K. Rowling's *Harry Potter* series because of its large canon text source and active fandom (over 100,000 stories on Archive of Our Own). We collect all complete *Harry Potter* fanfiction stories in English posted by November 2018, tokenized and with stopwords removed.

We focus on 5 main characters in the Harry Potter series who are represented in a variety of canon and non-canon romantic pairings. We narrowed consideration to pairs of these characters that are listed as romantic in at least 1000 stories. The selected 6 pairings among 5 characters are listed in Table 7.1. Note that no pairings are between two female characters; fanfiction has a strong historical trend of pairing male characters together (Tosenberger, 2008; Fazekas, 2014) yet lacks widespread representation of lesbian relationships. Many stories contain multiple romantic relationships, including relationships with more than two characters, but we narrow our focus to two-character relationships.

For each character pairing, we sampled 900 stories containing at least 5 paragraphs with both characters mentioned or at least one quote spoken between the characters. Ninety percent of the stories in this set are taken as a training set, with the rest as a test set. We also sampled 332 stories per pairing for a validation set for tuning. Multiple pairings can be drawn from the same stories, but no stories occur in both training and test sets. Dataset statistics are presented in Table 7.2.

[1] `https://www.vox.com/2016/6/7/11858680/fandom-glossary-fanfiction-explained`

	Selected values
Characters	Harry Potter, Hermione Granger, Ron Weasley, Draco Malfoy, Ginny Weasley
Pairings	Draco/Harry, Hermione/Ron*, Draco/Hermione, Ginny/Harry*, Harry/Hermione, Harry/Ron

Table 7.1: Selected characters and character pairings from the *Harry Potter* series. Starred relationships are romantic in the canon text.

	Number of instances
Train	4,866
Validation	1,992
Test	534

Table 7.2: Number of instances per data split.

7.4.2 Prediction task

For each character pairing within each story, we train models to distinguish three separate binary labels. The first task, **Canon** match, is our primary focus. The other two tasks, **Romantic** and **M/M**, allow us to confirm that we are actually able to perform the first task instead of capturing other related phenomena.

1. **Canon**: Whether a relationship between two characters has changed from canon with respect to being romantic. If the pairing is romantic in canon and romantic in the story, or non-romantic in both canon and the story, then this value is true. Otherwise, it is false and the author has changed the relationship. Training set label skew: 44% changed, 56% same as canon.
2. **Romantic**: If a pairing in a story is romantic, regardless of canon, then this value is true. Predicting romance between characters does not necessarily capture anything about shifts in framing from canon, so we include this as an auxiliary prediction task. Training set label skew: 27% romantic, 73% non-romantic.
3. **M/M**: If the two characters in the pairing are male, regardless of romance, then this value is true, otherwise it is false. Note that pronouns are removed as stopwords for all tasks. Predicting gender could serve problematically as a proxy for divergence from canon since most fanfiction relationships are male-male pairings and most in mainstream media are female-male. Training set label skew: 33% male/male, 64% other.

As an example, the characters of Harry and Draco in our Harry Potter data are frequently paired together as lovers in fanfiction but are enemies in the original series. Such stories would

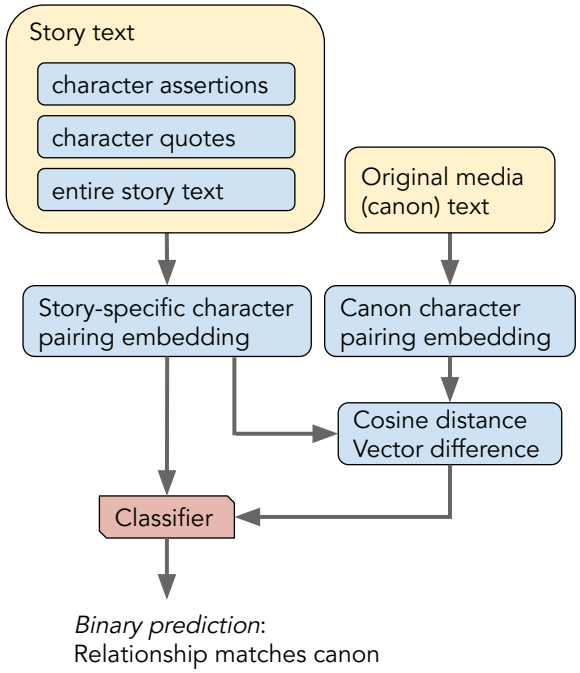

Figure 7.1: Computational model overview.

be annotated "false" for the **Canon** value and "true" for both **Romantic** and **M/M**. Fanfiction stories that pair the characters Ron and Hermione in a romantic relationship, just as they are in canon, would be annotated "true" for **Canon**, "true" for **Romantic**, and "false" for **M/M**.

Gold labels for all three tasks come from author-provided metadata in Archive of Our Own, which has an extensive and well-used tagging system (Fiesler et al., 2016). Canon relationships are determined from the canon text.

7.4.3 Computational model

Each prediction instance is a specific character pairing's overall representation within a story. We construct embedding representations for character pairings within each story, which are used as input to a logistic regression model. Though better performance could likely be obtained with more powerful models, our goal is to evaluate the expressiveness of just the unsupervised embedding space representations, not model architectures. These approaches more readily generalize to settings with fewer or more indirect labels. An overview of the computational modeling approach is shown in Figure 7.1.

First, text that relates to the specific pair of characters is extracted from the story using the narrative processing pipeline (Chapter 6). Then, an embedding for the extracted text is built. An embedding is also constructed for the character pairing from the canon text. We investigate whether differences between the embedding of the character pairing in the story and the embedding in canon can capture differences in intertextual framing. To test this, we include cosine distance and the vector difference between these embeddings as features.

We compare the performance of alternative representations along three dimensions: which

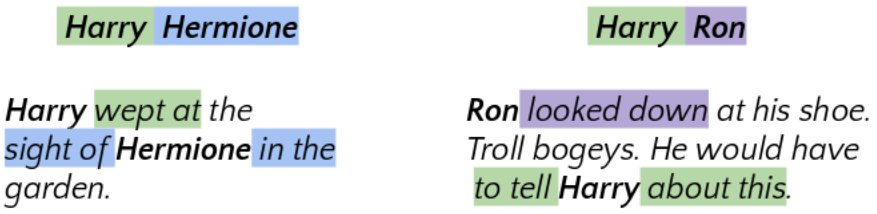

Figure 7.2: Example of fanfiction story character pairing representations from assertions. For each character, TF-IDF weighted embeddings for context words are averaged. Character representations are then concatenated.

text is extracted from the story (assertion, quotes, or the entire story), which word embedding space is used, and whether we include the explicit comparison between canon and fanfiction.

Character Pairing Embeddings in Fanfiction and Canon

To construct representations for character pairings within specific stories, we take a TF-IDF weighted average of word embeddings in the input of the entire story, character assertions, or character quotes.

For assertions (not quotes), we apply a context window around the coreferenced mentions of each character. To find the optimal window size, we tested embeddings averaged over context windows of 5, 10, 25 and 50 words both before and after character names. Words that overlap in multiple context windows are counted multiple times, with the intuition that terms close to both character mentions are important context for the relationship. Performance on canon prediction for a window size of 10 words was highest on 5-fold cross-validation on the validation set, so we select a 10-word window.

To construct representations of each pairing in canon, we take TF-IDF weighted average embeddings from 10-word windows before and after character names in paragraphs within the *Harry Potter* novels where both characters are mentioned. Coreference is less necessary in the case of canon since names are frequent enough in book-length texts.

We evaluate the following alternative representations for pairings in each story:

- **Unigrams**: TF-IDF weighted bag-of-words unigram features over the context.
- **Embeddings**: **Base**, **Shared**, and **Aligned** embedding spaces, described in Section 7.4.3.
- **Embeddings + canon difference**: In addition to the fanfiction story-specfic pairing embedding, we compute cosine distance and vector difference (vector subtraction) from the pairing embedding in canon.

Input Text for Characterization

We expect signals of changes to character relationships to be located in text that refers to those characters. To test this, we compare using the whole story to using extracted assertions and quotes where both characters in the selected pairing interact.

We remove all Harry Potter character names[2] and pronouns, as we wish to learn representations that generalize beyond specific characters and relationships to dissuade the model from lexical memorization (Levy et al., 2015). We attribute quotes and assertions from the narrative pipeline to our selected characters by matching character clusters that contain the first name of the selected character.

The text segments used in experiments are:

- **Entire story**: All words, not including character names and pronouns.
- **Character assertions**: Assertions attributed to either character in the pairing, in which the other character is mentioned.
- **Character quotes**: Quotes attributed to one character in the pairing, either a) spoken in reply to a quote from the other character in the pairing or b) with a mention of the other character appearing in the quote or in the immediately preceding or following paragraphs.

Embedding Space Construction

We explore three approaches for generating embeddings that can be compared between canon texts and derivative stories:

1. **Base.** As a base word embedding space, we trained FastText embeddings (Bojanowski et al., 2017) on a collection of 5130 science fiction and fantasy e-books[3], the closest genre to the Harry Potter series.
2. **Shared.** Starting from these pre-trained base embeddings, we trained a shared embedding space using the concatenated fanfiction and canon corpora.
3. **Aligned.** From the pre-trained base embeddings, we trained separate spaces for canon and fanfiction and then align them to a unified space. Following Mikolov et al. (2013), we learn a linear transformation using the top 1500 verbs in the canon text that are also present in the fanfiction corpus as seed words.

7.4.4 Results

Our first set of experiments test which input text provides the strongest signal of intertextual changes to relationships. We also test whether the inclusion of a distance value or difference vector between canon and fanfiction is informative in distinguishing the nature of an intertextual reference, i.e. whether the story is presenting the relationship similarly or differently from canon. Results are reported using the **Shared** embedding space.

Focusing on text that is relevant for characterization (quotes and assertions from the informal narrative processing pipeline) performs substantially better at capturing divergence from canon than using the entire story text (Table 7.3).

Also evident in Table 7.3 is that difference vectors improve results across all input ($p < 0.05$, McNemar's test). This suggests that vector differences in lexical-semantic embedding spaces

[2]Downloaded from `https://harrypotter.fandom.com/wiki/Category:Individuals`

[3]`https://github.com/soskek/bookcorpus`

	Prediction task		
Feature set	**Canon**	**Romantic**	**M/M**
Entire story (S)	54.82	66.36	65.24
+cosine distance	54.82	66.36	65.24
+vector difference	55.78	67.00	65.03
Assertions (A)	64.76	77.29	80.31
+cosine distance	64.28	77.64	80.52
+vector difference	67.86	77.83	**90.10**
A + Quotes (Q)	64.35	76.47	79.95
+cosine distance	64.35	77.32	80.48
+vector difference	**68.52**	**77.88**	88.82

Table 7.3: Prediction accuracies across text input types. All results are using the **Shared** embedding space.

between our canon and fanfiction representations carry value for approximating the relationship differences measured by our task.

Quotes extracted by this method only appear in around 60% of the stories in our experimental dataset, so we only report results in Table 7.3 on quotes in combination with assertions. On a subset of the data where quotes are present, they perform slightly better than assertions for the **Canon** prediction task (66.7 accuracy over 66.0), a pattern that is also evident in results on the full dataset. This suggests that much of the relevant relationship characterization comes from quotes between characters, though signals coming from both quotes and assertions complement each other.

Table 7.4 shows prediction results across different embedding space construction methods. These results are with canon-fanfiction difference vectors concatenated with story pairing embeddings, the highest-performing setting (Table 7.3). The shared embedding space was found to give the highest performance on canon prediction, with a significant improvement in performance ($p < 0.05$, McNemar's Test) with the A+Q shared space compared with A+Q with the background space.

The embedding space alignment technique allows widely different embedding spaces to be cast in the same space for comparison, perhaps preserving more of the differences between the spaces. However, in our case, this approach only gives us a minor improvement on the **Romantic** task compared to simply training embeddings on a shared space of canon and fanfiction text. Though relationship presentation could be vastly different to human readers from canon to fanfiction, such differences may not be strong enough in learned embedding space to warrant the value of explicit alignment.

Results on the two corollary tasks, **Romantic** and **M/M**, pattern differently from our main canon prediction task (Table 7.4). This suggests that our representations are capturing an element of canon divergence in relationship framing beyond that of simply predicting romance or gender.

		Prediction task		
Approach		**Canon**	**Romantic**	**M/M**
Majority	–	52.40	73.16	75.72
Unigram	F	55.30	65.08	65.56
	A	62.04	73.61	70.40
	A+Q	63.00	75.74	68.95
Base	A	64.44	77.58	85.50
	A+Q	64.22	77.44	85.50
Shared	A	67.86	77.83	**90.10**
	A+Q	**68.52**	77.88	88.82
Aligned	A	65.06	**78.07**	82.64
	A+Q	64.35	76.63	79.63

Table 7.4: Prediction accuracies across different embedding space approaches and feature sets. All the embedding-based approaches (**Base**, **Shared**, and **Aligned**) include the vector difference between the fanfiction and canon pairing embeddings. Results for the entire story (S) are comparably poor for approaches other than unigrams so are not displayed.

7.4.5 Qualitative analysis

To see what difference in relationship portrayal our best-performing model on the **Canon** task captures, we manually examined successful and unsuccessful classifications on the training set. From these examples, it appears our model is more likely to predict a relationship change if there is more emotionally intense language used to present the characters. Conversely, if the model predicts no change in the relationship from canon, there is less emotional intensity–even if the relationship is still romantic. If the romance is already "known" from canon, it may be more likely to fade into the background.

Melodramatic emotion is frequent in examples that our model correctly predicts as having a romantic relationship that was not present in canon. One such example with a romance between the characters of Draco and Harry, who are not romantic in canon, features dramatic, emotional descriptions between the characters:

> "Tears dripped from Harry's chin, and Draco felt the heat of them against his cheek."
> *(modified for author privacy)*

But when our model correctly predicted relationship changes to *not* being romantic, we also found examples of emotional intensity around mentions of these characters. For example, in one story where the relationship between the canon couple of Harry and Ginny is not romantic, the characters nonetheless are ex-lovers and confidantes in emotionally intense situations. The characters are freed from a jail and divulge their current crushes to each other. Words such as "attacker", "grasp", "cringed", and "engagement" appear in close proximity to mentions of them both.

This focus on emotional intensity also seems to have led our model astray in some cases when

relationships between the characters were unchanged from canon. In a story where the relationship between the characters Harry and Hermione remained non-romantic, our model nonetheless predicted it had changed to being romantic possibly due to scenes where they are both present at a ball. These scenes are described with words such as "twirled", "giggled" and even "kiss"–but not between these two characters. This points to our model's limitation in not precisely recognizing which characters are interacting, but relying on the surface context provided by word embeddings.

Another failure case was with a lack of emotional intensity around mentions of the character pairing when our model falsely predicted no change from canon. For example, one story where canon couple Ron and Hermione are not romantic focuses on other topics, such as student exams. Placed in the background, our model may expect this relationship to remain romantic as it was in the canon.

7.4.6 Lexicon analysis

To investigate the hypothesis that our model's predictions correlate with an intensity around the characters in relationships that changed from canon, we applied the NRC VAD lexicon (Mohammad, 2018) to paragraphs that contain both characters. In the training set, we find that mean values for valence, arousal and dominance are significantly higher ($p < 0.05$, *t*-test) for instances where our model predicts the relationship changed from canon. Words with higher sentiment, greater arousal and power are often those with higher emotional intensity; these terms surround character pairings when our model predicts a change.

This is in contrast to the gold-standard training set labels, where differences in means are in the same direction for valence and arousal as the predictions but are not statistically significant. Although this difference as captured by the lexicon is not pronounced in the data, our model reaches a level of success by focusing on character relationships with emotional language. Since fanfiction is generally known to have more emotional language, it could be the case that our model predicts that relationships that exhibit more of that stereotypical "fanfiction" portrayal have been changed from canon in a way that relationships unchanged from canon do not. Note that this is simply from logistic regression over the embeddings learned in an unsupervised fashion. For other domains, our methods would likely find other, domain-specific differences between source and derivative texts.

7.4.7 Generalizability

The paradigm we present could be useful in other domains with source and derivative texts, such as news stories from various outlets or social media comment threads. In our domain, we take advantage of labels that explicitly mark changes in character romantic relationships for individual texts in a supervised prediction task. In other domains, such as social media platforms, available labels may serve as a proxy rather than explicitly marking the underlying change. For example, a Reddit comment posted by a certain user may shift the portrayal of entities closer to the values of the communities that user participates in, but it is not guaranteed. However, insights from our evaluation of unsupervised embedding techniques could still be valuable in these other settings.

7.5 Exploration of character framing through visualized embeddings

In this section, we compare multidimensional representations of character framing across multiple stories at a time using a visualization. We use our representation approach to investigate divergences in representation of individual characters in fanfiction. When we were concerned with a specific decision regarding framing of a relationship (in the previous experiment described above), we aimed to construct a representation that was not specific to the characters involved, or to any specific story, but instead was related to ways in which authors within a community signal their framing regarding a cross-cutting issue, such as whether a relationship represents a divergence from canon or not. Thus, in that case, our representations did not include the exact name of the characters involved, but instead represented the text within a window around the mentions. Here we are specifically interested in the framing of a specific character within a specific story. For each of those framings of the same character, we want to visualize how they relate to one another and to the framing in canon. For this exploration into character vector modeling and visualization, we apply the contextualized vector approach described in the next section to the names of characters.

7.5.1 Contextualized word embeddings

The approaches for text representation described in the previous experiments provide embeddings for each word type in our vocabulary; we later take a weighted average over extracted text to represent the context in which a character or relationship is presented (see Section 7.4.3). As an alternative, here we consider directly learning contextualized embeddings that provide a unique embedding for every instance of a word in a text, similar to the ELMo model of Peters et al. (2018). To initialize these vectors, we train canon and fanfiction word vector spaces separately from a shared background space and use vector space alignment techniques to map terms to the same vector space (Mikolov et al., 2013).

In order to capture the representation of a word within its textual context, we train a recurrent neural network-based language model that predicts a word from its previous context, and use the learned hidden state of a word passed through that neural network as the context-specific representation of that word. To differentiate the contextualized representation of a character across stories, for each character within each story we average contextualized representations of all of that character's mentions. This gives us a representation of the character that can be compared across stories.

7.5.2 Data

For data, we again use fanfiction from Archive of Our Own due to its volume of stories and extensive metadata system. The language model component of the contextualization approach requires shorter texts due to computational restraints, so we set a smaller word limit for this exploration and scrape Harry Potter fanfiction stories in English with 1000-5000 words, totaling 42,792 stories.

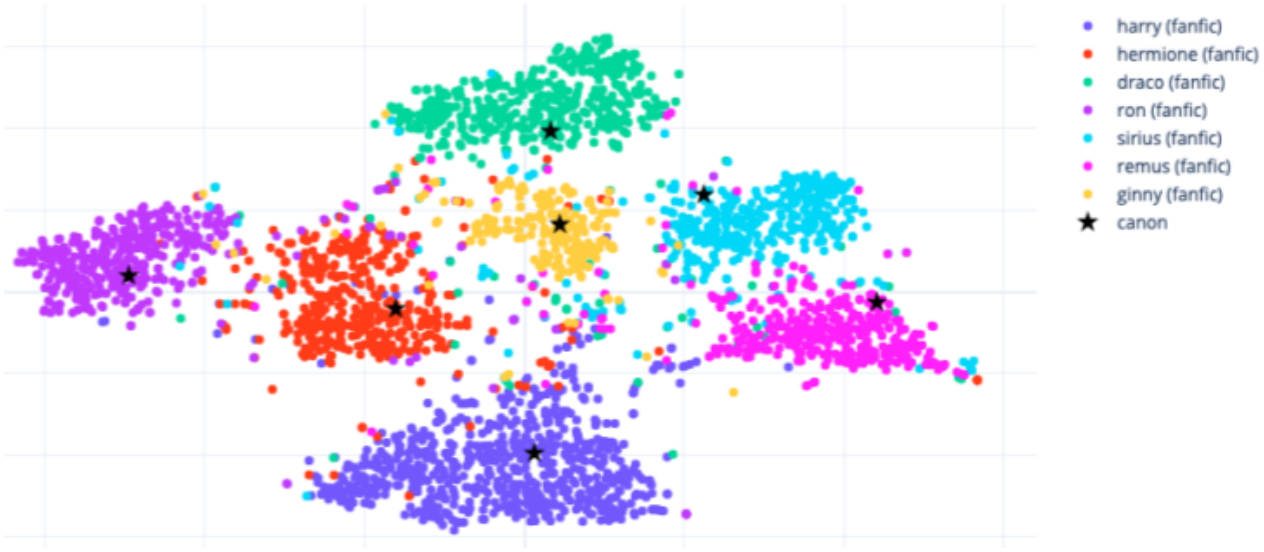

Figure 7.3: Character name vector visualizations. The closest star to each cluster is the vector for that character learned from the canon (original media). Credit to Qinlan Shen for this visualization.

7.5.3 Visualization

We reduce the dimensionality of the learned story-specific characters vectors with t-SNE (van der Maaten and Hinton, 2008) to allow visualization. We then plot the learned vectors for characters across fanfiction stories and representations learned from canon.

Visualizing the learned representations for seven main character names in 1000 randomly sampled fanfiction stories from our dataset, we see expected divergence from canon with characters known to be significantly transformed by fanfiction (Figure 7.3). For example, the canon representation for 'Draco' is near the edge of the cluster formed by fanfiction representations. Draco Malfoy is a villainous character often positively viewed as a "bad boy" romantic partner for Harry or others in fanfiction. Similarly, fanfiction representations for 'Remus' and 'Sirius' differ significantly from canon representations. Fanfiction that features these characters are often set in a time previous to the novels, so a significant difference in contextual representations also makes sense here. The canon representation for Harry, by contrast, is more central to the cluster of fanfiction representations.

We also find evidence that these representations capture variation among fanfiction focused on different relationships. When fanfiction pairs characters with different partners than in canon, we see evidence of different presentations of their original canon partner. For example, representations for Ginny Weasley, who is paired with Harry Potter in canon, vary between fanfiction stories that feature this canon pairing and those that feature Harry with Draco (Figure 7.4).

In Figure 7.5, we see separate patterning for Ron Weasley across different relationships. Ron is paired with Hermione Granger in canon, but fanfiction often pairs Hermione with Draco.

These could reflect "antagonistic" attitudes toward these characters taken by fans that omit them from relationships with well-liked characters. There is qualitative evidence for this in the

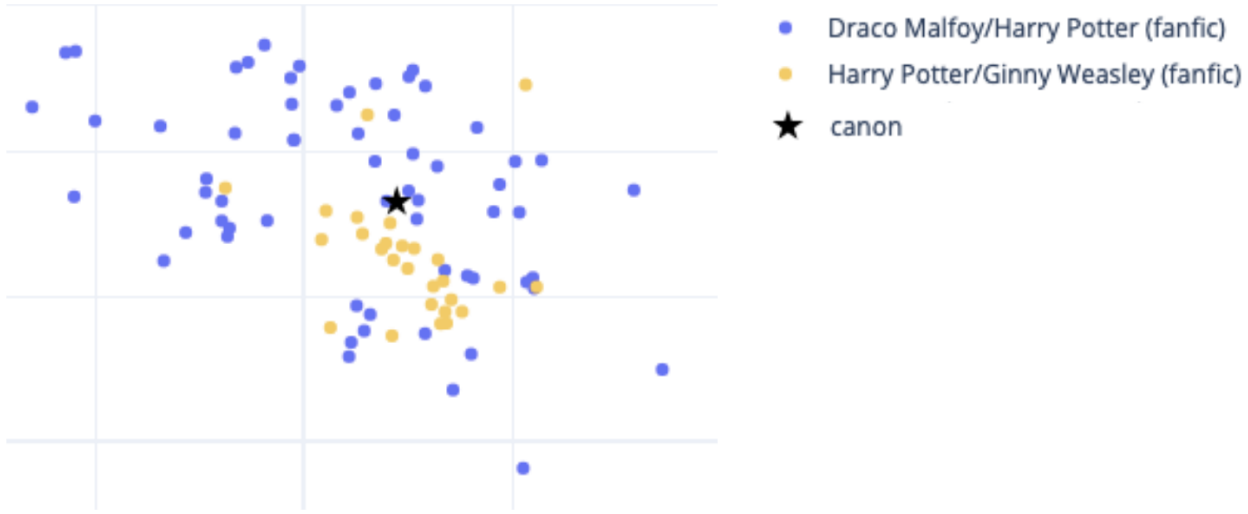

Figure 7.4: Representations for ‘Ginny’ in fanfiction and canon, colored by relationship. A few outliers are omitted.

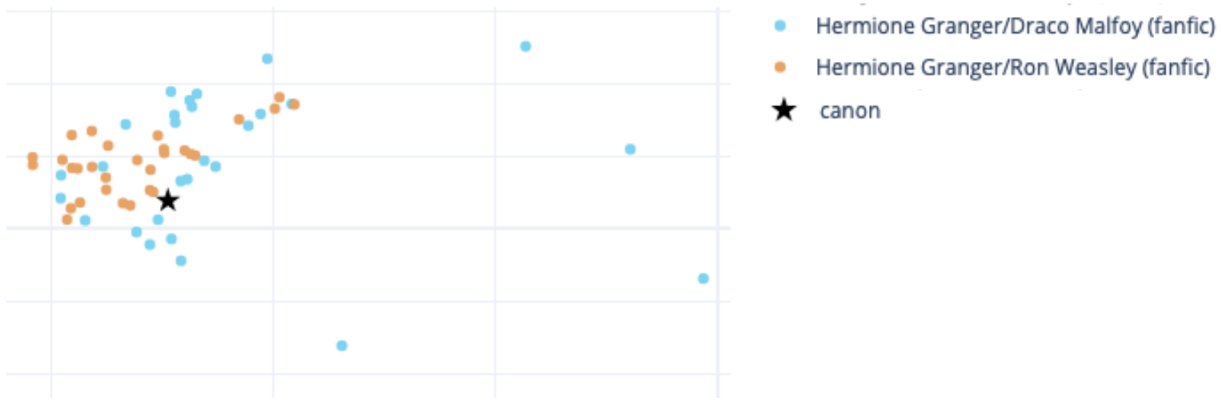

Figure 7.5: Representations for ‘Ron’ in fanfiction and canon, colored by relationship. A few outliers are omitted.

fan trope of 'Ron the Death Eater'[4], which turns a popular dislike of Harry's friend Ron into a casting of the character as evil. Further investigation would be needed to conclude this more concretely.

7.6 Conclusion and future work

Changes in the portrayal of entities from source to derivative texts can reflect the social differences between the communities that produced those corpora. In this work, we take first steps toward capturing such shifts. We focus on a domain with explicit labels for a new task: predicting character relationship changes in fanfiction. We find that focusing on text related to characterization, as well as using vector differences between source and derivative representations in word embedding space, is effective in this task. With a qualitative analysis, we find that our model is likely picking up on emotional emphasis in fanfiction around characters whose relationships have changed from the original source. In our exploration of character divergences, we find that learned representations for characters align with qualitative understandings of characters known to often be represented with different characteristics and in different situations in fanfiction than in the original stories. We also find evidence that these representations capture differences in character positioning depending on the relationships presented in different fanfiction stories.

Methodological improvements in future work could include attention-based approaches to automatically identify important distinctions between source and derivative corpora.

We assume that character relationships are stable in stories and canon text; future work could incorporate or predict changes in relationships based on the contexts and representations we present.

We plan to apply these methods to track influence and changes in more formal fiction, such as in folk tales or modern fiction. These methods could also be applied to settings outside of fiction where texts have clear relationships to prior texts (see Section 7.4.7). On social media, these methods could be used to "recontextualize" content that is lifted out of context to express a certain attitude toward speakers. In aggregate, such approaches could trace framing in derivative news stories or differing attitudes of social media communities.

[4] `https://tvtropes.org/pmwiki/pmwiki.php/Main/RonTheDeathEater`, accessed 31 August 2019.

Chapter 8

Fanfiction and the LGBTQ Social Movement

8.1 Abstract

Online communities organized around shared interests and hobbies often become sites for political and social action. In this chapter, we consider the relationship between an LGBTQ-friendly online fanfiction community and social movements for LGBTQ rights. We use quantitative techniques on a large corpus of fanfiction from 2010-2020 to see if trends in the portrayal of LGBTQ characters relate to events and trends in the LGBTQ social movement. We find that the representation of characters in same-gender relationships increases after the 2015 US Supreme Court marriage equality ruling, and that this time period also shows an increase in the use of marriage and relationship terms around LGBTQ characters in fanfiction. Fanfiction communities with higher estimated proportions of LGBTQ fans match issues in LGBTQ social movements more closely. These findings point to the potential of creative work from online communities as a rich, though noisy, indicator of shifts in social movements around identity.

8.2 Introduction

Many online communities organize around practices such as gaming or knitting that seem disconnected from social movements and politics. Yet these communities have increasingly played a role in offline social and political action. Examples of this include the "hacktivism" of Anonymous and the relationship between GamerGate and the alt-right. The relationship between these interest-based groups and offline social movements remains open to study. In this work, we present a case study of such connections between an online fanfiction community and LGBTQ social movements.

Fanfiction is fan-written fiction, sometimes referred to as *fic*, that creatively transforms original media such as TV shows, movies, books, and video games. Fanfiction writers, many of whom identify as LGBTQ (Dym et al., 2019), often use fanfiction to explore same-gender relationships and gender expressions outside societal norms. This has led many researchers to approach fanfiction as a "queer space" (Lothian et al., 2007; Ng, 2017), which is evident in fanfiction tags

and tropes such as "everyone is gay" and "het is ew"[1]. A culture of distributed mentoring within fanfiction (Evans et al., 2017) builds community around writing while also empowering LGBTQ authors and representations of LGBTQ characters. The lack of LGBTQ representation in mainstream fiction is a main draw for many fanfiction readers. In the words of one reader, "it's about high quality stories with LGBTQ characters. Fanfiction has so much more representation than actual books and it's fantastic."[2]

At the same time, online fanfiction has developed amidst the broader LGBTQ social movement around issues such as same-gender marriage and transgender rights. Fans themselves have engaged in activism for the continuation of certain shows and more queer characters in mainstream media (Navar-Gill and Stanfill, 2018). How does the community-building activity in fanfiction relate, if at all, to offline social action? Are changes in focus over time within the LGBTQ social movement reflected in the portrayal of LGBTQ characters in fanfiction? If so, fanfiction may be a source for tapping into the progression of views and issues within the LGBTQ movement, and point toward the possible use of online creative work as a grassroots measure of social change.

There has been much qualitative research on fanfiction (Jenkins, 1992; Tosenberger, 2008; McClellan, 2014). Here we wish to capture broad trends in the portrayal of LGBTQ characters with quantitative techniques over a large corpus of fanfiction posted 2010-2020 on Archive of Our Own[3]. We relate these trends to one of the most salient LGBTQ social movement turning points in the 2010s, the US Supreme Court marriage equality decision in 2015, before more broadly examining relationships between trends in fanfiction and trends in news on LGBTQ issues.

8.2.1 Hypotheses

Our main hypothesis is that trends in the representation of LGBTQ characters in fanfiction correlate with events and issues in the LGBTQ social movement. This analysis over historical data does not allow testing for causal relationships. We thus adopt a quasi-experimental design and look for convergent evidence across related, more specific, hypotheses:

- **H1:** We expect to see changes in the representation of LGBTQ characters in fanfiction around the 2015 US Supreme Court marriage equality ruling, a pivotal event in LGBTQ social movements from 2010-2020. We hypothesize that these changes will match known shifts in the LGBTQ social movement regarding this event. Specifically, we hypothesize a) more focus on same-gender married relationships around the event and b) a shift in focus to transgender issues after marriage equality.
    - **H1b:** Fanfiction appeals to fans across a wide variety of *fandoms*, fan communities organized around particular media series. We expect to see a higher degree of sensitivity to the above issues in the LGBTQ social movement from a set of fandoms with greater LGBTQ fan participation.

[1] `tvtropes.org`

[2] Excerpt from this Tumblr post was modified for privacy.

[3] `archiveofourown.org`

- **H2:** Marriage equality is just one significant event in the LGBTQ social movement, which has also focused on issues such as healthcare, transgender rights, and intersections with race and class. We expect that trends in the representation of LGBTQ characters in fanfiction correlate with trends in the issues that the LGBTQ social movement has focused on from 2010-2020.
    - **H2b:** We again expect to see these correlations more pronounced for fandoms with higher estimated LGBTQ fan representation.

8.3 Fanfiction and fan Activism

From its modern start in fanzines, fanfiction has been known for same-gender relationships, particularly gay male relationships such as the classic Kirk/Spock pairing from *Star Trek*. Today in vast online archives, representations of characters with diverse gender and sexual identities within fanfiction can be a resource for fans exploring their own gender and sexual identities in a space remarkably free of queerphobia (Dym et al., 2019). In this way, fanfiction as a community develops support and solidarity around an interest-based "affinity space" (Gee, 2004).

Increasingly, fans recognize their own power as media consumers and have greater expectations for both queer subtext in media and actual representation of queer characters (Ng, 2017; Navar-Gill and Stanfill, 2018). This "fan activism" often maps the values expressed in the fictional worlds beloved by fans into real-world political concerns. Unlike a community-building focus where the goal is to build connections for support and solidarity, fan activism may expect to challenge and change mainstream media or dominant culture. For example, organizations such as LGBT Fans Deserve Better have organized social media campaigns to direct messages to TV show producers about the representation of queer characters (Navar-Gill and Stanfill, 2018). Similarly, the Harry Potter Alliance uses a framework of values drawn from the *Harry Potter* books to organize more traditional political and social actions, such as phone-banking against anti-gay marriage legislation (Jenkins, 2015). We investigate if effects of this outward-focused, political engagement with LGBTQ issues can be detected in fanfiction.

8.4 LGBTQ social movements

LGBTQ social movements have been working toward "respect for the dignity of sexually marginalized identities, lifestyles, and subcultures" for decades, both in the US and around the world (Marche, 2019). From the moderate 1950s homophile movement to radical uprisings beginning with the Stonewall Riots of 1969, there have been "assimilationist" and "liberationist" tendencies in the movement (Stulberg, 2018). Pushes to reform existing social institutions, such as marriage, are often thought of as assimilationist demands, while liberationist parts of the movement more often focus on creating alternative institutions that do not rely on state recognition.

The LGBTQ social movement has spanned political, judicial, and cultural action. In this chapter, we examine the relationship between this action and a community that is not explicitly political. Social media and online organizing have enabled social movements with new communication infrastructure for online and offline action (Choi et al., 2020). Fanfiction may influence

the shaping of "cultural frames" (Andersen and Andersen, 2017) that enable the capacity of social movements. Stulberg (2018) notes that emotional connections to LGBTQ-focused art and culture can change "hearts and minds", as well as shift values and challenge notions of identity. These cultural shifts are often necessary for political action. Fanfiction may shape these frames to normalize the visibility of sexual and gender minorities, such as asexual or pansexual people, or to destabilize cis- and heteronormative assumptions about popular characters in fandom.

Art and popular culture, including online communities based on media consumption, are "central to political mobilization for LGBTQ social change" (Stulberg, 2018). This project investigates the connection between building community and social action by testing for a discernable relationship between fanfiction and LGBTQ social movements.

8.5 Data

We choose fanfiction from Archive of Our Own (AO3), one of the largest online fanfiction platforms. On AO3, writers tag their stories with rich metadata, including the names of characters, romantic relationships, fandom/s, and any other free-form tags (Fiesler et al., 2016). Archive of Our Own originated in the United States and likely has a US focus, but the site attracts fanfiction readers and writers from across the globe. See Chapter 3 for more information on this fanfiction and Archive of Our Own. We choose the time frame of 2010 through 2020, enough time for Archive of Our Own to mature since its establishment in 2007.

We wish to characterize broad trends in fanfiction written in English, so we sample fanfiction across a variety of fandoms. We start with the 40 most popular fandoms on AO3 as of January 2021 and filter to those that have significant activity throughout the 2010-2020 time period.

8.5.1 High- and low-LGBTQ datasets

We hypothesize that fandoms with more LGBTQ writers are more likely to match trends in the LGBTQ social movement.

To test this, we select 10 fandoms with high estimated LGBTQ fan representation and 10 fandoms with low estimated LGBTQ representation. To estimate fan demographics for fandoms, we start with a November 2018 data snapshot of Tumblr blog descriptions presented in Chapter 4. Tumblr is a blogging and social media platform popular with fans and known for an overlap with AO3 users (Fiesler and Dym, 2020). In these self-descriptions, Tumblr users often list their favorite fandoms as well as demographic characteristics such as age and pronouns (Chapter 4; Oakley, 2016). To estimate LGBTQ fan proportions, we calculate percentages of users in this Tumblr blog description dataset who list both a term related to a specific fandom and an LGBTQ identity label from a bank of regular expressions described in Chapter 4. We also triangulate these fan demographics with community surveys completed by fans and advertised on platforms such as Reddit. These surveys vary widely with respect to methodology and demographic categories, so we use them simply to verify the coarse ranking that we use for our main dataset split.

We seek fandom sets that maximize the difference in LGBTQ fan representation while keeping other factors, such as the age of fans, genre, and medium, as similar as possible. To select fandoms for each set, we started with the highest- and lowest-ranking fandoms in estimated

LGBTQ representation and manually adjusted to evenly balance other attributes. Fandoms in each dataset are listed in Table 8.1. The percentage of estimated LGBTQ fans averaged across all 10 fandoms is 17.7% for the high-LGBTQ set and 10.6% for the low-LGBTQ set. Note that these figures should not be interpreted as accurate demographic estimates, since fans on Tumblr differ from those in AO3. These numbers simply indicate a large difference in LGBTQ representation between these sets. Community surveys of fandoms in the high-LGBTQ dataset had consistently higher LGBTQ representation than low-LGBTQ fandoms for almost all fandoms.

Dataset	**High-LGBTQ**	**Low-LGBTQ**
Fandoms	Homestuck	The Walking Dead
	Star Trek	Shadowhunters
	Dragon Age	Song of Ice & Fire
	Buffy	Teen Wolf
	JoJo's Bizarre Adventure	Naruto
	Pokemon	JRR Tolkien works
	Danganronpa	Percy Jackson
	Glee	Harry Potter
	Fire Emblem	Attack on Titan
	Hannibal	DC Comics
Avg. LGBTQ %	17.7	10.6
Avg. fan age	23.0	22.8
Avg. fandom age	21.8	31.2
Avg. # stories	18,420	28,762
Genres	fantasy	fantasy
	manga/anime	manga/anime
	drama	drama
	sci-fi	superhero
	historical	mythology
Media	TV	TV
	movies	movies
	comics	comics
	videogames	videogames
	webcomics	books

Table 8.1: Fandoms in the high-LGBTQ and low-LGBTQ datasets with average dataset values estimated from Tumblr blog descriptions. 'Fan age' is estimated age of fans, while 'fandom age' is the average age of the media franchise.

The mean age of fans, estimated from Tumblr blog descriptions, is very similar: 23.0 in the high-LGBTQ set and 22.8 in the low-LGBTQ set. There are, however, significantly more stories per fandom in the low-LGBTQ set than the high-LGBTQ set. We sample uniformly from all fandoms in analyses, but it is worth considering that the low-LGBTQ fandoms seem to have more mainstream popularity than the high-LGBTQ set. There are also slight differences in genre

and medium. Both sets contain fandoms with fantasy, manga/anime, and drama storylines, but the high-LGBTQ set contains more sci-fi and the low-LGBTQ set has more superhero and book-based fandoms.

Our main dataset consists of stories from these 20 fandoms that are written in English, marked as "complete", and contain 1000-5000 words. We choose these word limits, which captures the vast majority of stories, for efficiency in processing story text for character analysis.

8.5.2 LGBTQ fanfiction tags

To track the representation of LGBTQ characters at a high level, we identify a number of fanfiction *tags*, specific combinations of metadata provided by authors about each story, relevant to LGBTQ issues. The tagging system is commonly used on Archive of Our Own (Fiesler et al., 2016); almost all fanfiction stories in our dataset contain tags marking information, such as characters and relationships. We consider tags that relate to LGBTQ issues, as well as two controls unrelated to the movement. One control tag is 'hurt/comfort', a popular genre tag. The other is a binary indicator of whether the story has the most frequent free-form tag in the fandom that a story belongs to. Tags considered are listed in Table 8.2. We label stories with tags based on regular expression matches in author-provided metadata.

Tag	Field	Description
trans	free-form tag	Trans-related tags such as *transgender* or *trans character*
F/F	relationship type	Female/female relationships
M/M	relationship type	Male/male relationships
F/M	relationship type	female/male relationships
Multi	relationship type	Multiple relationships or relationships with more than 2 characters
Other	relationship type	Relationships with other types of genders (e.g. with non-binary characters)
same-gender marriage	free-form tag, relationship type	Having a wedding or marriage free-form tag + an M/M or F/F relationship type tag, or explicitly mentioning gay marriage
different-gender marriage	free-form tag, relationship type	Having a wedding or marriage free-form tag + an F/M relationship type tag
hurt/comfort	free-form tag	A control tag unrelated to the LGBTQ movement
top fandom tag	free-form tag	Baseline: whether a story includes the most popular tag in its fandom

Table 8.2: Tags used to measure LGBTQ character representation in fanfiction.

The proportion of stories in each dataset labeled with 'trans' or 'same-gender marriage' tags, binned every 3 months, are plotted in Figure 8.1.

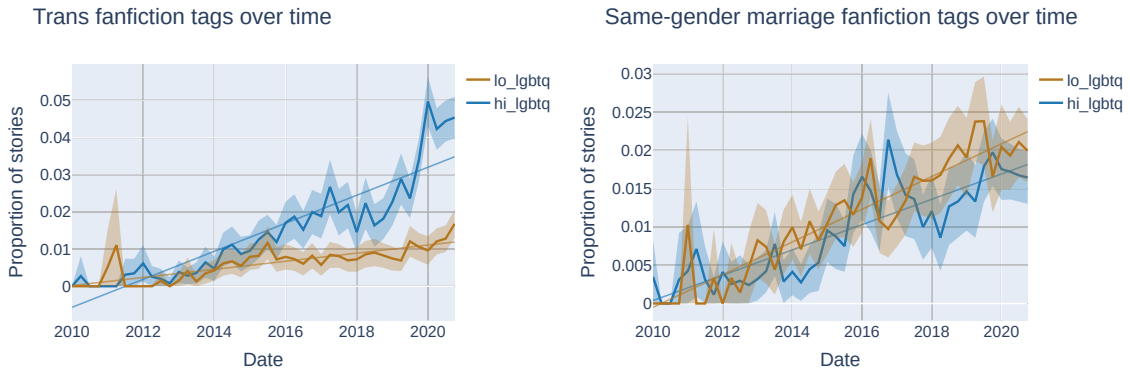

Figure 8.1: Proportions of fanfiction stories (in bins of 3 months) that display trans and same-gender marriage tags. Regression lines and 95% confidence intervals are shown.

8.5.3 Events in the LGBTQ social movement

We first look for changes in the representation of LGBTQ characters in fanfiction around events in the LGBTQ social movement. To identify the most important events in LGBTQ social movements from 2010-2020, we looked for events that were commonly included in timelines of the LGBTQ social movement. We collected seven timelines from popular online sources and organizations, including PBS, InfoPlease, BusinessInsider, the American Psychological Association, and a Wikipedia timeline (accessed November 2020). Our goal was not to rely on any particular source, but to note events across multiple sources.

Events that occurred in a majority of timelines were the 2010 repeal of "Don't Ask, Don't Tell" in the US military and two same-sex marriage judicial events. The first was the 2013 Supreme Court decision striking down California's same-sex marriage ban, Proposition 8, and the federal Defense of Marriage Act. The second was the landmark Supreme Court *Obergefell v. Hodges* case that legalized same-sex marriage throughout the United States in 2015.

The centrality of the mid-2010s marriage equality issue to LGBTQ social movements is confirmed by sociologists and legal scholars (Ghaziani et al., 2016; Andersen and Andersen, 2017; Stulberg, 2018). We thus choose to first look at the influence of US marriage equality in 2015 as a transformational change in the LGBTQ social movement. Apart from its national significance, this achievement also pragmatically shifted resources within LGBTQ activist organizations (Andersen and Andersen, 2017). Entire organizations dedicated to advocating for marriage rights, having achieved their goal, closed their doors or shifted focus. Marriage equality has been critiqued as an assimilationist goal benefiting mostly urban, white, gay men (Stulberg, 2018). After marriage equality was achieved in the US, scholars note a shift in focus for both pro- and anti-LGBTQ rights activists toward transgender issues (Andersen and Andersen, 2017; Stulberg, 2018).

8.5.4 Topic shifts in the LGBTQ social movement

To gain a sense of the relationship between character representation in fanfiction and the LGBTQ social movement at a finer level than a single event, we measure trends in a corpus of LGBTQ news from mainstream sources. This corpus of LGBTQ news consists of mainstream news articles from January 1, 2010 through December 31, 2020 from the News on the Web (NOW) corpus[4], filtered to articles that contain terms from a list of LGBTQ identity terms expanded from the list used by Mendelsohn et al. (2020). This LGBTQ news corpus contains 213,921 articles from English-speaking countries, mainly from the US.

To get a sense of which issues are represented in LGBTQ news over time, we use the Structural Topic Model (Roberts et al. 2014; STM). STM incorporates document covariates so that in our case, the prevalence of news topics is estimated to vary with the date articles are published. STM fitted with 20 topics and with the publish date of the news article as a covariate yielded coherent topics. These topics, along with highly ranked words for teach topic are shown in Table 8.3. Words in articles were lowercased, tokenized, and stemmed.

To estimate the prevalence of LGBTQ news topics over time, we fit a B-spline regression with the proportion of each document that contains a topic (from STM) as the dependent variable and the date (in addition to any other covariate of interest) as the independent variable, binned every 3 months (see Roberts et al. 2019). We see a general increase in the topic relating to trans issues, which Mendelsohn et al. (2020) also find in a corpus of *New York Times* articles in a period from 2010 to 2015.

8.6 Summary of analyses

We test our hypotheses with multiple regression analyses over different measures of trends in both fanfiction and LGBTQ social movements. We test for significant shifts in dependent measures of fanfiction from independent measures of LGBTQ movements and fandom demographics.

We perform four analyses, one relating each of two measures of trends in LGBTQ social movements to two measures of LGBTQ character representation in fanfiction (see Figure 8.2). First, we examine changes in fanfiction around 2015 US marriage equality, a pivotal turning point in the movement (**H1**). We then analyze finer-grained correlations between LGBTQ social movement topics measured from news (**H2**). For each analysis, we test for relationships using LGBTQ character representation in fanfiction tags and fanfiction story text, seeking convergent evidence.

8.7 H1: Effects of 2015 US marriage equality

We first test if the 2015 US Supreme Court marriage equality ruling is associated with changes in LGBTQ character representation in fanfiction. To do this, we use regression analyses that

[4] `https://www.corpusdata.org/now_corpus.asp`

Topic	Top terms
common words	realli, mayb, think, thing, guy
family	mother, parent, daughter
education, work	student, invest, program, fund
trans/queer/feminism	tran, gender, women, male
social media	cooki, content, user, facebook
ideology, Nazism	nazi, hitler, intellectu, ideolog
music	album, song, lyric, band, guitar
criminal justice	alleg, investig, prison, sentenc
law, same-gender marriage	legal, suprem, constitut, law
pride, black identity	pride, parad, rainbow, orlando
TV	episod, comic, hbo, sitcom
American politics	biden, trump, republican, clinton
British/Irish politics	ireland, turnbul, brexit, tori
travel	restaur, beach, beer, meal, wine
Global South	malaysia, nigeria, singapor
movies	oscar, film, filmmak, cinema
HIV, healthcare	hiv, infect, patient, diseas, virus
sports	footbal, athlet, nba, nfl, player
Christianity	pope, vatican, cathol, bibl
India, web artifacts	ltpgt, href, lta, src, mumbai

Table 8.3: Topics from STM over mainstream news articles on LGBTQ issues 2010-2020, ranked by overall prevalence in the corpus. Topic names were given based on top-ranked lemmas in each topic based on the FREX ranking (Roberts et al., 2019). Note that 'marri' appears soon after the top 4 terms for the law/same-gender marriage topic.

relate time period as an independent variable with dependent variables of LGBTQ character representation in fanfiction tags and story text.

8.7.1 Analysis of marriage equality and fanfiction tags

First, we investigated if the distribution of fanfiction LGBTQ tags changes before and after US marriage equality. We set up a logistic regression analysis with the presence of an LGBTQ tag on a story as the dependent measure and include whether a story was published before or after US marriage equality as an independent variable.

Our dataset for this regression consists of fanfiction stories published one year before to one year after US marriage equality on June 26, 2015. We sample 799 stories (the size of the smallest fandom set in this time period) from each of the 20 fandoms. To include all tags in one regression, we repeated each story for each LGBTQ-related tag, and marked the instance as `TRUE` if that tag was present for that story and `FALSE` if not. This is the outcome measure predicted in the regression. We include independent variables displayed in Table 8.4, as well as

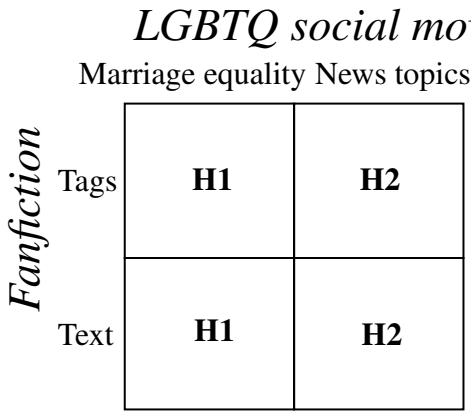

Figure 8.2: Summary of analyses. Each axis represents different measures of activity from fanfiction or the LGBTQ social movement. Within each cell is the hypothesis corresponding to that analysis.

Effect	Type	Description
after event	binary	Whether a story is published after marriage equality
dataset	binary	High- or low-LGBTQ fandom set
fandom	categorical	Fandom of the story, nested within dataset
tag	categorical	Specific tag from Table 8.2
dataset * tag	interaction	
after event * tag	interaction	
after event * dataset	interaction	
after event * dataset * tag	interaction	

Table 8.4: Predictors included in logistic regression analysis of US marriage equality and fanfiction LGBTQ tagging.

interactions between these variables, as predictors.

Free-form tagging behavior overall on AO3 increases over time. To control for this trend, we include average monthly tag use as a covariate in the logistic regression analysis.

Results

From this logistic regression analysis, we find a significant increase in same-gender marriage tags after the event (0.16 coefficient difference, $p = 0.03$). This is higher than the control tag of hurt/comfort, which decreases slightly in frequency after US marriage equality. This provides evidence for a change in fanfiction around this salient event in the LGBTQ social movement. Plotting the probabilities of same-gender marriage tagging over time in Figure 8.1, we see that this effect does not come from a spike directly after the US Supreme Court decision in June 2015, but instead a steady rise from 2015 to 2016.

Examining stories included in this analysis after US marriage equality, we find many representations of LGBTQ married relationships. Some seem to be intentionally celebratory stories;

for others it is not clear. For example, one story from August 2015 was written in response to a honeymoon prompt that may have been inspired by the US marriage equality decision. The story centers on a honeymooning couple of popular male *Glee* characters. The scene is domestic, with the couple's fingers entwining as one character thinks "I don't know if I'll ever get used to my husband's small signs of love".

Shift from marriage equality to trans rights

We hypothesize a shift in focus from marriage equality issues to trans issues in fanfiction after marriage equality, following a similar shift in LGBTQ social movements Andersen and Andersen (2017); Stulberg (2018). From the logistic regression analysis, we find no significant change in 'trans' tag use after marriage equality. This does not support evidence of an immediate shift in focus to trans issues following a shift in focus in the movement. We do see an increase in trans tags over time, though this is change is much more dramatic for the high-LGBTQ fandom set (Figure 8.1). We also see a rise in the Other relationship tag, often an indicator of non-binary characters in relationships, in high-LGBTQ fandoms specifically.

This suggests the presence of a general shift toward trans issues in fanfiction, especially in high-LGBTQ fandoms, but not one that is concentrated around a specific event.

8.7.2 Analysis of marriage equality and fanfiction characters

In addition to a statistical effect on tagging, we expect to see an increase in LGBTQ characters portrayed in domestic, married life in fanfiction story texts after US marriage equality. To test this, we extract text relevant to character portrayals in fanfiction stories and apply a topic model to learn trends in character representation and graph such changes over time.

We use the NLP fanfiction processing pipeline described in Chapter 6 to extract which characters are present in stories and the text portraying those characters. This pipeline provides a list of characters for each story, as well as associated character mentions, actions and attributes, and quotes for each character. Character mentions are any mention, including pronouns, resolved to refer to a particular character. Character *actions* and *attributes* are descriptions of characters and activities that character is portrayed as doing. Specifically, these are verbs for which a character is subject or object, as well as adjectives and appositives whose head word is a character mention.

To infer which characters are placed in straight or LGBTQ relationships, we use author-provided tags of romantic relationships present in the story and gender inferred from pronouns in stories. We match characters in tagged relationships to characters identified by the processing pipeline if there is a sufficient name overlap, then infer character gender (female, male, or non-binary) from the pronouns identified as character mentions by the pipeline. We do not consider characters that are not in relationships annotated by fanfiction story authors and only consider relationships between two characters that are not in any other relationships. A relationship is labeled 'straight' if one character is inferred female and the other is inferred male and 'LGBTQ' if both characters are the same gender or one is non-binary. Note that either character in a relationships where one character is female and the other is male could be trans and the relationship is still labeled 'straight', a possible shortcoming. We process the text of fanfiction stories from the 20 selected fandoms published within our time frame, January 1, 2010 through December 31,

Topic	Top FREX terms
found/wonder	found, wonder, recal, recogn
cognition	say, think, tell, know, sit
quotatives	said, ask, repli, nod, shook
people	someon, peopl, happen, new, day
movement/expression	smirk, lean, older, slid, younger
body parts	bodi, lip, along, tongu, soft
feel/hold	feel, hold, bite, touch, hear
sleep/eat	sleep, eat, cough, feed, punch
love/hate/marry	love, die, miss, marri, deserv
sex	moan, gasp, insid, whine, felt

Table 8.5: Topics from STM over fanfiction character actions and attributes, ranked by overall prevalence in the corpus. Topic names were given based on top-ranked lemmas based on the FREX ranking (Roberts et al., 2019).

Topic	Top FREX terms
fillers	okay, gon, wanna, shit, fuckin
face/gaze	gaze, shoulder, forehead, eye
marriage/relationships	marri, ship, dear, famili, maker
like/hate	say, hate, person, peopl, like
reactions	said, repli, blush, sat, nod
sword/dragon	sword, upon, dragon, fear, grace
sex	moan, cock, hip, thrust, thigh
logistics	phone, car, text, class, school
sleep/health	sleep, wake, asleep, night, awak
special occasions	christma, cake, flower, chocol

Table 8.6: Topics from STM over fanfiction character quotes.

2020. Each document for topic modeling contains one character's quotes, actions and attributes in a particular fanfiction story. We uniformly sample 8100 such documents from each fandom (the lowest amount across fandoms).

To identify patterns in character representations, we again estimate STM to identify topics within these documents of character features. We include the date a fanfiction story is published, whether the character is in a straight or LGBTQ relationship, and whether the fandom is in our high- or low-LGBTQ set as covariates. STM estimated with 10 topics on separate datasets of character quotes and character actions and attributes yielded coherent topics. Top terms in the topics from STM estimated on character actions and attributes, using the FREX ranking (Roberts et al., 2019), can be seen in Table 8.5, and for quotes in Table 8.6. Articles were lowercased, tokenized, and stemmed before training. We removed words that appeared in fewer than 100 documents for actions and attributes, and fewer than 50 documents for quotes (which had a smaller set of documents).

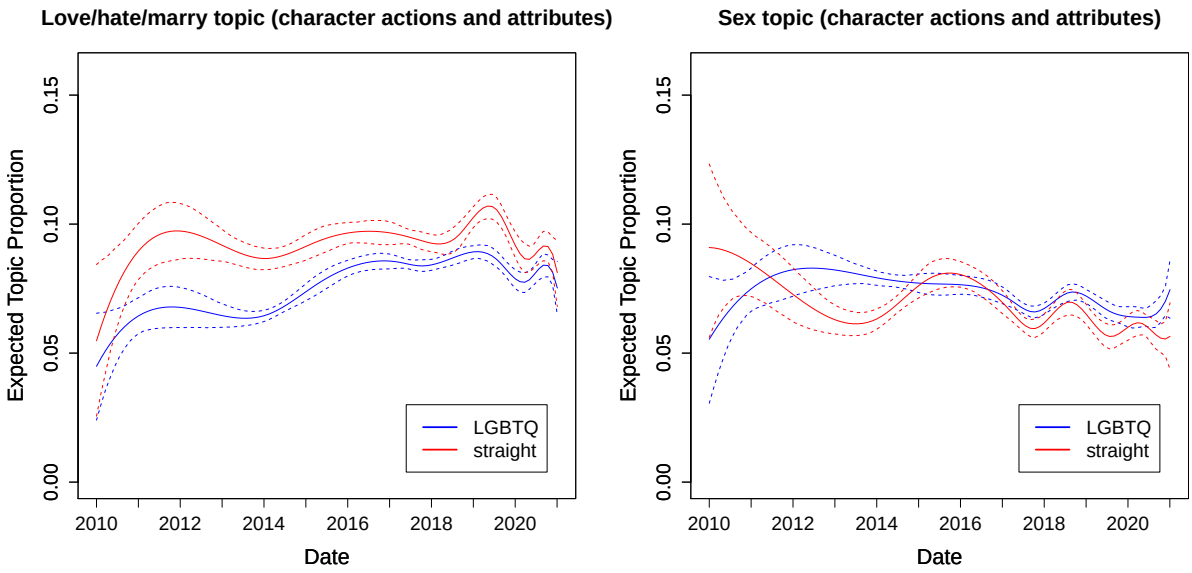

Figure 8.3: Topic proportions from character actions and attributes over time.

We estimate the effect of covariate values (publication date, type of relationship, and high-/low-LGBTQ dataset) on document topic proportions by fitting a B-spline regression predicting topic proportions. We then analyze the effects of covariates on topics that relate to relationships and marriage.

Results

In character actions and attributes, we find an increase from 2015-2016 in LGBTQ characters portrayed in the topic that includes terms related to marriage, as well as strong emotional relationship language such as "love" and "hate" (Figure 8.3). This rise brings the proportion of LGBTQ characters in this topic closer to the level for straight characters. This contrasts with the 'sex' topic, which declines in frequency for LGBTQ characters through 2016.

An example of a character representation high in the 'love/hate/marriage' topic is found in a *Shadowhunters* story about the engagement of two male characters. They are described with, "Who would have thought they would be the first lucky ones to be engaged? But here they were telling friends about the upcoming wedding."

In character quotes, we see a rise in prevalence for characters in LGBTQ relationships from 2015-2016 around US marriage equality for the 'marriage/relationships' topic, though this topic remains more common for straight characters (Figure 8.4). An example of this increase is a dialogue between a male/male pairing in a late 2015 *Star Trek* story: "You would get the visa with a job sponsor, or if we were married... If you're proposing, you should follow my people's protocol."

These trends provide convergent evidence for a rise in marriage-related portrayals of LGBTQ characters in AO3 after US marriage equality.

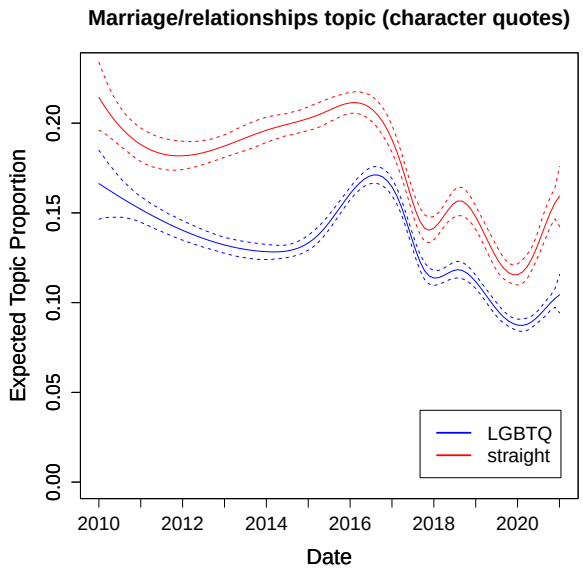

Figure 8.4: Topic proportions from character quotes over time.

8.7.3 H1b: High- and Low-LGBTQ Fandom Differences

We hypothesize that a relationship between trends in fanfiction and the LGBTQ social movement is more likely to be pronounced in fandoms with more LGBTQ fans. To test this hypothesis, we examine relationships separately for high- and low-LGBTQ fandom sets.

From the logistic regression analysis of tagging, we find a small but significant interaction between 'after event' and 'high-LGBTQ' (0.07 coefficient, $p < 0.01$). This suggests that the increase in same-gender marriage tagging in fanfiction from 2015-2016 is led by high-LGBTQ fandoms.

We also see a significant difference in the use of trans tags between fandoms, with a higher rate for high-LGBTQ fandoms (0.38 coefficient, $p < 0.01$). Plotting these tagging proportions over time (Figure 8.1), we see a dramatic spike in tagged trans characters in 2019-2020 for high-LGBTQ fandoms. Qualitatively investigating stories in this spike, we find no evidence of a particular fandom or trope driving this increase. Once again, we find evidence against an immediate shift to trans issues after marriage equality in fanfiction, though there is a gradual shift toward trans portrayals.

8.8 H2: Effects of topic shifts in the LGBTQ social movement

Tracking topic changes in the LGBTQ news corpus gives a measure of the visibility of issues related to the LGBTQ social movement over time, beyond just the marriage equality event. We hypothesize that these topic changes align with changes in LGBTQ character representation in fanfiction. To test this, we again perform regression analyses with dependent variables of LGBTQ character representation in fanfiction tags and text, and independent variables related to topic shifts in LGBTQ news.

8.8.1 Analysis of LGBTQ news topics and fanfiction tags

To test for associations, we set up a regression predicting LGBTQ tagging behavior over time from LGBTQ news topic proportions over time (see Data section for details on extracting tags and news topics). For estimating tagging behavior from 2010-2020, we randomly sampled 8600 stories (matching the smallest fandom) from each of the 10 fandoms in the high- and low-LGBTQ datasets.

To control for the overall increase in free-form tagging on AO3 from 2010-2020, we first calculate the residuals from a linear regression predicting the use of specific tags from the average monthly free-form tags per story, which rises steadily over time. We use these residuals, which represent the variance in specific tag use apart from the general increase in tagging over time, as the dependent variable in a linear regression model. The independent variable in this regression is topic prevalence in LGBTQ news over time. All features are normalized to have a mean of 0 and standard deviation of 1.

Results

We find a positive relationship between 'same-gender marriage' tagging in fanfiction and the 'pride, black identity' news topic (0.38 coefficient, $p = 0.01$). The expected LGBTQ news topic proportion for the 'pride, black identity' topic plotted alongside same-gender marriage fanfiction tag proportions over time is shown in Figure 8.5. This suggests a relationship between LGBTQ tagging practices and one of the most visible LGBTQ social movement events, Pride, as well as intersectionality with race. However, we do not find a significant relationship between same-gender marriage tagging in fanfiction and the highly related law/same-gender marriage news topic. This suggests that at the granular level of topic shifts in LGBTQ social movements from 2010-2020, same-gender marriage tagging in fanfiction is more related to cultural aspects like Pride than legal issues regarding marriage equality. We also find a negative relationship between 'trans' tagging in fanfiction and the 'trans/queer/feminism' news topic when factoring out the general tagging increase (-0.37 coefficient, $p = 0.01$). This suggests that transgender character representation in fanfiction does not align well with transgender issues in news.

8.8.2 Analysis of LGBTQ news topics and fanfiction characters

We hypothesize that shifts in the portrayal of LGBTQ characters in the text of fanfiction stories also reflect reflect topic shifts in the LGBTQ social movement. To test this, we perform a linear regression analysis predicting the proportion of topics in characters' quotes, actions and attributesn in fanfiction stories. We include the proportion of news topics during the publication date of a story, whether the story was from a high-LGBTQ fandom, and interactions between these variables as independent variables. Only a few of the 20 news topics correspond to topics in LGBTQ character portrayals, so we just include 'law, same-gender marriage', 'pride, black identity', and 'trans/queer/feminism' as features. All features are normalized to a mean of 0 and standard deviation of 1. Since we are interested in shifts in the portrayal of LGBTQ characters, we narrow our focus to characters in LGBTQ relationships in the same dataset of character features used in the analysis for **H1**.

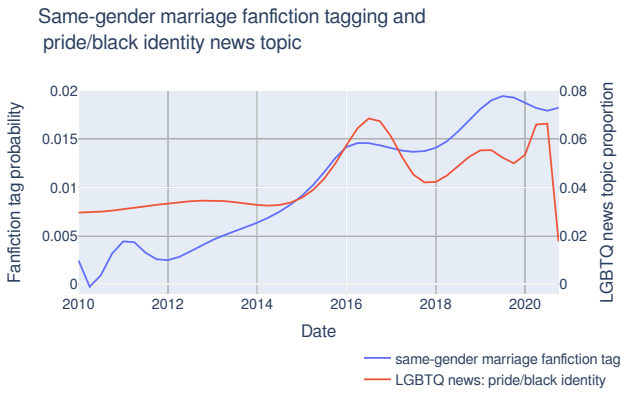

Figure 8.5: News topic and fanfiction tag probabilities.

Results For the 'law, same-gender marriage' news topic, we find a small but significant negative relationship (-0.09 coefficient, $p < 0.01$) with the marriage topic in quotes from LGBTQ characters, and a non-significant relationship from actions and attributes. This suggests that at this more granular level across the entire 2010-2020 time period, we do not find a correlation between marriage representations in fanfiction and legal-focused marriage talk in LGBTQ news.

However, there is a small but significant positive relationship between the marriage topic in character quotes and the 'pride, black identity' news topic (0.07 coefficient, $p < 0.01$). Just as with tagging behavior, this provides evidence for a relationship between culturally relevant LGBTQ social movement news, such as Pride, and representations of same-gender marriage in fanfiction. This relationship does not hold with the 'love/hate/marriage' topic from character actions and attributes, however, which has a small but significant negative relationship with the 'pride/black identity' topic in news (-0.08 coefficient, $p < 0.01$). This could be due to that topic having strong negative relationship terms (such as 'hate') that are different from Pride. It could also simply indicate an unevenness in this relationship between same-gender marriage representations and Pride news.

8.8.3 H2b: High- and low-LGBTQ fandom differences

We again test whether these trends are more pronounced for fandoms with high estimated LGBTQ fan representation. We find that the relationship between the 'same-gender marriage' tag and the 'pride, black identity' news topic is significant for the high-LGBTQ dataset (0.50 coefficient, $p < 0.01$) but not for the low-LGBTQ dataset. This provides convergent evidence that high-LGBTQ fandoms are more responsive to LGBTQ social movement trends, and in particular cultural events such as Pride.

Analysis of news topics and fanfiction text, however, gives a more mixed indication of this relationship. We find a negative relationship with the interaction term of high-LGBTQ fandoms and the 'pride, black identity' news topic in predicting fanfiction marriage quotes (-0.49 coefficient, $p < 0.01$), but a positive relationship with actions and attributes (0.25 coefficient, $p < 0.01$). This suggest that Pride, particularly for high-LGBTQ fandoms, is more likely to be

detected in tags than in how LGBTQ characters are voiced, though there is an effect on how these character are described.

8.9 Discussion

Overall, we find evidence for a positive relationship between LGBTQ character representations on AO3 and trends in LGBTQ social movements. We find evidence supporting our first hypothesis, that there is a notable difference in LGBTQ content in fanfiction around one transformative event in the LGBTQ social movement, US marriage equality. Specifically, we see increased tagging of marriages and weddings in stories with same-gender relationships. In the portrayal of LGBTQ characters in story texts, we also find an increase in marriage-related terms. However, the LGBTQ trends in fanfiction do not appear to dramatically or simplistically reflect changes in the LGBTQ social movement. We do not find evidence for a shift to transgender issues after US marriage equality in fanfiction, as seen in LGBTQ social movements. Instead, we observe a more gradual shift to trans issues over the 2010-2020 period, with fandoms with high-LGBTQ fan representation leading the way.

We also find evidence supporting our second hypothesis that trends in LGBTQ character portrayal in fanfiction correlate with trends in the LGBTQ social movement. However, at this more granular level of LGBTQ social movement trends from 2010-2020, we find a positive correlation between same-gender marriage tagging in fanfiction with cultural events such as Pride, but not with more law-related marriage equality news. This could point to fanfiction's relationship with cultural aspects of the LGBTQ social movement, and more broadly as a part of shifting cultural frames that lay the groundwork for social action (Stulberg, 2018).

Beyond this particular social movement, these findings suggest a role for creative works, such as fiction, as a source of grassroots data reflecting the cultural aspects of social movements. Using creative artifacts from online communities, like fanfiction, is certainly noisier than surveys, interviews, or traditional measures of social movements or attitudes. But such artifacts can reflect the rich, natural expressions of a specific time in a movement.

8.9.1 Ethics

Anonymity and privacy are very important to fanfiction communities (Fiesler et al., 2016; Dym et al., 2019). Fanfiction authors often prefer consent before using their work in research (Dym and Fiesler, 2020). This is not possible for quantitative work on fanfiction at a large scale, but it emphasizes that such work must be held to high standards of privacy. All fanfiction examples in this paper were modified to hinder finding original sources. Author handles were not saved with datasets, and no analysis was done at the level of individual authors. For ethical scraping, we communicated with Archive of Our Own and respected their guidelines for collecting data for this project.

There is a history of research on LGBTQ issues that harmfully biologizes or "Others" people who identify as LGBTQ, and this extends to computer science (Wang and Kosinski, 2018). Our work attempts to study patterns within an LGBTQ-friendly online community to better understand how positive connections are made with social action outside the online community.

8.9.2 Limitations

Confounds such as the general rise in AO3 tagging behavior and the idiosyncrasies of specific fandom communities may also affect LGBTQ character representation in fanfiction. To control for these, we uniformly sampled across many fandoms and included covariates in regressions. However, other influences may still have affected our results.

We use pronouns to infer character gender, but these are not always aligned in expected ways and can vary within stories. We also restrict our character analysis to characters in relationships with only one other partner in each story. This removes characters who change partners in the course of a story and disproportionately removes characters in polyamorous relationships. We hope to extend this analysis to include these relationships in future work.

8.10 Conclusion

Using quasi-experimental methods on observational data, we find evidence of a positive correlation between events and topics in the LGBTQ social movement from 2010-2020, and LGBTQ character representation in a large corpus of fanfiction. These correlations between issues in fanfiction and LGBTQ social movements support a view that this interest-based community is connected to social movements offline, especially with cultural issues such as Pride. These findings also suggest that creative works may be a source for tapping into the progression of views and issues within social movements. With this relationship established, we hope to investigate more specifically if attitudes around LGBTQ issues lead or follow the broader movement in future work.

Chapter 9

Conclusion

This thesis provides a framework and set of methodology for answering research questions about the relationship between identity presentation and social interaction in online communities at a large scale (see Section 1.1). The first step in this framework is extracting theory-driven features of identity presentation in online communities. We then relate those identity presentation features to social interactions and outcomes. Using these interactions as a lens, we can see how the presentation of identity is valued in online communities or even reflects offline contexts. This helps us understand how identity presentation "matters" in these spaces (Chapter 4) or how identity portrayals are changing over time in online communities (Chapter 8). This framework also demonstrate the applicability of a conception of identity as dynamic and contextual in computational social science and NLP.

In this chapter, I summarize how projects implemented this framework in both identity feature extraction (Section 9.1) and the relation of those features to social interactions and outcomes (Section 9.2). I then discuss implications of this thesis and recommendations for both NLP researchers (Section 9.3.1) and other computational researchers studying online communities (Section 9.3.2).

9.1 Features of identity presentation

In this thesis, I bring the perspective of identity emerging in discourse and interaction from the humanities and social sciences into computational research and NLP. I argue for a theory-driven approach to identity in language as a contextual display rather than a static attribute or characteristic of people. Projects in this thesis incorporate the social construction of identity mainly in how features for identity are extracted.

We challenge the default paradigm for such researchers to simply encode the identity of speakers (inferred or obtained in any way) as another static attribute to relate to language use. Instead we argue for representations of identity labels (and speakers) that change with the time, situation, and context. How was this accomplished in practice in this thesis?

First, we took into account social media users' *choices* to present identity labels. Not to provide such labels is also a choice users make. Instead of treating that as missing data, we capture whether or not users present labels as a choice in Chapter 4. We also take into account

users' open-ended choices of identity labels and dimensions of identity that are relevant to them, instead of defining identity categories of interest *a priori*, which may not be relevant to users on the platform. In Chapter 4 we bootstrap identity categories and labels that are popular on Tumblr, while in Chapter 5 we identify communities emergent in network structure.

Second, we looked specifically at how identity portrayals change. We used and developed NLP tools to capture aspects of identity change in narrative text. With data from fanfiction, we first develop a pipeline to extract text segments and features about the portrayal of character identity (Chapter 6). Using this pipeline, along with author-provided metadata and tagging behavior, we extract which characters are presented in text in specific types of relationships (Chapters 7 and 8) and how the associations made with characters of different genders and sexualities change over time (Chapter 8). This thesis also demonstrates that NLP representations of text, such as embeddings for words and phrases, can be effective in capturing aspects of identity change. For example, in the case of fanfiction we found word embeddings representing the context of text character portrayal to be effective in predicting relationship change from an original source text (Chapter 7).

9.2 Identity presentation and social interaction

With data from both social media and online narrative, we relate this identity presentation to social outcomes and interactions in order to understand the effects of identity presentation. In Tumblr, we find evidence for identity alignment between users relating to content propagation (Chapter 4). We also find a significant but small association of network community alignment on content propagation (Chapter 5). Using content propagation as a signal allows us to observe the construction of social solidarity or alignment that comes from identity presentation in language. In fanfiction, we use measures outside fanfiction text, such as metadata indicating a change in a relationship or events related to the LGBTQ social movement, to discover how values around identity are signalled within the text of stories. We find that NLP word embedding-based representations of character portrayal are able to predict the changes in romantic relationships that give fanfiction such high representation of LGBTQ characters (Chapter 7). We also find that patterns in the portrayal of characters of different genders and sexualities correlate with events and topics over time in the LGBTQ social movement offline (Chapter 8).

9.2.1 Challenges in relating identity presentation to social interaction

This thesis offers lessons learned for those hoping to use NLP and other computational methods for investigating identity presentation in online communities. With data from Tumblr, we set out to design a paradigm that allowed measuring reactions to identity presentation by modeling the choice of sharing content from one user over another. However, this paradigm did not end up being the best setting for seeing this effect in isolation, as experiments with a larger dataset showed only a small effect of identity and community alignment (Chapter 5). Rather counter-intuitively, the more implicit portrayals of identity in fanfiction story text were more fruitful for discovering relationships between identity presentation and social interaction and outcomes.

There are surely many reasons for this, but I will draw attention to two that are methodological: sampling and confounds. In the case of Tumblr, we sampled broadly from users above a threshold of activity since there are no clear indications of factors such as community membership. This may have contributed to the lack of the ability to isolate effects from identity presentation from issues such as topic, content, and interests more generally. The conflation of topic and features of content with identity and communities on Tumblr was challenging. In Chapter 5, we recognized that some of the content hashtag topics that were most informative to content propagation prediction indicated communities. This conflation of identity and community membership expressed through content is a feature of Tumblr which can draw in users who may not want to be direct about their identity, especially if it is marginalized or stigmatized. Such properties made it difficult to disentangle features of identity presentation from features of the content, and we had to shift focus to understanding how these worked in concert on the platform.

In contrast, the tags that indicate which fandom a fanfiction story belongs to offered a clearer demarcation of groups for sampling. This allowed us to sample uniformly across fandoms and more carefully control for this important factor affecting content. In many ways, identity was also more clear in fanfiction, especially in the tags placed on stories by authors. This could be in part because there was less at stake: authors were not presenting their own identities, but those of characters. With a perspective from the social construction of identity, all associations with identity are constructed in interaction and associations made with characters can be just as relevant to study as associations made with identity self-presentation. This also suggests that indirect identity presentation of others, or through others, may be more clear and easier to separate from confounding factors, perhaps especially for those with marginalized identities.

9.3 Implications and Recommendations

9.3.1 Recommendations for NLP

At a high level, our recommendation from this thesis is that NLP work should not take attributes of identity as "given" or natural. Informed by the social construction of identity, we do not assume that there are expectations for behavior based on a stable set of identity variables defined by researchers. Assuming that identity is the source of variation in behavior casts identity as natural, inherent, personal, and stable. Instead, we analyze identity presentation as the impression a person gives through a set of behaviors in a particular social context. This allows flexibility and agency in if, how, and in what ways people present their identities. We also find that it better matches data from linguistic and social interaction in online communities.

With this change in perspective about the nature of identity, NLP should consider:

- the social, cultural, and historical contexts that give rise to certain identity categories and distinctions,
- which aspects of identity are relevant to behavior and language production in particular situations, and
- what happens when people do not cleanly fit such identity categories.

What does this look like when enacted in practice? This thesis is an attempt to illustrate how this

perspective can be incorporated into computational studies. But what recommendations does this work imply for others in NLP working on different questions and datasets involving identity and language? This section provides specific recommendations along these lines.

Be specific about what is needed from identity types. Identity is multifaceted. Instead of viewing identity as the existing, natural trait of a person that is to be discovered or labeled from behavior (as user attribute inference does), it is more consistent with social theory to think of identity as a particular lens that a researcher takes on people. Different lenses, or in this case dimensions of identity, will be more or less useful for explaining social or linguistic behavior in different situations for different research questions. Thus, it is important to be specific about what aspect of identity is the lens in a particular project. In their argument for a critical view of race as socially constructed in computational work, Hanna et al. (2020) provide a starting place for considering the multiple dimensions of identity and tying these to which sorts of studies they fit with. Their list of dimensions of race includes:

- Self-identification, either open-ended or from a closed set of options
- Perceived: what identity category others would say you are, either from a single observation or from interaction. They also consider what a person might say others would perceive them as.
- Phenotype: these are more "objective" (measurable) characteristics of the body, such as skin shade, hair texture, eye color and other physical features.

Some of these may seem to contradict a view of identity as socially constructed, especially the physical dimension of phenotype. Using such a perspective demands a consideration of the harm caused historically by those practicing scientific racism, phrenology, and other pseudosciences that claim that particular traits originate in the bodies of those judged to be inferior. But adopting the social construction of identity does not mean disregarding the saliency that belief in a physical or biological nature of identity has in societies. People are treated differently based on the identities that others perceive them to "naturally" be. Focusing on perceived identity or even phenotype may be useful when measuring discrimination and bias. These perceptions are culturally and historically specific, but within clearly described boundaries it may be helpful to quantify them with measurements of physical features such as how dark someone's skin is. If what is relevant about someone's treatment is their perceived identity, then that is what a researcher should be attempting to measure.

The growing field of fairness, accountability and transparency in AI commonly attempts to measure bias that machine learning systems reproduce in data, methods, or outcomes. In this case, the relevant lens of identity is what these systems "see" or infer about users. For example, Sap et al. (2019) and Davidson et al. (2019) establish that hate speech classifiers predict lexical items and phrases representative of African American English to be more hateful than phrases from white-aligned English. In this case, what is relevant to a machine learning system is not what someone's "true" race is, because that does not exist and is always culturally defined, but rather their expressed language, which is associated with being African American in the United States or other particular contexts. People who are African American often do not use African American English, just as everyone modifies their language depending on the context. So self-described or perceived identity would not be as good a fit to measure bias against

African American English. Self-described race, perceived race, or how others perceive one's own race may be more relevant in a study of how race interacts with gender or occupation in business meetings or emails (see Prabhakaran et al. (2014) for similar work on perceived gender in emails).

As shown in this thesis in work on Tumblr (Chapters 4 and 5), the mode of identity expression also plays a role. NLP work reinforces a view of identity as personal and natural when it seeks to infer users' identities through multiple means or any means necessary without regard to the mode of presentation. An example of paying attention to the mode of identity presentation is the work of Wang and Jurgens (2018). They find differences in reactions of support online according to not only gender, but the mode of gender expression, revealed through a name or by using language that is associated with particular genders. If what is relevant is users' own self-expressed identities, look at the specific terms used by users. We found specific combinations of terms to be relevant in predicting content propagation on Tumblr. More broadly, it is well-established that particular terms used for identity hold particular associations, as Mendelsohn et al. (2020) and other have found around the term 'homosexual', for example.

Consider cases of those who do not fit the identity categories a researcher provides. Though we attempt to avoid assigning predefined categories or dimensions of identities (such as gender, race, sexual orientation, etc.) in the projects in this thesis, many projects require such categorization. These include projects measuring bias and discrimination against particular groups, where a definition of who is in the group and who is out of the group is necessary. Such a discrete classification of individuals into groups contrasts with the social construction of identity, which argues for an understanding of individual identity as fluid and contextual. But this tension is not new to social science, particularly areas concerned with social activism. An "essentialist", or biological/natural view of identity is often adopted to illustrate discrimination based on that identity. This is often termed "strategic essentialism" (Spivak, 1988; Bucholtz and Hall, 2004). However, this adoption is not easy or straightforward. There are many controversies over the boundaries of groups in social movements. In a classic case, there has been much controversy over who gets to count as a "woman" in second-wave feminism, which often denied access to transgender women.

Work in computational social science and NLP often implicitly or explicitly involves decisions about who is in a group and who is not. For example, Huang et al. (2020) bases decisions about who is Latinx based on a facial recognition algorithm on Twitter profile pictures, instead of other widely adopted criteria such as speaking Spanish and being from the Americas. The recommendation from this thesis is to at least acknowledge, and at best center, the cases of those who do not fit cleanly into whatever categories are determined. If a researcher has discrete categories for ethnic groups of African Americans and Latinx people, what about the large population of Afro-Latinx people in the United States? In a study of bias based on perceived identity, language from Afro-Latinx people may lead to a perception as Latinx, while a profile image may lead to a perception as black. Such cases illustrate the importance of a researcher's choice of identity as a lens through which to view people and their behavior online, or how people are treated based on others viewing them through the lens of identity. A computational project that centers the experiences of people who do not fit categories cleanly may specify those cases and look

more carefully at how technology interacts with their experiences. For example, are people from Spain cast as "white" or European in some online community contexts, while they are "Latinx" in others? How are the vast diversity of ethnicities, nationalities, perspectives and experiences of "*Latinidad*" flattened in US-centric online contexts and how do Latinx users respond or challenge this erasure? With gender, how are non-binary individuals considered in systems such as computer vision gender prediction algorithms (Keyes, 2018)?

Part of considering the cases of those who do not fit a researcher's categories is recognizing those who do not present a particular identity at all. Instead of treating those who do not give themselves labels as missing data, consider how a lack of self-disclosure may affect behaviors and reactions of others online (as we found that not presenting pronouns had an effect on content propagation in Chapter 4). For example, if a computational project is considering how the presentation of femininity relates to interaction in software development, contrasting with those who present no gender may draw this contrast into relief, as Vedres and Vasarhelyi (2019) find with open-source development on GitHub. If other users cannot tell who is a woman on a platform, it is not helpful to try to infer gender through a variety of means in an effort to learn some sort of "natural" association between social interaction and being a woman.

Consider new research questions that look into how identities are changing, forming, or developing associations. The two sections above largely describe recommendations for existing, established areas of NLP and computational social science, such as bias, fairness, and ethics. Much of this work currently treats identity as given and looks for how behavior varies around those fixed identities. But adopting the social contruction of identity leads to new research questions as well.

My recommendation for the field is to explore how identities themselves are changing and being shaped in online platforms. This would include how new identities, new terms for identities, and new associations for well-established identities are formed online as a process, not an end result. For example, how do people align themselves or disalign themselves with groups such as "techie" on help forums, or how are the meanings and associations with terms like "stan" or "terf" negotiated in social media? See Section 10.2.1 for a description of possible lines of inquiry in this space.

9.3.2 Implications for human-computer interaction (HCI) and the study of online communities

The main contribution of this thesis is a framework (described in Section 1.1). This framework includes methods and a demonstration of those methods to integrate social constructionism into computational and quantitative investigations of identity in online communities. Specific findings of how identity operates in online communities are not the main goal of this thesis. In the presented projects, we often find evidence from computational and quantitative techniques on large datasets that matches what has been found in smaller datasets from qualitative research that incorporates the social construction of identity. These findings validate that our methods can be used to capture identity construction at a large scale. For example, in Chapter 5 we find evidence that communities on Tumblr indicated by hashtags have more of a relationship with content prop-

agation than do communities based on network connections. This corroborates evidence from interview studies with LGBTQ Tumblr users who experienced community unevenly, indirectly, and through content more than through direct connections with other users (Byron et al., 2019). In Chapter 7, we find that visualizations of embedding representations for characters show trends in fanfiction character divergence from source texts that match what is known from qualitative work about which characters fanfiction writers change.

In addition, we demonstrate how our framework can be used to find new insights into identity in online communities that are not known from prior qualitative work. These findings start with the specific online communities of Tumblr and fanfiction. We discuss implications of these findings in the rest of this section. See Section 10.1 for comments on how these may generalize to other settings.

Implications from findings in Tumblr On Tumblr, self-presentation of identity labels is quite straightforward in blog descriptions. In such a setting we find that identity alignment plays a role in shaping the content that propagates through the social media platform. From Chapter 4, we find evidence for an extension of homophily in the sense that users who present similar attributes are not only more likely to be linked in a network but are also more likely to share content from each other. Crucially, identity-based similarity impacts the flow of information and content on the platform. At a large scale, this means content that is shared widely is slightly more likely to have been spread from those with shared identities. It also suggests that users who are able to connect over shared identities on Tumblr are at a slight advantage in sharing their content.

A question for follow-up work arises from this finding: does the content that is shared between users who align in self-presentation match what the users share in common? For example, are pairs of users who present as Voltron fans more likely to share any content, or content specific to Voltron? If the content is likely to match, this would mean that content that is produced and shown to the most people is more likely to be related to the shared identities that people connect around on Tumblr. Patterns when the content does *not* match the shared identity of users could show how participation in one group can introduce users to content from other groups. For example, fanfiction is a community of readers and writers that also has built an LGBTQ-friendly space where issues of minoritized genders and sexualities are commonly discussed (as explored in Chapter 8).

A link between identity and content sharing also suggests a mechanism for echo chambers forming. Content that rises to prominence on Tumblr is more likely to be what people with shared identities have connected over and thus may reinforce a particular perspective from that shared identity. But the assumed negativity of echo chambers is in question here. From a standpoint of promoting equity and social justice, building solidarity around shared, marginalized identities on Tumblr is not harmful. Echo chambers are often associated with political polarization or contributing to a lack of diversity of ideas. Follow-up work could address this more directly. Is there a diversity of ideas around these identity issues in content (what it means to be progressive, or a Star Wars fan, or asexual, for example)? Or does the interaction around shared identity on Tumblr and in online communities lead to one opinion that stifles diversities of perspectives? Such an investigation would nuance the concept of "echo chambers" online for groups without relative power in society.

For researchers in human-computer interaction, the link between self-presentation and content propagation has implications for platform design affordances. Affordances in online communities for self-presentation can have an impact on the connections that people make on the platform–and the content that propagates through these platforms. Platform designers should keep this in mind when structuring affordances for self-presentation. If they do not want as much polarized political content spreading, for example, they may not want to allow political self-expression for members to be prominent or visible.

Chapter 5 interrogates the finding from Chapter 4 that shared identity is associated with content propagation by investigating effects of community-based identity. Findings in Chapter 5 temper the idea that connected communities are driving content propagation in a straightforward way, since there is only a small effect of network community on reblogging. This finding leads to further questions about the degree to which behavior on Tumblr is driven by communities based on content and not user connections. Are Tumblr users often reblogging content that shares their interests, but from users with little direct connection to their community? Are there many network communities that are organized around similar content, but are situated in different parts of the social network? Further evidence that communities on Tumblr are mainly around content would nuance claims that anonymity online can make spaces more prone to identity-based attacks on those with marginalized identities, as those without direct connections may still be crucial parts of the community as they interact with shared content. For human-computer interaction researchers, this may point to opportunities to strengthen direct user-user connections on platforms like Tumblr where this does not seem to be a driving force for content propagation. On the other hand, users may not want more of these direct connections. If so, this would point to the importance of building affordances to connect around content, especially in communities with large numbers of people with marginalized identities.

The uneven impact of community found in Chapter 5 could also suggest that uniquely individual identities and identity alignments relate to content propagation. Out of multiple presented identities, do people share content where they relate to someone around a particular identity, then share other content when they relate to someone else around a different identity? If so, this would provide evidence for a view of content propagation as individual expression rather than the work of communities. Further investigations could explore if individual users are connecting with multiple users over multiple dimensions of their identities, or if their reblogging behavior largely matches that of communities. In this case, perhaps the communities we defined in Chapter 5, which had a small effect on reblogging, are not the ones most clearly associated with this behavior.

Such follow-up work would clarify and nuance the findings from this thesis about how identity operates on Tumblr.

Implications from findings in fanfiction In the narrative context of fanfiction, identity presentation of characters is less straightforward. To handle this, we first built a processing pipeline and methodology to capture character portrayal and change from source texts (Chapters 6 and 7). We first validated that our models match what is known qualitatively about which characters fanfiction authors change. Then as a new contribution, we found emotionally intense language to be a marker of fanfiction writers changing character relationships from source media. Fur-

ther investigations could explore whether there is evidence for emotionality as a broader value in fanfiction for authors transforming original work.

In Chapter 8, we find a link between LGBTQ character portrayal in fanfiction and trends in LGBTQ social movements. More broadly, this indicates that the indirect portrayal of identity in fanfiction could draw attention to LGBTQ social movement issues. This illustrates how identity representations that are not the focus in an interest-based community have the potential to engage people with an offline identity-based social movement. Next steps would include analyzing whether fanfiction leads the social movement in trends, and how fanfiction could act as a noisy indicator of where the movement is headed. Another branch of inquiry could work toward identifying the mechanisms at the micro-level by which this interest-based community relates to the social movement. Are these links largely through timely fanfiction writing prompts, for example, or representative of broader trends in fanfiction to engage with current events? Looking into the inspiration for writing fanfiction, where stated, and surveying fanfiction's relationship with other current events could address these questions.

HCI researchers may be interested in how fanfiction readers and writers are affected by the trends identified in this thesis. From the area of discourse analysis, we know that pieces of writing are continually in conversation with prior pieces of writing and discourse (Johnstone, 2018). But how does this affect individual readers and writers as they engage in the fanfiction community? Do writers learn to reproduce particular character relationships, types, and framings that relate to the LGBTQ social movement as they are introduced to content exhibiting these patterns? Longitudinal studies of authors' writing and interview studies with writers may start to answer these questions and would speak to the potential social impact of fanfiction in shaping views.

Identifying fanfiction stories and trends that relate to the LGBTQ social movement could enhance the tagging systems of fanfiction archives with recommendations for stories to include in collections. If users are interested in LGBTQ social movements or LGBTQ identity more generally, suggested tags and collections of stories could help introduce new users to this content.

Potential for new findings With our validation on fanfiction and Tumblr, the framework and methodology proposed in this thesis could be used to find new insights about identity in online communities. The tools developed for processing fanfiction in Chapters 6 and 7 could be used to see changes in character representations from original media within and across fandoms. Identifying patterns in what characters and aspects of identity fanfiction authors change suggests what fanfiction values about identity that is missing in mainstream media. Content propagation could be used as a common lens for identifying values placed on self-presented identity beyond just Tumblr. Are connections over shared identities a factor in content sharing and virality on Twitter, which often focuses more on text content? What forms of identity-based solidarity are associated with content spread on political right-wing extremist platforms such as Gab? If users are "rewarded" by having content that relates to particular aspects of their identity shared, this may strengthen these aspects of their identity. Content related to shared identities will also be more likely to be seen by others on the platform as it is spread. Such findings, then, could play a role in understanding what forms of identity are encouraged by engagement on online platforms.

See Section 10.1.1 for further comments on applying the framework from this thesis to other settings.

Chapter 10

Future Work

This chapter discusses how the framework introduced in this thesis may be applied to other settings and what future directions in NLP and computational social science it lays the groundwork for. First, I comment on how the framework and findings from this thesis may generalize and apply to settings outside the online communities of Tumblr and fanfiction. Next, I elaborate on extending the work in this thesis in three directions to deepen understanding of identity in online communities. Finally, I discuss methodological challenges that are critical to address for this future work to be successful.

10.1 Generalization

This thesis looks at two particular contexts: Tumblr and fanfiction. What can be said about how the methods and findings about identity presentation in language may generalize to other settings? In this section, I first discuss how to apply the framework of analysis presented in this thesis to discover relationships between identity presentation and social interaction in other contexts. I then discuss the potential for findings from this thesis to generalize outside Tumblr and fanfiction.

10.1.1 Applying the framework of analysis in other settings

The framework for projects in this thesis (see Section 1.1) can be applied to study the relationship between identity presentation and social interactions and outcomes in other settings, even settings offline. Two variables are needed: measures of identity presentation and a measurable social outcome or interaction. In this framework, we model how measures of social interaction are predicted by, or are regressed on, identity presentation. Enough data for testing associations between these variables with statistical and machine learning methods is thus necessary.

The variables chosen should give evidence for the choices people make in presenting their identities and interacting in a community. These choices should relate to the values held by people in particular communities. This framework is a means to identify the values around identity that are central to communities, as well as effects of the presentation of identity in communities.

In this section, I comment on the choice of these variables (both identity presentation and social outcome) and considerations of the setting in which data is collected.

Features of identity presentation A measure of identity presentation is the first requirement in this framework. Self-presented identity information from users is one possibility, but not the only one. In fact, if user profiles are never visited or rarely visible, they may have little potential for impact on interaction. Drawing on the social construction of identity, we are not interested in users' "true" identity values, but instead how users present themselves *or others* in discourse. A site like Reddit may offer few self-presentation features (though flair is one indicator), but there is no shortage of talk *about* identity. Such talk constructs associations and attitudes toward identity. For example, how are attitudes about identities expressed or challenged on Reddit, and how does this relate to practicing such as upvoting or commenting (see LaViolette and Hogan, 2019)? Looking at what is not said, but is nevertheless constructed or reinforced indirectly may also be useful. For example, social psychologists and communication scholars have long noted the associations between being "American" and being white (Devos and Banaji, 2005; Gavrilos, 2010). How are assumptions about default or prototypical Americans being white reinforced or challenged on r/AskAnAmerican?

Researchers must pay attention to what affordances are allowed for the self-presentation of identity in the context of the community they are studying. One potential study would clarify how affordances for self-presentation affect what computational social scientists find to be associated with identity in different communities. Researchers could infer user identity on social media differently based on a variety of methods (presented names, profile images, self-report, etc.) and measure differences in the associations learned between identities and language used in posts on the site. For example, what words do we find associated with particular genders when we infer gender based on name versus based on self-report (which allows some users to opt out of this inference)? A view of identity as natural and innate would not be able to explain such differences in associations based on how identities are presented. This would provide evidence for the importance of the social construction of identity in computational social science research and the need to consider not only how identity is being displayed in online settings, but if and in what ways it is being displayed (Nguyen et al., 2014, 2016).

Similarly, statistical associations with identities would likely vary across platforms with different cultures and self-presentation affordances. What terms are associated with presenting as black on Twitter or Tumblr, which have free-form bio boxes versus Facebook's drop-down menus? Such a study could illustrate how platform affordances are part of the contexts that shape how identity is perceived on platforms, particularly informative to researchers in human-computer interaction.

Social outcome or interaction A measurable social outcome or interaction is also needed in this framework to relate to identity presentation. Examples of this are records of content sharing, upvoting, responding in comments, or agreeing in conversation. For the projects in this thesis, we selected social measures that were central practices of the community and/or reflected a particular property of the community that interests researchers. Reblogging on Tumblr is a vital practice in the community, as a vast majority of posts on Tumblr are reblogs of other posts. The

transformation of character relationships in fanfiction (Chapter 7) is a key aspect of fanfiction writers' main practice of transforming original media. Connections with the LGBTQ social movement (Chapter 8) were of particular interest to us in studying representations of gender and sexuality in fanfiction.

Measures of social interaction that reflect important community practices can then be used as a proxy for community values. Relating these to features of self-presentation enables researchers to see how these values are distributed among different forms of self-presentation.

Setting: online and offline contexts This framework requires enough data to measure relationships between social outcome and identity variables. Recordings of interaction from online communities are generally large and often richly structured, which makes the framework easier to apply in those settings. But there is nothing stopping a researcher from looking at the influence of identity presentation alignment in recorded telephone conversations, for example, or minutes taken in face-to-face business meetings. Care would have to be taken to account for the differences in modalities available to the language users in each setting, particularly in offline settings where the full physicality of the interaction is difficult to capture in recorded data. There is also the risk for more unrecorded confounding variables in such settings.

But using data from online settings does not absolve researchers of the responsibility of carefully considering the cultural context in which their data is situated. Data from online recordings should not be treated as a generic expression of people with different identities interacting, as user attribute inference research assumes. Instead, it should be treated as interaction that occurs within the design affordances and cultural norms of specific online communities and contexts. For example, Chapters 4 and 5 draw on theories of "networked counterpublics" (Renninger, 2015) for interpreting data and results in the context of Tumblr. In Chapters 7 and 8, theories of participatory culture, fan activism, and fanfiction as a female-centered queer space inform the research questions we consider, the variables in the methodology, and the interpretation of the results.

10.1.2 Generalization of findings

We have just started this work with this thesis. Specific findings in other settings may vary depending on how different such settings are from Tumblr and fanfiction. For example, the association between giving pronouns and sharing content in Tumblr may be specific to Tumblr's culture of emphasis on conceptions of gender, as well as strong transgender and non-binary representation. Just as responses to gender on surveys differ across online contexts (Jaroszewski et al., 2018), researchers may not see this effect in contexts where gender is considered differently.

I would expect identity alignment to be an informative signal for predicting content propagation in settings outside Tumblr, but that also may be more specific to Tumblr's emphasis on identity and social justice. Caring about marginalized identities may influence how much users respond or consider the presented identities of users posting content. That is not to say that an effect of identity alignment on content propagation may be more prominent in spaces that lean heavily left or right politically. Users' presented identities may actually be more relevant in a mixed space such as Reddit's r/politics. When there are a variety of perspectives, whether or

not a user espouses the same perspective may make a user more likely to share or upvote their content. But in an online space where the identity of users is backgrounded a bit more–even in a space like Archive of Our Own where user profiles are generally not extensive–this effect would be less likely to be present.

Our finding in Chapter 8 that representations of LGBTQ characters in fanfiction are related to the events and shifting focuses of the LGBTQ social movement may generalize to other online communities that relate indirectly to social movements. For example, online gaming communities have been linked to right-wing social movements. Do we see similar associations in talk about specific games or characters that relate to right-wing activism offline? Particularly interesting would be work that looks at whether these online communities generally *lead* offline social movements in shifts in focus.

10.2 Future directions

In this section, I outline three areas for future investigation of identity in online communities that are informed by cross-cutting themes from this thesis.

First, we found that computational and NLP tools can be adapted to capture the dynamic, contextual presentation of identity in language. We saw this with simpler approaches in Chapter 4, encoding exact labels used for self-presentation on Tumblr, as well as more sophisticated methods for representing the portrayal of characters in narrative (Chapters 6 and 7). However, these approaches just capture the end result, or a snapshot, of the presentation of identity in context. From the social construction of identity, we take the view that notions of identity are being formed in interaction in these contexts as well, which inspires a future direction of research. How can we study the *formation* of identities and identity associations in online communities at a large scale?

Second, we found that the form of identity presentation plays a role in its relationship to social interactions and outcomes. From Chapter 4, we saw complementary effects of identity presentation across text and visual signals on content propagation in Tumblr. In Chapter 8, we saw that it was not simply if characters are tagged as queer in fanfiction that related to the LGBTQ social movement, but how they are portrayed in the text. This leads to a further area of investigation looking specifically at how the effects of identity presentation varies across multiple modalities.

Finally, findings from this thesis demonstrated that identity presentation relates to social interactions within and outside online communities. This includes self-presentation relating to content sharing on Tumblr (Chapter 4), and the portrayal of characters in fanfiction relating to the LGBTQ social movement (Chapter 8). These findings lay the groundwork for investigating the impact of identity on social interactions that have negative impacts on users: issues of discrimination, bias, and power.

In the rest of this section, I discuss each of these future directions. I comment on how this thesis can be useful as a starting point and then generate research questions and outline potential projects.

10.2.1 Identity formation

The social construction of identity inspires new research questions apart from the tired task of user attribute inference. In particular, this theory shifts focus to questions of how identity is contextual, dynamic, and emergent in interaction. In this thesis, we start along this path by demonstrating methods for incorporating the social construction of identity into computational research. Features in our projects usually encode the results of processes of identity formation. For example, the Tumblr data used in Chapters 4 and 5 is a snapshot of interaction in particular time period used to find associations with self-presentation. Similarly, a snapshot of fanfiction data in Chapter 7 is used to capture how story text reflects changes in character identity that authors made from the original media. These data are endpoints after identity change has occurred, or snapshots of how identity relates to interaction in a particular moment while associations with identities are being shaped in online communities.

Future work could use computational methods to investigate specifically the *process* of identity formation, as Chapter 8 begins to do by capturing associations made with LGBTQ characters in fanfiction over time. This would include investigating how associations made with identity terms or types are reinforced or challenged over time in discourse.

For example, do associations with newer terms such as "terf" stabilize over time or shift in predictable ways? What about the associations made with reclaimed terms such as "queer" or (unevenly reclaimed) "dyke"? This could draw on NLP work that has considered semantic change over time (Hamilton et al., 2016). Labels for identity hold meaning for people's lives; our sense of self is expressed by identifying with certain terms and distancing from others. With data from interactions in online communities, we could identify the contexts and circumstances that put certain terms on a path toward reclamation or degradation as members of certain identity groups identify with terms or distance themselves from terms. Understanding these dynamics is a step toward monitoring and shifting online communities toward greater equity and freedom of expression of marginalized identities.

At a smaller scale, how are associations made in the context of conversations? Just as there is NLP research on identifying authority claims, classifiers could be trained to identify in-group claims of identification with identities (such as "we don't really do that on our side of the tracks") or distancing ("hipsters love their aioli, but I like plain mayo"). How are such claims distributed and debated in online forums and conversations? For example, are users more likely to identify with groups after others have as well? Are users more likely to move from self-identification with an in-group to rejecting an out-group, or vice versa? Such research could work toward understanding group dynamics in discussions of identity and how interactions online relate to people's self-conceived identities.

The fanfiction processing pipeline (Chapter 6) and NLP tools such as those used in Chapter 8 would be useful to investigate the formation of associations made with queer identity in fanfiction:

- What are the changing associations made with characters in queer relationships over time in fanfiction? Chapter 8 measures associations between the portrayal of queer characters and external events and news topics in the LGBTQ social movement. But the work presented in that project mainly considers how the portrayal of queer characters changes around topics such as marriage, not with other possible associations over time. Textual

associations can be viewed as a proxy for how associations with queerness are being constructed, or at least reinforced, in fanfiction, lending more nuanced insight into this online "queer space". An exploratory topic analysis of the contexts in which queer characters are presented in over time in fanfiction would be a first step. Then themes in portrayal could be compared with stereotypes about gay men and lesbians, as Fast et al. (2016) do for gender stereotypes in online original fiction.

- Inspired by the broader diversification of LGBTQ identities over time, we would expect to see a diversification of representations of queer characters in fanfiction. This hypothesis could also be tested across a number of fandoms using tools presented in this thesis.
- How do the changing representations of characters in mainstream media affect portrayals in fanfiction? Does a character coming out as queer in canon lead to even more excitement and shifts toward canon queer representations of these characters, or a shift away from these characters now that there is not that particular gap that fans want to write about? Case studies of fandoms with characters who come out as queer in canon would illuminate the relationship between media and fandoms and the "message" to mainstream media about what is lacking with respect to LGBTQ representation. These studies would investigate how the context of the original media story changes the associations made with queer identity in fanfiction.

10.2.2 Multimodality

This thesis examines the complementary impact of text, visual, and network alignment on factors such as content propagation (Chapters 4 and 5) in online communities, but it primarily focuses on identity presented in text. However, text is just one mode from which an impression of someone's identity is formed. To users, of course, there likely is little difference to someone presenting themselves as a *Star Wars* fan with a Rebel Alliance Star Wars insignia or by listing 'Star Wars fan' in text. A user would simply notice that that person is fan. But there could be a difference in the salience of the self-presentation or how much it demands insider knowledge to decipher.

Future work should investigate how these modes of self-presentation interact and/or have different effects on social outcomes. For example, does visual self-presentation have a stronger impact on outcomes like content propagation? Such research could have implications for the design of online communities to amplify or mitigate connections over self-presentation. It also could contribute to knowing what kinds of self-presentation are most salient and thus may have more of an impression on others, at least in an online context.

There are methodological challenges to address when incorporating multiple modes. Text is easily decomposable into segments (words or phrases) with individual meaning, whereas such decomposition is more challenging with image or network data, which can hinder interpretability. Powerful neural representations for each of these modes can be learned, but these also have challenges with interpretability, especially in interactions among text, images, and network data. Future work could focus on adapting interpretable multimodal machine learning methods for computational social science questions. Existing techniques could be surveyed for what insights they allow and evaluated in areas where findings about the differences between visual and textual

information have already been established, such as the virality of image-based content. The best-performing methods could then be used in these areas where we don't understand how different modes contribute to social interaction, taking care to discuss how differences in the context may affect the results from different methods.

10.2.3 Impact: Bias and power

This thesis does not directly focus on issues of power or material consequences for users of online platforms, instead seeking to understand trends associated with identity presentation in online communities. Furthermore, some critics of the social construction of identity perspective levy that it falls short in addressing the material implications of identity in society. For example, it is difficult to explain persistent anti-black state violence (Saucier and Wood, 2016) or the denial of healthcare and housing to transgender people (Stryker, 2008) with a conception of identity as dynamic and presented with agency in context. Though the framework presented in this thesis is helpful for basic understanding of how identity operates in online contexts, more could be done to address the pragmatic, identity-based issues that users face in these contexts. An important direction for future work is to address harassment, racism, sexism, queerphobia, and other identity-based issues that users face online.

In this section, I outline projects that expand upon work in this thesis to directly address these issues.

Responses to identity presentation How do responses that users receive in online communities depend on their self-presentation? Chapter 4 outlines how self-presentation relates to content sharing on Tumblr, and Wang and Jurgens (2018) find evidence for a difference in support based on the presentation of gender in several online contexts. The framework in this thesis, relating identity presentation to social outcomes, could be used to study how self-presentation relates to negative identity-based treatment. This could include the amount of harassment or other abuses faced by those who present marginalized gender identities (such as non-binary or transgender) in online contexts. Instead of treating those identities as innate to users, how do responses change with more or less self-presentation of those identities?

Intersectionality Chapter 8 reveals trends in associations with LGBTQ characters in fanfiction, such as correlations with the LGBTQ social movement. But this work only considers a few aspects of identity, such as the sexuality of characters, and considers these aspects in isolation. Similarly, work with Tumblr data in Chapter 4 considers all popular forms of self-presentation, but again largely considers impact on content propagation separately across gender, age, nationality/ethnicity, etc.

However, identity is multifaceted and people are treated differently based on the particular intersections of identity dimensions (Crenshaw, 1989). Researchers hoping to address identity-based issues faced by users need to foreground this intersectionality. How are associations with identity in online communities made at these intersections of identity types? How does the representation of sexuality and gender interact in gay male relationships in fanfiction, for example, which many say has a tendency to reproduce heterosexual gender roles? Fanfiction and fandom

in general is known as a white-centered space. Representations of characters often center whiteness in their lack of attention and degradation of characters of color, though there are challenges to this white supremacy in the space as well (Fazekas, 2014; Messina, 2021). How does race interact with the trends we already see represented in Chapter 8, such as increased eroticization with queer characters? The fanfiction processing pipeline presented in Chapter 6 would be a valuable resource in answering such questions with fanfiction data at a large scale.

Support when experiencing discrimination Through fanfiction, many LGBTQ writers access community support and construct identity narratives (Dym et al., 2019). Many stories also relate to authors' own experiences of discrimination, queerphobia, and mental health challenges. Again moving closer to the experiences of users, what can be learned about the practices of creative writing in community to work through this discrimination? How would these stories inform what is most important in preventing this discrimination in the first place? Trends in depictions of queerphobic interactions and violence in fanfiction portrayed against characters of certain identities may provide insight into which forms of discrimination are most salient to fanfiction writers or unaddressed and unacknowledged in mainstream media.

Movements restricting identity expression The work in this thesis focuses on Tumblr and fanfiction, communities associated with a general celebration of marginalized identities. Chapter 8 investigates the connection between fanfiction and political action in LGBTQ social movements. The methods used in that project can be used to study online communities associated with ideologies against the expression of marginalized identities as well. Like fanfiction, online gaming communities also center around shared interests and practices. Certain gaming communities are known to be associated with radical right-wing movements. Portrayals of marginalized identities in these communities could be correlated with events and visibility of issues in news with methods similar to those used in Chapter 8. Similarly, datasets of public police Facebook and social media posts could be processed to visualize associations made with marginalized identities over time, and how that relates to reports of police brutality and developments in the Black Lives Matter movement.

Commodification of creativity Finally, Tumblr is a place hosting extraordinary online creativity in which users often sell art or raise money for things like transgender healthcare. How does this interact with the capitalist power structures of platforms such as Tumblr? How does the need for platforms to make money through advertising select or constrain identity presentation on the site? How is creative expression of marginalized communities commodified, even through promotional material or pulling in advertising? Official messaging from Tumblr and other platforms could be analyzed at a large scale for their representations of creativity from marginalized communities, and how this contrasts with the creativity and agency from these communities themselves.

These projects involving power, harm, access, and well-being would progress NLP and computational social science toward issues with relevance to users.

10.3 Challenges and future directions in methodology

In this section, I outline methodological challenges to be addressed in NLP and computational social science, often inspired by limitations identified in this thesis. Specifically, I discuss pseudo-causal methodology and interpretability before moving to a broader discussion of the tension between quantitative methodology and critical approaches.

10.3.1 Pseudo-causal methodology and methods for controlling confounds

Relationships between identity presentation and social variables inevitably involve potential confounding variables. We attempted to control for as many of these as possible in this thesis, often with sampling techniques. For example, we sampled negative examples within a short time frame of positive examples of reblogged posts on Tumblr to increase the likelihood that users saw the negative examples. We also sampled across a variety of fandoms in the work on fanfiction to mitigate effects that are unique to specific fandoms. But there are always more factors to consider; we do not perfectly capture the experience of choosing to reblog content on Tumblr or writing LGBTQ characters in fanfiction.

Methodology that attempts causal inference from observational data in quasi-experimental designs would be helpful to more carefully control these confounds and build a better case for findings. Such pseudo-causal methodology includes propensity score matching, which we used to estimate the effect size of community alignment beyond potential confounds in Chapter 5. Other causal techniques could be used to test the relationship between LGBTQ character representation in fanfiction and events in the LGBTQ social movement (Chapter 8). For example, difference-in-difference estimation could be used to estimate the proportions of LGBTQ characters and topics in the counterfactual case that US marriage equality had not occurred in the summer of 2015.

Many pseudo-causal techniques originated in social science fields such as econometrics and public health to estimate the effects of interventions. The challenge for NLP researchers is to adapt these to be used with text. NLP methods usually extract numeric features from text as a first step. Care must be taken to acknowledge possible violations of statistical assumptions about these variables, such as the problem of inducing dependence between the numeric features and treatment in an analyst-induced SUTVA violation (Egami et al., 2018). Separating datasets for estimating these numeric features (such as character topic features with STM used in Chapter 8) and using them for causal inference is one way of addressing this violation.

In Chapters 4 and 5, the integration of content and self-presentation on Tumblr posed a methodological challenge. It is difficult to isolate the effects of self-presentation on content propagation. We were most successful when we acknowledged the role of both content and self-presentation in framing identity and community on Tumblr in Chapter 5 and tested for the influence of both on content propagation. In this case, directly operationalizing what was once considered a confound (interests and content types) as yet another marker of the treatment (identity) reframes the experimentation.

Future work could look specifically at this integration between identity and online content. Social media arose during a time of postmodern interest-based affinity groups becoming a larger part of people's identities (Gee, 2004). How do online content and interests relate to people's

notions of themselves and others, and can we see this develop over time? Tumblr's culture of reblogging with replies could be one data source for looking at this relationship. These replies could shape how people of certain identities are thought to react to types of content (like "that's such a hipster thing to say"). This could be examined longitudinally: if users engage with identity-relevant content (as the treatment), does that result in changes in how users present themselves over time?

Multiple and mixed methods would also provide stronger evidence for findings from observational data. Interview studies or even randomized, controlled experiments exposing participants to content and then asking about self-presentation before and after would help build a case for the causality of this link. Such projects could help determine how engagement with content online affects personal notions of identity.

More broadly, methods to move NLP closer to pseudo-causal experimentation, as well as using multiple methodologies to verify findings, lend trustworthiness and reproducibility to NLP for computational social science. In many cases, looking to text analysis in fields such as economics and political science can be helpful (Grimmer and Stewart, 2013).

10.3.2 Interpretable, structured NLP representations for social science

Existing NLP representations are built for performance in machine learning tasks, not capturing particular notions of desired social context for social science investigations. In Chapter 7 we ran up against fundamental limitations in the shallow representations currently prominent in NLP. Our word embedding-based representations for classifying character relationships did relatively well in capturing whether authors changed relationships, but they often relied on cues like genre changes and lexical differences instead of the desired change in personality or circumstance of particular characters. Representations that are more structured around what researchers are attempting to capture may work better for these kinds of scenarios. Recent methods for interpretability in machine learning may be helpful here, such as concept bottleneck networks (Koh et al., 2020). These predict human-understandable categories at an intermediate level and restrict the information for the final prediction to just those concepts and categories. Using such structured representations avoids having to speculate about what embeddings are using to make predictions with probes or visualizations. Structured representations may lose some predictive power in machine learning tasks or require more annotation, but this predictive signal is less important for the goals of computational social science in characterizing trends and associations between variables (Wallach, 2018).

The trade-off between interpretability and expressiveness is an important challenge for the field. More interpretable vector representations of text available for computational social science work, such as topic models or the Structured Topic Model used in Chapter 8, often do not encode information from lexical context as well as embeddings, either traditional or contextualized. Directing particular information such as gender into certain dimensions, as in Zhao et al. (2018), is one starting point to create more interpretable embeddings. There is a possibility in rebooting doc2vec-style embeddings to ensure learning across text spans that are related in ways that the analyst cares about, such as extralinguistic attributes, identities portrayed, or situational context. The CLS token in BERT embeddings, though it was created for a machine-learning focused purpose, could also be extended to represent other spans of text related in a way that the analyst

controls.

All textual context is not equal for questions of computational social science. Chapter 7 illustrated the issue that using embeddings for computational social science means often means a desire for representations that encode "social" context such as ideological terms around politicized issues like gun control (An et al., 2018). But what is learned includes semantic and even syntactic level information mixed in. For example, embeddings for "Republican" would likely have a close neighbor of "Democrat" since these are both semantically American political parties, though carry opposite social connotations in a project on political polarization. How can we as NLP and computational social science researchers encourage representations that focus on social context? Giving low weight to similarities based on syntax (words having the same dependency parse label) or lessening attention on common words whose presence emphasizes this common syntactic or semantic context are possible approaches. These representations could be evaluated on their ability to capture differences in the portrayal of LGBTQ characters across different fandoms or known ideological differences on Reddit, for example.

10.3.3 Redeeming quantitative and computational methods for critical approaches

The social construction of identity and other critical approaches are not easily incorporated with NLP and other computational approaches. This thesis provides a framework and demonstration for doing so, and we find that flexible representations from NLP can adapt quite readily to capture changes in character identity portrayal, for example. But significant challenges remain. This section discusses these potential pitfalls in mating critical theories with computational approaches, especially when attempting to address issues of discrimination, bias, and power.

Some critique the ability of computational tools and investigations to help mitigate identity-based oppression, which is often enabled by computational algorithms. In the words of Audre Lorde, "the master's tools will never dismantle the master's house". This critique is quite relevant. It challenges those of us who have been taught to implement computational systems to think about how they are generally used to serve "the master" of neoliberal, transnational capitalism or other oppressive systems. Researchers from philosophy of technology remind us that technology is not simply a neutral tool, but can seem to have agency (Kroes and Verbeek, 2014). For example, technology makes certain things easier for us to do, which makes us more likely to do them. Technology developed in the service of large companies enhances their ability to consolidate control and make more money, a goal that usually takes precedent over values such as equity, social justice, or individual agency.

To address this critique of using computational methods to reduce the harms of problems often enabled by computational methods, we can supplement such tools with other techniques that more fully allow for the individual agency and complexity of identity. We attempt to do this with qualitative investigation of our findings in Chapters 4, 7, and 8, but perhaps starting with the qualitative methodology would lead to fruitful investigations.

There is still a distrust of statistical and computational tools in the humanities. This is in part because of the history of quantification, which is rooted in post-Enlightenment, large-scale classification of people. Such methods were also often imported inappropriately from the physical

sciences to ascribe a view-from-nowhere reductionism to make differences in races or genders "natural" in order to reinforce the view from those in power.

Furthermore, computational and statistical approaches are used to identify larger trends, which assumes that what is most common in the dataset matters most. When this data relates to people, these larger patterns can obscure the experiences of those who are fewer in number and may have less power. As an example of broadening inclusion of the tools we present in this thesis, future work should test our fanfiction pipeline described in Chapter 6 on its ability to correctly associate non-binary pronoun mentions with characters (see Cao and Daumé III, 2020). The experiences of those with less power can more easily be foregrounded in qualitative work. The social construction of identity is in part a reaction against the quantification of people, and so it is difficult to adapt quantitative, computational, and NLP methodology to fit this perspective.

But I do not think it is impossible. This thesis attempts to start down this path: developing methods and frameworks that are more suitable to theories arising from qualitative research, such as the social construction of identity. This work must be done carefully and critically. An important first step is acknowledging the past harmful history of such methodology. This often entails a historical reckoning with who has benefited from these methods and for what reasons they have been developed. It is vital to reclaim an equity and justice-focused purpose for doing computational work. One example of this is the work of Keith et al. (2017), who develop tools for identifying civilians killed by police in news in order to further transparency and provide statistics for activism. Much other work in NLP and machine learning bias and fairness attempts to measure harmful outcomes and bias from computational systems (Wilson et al., 2021).

Along with their ability to capture broad trends, computational tools can also be used to identify rarities and focus on particular subgroups. This approach is common in digital humanities, where researchers often pay attention to the instances where classifiers fail as a measure of artifacts that do not fit particular molds (So et al., 2019). This is in opposition to trying to model trends as broadly as possible, with as much data as possible, to capture "universal" phenomena. Using statistical and computational methods in smaller, more carefully controlled, domains, such as smaller Tumblr subcommunities, allows focusing in on rarer cases or the experiences of particular groups that may be washed out in larger datasets. A systematic review of which methods for learning associations from texts in such small-data domains are most effective (topic models, traditional embeddings, techniques with contextualized embeddings, etc.) would be a good first step in this direction. Methods from low-resource NLP and corpus linguistics, which both often use smaller and/or carefully selected datasets, may be useful. A focus on interpretability can also avoid reductionism, as opaque vector representations often represent only what is most common and can obscure bias.

Being critical of the reasons NLP technology is often developed and creative in repurposing it is essential for computational work toward equity and resisting oppressive systems. Along with pseudo-causal methodology and a focus on interpretability, these developments in methods for NLP and computational social science would shift the field closer to these aims.

Bibliography

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013a. Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland. In *International Joint Conference on Natural Language Processing*. pages 1202–1208. 6.2, 7.1

Apoorv Agarwal, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2013b. SINNET: Social interaction network extractor from text. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 33–36. https://www.aclweb.org/anthology/I13-2009. 7.1

Brenda J Allen. 2010. *Difference matters: Communicating social identity*. Waveland Press. 2.1.2

Graham Allen. 2011. *Intertextuality*. Routledge. 7.2

Mariana SC Almeida, Miguel B Almeida, and André FT Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pages 39–48. 6.3.3

Nora Alrajebah, Leslie Carr, Markus Luczak-roesch, and Thanassis Tiropanis. 2017. Deconstructing Diffusion on Tumblr: Structural and Temporal Aspects. In *Proceedings of the 9th ACM Conference on Web Science*. pages 319–328. 4.3.1

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. pages 2450–2461. 10.3.2

Ellen Ann Andersen and Ellen Ann Andersen. 2017. Transformative Events in the LGBTQ Rights Movement. *Indiana Journal of Law and Social Equality* 5(2). 8.4, 8.5.3, 8.7.1

Benedict Anderson. 2006. *Imagined communities: Reflections on the origin and spread of nationalism*. Verso Books. 5.3

Jaime Arguello, Brian S Butler, Elisabeth Joyce, Robert Kraut, Kimberly S Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pages 959–968. 5.3

Albert Atkin. 2013. Peirce's Theory of Signs. In *The Stanford Encyclopedia of Philosophy*. Summer 2013 edition. https://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics. 2.1.1

Rose Attu and Melissa Terras. 2017. What People Study When They Study Tumblr: Classifying Tumblr-related Academic Research. *Journal of Documentation* 73(3):528–554. 3.1.3

JL Austin. 1962. *How to Do Things with Words*. Harvard University Press. 2.1.1

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pages 59–66. 4.8

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014a. Gender and variation in social media. *Journal of Sociolinguistics* 18(2):1–46. https://doi.org/10.1111/josl.12080/abstract. 2.2.4

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*. pages 44–54. 6.3.1, 6.4

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 2138–2144. https://doi.org/10.18653/v1/n19-1220. 7.2

David Bamman, Ted Underwood, and Noah A. Smith. 2014b. A Bayesian Mixed Effects Model of Literary Character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 370–379. 3.2.3, 6.2, 6.2, 7.1, 7.2

Liad Bareket-Bojmel, Simone Moran, and Golan Shahar. 2016. Strategic Self-presentation on Facebook: Personal Motives and Audience Response to Online Behavior. *Computers in Human Behavior* 55:788–795. 2.2.2

Rusty Barrett. 2014. The Emergence of the Unmarked. In Lal Zimman, Jenny Davis, and Joshua Raclaw, editors, *Queer Excursions: Retheorizing Binaries in Language, Gender, and Sexuality*. https://doi.org/10.1093/acprof. 2.1.1

Peter Barry. 2002. *Beginning Theory: An Introduction to Literary and Cultural Theory*. Manchester University Press, 2nd edition. 2.1

Abeba Birhane and olivia guest. 2021. Towards decolonising computational sciences. *kvinder, køn & Forskning* 29(2):60–73. http://arxiv.org/abs/2009.14258. 2.2, 2.2.1

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022. 5.5.1

Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10). https://doi.org/10.1088/1742-5468/2008/10/P10008. 5.5.1

Katherine Bode. 2020. Why you can't model away bias. *Modern Language Quarterly* 8(1):95–124. https://doi.org/10.1215/00267929-7933102. 2.2.3

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5:135–146. 1

Wayne C Booth. 1961. *The Rhetoric of Fiction*. University of Chicago Press. 7.1

Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its*

consequences. MIT Press. 2.2.2

danah boyd. 2007. Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. *MacArthur Foundation Series on Digital Learning – Youth, Identity, and Digital Media* 7641(41):1–26. 2.2.2

danah boyd. 2010. Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self*, Routledge, pages 47–66. 5.3

Simone Browne. 2015. *Dark matters: on the surveillance of Blackness*. Duke University Press. 2.2.1

Bertram Bruce. 1981. A social interaction model of reading. *Discourse Processes* 4(4):273–311. 6.2, 7.1

Amy Bruckman. 2002. Studying the Amateur Artist: A Perspective on Disguising Data Collected in Human Subjects Research on the Internet. *Ethics and Information Technology* 4(3):2017–231. 4.9

Mary Bucholtz and Kira Hall. 2004. Theorizing identity in language and sexuality research. *Language in Society* 33:469–515. 9.3.1

Mary Bucholtz and Kira Hall. 2005. Identity and Interaction: A Sociocultural Linguistic Approach. *Discourse Studies* 7(4-5):585–614. 1.2, 2.1.1, 2.2.2, 4.5.1, 5.1

Mary Bucholtz and Kira Hall. 2010. Locating Identity in Language. In Carmen Llamas and Dominic Watt, editors, *Language and Identity*, Edinburgh University Press, Edinburgh, pages 18–28. 1.2

Liam Bullingham and Ana C Vasconcelos. 2013. The Presentation of Self in the Online World: Goffman and Study of Online Identities. *Journal of Information Science* 39(1):101–112. 2.1.2

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Association for Computational Linguistics*. volume 146, pages 1301–1309. https://doi.org/10.1007/s00256-005-0933-8. 2.2.1

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge. 2.1.1, 2.1.2, 2.1.2

Paul Byron, Brady Robards, Benjamin Hanckel, Son Vivienne, and Brendan Churchill. 2019. "Hey, I'm Having These Experiences": Tumblr Use and Young People's Queer (Dis)connections. *International Journal of Communication* 13:2239–2259. 5.1, 5.7, 9.3.2

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 4568–4595. https://doi.org/10.18653/v1/2020.acl-main.418. 2.2.2, 10.3.3

Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, pages 3213–3226. 3.1.3

Yi Chang, Lei Tang, Yoshiyuki Inagaki, and Yan Liu. 2014. What is Tumblr: A Statistical

Overview and Comparison. *ACM SIGKDD Explorations* 26(1):21–29. 3.1.3, 4.3.1, 4.4.1

Anne H Charity Hudley, Christine Mallinson, and Mary Bucholtz. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language* 96(4):e200–e235. https://doi.org/10.1353/lan.2020.0074. 2.1.1, 2.1.2

Snigdha Chaturvedi, Shashank Srivastava, Hal Daumé, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*. pages 2704–2710. 7.1, 7.2

Rex H-G Chen, CC Chen, and Chi Ming J Chen. 2019. Unsupervised cluster analyses of character networks in fiction: Community structure and centrality. *Knowledge-Based Systems* 163:800–810. 7.1

Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A Comparative Study of Demographic Attribute Inference in Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*. 2.2.1

Justin Cheng, Lada A Adamic, P Alex Dow, Jon Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted? *Proceedings of the 23rd International Conference on World Wide Web* pages 925–935. https://doi.org/10.1145/2566486.2567997. 4.4.1

Judeth Oden Choi, James Herbsleb, Jessica Hammer, and Jodi Forlizzi. 2020. Identity-Based Roles in Rhizomatic Social Justice Movements on Twitter. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*. volume 14, pages 488–498. 8.4

Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proceedings of the 5th International Conference on Digital Health*. ACM, pages 43–50. 3.1.3

Cuong Xuan Chu, Simon Razniewski, and Gerhard Weikum. 2020. EntyFi: Entity typing in fictional texts. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM 2020)*. pages 124–132. https://doi.org/10.1145/3336191.3371808. 3.2.3

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. volume 1, pages 1405–1415. 6.5.1

Nikolas Coupland. 2007. Sociolinguistic Resources for Styling; Styling Social Identities. In *Style: Language Variation and Identity*, Cambridge University Press. 1, 1.2, 2.1.1, 2.2.1

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *Feminist Legal Theory: Readings in Law and Gender* pages 57–80. https://doi.org/10.4324/9780429500480. 1, 2.1.2, 10.2.3

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, pages 25–35. https://doi.org/10.18653/v1/W19-3504. 9.3.1

Michael A. Devito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. "Too Gay for Facebook": Presenting LGBTQ + Identity Throughout the Personal Social Media Ecosystem. In

Proceedings of the ACM on Human-Computer Interaction, Vol. 2 - CSCW. 4.2, 5.7

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*. 6.5.1

Thierry Devos and Mahzarin R. Banaji. 2005. American = White? *Journal of Personality and Social Psychology* 88(3):447–466. https://doi.org/10.1037/0022-3514.88.3.447. 10.1.1

Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT Press. 2.2.2

Johanna Drucker. 2011. Humanities approaches to graphical display. *Digital Humanities Quarterly* 1(5):1–21. 2.2.3

Brianna Dym, Jed R. Brubaker, Casey Fiesler, and Bryan Semaan. 2019. "Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–28. https://doi.org/10.1145/3359256. 6.2, 6.6, 8.2, 8.3, 8.9.1, 10.2.3

Brianna Dym and Casey Fiesler. 2020. Ethical and privacy considerations for research using online fandom data. *Transformative Works and Cultures* 33. 8.9.1

Penelope Eckert. 2000. The Social Order of Belten High. In *Language Variation as Social Practice: The Linguistic Construction of Identity in Belten High*, Wiley. 2, 2.1.1

Penelope Eckert. 2012. Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology* 41(1):87–100. 2.1.1

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to Make Causal Inferences Using Texts. *Stat* 1050:6. 10.3.1

Sarah Evans, Katie Davis, Abigail Evans, Julie Ann Campbell, David P Randall, Kodlee Yin, and Cecilia Aragon. 2017. More Than Peer Production: Fanfiction Communities as Sites of Distributed Mentoring. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* pages 259–272. 6.2, 8.2

Norman Fairclough. 1992. *Discourse and Social Change*. Polity Press. 7.1, 7.2

Quan Fang, Jitao Sang, Changsheng Xu, and M. Shamim Hossain. 2015. Relational user attribute inference in social media. *IEEE Transactions on Multimedia* 17(7):1031–1044. https://doi.org/10.1109/TMM.2015.2430819. 2.2.1

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the 10th International Conference on Web and Social Media (ICWSM)*. pages 112–120. 3.2.3, 7.3, 10.2.1

Angela Fazekas. 2014. *Queer and Unusual Space: White Supremacy in Slash Fanfiction*. Master's thesis, Queen's University. 3.2.2, 7.4.1, 10.2.3

Rita Felski. 2020. *Hooked: Art and Attachment*. University of Chicago Press. 6.2

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2.2.2

Casey Fiesler and Brianna Dym. 2020. Moving Across Lands: Online Platform Migration in Fandom Communities. In *Proceedings of the ACM on Human-Computer Interaction*. volume 4, pages 1–25. https://doi.org/10.1145/3392847. 8.5.1

Casey Fiesler, Shannon Morrison, and Amy S. Bruckman. 2016. An Archive of Their Own. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. pages 2574–2585. https://doi.org/10.1145/2858036.2858409. 6.4, 6.6, 7.4.2, 8.5, 8.5.2, 8.9.1

Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media and Society* 4(1):1–14. https://doi.org/10.1177/2056305118763366. 4.9, 6.6

Marty Fink and Quinn Miller. 2014. Trans Media Moments. *Television & New Media* 15(7):611–626. 3.1.3, 4.2

Mark Fisher. 2009. *Capitalist Realism: Is There No Alternative?*. Zero Books. 2.1.2

Michel Foucault. 1990. *The history of sexuality: An introduction*. Vintage Books. 2.1.2, 2.1.2

Dina Gavrilos. 2010. Becoming '100% American': negotiating ethnic identities through nativist discourse. *Critical Discourse Studies* 7(2):95–112. https://doi.org/10.1080/17405901003675398. 10.1.1

James Paul Gee. 2004. *Situated language and learning: A critique of traditional schooling*. Psychology Press. 8.3, 10.3.1

James Paul Gee. 2011. *An Introduction to Discourse Analysis: Theory and Method*. Routledge. 4.3, 5.1

Amin Ghaziani, Verta Taylor, and Amy Stone. 2016. Cycles of Sameness and Difference in LGBT Social Movements. *Annual Review of Sociology* 42. 8.5.3

Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT Press. 2.2.2

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday. 2.1.2, 2.1.2, 2.2.1

Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine Runting Shi, and Dawn Song. 2014. Joint Link Prediction and Attribute Inference using a Social-Attribute Network. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(2):27. 2.2.1, 2.2.2

Lesley Goodman. 2007. Disappointing fans: Fandom, fictional theory, and the death of the author. *The Journal of Popular Culture* 48(4):662–676. 3.2.2

Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, and Ananth Nagarajan. 2015. Gender and interest targeting for sponsored post advertising at Tumblr. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 1819–1828. 3.1.3

Melanie C. Green, Timothy C. Brock, and Geoff F. Kaufman. 2004. Understanding media enjoyment: The role of transportation into narrative worlds. *Communication Theory* 14(4):311–327. 6.2

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3):267–297. 10.3.1

Ian Hacking. 1986. Making up people. In T.L Heller, M. Sosna, and D.E. Wellbery, editors,

Reconstructing Individualism, Stanford University Press, Stanford, CA. 2.1.2

Oliver L Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2021. Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies* 21(3):345–361. https://doi.org/10.1080/14680777.2019.1678505. 5.2, 5.6

Oliver L. Haimson and Gillian R. Hayes. 2017. Changes in Social Media Affect, Disclosure, and Sociality for a Sample of Transgender Americans in 2016's Political Climate. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. pages 72–81. 3.1.3

Kira Hall. 2005. Intertextual sexuality: Parodies of class, identity, and desire in liminal Delhi. *Journal of Linguistic Anthropology* 15(1):125–144. 7.2

David M Halperin. 1990. *One hundred years of homosexuality: And other essays on Greek love*. Psychology Press. 2.1.2

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. page 2116. 10.2.1

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* pages 501–512. https://doi.org/10.1145/3351095.3372826. 2.2, 2.2.1, 2.2.2, 9.3.1

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. volume 1: Long Papers, pages 1312–1320. 6.5.2

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 770–778. 4.5.2

Marti A Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64. 6.3.4

Aurelie Herbelot, Eva von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. pages 45–54. 1

Serena Hillman, Jason Procyk, and Carman Neustaedter. 2014. alksjdf lksfd: Tumblr and the Fandom User Experience. *ACM Conference on Designing Interactive Systems* pages 1–10. https://doi.org/10.1145/2598510.2600887. 5.2

Bernie Hogan. 2010. The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. *Bulletin of Science, Technology & Society* 30(6):377–386. 4.2, 4.3

Michael Holquist. 2003. *Dialogism: Bakhtin and his world*. Routledge. 7.2

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *12th International Conference on Language Resources and Evaluation, Conference Proceedings (LREC 2020)*. May, pages 1440–1448. 1, 9.3.1

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1534–1544. 3.2.3, 6.2, 7.1, 7.2

Samantha Jaroszewski, Danielle Lottridge, Oliver L Haimson, and Katie Quehl. 2018. "Genderfluid" or "attack helicopter": Responsible HCI practice with non-binary gender variation in online communities. In *Proceedings of the Conference on Human Factors in Computing Systems*. pages 1–14. https://doi.org/10.1145/3173574.3173881. 10.1.2

Henry Jenkins. 1992. *Textual Poachers: Television Fans and Participatory Culture*. Routledge. 3.2.2, 6.2, 7.3, 8.2

Henry Jenkins. 2015. "Cultural acupuncture": Fan activism and the Harry Potter alliance. In *Popular media cultures*, Springer, pages 206–229. 8.3

Thorsten Joachims. 2002. Optimizing Search Engines using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 133–142. 4.4.1

Barbara Johnstone. 2010. Locating Language in Identity. In *Language and Identities*, Edinburgh University Press. 4.3

Barbara Johnstone. 2018. Prior texts, prior discourses. In *Discourse Analysis*, Wiley, chapter 6. 3rd edition. 9.3.2

Ann Jones and Jenny Preece. 2006. Online communities for teachers and lifelong learners: a framework for comparing similarities and identifying differences in communities of practice and communities of interest. *International Journal of Learning Technology* 2(2/3):112. https://doi.org/10.1504/ijlt.2006.010615. 5.3

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* 8:64–77. https://doi.org/10.1162/tacl_a_00300. 6.3.1

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 5803–5808. https://doi.org/10.18653/v1/D19-1588. 6.5.1

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Writer profiling without the writer's text. *International Conference on Social Informatics* pages 537–558. https://doi.org/10.1007/978-3-319-67256-4_43. 2.2.1

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (Male, Bachelor) and (Female, Ph.D) have different connotations: Parallelly Annotated Stylistic Language Dataset with Multiple Personas. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. pages

1696–1706. https://doi.org/10.18653/v1/d19-1179. 2.2.1

Deborah Kaplan. 2006. *Fan fiction and fan communities in the age of the internet: Construction of fan fiction character through narrative*. Cambridge University Press. 3.2.2

Anna Kasunic and Geoff Kaufman. 2018. Learning to Listen: Critically Considering the Role of AI in Human Storytelling and Character Creation. In *Proceedings of the First Workshop on Storytelling*. pages 1–13. 6.2

Katherine A. Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. 2017. Identifying civilians killed by police with distantly supervised entity-event extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 1547–1557. https://doi.org/10.18653/v1/d17-1163. 10.3.3

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection. *CEUR Workshop Proceedings* 2125. 3.2.3, 6.2

Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. In *Proceedings of the ACM Conference on Human-Computer Interaction*. volume 2 (CSCW). https://doi.org/10.1145/3274357. 2.2.1, 2.2.1, 9.3.1

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017a. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*. 4.7

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, Wincing Leia, and a Seriously Great Friendship: Learning to Classify Emotional Relationships of Fictional Characters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. pages 647–653. 3.2.3, 6.2, 7.3

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017b. Investigating the Relationship between Literary Genres and Emotional Plot Development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. pages 17–26. 3.2.3

Sunghwan Mac Kim, Qiongkai Xu, Lizhen Qu, Stephen Wan, and Cecile Paris. 2017c. Demographic Inference on Twitter using Recursive Neural Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* pages 471–477. https://doi.org/10.18653/v1/P17-2075. 2.2.1

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*. 4.4.3

Pang Wei Koh, Thao Nguye, Yew Siang Tang, Stephen Mussmann, Emma Pierso, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*. pages 5294–5304. 10.3.2

Zornitsa Kozareva and Makoto Yamada. 2016. Which Tumblr Post Should I Read Next? In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pages 332–336. 3.1.3

Siddharth Krishnan, Patrick Butler, Ravi Tandon, Jure Leskovec, and Naren Ramakrishnan. 2016. Seeing the forest for the trees: New approaches to forecasting cascades. In *Proceedings of the 2016 ACM Web Science Conference (WebSci 2016)*. pages 249–258. https://doi.org/10.1145/2908131.2908155. 4.4.1

Julia Kristeva. 1986. Word, dialogue and novel. In Toril Moi, editor, *The Kristeva Reader*, Basil Blackwell, Oxford. 7.2

Peter Kroes and Peter-Paul Verbeek. 2014. *The Moral Status of Technical Artefacts*. Philosophy of Engineering and Technology. Springer Netherlands. https://doi.org/10.1007/978-94-007-7914-3. 10.3.3

William Labov. 1963. The social motivation of a sound change. *Word* 19(3):273–309. 2.1.1

William Labov. 1972. The Social Stratification of (r) in New York City Department Stores. In *Sociolinguistic Patterns*, University of Pennsylvania Press. 2.1.1

Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 3, pages 1–11. https://doi.org/10.18653/v1/w17-1601. 2.2.2

Jack LaViolette. 2017. Cyber-metapragmatics and alterity on reddit.com. *Tilburg Papers in Culture Studies* (196). https://research.tilburguniversity.edu/en/publications/cyber-metapragmaticsand-alterity-on-redditcom. 2.1.1

Jack LaViolette and Bernie Hogan. 2019. Using platform signals for distinguishing discourses: The case of men's rights and men's liberation on Reddit. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*. pages 323–334. 2.2.2, 10.1.1

Robert Brock Le Page and Andrée Tabouret-Keller. 1985. *Acts of Identity: Creole-Based Approaches to Language and Ethnicity*. Cambridge University Press. https://books.google.com/books?id=cbQ8AAAAIAAJ. 2.1.1

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*. pages 28–34. 6.5.1

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 472–479. https://www.aclweb.org/anthology/P07-1060. 7.2

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 188–197. 6.3.1, 6.5.1

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 497–506. 7.2

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pages 970–976. 7.4.3

Aaron W. Li, Veronica Jiang, Steven Y. Feng, Julia Sprague, Wei Zhou, and Jesse Hoey. 2019. ALOHA: Artificial Learning of Human Attributes for Dialogue Agents. In *Proceedings of AAAI*. 2.2.1

Alexis Lothian, Kristina Busse, and Robin Anne Reid. 2007. Yearning void and infinite potential: Online slash fandom as queer female space. *English Language Notes* 45(2). 3.2.2, 6.2, 7.1, 8.2

Guillaume Marche. 2019. *Sexuality, Subjectivity, and LGBTQ Militancy in the United States*. 8.4

Travis Martin, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. 2016. Exploring limits to prediction in complex social systems. In *25th International World Wide Web Conference, WWW 2016*. pages 683–694. https://doi.org/10.1145/2872427.2883001. 5.6

Ann McClellan. 2014. Redefining genderswap fan fiction: A Sherlock case study. *Transformative Works & Cultures* 17. 6.4, 8.2

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence* 3(55). 8.5.4, 8.5.4, 9.3.1

Cara Marta Messina. 2021. *The Critical Fan Toolkit: Fanfiction Genres, Ideologies, and Pedagogies*. Ph.D. thesis. https://www.proquest.com/docview/2532587933. 10.2.3

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* . 3, 7.5.1

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11):39–41. 6.3.1

Smitha Milli and David Bamman. 2016. Beyond Canonical Texts : A Computational Analysis of Fanfiction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 2048–2053. 3.2.2, 7.3

Ryan M Milner. 2013. Pop Polyvocality: Internet Memes, Public Participation, and the Occupy Wall Street Movement. *International Journal of Communication* 7:2357–2390. 3.1.3

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: (Long Papers)*. pages 174–184. 7.4.6

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 632–642. https://doi.org/10.18653/v1/P16-1060. 6.5.1

Felix Muzny, Michael Fang, Angel X. Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. volume 1, pages 460–470. https://doi.org/10.18653/v1/e17-1044. (document), 6.3, 6.3.2, 6.5, 6.5.2, 6.5.3, 6.5.3, 6.6

Annemarie Navar-Gill and Mel Stanfill. 2018. "We Shouldn't Have to Trend to Make You Listen": Queer Fan Hashtag Campaigns as Production Interventions. *Journal of Film and Video* 70:85–100. 8.2, 8.3

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference*. 4.2, 4.3.1, 4.5.3

Eve Ng. 2017. Between text, paratext, and context: Queerbaiting and the contemporary media landscape. *Transformative Works and Cultures* 24:1–25. 8.2, 8.3

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42(3):537–593. https://doi.org/10.1016/j.jksus.2015.08.001. 2.2.1, 2.2.4, 10.1.1

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How Old Do You Think I Am?" A Study of Language and Age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. pages 439–448. 2.2.1

Dong Nguyen, Dolf Trieschnigg, and Theo Meder. 2014. TweetGenie: Development, Evaluation, and Lessons Learned. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* 2(1):62–66. 2.2, 2.2.1, 2.2.1, 2.2.4, 10.1.1

Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 798–808. 7.2

Abigail Oakley. 2016. Disturbing Hegemonic Discourse: Nonbinary Gender and Sexual Orientation Labeling on Tumblr. *Social Media + Society* 2(3):1–12. 2.1.2, 3.1.1, 3.1.2, 4.2, 4.5.1, 5.2, 8.5.1

Elinor Ochs. 1992. Indexing Gender. In Alessandro Duranti and Charles Goodwin, editors, *Rethinking context: Language as an interactive phenomenon*, Cambridge University Press, chapter 14, pages 335–358. 2.1.1

Tim O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference (EMNLP-CoNLL 2012)*. pages 790–799. 6.3.2

Michael Omi and Howard Winant. 2014. *Racial Formation in the United States*. Taylor & Francis. 2.1.2

Zhao Pan, Yaobin Lu, Bin Wang, and Patrick Y.K. Chau. 2017. Who Do You Think You Are? Common and Differential Effects of Social Self-Identity on Social Media Usage. *Journal of Management Information Systems* 34(1):71–101. https://doi.org/10.1080/07421222.2017.1296747. 5.3

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. 6.5.1

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*. pages 2227–2237. 7.5.1

Susan U Philips. 1998. Language ideologies in institutions of power: A commentary. *Language ideologies: Practice and theory* pages 211–225. 2.1.1

Mario Piergallini, A Seza Doğruöz, Phani Gadde, David Adamson, and Carolyn Rosé. 2014. Modeling the Use of Graffiti Style Features to Signal Social Relations within a Multi-Domain Learning Paradigm. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pages 107–115. 2.2.2

Vinodkumar Prabhakaran, Emily E Reid, and Owen Rambow. 2014. Gender and Power: How Gender and Gender Environment Affect Manifestations of Power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. pages 1965–1976. 9.3.1

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. page 1–40. 6.3.1, 6.5.1

Deborah A Prentice, Dale T Miller, and Jenifer R Lightdale. 1994. Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. *Personality and Social Psychology Bulletin* 20(5):484–493. 5.3

Christopher Pullen and Margaret Cooper. 2010. *LGBT identity and online new media*. Routledge. 1

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. pages 492–501. 6.5.1

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn A Walker. 2017. Modelling Protagonist Goals and Desires in First-Person Narrative. In *Proceedings of the SIGDIAL 2017 Conference*. pages 360–369. http://arxiv.org/abs/1708.09040. 3.2.3

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. *CEUR Workshop Proceedings* 1866. 2.2.1

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. pages 2289–2299. 3.2.3

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 627–633. 6.5.1

Sravana Reddy and Kevin Knight. 2016. Obfuscating Gender in Social Media Writing. In

Proceedings of the First Workshop on NLP and Computational Social Science. pages 17–26. 2.2.1

Bryce J Renninger. 2015. "Where I can be myself ... where I can speak my mind" : Networked counterpublics in a polymedia environment. *New Media Society* 17(9):1513–1529. https://doi.org/10.1177/1461444814530095. 2.2.2, 5.2, 5.3, 10.1.1

Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. stm: An R package for structural topic models. *Journal of Statistical Software* 91(2). https://doi.org/10.18637/jss.v091.i02. (document), 8.5.4, 8.3, 8.5, 8.7.2

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58(4):1064–1082. https://doi.org/10.1111/ajps.12103. 8.5.4

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society* 46(5):621–647. https://doi.org/10.1017/S0047404517000562. 2.1.1

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y. 4.5.2, 4.8

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 1668–1678. https://doi.org/10.18653/v1/P19-1163. 9.3.1

Kai Sassenberg. 2002. Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats. *Group Dynamics* 6(1):27–37. https://doi.org/10.1037/1089-2699.6.1.27. 5.3

P Khalil Saucier and Tryon P Wood, editors. 2016. *Conceptual aphasia in black: Displacing racial formation*. Lexington Books. 2.1.2, 10.2.3

Bambi B Schieffelin, Kathryn A Woolard, and Paul V Kroskrity. 1998. *Language ideologies: Practice and theory*. Oxford University Press. 2.1.1

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*. volume 86, pages 199–205. https://doi.org/10.1155/2015/862427. 2.2.1, 2.2.1

Eve K Sedgwick. 1990. *Epistemology of the Closet*. University of California Press. 2.1.2

Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of Social Identity Theory to Research and Design in Social Computing. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. volume 2, pages 1–33. 5.1

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Com-*

puter Vision. pages 618–626. 4.7

Bingyu Shen, Christopher W. Forstall, Anderson De Rezende Rocha, and Walter J. Scheirer. 2018. Practical text phylogeny for real-world settings. *IEEE Access* 6:41002–41012. 3.2.3, 7.2

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language and Communication* 23(3):193–229. https://doi.org/10.1016/S0271-5309(03)00013-2. 2.1.1

Matthew Sims and David Bamman. 2020. Measuring information propagation in literary social networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 642–652. https://doi.org/10.18653/v1/2020.emnlp-main.47. 6.3.2

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 3623–3634. https://doi.org/10.18653/v1/P19-1353. 6.2

Richard So, Hoyt Long, and Yuancheng Zhu. 2019. Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000. *Journal of Cultural Analytics* pages 1–30. https://doi.org/10.22148/16.031. 2.2.3, 10.3.3

Gayatri Chakravorty Spivak. 1988. *Subaltern studies: Deconstructing historiography*. 9.3.1

Nancy Springer. 2007. *The Case of the Missing Marquess: An Enola Holmes Mystery*. Enola Holmes Mystery. Sleuth/Puffin Press. 7.1

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. pages 1–14. https://doi.org/10.1145/3313831.3376315. 2.2.2

S Stryker. 2008. *Transgender History*. Seal Press. 10.2.3

Susan Stryker. 2004. Transgender Studies: Queer Theory's Evil Twin. *GLQ: A Journal of Lesbian and Gay Studies* 10(2):212–215. https://doi.org/10.1215/10642684-10-2-212. 2.1.2

Lisa M Stulberg. 2018. *LGBTQ social movements*. John Wiley & Sons. 1.2, 8.4, 8.5.3, 8.7.1, 8.9

Jeanna Sybert. 2021. The demise of NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media and Society* https://doi.org/10.1177/1461444821996715. 5.6

Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social Science Information* 13(2):65–93. 2.1.3, 5.1

Stefano Tardini and Lorenzo Cantoni. 2005. A Semiotic Approach to Online Communities: Belonging, Interest and Identity in Websites' and Videogames' Communities. *International Conference e-Society* (March):371–378. https://www.researchgate.net/publication/266218884. 5.3

Yla R. Tausczik, Laura A. Dabbish, and Robert E. Kraut. 2014. Building loyalty to online communities through bond and identity-based attachment to sub-groups. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* pages 146–157. https://doi.org/10.1145/2531602.2531688. 5.3

Katrin Tiidenberg. 2014. Bringing Sexy Back: Reclaiming the Body Aesthetic via Self-shooting. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 8(1). 3.1.3

Katrin Tiidenberg. 2016. Boundaries and conflict in a NSFW community on Tumblr: The meanings and uses of selfies. *New Media Society* 18(8):1563–1578. https://doi.org/10.1177/1461444814567984. 5.2

Catherine Tosenberger. 2008. Homosexuality at the Online Hogwarts: Harry Potter Slash Fanfiction. *Children's Literature* 36(1):185–207. 6.2, 7.3, 7.4.1, 8.2

William E Underwood. 2016. The life cycles of genres. *Journal of Cultural Analytics* 1(1). https://doi.org/10.22148/16.005. 7.1

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605. https://doi.org/10.1007/s10479-011-0841-3. 7.5.3

Balazs Vedres and Orsolya Vasarhelyi. 2019. Gendered behavior as a disadvantage in open source software development. *EPJ Data Science* 8(1). 2.2.2, 9.3.1

David Vilares and Carlos Gómez-Rodríguez. 2019. Harry Potter and the action prediction challenge from natural language. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*. volume 1, pages 2124–2130. 3.2.3

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring Latent User Properties from Texts Published in Social Media. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*. pages 4296–4297. 2.2.1

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359(6380):1146–1151. 4.3.1

Anthony Wall. 1984. Characters in Bakhtin's Theory. *Studies in 20th Century Literature* 9(1):2334–4415. 6.2

Hanna Wallach. 2018. Viewpoint: Computational social science ≠ computer science + social data. *Communications of the ACM* 61(3):42–44. https://doi.org/10.1145/3132698. 10.3.2

Yilun Wang and Michal Kosinski. 2018. Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images. *Journal of Personality and Social Psychology* 114(2):246–257. https://doi.org/http://dx.doi.org/10.1037/pspa0000098. 1, 2.2.1, 8.9.1

Zijian Wang and David Jurgens. 2018. It's Going to be Okay: Measuring Access to Support in Online Communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 33–45. 2.2.2, 9.3.1, 10.2.3

Michael Warner. 2002. Publics and counterpublics. *Public Culture* 14(1):49–90. https://doi.org/10.1215/08992363-14-1-49. 5.3

Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. https://doi.org/10.1145/3442188.3445928. 10.3.3

Daniel Xie, Jiejun Xu, and Tsai-Ching Lu. 2017. What's Trending Tomorrow, Today: Using Early Adopters to Discover Popular Posts on Tumblr. In *Proceedings of the 2017 IEEE International Conference on Big Data*. pages 2168–2176. 4.3.1, 4.5.3

Jiejun Xu, Ryan Compton, Tsai-Ching Lu, and David Allen. 2014a. Rolling through Tumblr: Characterizing Behavioral Patterns of the Microblogging Platform. In *Proceedings of the 2014 ACM Conference on Web Science*. pages 13–22. 3.1.3, 4.2, 4.6

Jiejun Xu and Tsai-Ching Lu. 2015. Inferring User Interests on Tumblr. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. pages 458–463. 3.1.3

Jiejun Xu, Tsai-Ching Lu, Ryan Compton, and David Allen. 2014b. Civil Unrest Prediction: A Tumblr-based Exploration. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. pages 403–411. 3.1.3

Carl Yang, Lin Zhong, Li-Jia Li, and Luo Jie. 2017. Bi-directional Joint Inference for User Links and Attributes on Large Social Graphs. In *Proceedings of the World Wide Web Conference (WWW '17)*. pages 564–573. https://doi.org/10.1145/3041021.3054181. 2.2.1, 5.3

Kodlee Yin, Cecilia Aragon, Sarah Evans, and Katie Davis. 2017. Where no one has gone before: A meta-dataset of the world's largest fanfiction repository. In *Proceedings of the Conference on Human Factors in Computing Systems*. pages 6106–6110. 6.2

Michael Miller Yoder and Barbara Johnstone. 2018. Unpacking a political icon: 'Bike lanes' and orders of indexicality. *Discourse Communication* 12(2):192–208. https://doi.org/10.1177/1750481317745753. 2.1.1

Robert J. Zeglin and Julie Mitchell. 2014. Using Social Media to Assess Conceptualizations of Sexuality. *American Journal of Sexuality Education* 9(3):276–291. 3.1.3

Lumin Zhang, Jian Pei, Yan Jia, Bin Zhou, and Xiang Wang. 2014. Do Neighbor Buddies Make a Difference in Reblog Likelihood?: An Analysis on SINA Weibo Data. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. pages 208–215. 4.2, 4.3.1

Qi Zhang, Yeyun Gong, Jindou Wu, Haoran Huang, and Xuanjing Huang. 2016. Retweet Prediction with Attention-based Deep Neural Network. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. pages 75–84. 4.2, 4.3.1

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 4847–4853. 10.3.2