Computational Models of Identity Presentation in Language

Michael Miller Yoder

December 13, 2019

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: Identity, Computational Modeling of Identity, Identity and Language, Computational Sociolinguistics

Abstract

Language use varies across demographic, social, and cultural distinctions among speakers. Consequently, many researchers in computational sociolinguistics have built models of how language use reflects latent, stable identities of language users. However, researchers in sociolinguistics, linguistic anthropology, and discourse analysis posit that language also *constitutes* these very identities. Identity, which we define as the positioning of self and others in interaction, is in part reproduced, challenged, and performed in language. This thesis attempts to apply this perspective in computational tools that investigate not only how the identity of language users affects language, but how language positions the identity of the speaker and others. We explore identity positioning in language along several dimensions, from the effects of explicit self-labeling on social media to the implicit framing of gender and sexuality in narrative. We investigate each of these with machine learning and methods drawn from computational linguistics for use on large datasets of linguistic and social interaction.

First, we examine the social effects of self-presented identity labels on Tumblr, a social media site known for talk about identity. 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 these 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. Specific 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.

We then examine how identity positioning can take place more implicitly in representations of characters in narrative. For this we use a corpus of fanfiction, stories written by fans that expand or change original narratives from TV shows, comics, books and movies. We investigate a suite of computational approaches that use neural network-based word vectors to represent characters and relationships between them in narrative. In particular, we investigate the ability of such approaches to capture fanfiction writers changing the depiction of a relationship from its portrayal in the original, source narrative. To validate these approaches, we use a machine learning task predicting if fictional characters are paired in the same or different romantic relationships in a specific work of fanfiction compared to the original media on which it was based. We also construct visualizations of the learned representations of characters, and use these experiments and visualizations to analyze the extent to which contrasts in these representations reflect contrasts in the positioning

of characters between original and derived works. We find that these representations can capture elements such as the linguistic construction of a writer's attitude toward a character, or gender and sexuality changes in portrayed characters or relationships. However, some difficulty in separating out social phenomena from more surface-level features in language such as genre persists.

Taking this work a step further in a proposed project, we plan to examine changes in ideological positioning of identity types in fanfiction, specifically gender roles in relationships. Fanfiction has been considered a "queer space" for its considerable representation of same-sex relationships. But how are fanfiction writers depicting gender roles in both gay and straight relationships? Does the "queer space" of fanfiction transcend traditional gender roles in relationships, or does it redeploy heterosexual gender roles in gay and straight relationships? We seek to discover gender roles in relationships in fanfiction in this project. To do so without imposing a binary framework of gender inappropriate to the context, we first find roles played by characters in an unsupervised fashion, clustering text relevant to character portrayal across stories. We then quantify the effect of character gender on distributions of character roles–and how this varies from the original series to fanfiction, and across straight and queer relationships. Such computational tools allow us to study the extent to which fanfiction has created an alternative representation in language for gender along with sexuality–or discover that traditional gender roles are redeployed or even left intact in same-sex relationships.

Finally, we attempt to learn flexible, relational representations for identity labels with notions of similarity and difference that are relevant to observed user behavior. We again use a machine learning prediction task predicting user content sharing behavior, but this time to supervise the construction of an embedding space that positions vector representations of identity labels. To evaluate this learned representation space of identity labels, we compare with a feature-based approach in the content propagation task. Meanings and values of identity labels are relative to the communities in which they are used. To evaluate if our learned identity representations are capturing community-specific meanings, we explore how this space corresponds to community-defining identity characteristics of communities defined by network structure.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"One is not born, but rather becomes, a woman." — Monique Wittig

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 mere 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 emerging in social and linguistic interaction. 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. This may not always be directly stated, but is often an assumption that is realized in computational data and models. We 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. In turning the tables and acknowledging the role of language in *constructing* identity, we can build computational models that explore how identity changes, how new associations can be made with identity types, and how identity self-presentation itself can affect interaction. This is the goal of this thesis.

1.1 Definition of identity

From sociocultural linguists Bucholtz and Hall (2005; 2010), we adopt the "broad and open-ended" conception of identity as *the social positioning of self and other*. The idea of identity as *positioning* is a helpful metaphor in that the functional role of identity labeling is akin to drawing a map. These people are located over here in mental or conceptual space, others are different because of certain personal or group characterstics and are located in a different space. Note that such positioning necessarily foregrounds certain aspects of identity similarity and erases others. In the final proposed project (Chapter 8), our goal is to computationally position Tumblr users in a mathematical high-dimensional space that reflects the distinctions that affect user behavior.

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 inherently social. One's conception of self cannot exist alone, but is necessarily tied to the ideas of which others one is like, and which others one is unlike. 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 (Chapters 5 and 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 certainly the generalizable categories which social scientists most often try to correlate with behavior. However, care must be taken not to assume that identity is a sum or collection of these categories, as this 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).
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 often are 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 Penelope Eckert's work in a Detroit area high school in the 1990s 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 proposal. 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 also 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 because of 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 to gender in our proposed experiments investigating the presentation of gender in narrative (Chapter 7).

1.2 The author's identity

I adopt the view from anthropology 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, American (Midwestern), liberal, cisgender, male, and educated. (This document, of course, is an attempt to claim to be even more educated with the label of *PhD*.)

1.3 Task in this thesis

Prior computational work has largely viewed identity as a latent, personal, stable attribute that affects linguistic and other behaviors. In this thesis we take a view of identity as dynamic and emergent in interaction. We then build computational models that allow for that flexibility and relationality and use them to find patterns in how language users are positioning their own identities and those of others in large datasets of linguistic and social interaction.

We look at how personal identities are constructed in language and social interaction (Chapters 4 and 5), as well as how associations with broader identity types are made in narrative (Chapters 7 and 8).

Though it may be the most obvious approach, our task is not to study how people with specific identity types speak differently from others, as in classical variationist sociolinguistics or much of computational sociolinguistics (see Chapter 2). Such an approach assumes and even reinforces the "naturalness" of these categories and ignores their capability for change. Furthermore, if the categories of identity (usually demographic) are not relevant in particular social situations from which language data was drawn, they may not explain the language choices very well. Our goal is to discover patterns in how language is being used to construct speakers' own identity positioning or associations with specific identity labels. This allows the discovery of changing identities and locally specific identity labels.

1.4 Structure of this thesis

We present four computational analyses of identity construction in language in this thesis (Chapters 4, 5, 7, and 8). These 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, including 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.
- **Features.** The analyses also vary according to which linguistic and behavioral features are considered. Language features include identity labels to describe oneself on social media, text used to portray others in narrative, and hashtags used to label and make commentary on social media content. Behavioral/network features include social media users following each other and sharing each others' content. Finally, we use the structured metadata features of labels placed on stories.
- **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:

Experiment description	**Chapter**	**Data**	**Features**	**Identity focus**
Self-presentation and interaction in social media	4	social media	identity labels, content sharing	self, other social media users
Portrayal of characters & relationships in fanfiction	5	narrative	text descriptions of characters	characters, relationships
Positioning gender & sexuality in fanfiction	7	narrative	text descriptions of characters, story metadata	characters, identity types
Relational positioning of identity in social media	8	social media	network behavior, identity labels, content sharing	self, other social media users

Table 1.1: Comparison of experiments.

Chapter 2

Related Work

2.1 'Critical approaches' to the study of identity

2.1.1 Sociocultural linguistics

The study of self-presentation through linguistic representation has a long history. In classic sociolinguistic work, Labov (1972) relates social class to the pronunciation of post-vocalic *r* among retail workers in New York. 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). Online settings are a natural fit to this interactional perspective; without a physical presence, individuality can only be constructed in discourse.

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. In this work, we examine only the "performed" self in Tumblr blogs. There is evidence that in Tumblr, as in many online spaces, users feel free to present themselves in ways that are often stigmatized in other contexts. Working with Tumblr blog descriptions, Oakley (2016) found that LGBTQIA 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.

What separates this work from most other analyses of self-presentation online is a focus on the effects of identity self-presentation on interaction, specifically on how content is propagated. Though predicting identity labels in user attribute inference is more common as a task in related computational work (Chen et al., 2015), 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. We take a similar approach but specifically quantify self-presentation effects on content propagation on Tumblr.

2.1.2 Gender studies and queer theory

[Butler, Sedgwick, Halperin, Foucault, relation to speech acts and Austin, Althusser example of "Hey you" from police officer]

2.1.3 Critical race theory

[Intersectionality, performance of race]

2.1.4 Social identity theory

[Briefly, consulting Seering and Kaufman and original sources]

2.2 Computational approaches to the study of identity

[Including Kenny Joseph's work, dissertation]

2.2.1 User attribute inference

2.2.2 Computational sociolinguistics

[The field that this work fits into. There has been work that givens more attention to the construction of social context, dynamism instead of immediately assuming latent, personal attributes.]

[Relation to the computational study of community (Laviolette and Hogan 2019), presentation of particular people in news (Field et al. 2019)]

Chapter 3

Data

3.1 Social Media: Tumblr

3.1.1 Background

As a result of its affordances and culture, Tumblr provides an ideal platform for studying the relationship between identity and content propagation. As of December 2018, Tumblr had 167 billion posts over 450 million blogs.[1] The site describes itself as “a place to express yourself, discover yourself, and bond over the stuff you love” and “where your interests connect you with your people.”[2] Additionally, 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).

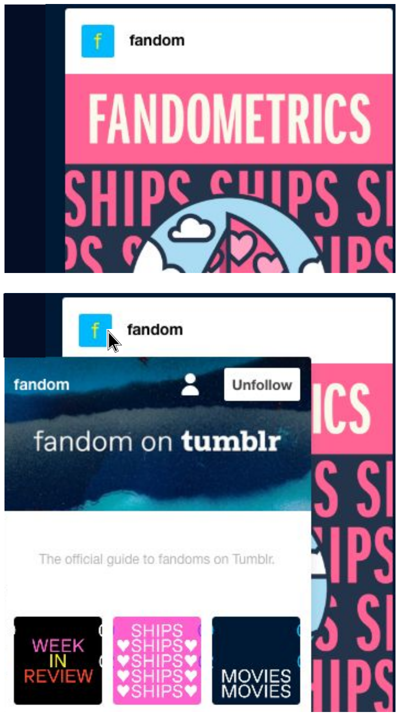
The main form of content propagation on Tumblr occurs through the reblogging of posts. A user can choose to reblog content, which then appears as a post on one of their blogs. In addition to simply propagating the content being reblogged, the user can expand on the reblogged content by adding responses 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).

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 would choose whether to reblog posts (e.g. when viewing personal blogs or when posts appear on a user’s individualized dashboard) In this section, we detail such affordances and explain our decision to focus on blog descriptions as a first step.

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

[2] `https://itunes.apple.com/bn/app/tumblr/id305343404?mt=8`, accessed January 14, 2019.



The official guide to fandoms on Tumblr.

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

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 that blog. This is one of the motivating factors for the influence of self-presentation on sharing content: users know that content they reblog will be posted on their blog and becomes part of their own self-presentation.
- **Other blog aesthetics**. Tumblr users can choose from a myriad of "themes" to customize the appearance of their blogs. Aesthetic choices can 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). However, our analysis 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.

Though all available identity presentation affordances can display important self-presentation information, we decided to focus on blog descriptions as a first step in the analysis of self-presentation in interaction on Tumblr. Textual blog descriptions are relatively well-used and are can be processed straightforwardly for automated analysis. In a sample of 1 million blogs that made at least 10 reblogs from June to December 2018, 61.2% had filled in blog descriptions.

To explore whether blog descriptions are a good choice in comparison with other alternatives like profile images, we sampled profile images for 810,800 blogs from this set of 1 million (the remainder could not be accessed, likely due to account deletion). Of the blogs we could access, 69.6% provided a non-default profile image. However, characteristics such as age, gender, and ethnicity cannot easily be determined using automatic classification over profile images, since most profile images do not include facial information about the owner of the blog. In a sample of 1000 profile images annotated by the authors from blogs in the same description sample set, only 29.3% supplied a profile image of a person who is likely the user, whereas 42.5% supplied another type of image. This differs from the conventions in many other social media sites such as LinkedIn, Facebook, Twitter, and Instagram, on which Zhong et al. (2017) perform automated profile image analysis. Profile images undoubtedly display information about interests and fandoms, for example, but these distinctions (such as which relationship in an anime a blog supports) are much more difficult to automatically determine than with blog descriptions, where these distinctions are simply stated. Usernames, on the other hand, are a more impoverished linguistic signal of self-presentation, as they are often much shorter than blog descriptions and with more opaque information, such as idiosyncratic spellings for uniqueness.

For these reasons, we select blog descriptions as a first step in our self-presentation analysis.

3.1.3 Prior work

Even though Tumblr has been the focus of much less research than the related social media sites Twitter and Facebook, it has been the focus of a wide variety of research (Attu and Terras, 2017). This research can roughly be divided into social-scientific research, often focusing on identities expressed on the platform, and 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 seeks to illuminate the effects of identity self-presentation on interaction, and brings computational tools to bear on a social-scientific investigation.

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 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

3.2.1 Background

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, there are many cases when terms' precision and efficiency warrant inclusion; 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 need special terms. 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. **fics**: individual fanfiction stories.

3.2.2 Prior work

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 culture' (Jenkins, 2003). Currently, there is expanding interest in quantitative exploration of fanfiction in part due to the enormous volume of stories accessible on the Internet; over 4.5 million stories are hosted on the fanfiction website Archive of Our Own[3] alone.

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 as a 'queer space' (Lothian et al., 2007; Fazekas, 2014).

[Include other work such as Dym et al. (2019)]

[Practice of genderswap, study of that McClellan (2014).]

3.2.3 NLP Work on Fiction

Many computational tools and studies have focused on fiction and characterization (Iyyer et al., 2016; Rahimtoroghi et al., 2017; Rashkin et al., 2018). Our text extraction pipeline builds on and extends prior work in the field of NLP (Bamman et al., 2014). In particular, Bamman et al. (2014) developed BookNLP, an NLP pipeline for character identification, coreference, and feature extraction in novels. We adapt this pipeline to work more specifically with fanfiction for our purposes (see Section 5.3.1).

Additionally, Kim et al. (2017) model relationships between characters with lexicons related to basic emotions, and Kim and Klinger (2019) manually annotate types of emotional relationships between characters in 19 fanfiction stories. In contrast, we focus on capturing how relation-

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

ships are presented differently from canon, a question that lends itself to a much larger corpus of fanfiction.

Directly relevant to our project in Chapter 7, 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.

Part I

Current Work

Chapter 4

Self-presentation and Interaction on Social Media

4.1 Introduction

Tumblr, a blogging and social media platform, provides a unique space for individuals to both share content and express particular identities. Tumblr users can create, propagate, and respond to content related to their interests, lifestyle, or social circles across a variety of media, including text, images, and videos. Sharing others' 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).

Most existing studies of content propagation on social media rely primarily on content and network features. However, content propagates through a social network from individual decisions to share others' posts. At this local level, a choice to share content could be affected by the self-presentation of the user posting content and the user choosing to share that content.

Users on Tumblr each have a personal blog, an individualized artifact that reflects a user's identity (Hogan, 2010). Once a user chooses to reblog a post, it becomes visible on their blog to their followers. Content that users choose to reblog on Tumblr can thus be seen as an incorporation of another user's self-presentation into their own, or at least that another user's content is worthy of being interacted with.

Furthermore, identity construction is uniquely part of the culture on Tumblr. Tumblr's focus on multimedia content, affordances for personalizing blog layout, and allowances for users to maintain multiple blogs without being tied to a real name have created a unique environment for users to express their identity 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 popular on Tumblr (Fink and Miller, 2014). This makes the identity positioning of users who create content even more 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.

To extract features related to self-presented identity on Tumblr, we use text blog descriptions. A common norm on Tumblr is for users to use the free-text blog description field in their profile to provide terms that can be viewed as identity labels such as 'girl', 'canadian', or 'intj' (Oakley, 2016).

Identity alignment on Tumblr can then be operationalized as lexical alignment between blog descriptions (which we refer to as *label alignment*) or as alignment between broader identity dimensions invoked by those specific labels (which we refer to as *category alignment*). For example, users who give identity labels as fans of a TV show may be more or less likely to reblog content from users indicating in their blog descriptions that they are fans of a different show. In the case of category alignment, users who provide any zodiac sign may be more or less likely to reblog content from users who also provide a zodiac sign, regardless of the specific label provided. Our intuition is that sometimes providing similar categories of identity, even if labels are different, may orient users to the platform in similar ways.

> **RQ 1:** What evidence do we find of an effect of identity alignment on content propagation at the label and/or category level?

To explore 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 blog descriptions. Using a learning-to-rank framework (Joachims, 2002), we investigate the effect of label and category alignment in this reblog prediction task. We find that incorporating identity-related features improves classification performance when used in addition to content-related features. This shows evidence that identity presentation is associated with content propagation.

However, the nature of this effect is not made precise by the above result: it could be features signifying similarities, differences, or some other interaction between specific labels and categories that are responsible for the predictive power achieved by the models used to explore RQ 1. We therefore ask a further research question:

> **RQ 2:** What is the nature of the impact of identity alignment features on content propagation?

To investigate this question, we examine the learned feature weights from logistic regression models trained with both content and identity alignment features. We find that similarities between categories and labels generally increase the likelihood of users reblogging, with some variability across identity categories and labels discussed.

4.2 Theoretical Motivation and Previous Work

4.2.1 Content Propagation

Content propagation and virality on social media have been well-studied (Naveed et al., 2011; Zhang et al., 2016; Vosoughi et al., 2018). Most existing analyses treat content propagation as a product of content features and network features. Naveed et al. (2011), for example, predicted whether a tweet was going to be retweeted based on features extracted from the text of the tweet. Zhang et al. (2014) examined how reciprocity in follow networks affected the likelihood of reblogging on Sina Weibo. On Tumblr, where content can take diverse forms (e.g. text, photos, videos, audio, etc.), previous work focusing on network-based features is more common. Chang et al. (2014) examined reblog chains on Tumblr based on structure of the follower network. Xie et al. (2017) centered their analysis on "early adopters" (the first users to reblog a post) in combination with content-specific features to determine whether a post will become a trending post on Tumblr.

While content-specific features provide an important signal as to whether a post will spread, our work extends upon this work by additionally viewing content propagation as a social interaction. Through this lens, we explore the relationship between users' identity presentation on Tumblr and the content they choose to propagate on their blogs. In the next section, we introduce and motivate the identity concepts that we investigate in our work.

4.2.2 Identity and Interaction

In this work, we conceive of online self-presentation as the construction of an artifact that reflects identity in the particular space of Tumblr (Hogan, 2010). In Tumblr, the primary artifact that members construct is one or more blogs housed within their account. This artifact, a Tumblr blog, contains both linguistic elements of self-presentation (e.g. a written description) and multimedia content (e.g. posts with images and videos).

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 either explicitly with labels (*i'm a super fangirl*) or more implicitly (*♥♥OBSESSED♥♥ with exo*) (Johnstone, 2010; Gee, 2011).

Label Alignment

The well-known property of homophily suggests that users with similar attributes are more likely to be linked within social networks. Gong et al. (2014), for example, find that user attribute inference improves performance for link prediction in Google+. However, the relationship between identity labels and affinity within a network might be predicted by more complex patterns than simple semantic similarity. Bucholtz and Hall (2005), for example, identify authenticity and institutionalized legitimacy as axes other than sameness/difference on which speakers position themselves. Piergallini et al. (2014) discussed challenges with modeling the complex alliance systems in street gang affiliation, as certain affiliations acted as both rivals and allies within specific contexts. These more complex patterns of label interaction may also be evident in the relationship between identity and content propagation on Tumblr; our extracted features task allow such combinations.

Category Alignment

Sociolinguists emphasize that different facets of identity are relevant in different situations (Eckert, 2000; Coupland, 2007). Facets of identity are apparent in Tumblr blog descriptions, as users fill in whatever aspects of identity they feel are relevant to their self-presentation, which can vary even across a single user's multiple blogs. The specific categories of identity users choose to provide (e.g. age, gender, interests, fandoms) are decided by the users themselves, as opposed to these fields being defined *a priori* by the platform or by researchers in a survey. In this case, the choice of category—regardless of what label is provided for that category—signifies how a user

Identity Category	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 4.1: Identity categories and label examples

is orienting toward or framing their identity in this space. If users provide the same identity categories in blog descriptions, it can signify that they are interested in being part of communities in which those categories are relevant in some way. For example, users who list gender and sexual orientation may be indicating their interest in content related to feminism or LGBTQIA issues, though their choice of labels may indicate opposing positions on these issues.

Our methods for extracting and categorizing self-presentation information from blog descriptions are described in the following section.

4.3 Methods

4.3.1 Operationalizing Identity

To determine any effect on content propagation from users' self-positioning, we annotate blog descriptions for specific identity labels and broader identity categories. In this work we grouped labels at two different levels, but this could be extended to other distinctions at different levels of granularity. We use a bootstrapping approach to both find labels that indicate identity and to group them into categories for automatic annotation.

To discover popular labels and categories without imposing our own notions of identity (Bucholtz and Hall, 2005), we first examine the values frequently placed in blog descriptions. Tumblr users commonly use freeform blog descriptions to list their values for certain identity categories, usually separated with some sort of delimiter (as an artificial example, *dani . bi . pnw . hufflepuff*). To bootstrap identity labels, we first identified common *n*-grams used between delimiters in these types of descriptions.

Bootstrapping delimiters We started with a small list of identity labels manually identified from blog descriptions that followed the delimiter convention. In a separate, larger set of blog descriptions, we searched for these labels and found all sets of characters in between these terms as potential delimiters. We then manually reviewed this list of potential delimiters and 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 aggregated short texts (maximum 25 characters) in between any of these delimiters on the larger set of blog descriptions. We limited the length of the texts between delimiters because long spans often indicated quotes or other unrelated material. Splitting these segments into tokens, we extracted and ranked identity label candidate *n*-grams by frequency, discarding stopwords and other terms that were not indicative of identity.

Choosing identity categories We also test the effect of self-presentation on the broader category level; for this we need to group labels into more abstract categories. Our goal was to select identity categories that met the following criteria:

1. Popular (and thus relevant for the users and communities involved).
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. We want categories that can be identified by a relatively, if not completely, closed set of values so that they can be accurately identified outside of a training set.

Guided by Bucholtz and Hall (2005), we then manually examined and grouped these terms into categories that encompassed broad demographics, ethnographic labels more specific to Tumblr (such as fandoms, interests, links to external content), and interactionally specific roles (personal descriptions, NSFW content warnings).

Sixteen categories were found from this process: age, gender, interests (e.g. art, music, anime), links to external content, personal descriptions (e.g. student, “trash”), personality types (e.g. infp, slytherin), pronouns, sexual orientation, weight, zodiac sign (e.g. virgo, leo), location, ethnicity/nationality, fandoms, roleplay, NSFW warnings, and relationship status (e.g. single, engaged, taken). We removed ‘NSFW warnings’, ‘roleplay’ and ‘links to external content’ as categories based on the second criterion, as they were more about blog content than the user posting content. We removed ‘personal descriptions’ based on the third criterion, having a very broad set of possible values. Finally, we removed ‘weight’, as this category has more to do with the sensitive topic of eating disorders than with identity. Our final list of 11 identity categories, with example labels, is shown in Table 4.1.

Annotating identity categories Two authors then manually annotated for these categories in a random sample of 1,200 blog descriptions, from which we pulled 100 samples for a development set and 100 for a test set. From our annotations, we built regular expressions to automate annotation. On the development set, we iteratively added or refined regular expression patterns for each category. Identity labels that indicated more than one category, such as ‘LGBT’, were listed as patterns for all possible categories indicated (both gender and sexual orientation for ‘LGBT’).

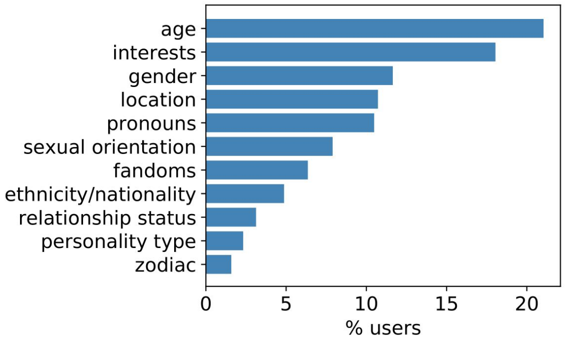


Figure 4.1: Proportion of users in our experimental dataset who present each identity category.

For a subset of identity categories, we compared our regular expression 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%); this is likely due to the small amount of available training data. Percentages of users in the training portion of our later dataset who present each category are shown in Figure 4.1 (see the **Experimental Dataset** section for details on this dataset).

4.4 Experiments

We designed an experimental paradigm to evaluate whether similarities and differences in the identity presentation of two users explains patterns in a user's decision to propagate the other's content (**RQ 1**). To test this hypothesis, we train logistic regression, SVM and neural network models on features that represent identity category alignment between users (both similarities, differences and other interactions) and look for changes in accuracy over a baseline of post content features. A significant increase in accuracy indicates that category and/or label alignment is relevant for predicting content propagation.

To explore the nature of this effect, we interpret feature weights learned from logistic regression models on this task (**RQ 2**). To isolate this effect for each identity category, we train logistic regression models on features specific to each category.

We treat the choice of whether to reblog 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 4.2). 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. Within such an environment, it makes sense to formulate a model that embodies a comparison, so we use a pairwise learning-to-rank modeling paradigm. We use this pairwise learning-to-rank paradigm to identify any effect of comparative identity presentation on choices to propagate content.

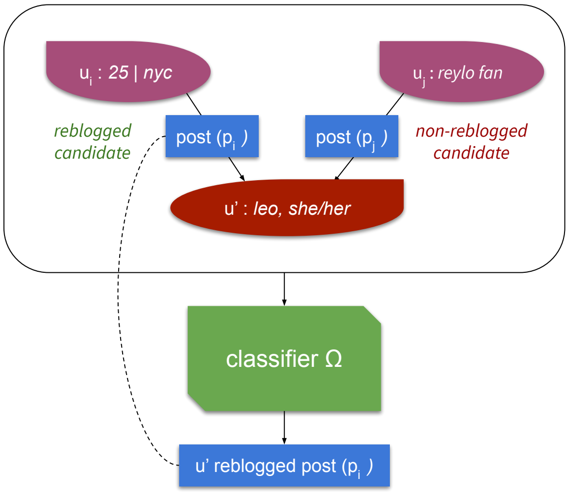


Figure 4.2: Illustration of the pairwise learning-to-rank formulation. From the perspective of user $u^{'}$, we want the classifier Ω to decide whether to reblog post p_i or p_j from different users $u_{i,j}$ that user $u^{'}$ is following.

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 $\Phi(p) : p \rightarrow \mathcal{F}$
   (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 posts represented in the comparison space. This ranking is between the two posts that a user chooses to reblog: one the user reblogged, the other they did not. While traditionally in RankSVM, SVM with a linear kernel was used as the choice for Ω, 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. However, this formulation leads to a few issues. While it is easy to determine which posts a user has reblogged, as they then appear on their blog, it is more difficult to provide a meaningful definition for posts that a user will not reblog. We can narrow down the set of posts a user will not reblog by considering all posts from blogs they have followed that they did not reblog. However, even filtering down non-reblogged candidates to posts from blogs they follow leads to heavily skewed data with the zero class as the majority; users on average reblog fewer than 1% out of all posts from blogs they follow. 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 more meaningfully.

4.4.2 Experimental Dataset

To construct the dataset of paired reblogged and non-reblogged posts for the learning-to-rank formulation, we first sample a set of 1,000 blogs[1], U, which have non-empty blog descriptions and which display at least 10 reblogs. For these 1,000 users, we find all users they have followed, F, that also have non-empty blog descriptions. From June 1, 2018 to November 30, 2018,[2] we collect reblogged posts made by blogs in U from blogs in F after their associated user began following the blog in F. In the case of reblog chains, in which content is reblogged multiple times, we take identity features from the most recent reblogger of the content. Features are thus extracted from the immediate user who the user follows. Our intuition is that this user's identity information is more relevant to the user choosing to reblog since the user follows that blog and will see the tags and comments that that user added to the post.

For each reblog in this set, we sample candidate non-reblogged posts to act as a comparison for our learning-to-rank framework. Since the details of Tumblr's dashboard ranking algorithm are not public, we use a heuristic based on the assumption that recent posts from followers will 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 the paired non-reblogged post was seen by the user, we select non-reblogged posts that were posted within 30 minutes of the 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 5 candidate non-reblogged posts, from unique blogs, for every reblogged post. This resulted in a set of 712,670 post comparisons.

Ranking labels were 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). Thus, 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.

[1] Hereafter, we will also refer to blogs 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.

[2] 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.

4.4.3 Features

We describe here the feature function $\Phi(p)$ that is applied to candidate posts p_i and p_j in each pairwise comparison. All features are normalized to have unit variance over the training set after generating comparison feature set C.

Content Features

Content features represent the type of post content, topic of the post and how popular the post has already been. These features likely capture much of the signal in content propagation; hashtags are similarly used by Naveed et al. (2011), and Xie et al. (2017) use the type of post as their content features. **Note Count**, a count of likes and comments, is an approximation of user popularity by early adopters, the approach also taken by Xie et al. (2017). In our experiments, we look for any additional signal provided by self-presentation alignment features above these baseline features.

- **Post Tags**: Hashtags attached to the post that a user is choosing whether to reblog. All post tags are lowercased and only tags that are used by more one than user are considered in the post tag vocabulary of 14,318 tags. Post tag features are binary variables indicating whether the post contains the tag in the vocabulary.
- **Note Count**: A count variable indicating the number of notes (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, and answer.

Category Alignment Features

Let $u^{'}$ be the user making the comparison and u_* be the user associated with the candidate post p_* when applying $\Phi(p_*)$:

- **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_*.
- **Directional Category Mismatch** $(c, u_*, u^{'})$: Directional version of **Category Mismatch** (c) indicating if u_* provided identity category c but not $u^{'}$.

Label Alignment Features

- **Count** (l_1, l_2, c): A count variable indicating the number of times $u^{'}$ used label l_1 and u_* used l_2 for the category c.
- **Label Match** (c): A count variable indicating the number of labels used by both $u^{'}$ and u_* in category c.

	LR	SVM	MLP
Content	62.69	62.60	64.42
Content + CA	64.72*	64.64*	69.46*
Content + LA	74.30*	74.15*	**80.06***
Content + CA + LA	**74.44***	**74.30***	79.99*

Table 4.2: Learning-to-rank accuracy with category and label alignment features. $*p < 0.05$ compared to the content features. A random baseline would give 50% accuracy.

- **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 Interactions** $(c, u^{'}, u_*)$: A binary feature indicating the presence of a specific label given by $u^{'}$ with a label given by u_*, for all label pairs within category c^3.

4.4.4 Model Hyperparameters

We train a logistic regression classifier with L2 regularization. The regularization constant of the model was chosen using grid search with 10-fold cross-validation on the training set from 10^{-4} to 10^4 on a base 10 logarithmic scale.

We train SVM models with a linear kernel due to the traditional use of linear SVM with RankSVM and the large size of our training set. The SVM models are trained with L2 regularization, with the regularization constant of the SVM models chosen using grid search with 10 fold cross-validation on the training set from .01 to 100 on a base 10 logarithmic 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 on each layer. The neural network model is trained with L2 regularization with a regularization constant $C = 10^{-4}$. We used the Adam optimizer with $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ to train the MLP model. The model was trained with early stopping, where 10% of the training data was randomly set aside as a validation set.

4.5 Results and Discussion

From improved accuracy over a baseline of content features, we find evidence that similarities and differences in self-presentation are associated with content propagation on Tumblr (Table 4.2). Referencing the learned feature weights on logistic regression models, we find that users who present similar identity categories and labels are largely more likely to reblog each other's content, with some exceptions. Details on these results are discussed below.

[3] All-label interactions were not considered as they would lead to a prohibitively large input dimensionality

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

4.5.1 Prediction Model Results (RQ 1)

Results on the reblog prediction task are shown in Table 4.2 for logistic regression, SVM, and MLP models. We find significant performance increases with the addition of category and label alignment features over content features, evidence for identity presentation being associated with content propagation on Tumblr.

To narrow our focus for interpretation and see if performance varies across categories, we explored the effects of more specific sets of features. We train separate logistic regression models with category features from each category, as well as separate models on label features from each category. Results on these models (which also include content features) are shown in Table 4.3. For each category, we see that label alignment features provide a greater increase in accuracy than category alignment features do. We also see significant improvements with label features over the baseline of content features for every individual category. The effects of category features, by contrast, are inconsistent across categories. Five out of the 11 categories do not show significant improvement over the content baseline, while 6 categories do improve with category alignment features: 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 (usually visual) content, and any zodiac label indicates an interest in astrology. Fandoms could be hypothesized as having a similar effect, but we did not see a significant effect of providing any fandom label. This could be because fandom-related content is so prevalent that only specific distinctions between fandom labels have any any noticeable affect on content propagation.

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 18-22, 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 a sexual orientation can serve as a proxy for belonging to the LGBTQ community.

Label alignment features significantly improve performance over content features (Table 4.3), 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 common gender labels, such as ‘male’ or ‘female’, does not express an ideological position, whereas giving pronouns is a more political signal of a position on gender issues.

The best configuration of features for logistic regression and SVM is a combination of the label and category alignment features. For the MLP model, however, the combination of the label and category alignment features does not offer greater performance than using the label alignment features by themselves. One possible reason for this is that much of the information about category alignment is built into the label alignment features. The greater capacity and the use of non-linear feature transformations in the neural label alignment model may allow it to implicitly reconstruct information about category alignment. Thus, adding the category alignment features in may not improve the performance of the model.

4.5.2 Interpretation of Learned Feature Weights (RQ 2)

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 categories and labels more likely to reblog each other’s content? Are users who present dissimilar categories less likely to reblog each other’s content? Or do more unexpected interactions play an important role? To investigate the nature of this effect, we consider the learned feature weights on the logistic regression models trained separately across categories (Table 4.3).

Category Alignment Interpretation For all categories except for pronouns, models learned positive weights on the **Category Match** feature. Models for sexual orientation (OR = 1.367), location (OR = 1.481), and interests (OR = 1.268) put the most weight on this feature out of any category alignment feature. 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-positioning. These users appear to have a preference for propagating content from other users that share that framing.

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 (OR = 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). What is signaled by framing one's identity in terms of pronouns cannot be known, though there have been calls from transgender activists for cisgender people to share pronouns to normalize the practice of not assuming gender.[4] Not giving pronouns may signal to a user that they may not share a view of gender that encourages listing pronouns, and we see that this has an effect on content propagation.

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 features in which the user making a decision to reblog does not provide a category, while the user they follow does. If a user deciding whether to propagate content does not place value in presenting an identity category, they are less likely to propagate content from a user who does.

Label Alignment Interpretation Models using only label alignment features placed positive weight on the **Label Match** feature in most categories. Similarly, most models learned negative weight on the **Label Mismatch** feature. Users who presented dissimilar identity terms in blog descriptions were generally less likely to reblog each other's content, though not in all categories. This gives evidence for matches in identity labels generally increasing, and mismatches decreasing, the likelihood of reblogging.

However, weights learned on specific label interactions (**Label Interactions** feature) are often higher in absolute value, and thus found to be more informative, than simple matches or mismatches. Looking more closely at these interactions, users who present labels that invoke similar interests in content often are more likely to reblog each other's content. 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 found to be more likely to reblog each other's content. For fandoms, one user listing 'star wars' and the other listing 'reylo', a popular character pairing in Star Wars, increased the likelihood of reblogging. With respect to interests, users who presented 'anime' were more likely to reblog content from those who presented 'design'. Users who presented 'gaming' as an interest were more likely to reblog content from 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 following those who list 'history'.

Other times specific interactions that were more likely indicative of shared values were more informative. For ethnicity/nationality, high weight was often placed on matches that included ethnicities such as 'latina' or 'black'. Alignment on specific ethnicities may be more emotional and political for users than listing a nationality, which makes specific label alignments for the

[4] `https://www.glaad.org/transgender/allies`

category of ethnicity/nationality more relevant. For sexual orientation, users presenting 'straight' were more likely to reblog others presenting 'straight' (OR = 1.25). Other specific interactions were also informative, such as those listing 'queer' as 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).

Examining feature weights for specific gender labels, we also found informative features related to potential shared values. Positive weight was learned on interaction features of trans-identified users reblogging users who give no gender terms (odds ratio = 1.23) and those who give other trans labels (OR = 1.06). The same holds for users giving 'non-binary' as a label: they are 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 rest of the extracted gender labels. Explicitly cis-identified users are slightly more likely to blog content from other cis-identified blogs (OR = 1.02), but the model gave slight negative weight to cis-identified users reblogging trans blogs (OR = 0.998). 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, shows a knowledge of the discourse around transgender issues. This negative weight, though small, shows that cis-identified users are slightly less likely to reblog content from trans blogs they follow.

There is a slight trend for some users who give female labels to reblog content from other female-identified users, such as those who give 'wife' more likely to reblog those who present 'girl' (OR = 1.08). Negative weight was learned on users who present female labels reblogging content from those who present no gendered labels. No gender-based patterns increasing likelihood of reblogging for users who use terms such as 'man', 'guy', or 'male' were found.

Counterintuitively, users who present similar locations are less likely to reblog each other's content (OR = 0.98), and dissimilar locations are more likely to reblog each other (OR = 1.30). This may point to a de-emphasis on location-based content or connections on Tumblr.

Users with aligned relationship status labels are more likely to reblog each other's content (**Label Match** feature, OR = 1.18), and this extends to labels that more implicitly align, such as users listing 'wife' being more likely to reblog those listing 'in a relationship' (OR = 1.40)

Matched labels for sexual orientation led to higher reblog likelihood (OR = 1.174), and the opposite for mismatched labels (OR = 0.915).

4.5.3 Possible Confounds

We operationalize self-presentation with blog descriptions as a first step. As mentioned in the **Self-presentation on Tumblr** section, Tumblr offers many opportunities for self-presentation. Thus, one might be concerned that affordances other than blog descriptions, such as profile images as an obvious choice, might interfere with the ability to draw conclusions from our analysis.

To investigate profile images as a potential confounding variable, we separated our test set predictions into three sets of instances, one where the two users who each posted the two candidate posts that are input to the ranking model both provided profile images, those where neither did, and those where only one did. Note that the identity self-presentation features are extracted only from blog descriptions, and thus distinctions related to profile images are invisible to the trained model. Therefore, if it were the case that profile images provide a signal that conflicted with that presented in blog description texts, we would expect to see lower performance from using just blog descriptions on instances where users provided profile images. In these instances, users could theoretically be responding to a signal from profile images that conflicts with or is at least different from the identity distinctions represented within the model.

We found the opposite: mean accuracy on instances where neither user provided a profile image was 73.2%, whereas mean accuracy for instances where one or both users provided a profile image was 75.2% and 75.0%, respectively, both significant improvements over the case with no users providing a profile image ($p < 0.01$). This would suggest that in fact, contrary to the concern, profile images appear to provide a consistent and reinforcing signal with blog descriptions. If profile images present similar identity distinctions, then even if the impact of self-presentation comes mainly from profile images and our blog description features are merely proxies, our results would still tell the same story, namely that self-presentation affects content propagation.

4.5.4 Limitations and Future Work

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

We use a regression-based analysis on this naturally occurring data to determine the effects of alignment on content propagation. While this approach allows us to make correlational conclusions, we cannot draw any conclusions regarding causality. Though we attempt to ensure a high likelihood that users will see both posts in our pairwise comparisons, Tumblr's own ranking of posts in users' dashboards would also have an effect that is difficult to measure. A larger sample of users may also provide a more comprehensive picture of the relationship between identity and content on Tumblr.

In future work, performance gains from the MLP could be further investigated in order to see which nonlinear combinations of features are informative to the model. For example, are there specific hashtags used in posts that in combination with blog description alignment increase likelihood of reblogging? This research could also be expanded to include a more detailed account of content propagation pathways. For instance, what identity features matter for the first reblog of a post as opposed to a reblog further along a reblog chain, and how do hashtags change as a post is reblogged?

4.5.5 Ethics and Privacy

Tumblr blog descriptions often 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 Tumblr blogs listed as public and available without a Tumblr log-in. All examples of blog descriptions given in this paper (except the staff Fandom account in Figure 3.1) have been fabricated from a mix of examples so as to not easily trace back to any individual blog (Bruckman, 2002). Though individual researchers did view examples of blog descriptions, 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. For extracting features, only labels occurring in more than one blog description were used to protect any users who used unique labels. Similarly, tags were only considered in aggregate, and we removed any tags used by only one blog for feature extraction.

We did not construct classifiers for identity category or label prediction. However, our classifiers did include these 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 available.

4.6 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 comparing self-presented identity categories and labels were informative for predicting content propagation between users. Investigating the nature of this effect, we found that users who presented similar identity categories and labels generally were 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 were more likely to reblog each other. As Tumblr has a particular emphasis on identity, it remains to be seen what effects hold in other social media contexts.

Chapter 5

Portrayal of Characters and Relationships in Fanfiction

5.1 Introduction

Fanfiction is a medium that offers a rich portrayal of a contemporary literary community's thought and values. It provides a space for amateur authors to take characters and settings from existing comics, TV shows, movies, and books and transform these works in virtually limitless directions. A popular practice since at least the 1970s (Tosenberger, 2008), fanfiction's current presence on online, public platforms presents an opportunity to study a crowd-sourced literary community where creative recontextualization occurs on a large scale. Researchers have identified quantitative patterns in fanfiction that set it apart from the media on which it was based, including a stronger focus on female characters (Milli and Bamman, 2016) and the common practice of pairing male characters in 'slash' romantic relationships (Lothian et al., 2007). From a text analytic perspective, fanfiction provides extensive opportunties for exploring issues of intertextuality.

In this paper we first make a methodological contribution, proposing word embedding-based computational tools for exploring intertextuality (and comparative text analysis more generally). We then illustrate how these tools can be used with a substantive contribution in terms of insights related to intertextuality in fanfiction.

Texts in general, but fanfiction in particular, are created in relationship to other texts; literary scholars use the term *intertextuality* to describe this inter-relatedness, where elements of a text are taken up, transformed and embedded in new texts (Worton and Still, 1991). Though a relationship to previous texts is present in some way in any text (Kristeva, 1986), this relationship is more explicit for certain text collections: news stories based on the same newswire, variations of folk tales told throughout history, re-tweets with associated comments in reply, or the focus of this study, fanfiction based on original media. Authors can introduce political perspective, inject humor, express a cultural identity, or update a narrative when reacting to and transforming other texts.

In this work, we develop methods to quantify the ways fanfiction writers portray characters and relationships so that divergences between the original media and fanfiction can be visualized and inspected. We build on a foundation of distributional semantic modeling approaches from natural language processing (NLP), which are well known to this Text and Discourse community. In particular, well-known methods such as Latent Semantic Analysis (LSA) (Landauer et al., 1998) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been used throughout this community to build vector-based representations that can be compared, visualized, and used as the basis for other forms of statistical modeling. In this paper, we explore the use of some newer contributions to this vector-based tradition that grow out of recent work in neural network modeling, where vector-based methods for approximating the semantics of natural language are often referred to as word embeddings. We use these embeddings to explore variation in how characters are presented across thousands of fanfiction stories. We add to contemporary methods of distributional semantic representations tools that are specifically designed to preprocess narrative texts and extract relevant portions for our analysis. By pairing strategic text extraction and representation, we demonstrate that insights into intertextuality can be obtained in large scale corpora.

Researchers in computational social science and digital humanities increasingly turn to word embeddings to capture and compare framing and attitudes toward concepts (Heuser, 2017; Grayson et al., 2017; Garg et al., 2018) with mixed results. The range of questions that may be asked about social meaning in language are broad, and often indirectly or subtly expressed in language (Nguyen et al., 2016). To test the ability of word embeddings to capture social meaning in intertextuality, we investigate a large corpus of fanfiction. *Fanfiction* is fiction centered on characters and settings from existing media (e.g. movies, TV shows, books, or comics) written by fans of the original work. Because of its derivative and transformative nature, the style and content of fanfiction is inextricable from its *canon* origin, the original media on which it is based. To study fanfiction effectively, we thus must understand how a fanfiction author's divergences from the canon material affect their text.

One of the primary forms of divergence in fanfiction is which characters are paired together in romantic relationships, often referred to as 'ships'. [1] Using a corpus of Harry Potter fanfiction, we set up a prediction task of determining whether a work of fanfiction, or *fic*, has depicted a relationship in the same way as canon. Attempting this task with various embedding-based text representations allows us to test whether word embedding spaces can effectively capture how fics differ from canon in their portrayal of relationships. We construct and validate representations along two dimensions: locating sections of text to build embeddings, and aligning or adapting embedding spaces.

The main methodological contributions of this paper are as follows: (1) a text extraction pipeline for preprocessing fanfiction, specifically including character coreference, exposition and quotes for each character, which provide a strong signal for predicting relationship framing, and (2) an empirical evaluation of word embedding alignment and adaptation approaches for capturing shifts from source texts.

In a further demonstration of these methods, we plot learned representations for characters against representations learned from canon text and find that these representations capture which characters are often presented by fanfiction writers similarly or differently than their canon coun-

[1] `https://www.vox.com/2016/6/7/11858680/fandom-glossary-fanfiction-explained`. Accessed 30 August 2019.

terparts. We also see character representations reflecting different portrayals of characters in terms of how they are cast in relationships in fanfiction stories. These measurements afford hypothesis formation regarding the aims of fanfiction writers as they portray of a world of their own creation.

5.2 Related Work

5.2.1 Comparative text analysis for computational social science

Recent approaches have used word embeddings for comparative text analysis in NLP. An et al. (2018) plot embeddings on axes between antonyms operationalizing either end of a conceptual spectrum and present differences in how concepts are discussed in online communities. Lin et al. (2018) design 'social word' vectors to capture differences in associations across texts from different languages and cultures. Comparing across 20th-century historical corpora, Garg et al. (2018) find that embedding spaces capture known social stereotypes. Using contextualized word embeddings, Field et al. (2019) examine representations of victims and perpetrators in news stories in the #metoo movement. Much of this work assumes that word embeddings are able to capture more abstract social meanings and associations in text. Our work tests this assumption with an empirical validation on the ability to capture differences in how relationships are presented in derivative works, considering how such works are intertextually related to original stories.

5.2.2 Intertextuality

Coined by literary theorist Julia Kristeva in the 1960s, the term *intertextuality* is used to refer broadly to the ways in which texts are necessarily understood in the context of other texts (Allen, 2011). This encompasses the process of what discourse analyst Norman Fairclough (2003) terms *recontextualization*: picking up a piece of text and placing it in a different context for reinterpretation in a different ideological framework. Also related is the *dialogic* approach of literary theorist Bakhtin, who approached texts as voiced with perspectives responding to or reproducing perspectives in other texts (Holquist, 2003).

Within the NLP community, Visser et al. (2018) use the term to describe a system to combine annotated corpora based on topical connections. Other NLP tasks, such as tracing the origin of specific pieces of texts on Wikipedia or in fanfiction (Shen et al., 2018), also relate to intertextuality. Sim et al. (2016), for example, trace the influence of amicus briefs on the language of US Supreme Court opinions with a Bayesian probabilistic model, while Niculae et al. (2015) trace quoting behavior of political media coverage. Our approach more directly examines how the transformation of abstract concepts, not just direct quotes, can be captured in derivative texts.

5.3 Methods

A large part of the contribution of this work is a description and evaluation of a suite of approaches for a) extracting text relevant to characterization in fanfiction and b) building embed-

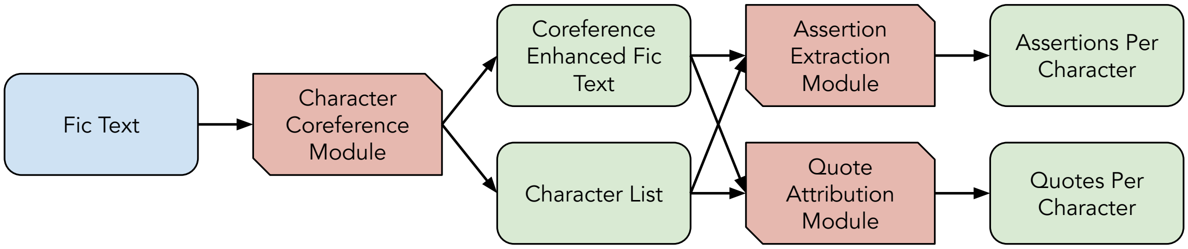


Figure 5.1: Fanfiction pipeline overview. From the text of a fic, the pipeline produces a list of character mention clusters with assertions and quotes for each character cluster.

ding representations for characters and relationships. We describe a variety of approaches appropriate for addressing different questions of intertextual framing and then later evaluate these approaches with both quantitative and qualitative investigations.

5.3.1 Text Extraction

In order to focus our analysis on portions of stories pertinent to relationship characterization, we build a publicly available pipeline to cluster character mentions and extract meaningful text segments for each character cluster.

We begin with the BookNLP toolkit (Bamman et al., 2014). BookNLP is an open-source toolkit building on the Stanford CoreNLP tools (Manning et al., 2014), designed for studies of narrative writing. It includes character name clustering, pronominal coreference resolution, and quotation speaker identification. However, we found poor performance when we applied BookNLP to the fanfiction we were studying, with many character clusters involving mentions from multiple characters.

For clustering of character names, BookNLP includes some heuristics which in our application led to ambiguity and incompleteness. In cases where one character name is a proper substring of another character name, BookNLP clusters them together as one single character. This does help in correctly clustering character names like, ‘Harry Potter’ and ‘Harry’ as one, however it fails to merge character names like ‘Harry Potter’ and ‘Mr. Potter’ into the same cluster and instead treats them as two different characters. Thus, BookNLP fails and gives ambiguous results when multiple characters are being referred by their similar last names. This happens mostly in cases when the story involves a big family. BookNLP also treats named entities labeled as organization names as characters, leading to confusion.

BookNLP provides verbs and adjectives related to characters, but we wanted larger spans of exposition and narrative related to characters to build embedding-based fic representations for characters. In addition, while processing quotes, BookNLP simply assumes the character mentioned closest to the quote is the speaker, while the quotes in fanfiction are often more complicated.

To improve performance on fanfiction and customize character feature extraction, we develop a publicly available fanfiction NLP pipeline. An overview of the pipeline is shown in Figure 5.1.

For each fic, we first perform character coreference. Here we integrate named entity recognition, entity coreference, and character name clustering in order to identify each mention of a character within a fic (including alternative forms of the name and anaphoric references) and associate them with a single standardized character name. This coreference is based on Stanford CoreNLP (Manning et al., 2014) with a few additional constraints and modifications. We start with CoreNLP for clustering these mentions (Clark and Manning, 2015) and additionally keep track of a cluster-level character name, which is the longest proper noun mention under 5 words after merging substrings with looser constraints than BookNLP. To increase cluster coherence, we track a cluster-level character gender, which can have values of *female, male* or *unknown/other*. This value is initialized to *unknown/other* until pronouns are added to the cluster and the gender can be decided accordingly. Once a cluster-level character name is determined for each mention, this standardized character name is inserted as a tag into the text after each mention, as well as being added to a list of characters for that fic. The list and augmented text is then used in the subsequent two branches of the pipeline.

After coreference, we extract assertion texts that mention a character. Here we use a very informal notion of an assertion, which is simply a topically coherent segment of text that mentions the character. We use a simple topic segmentation approach called TextTiling (Hearst, 1997) to segment the portions of the narrative with quotes removed. For each character, we then retain only those segments that mention the character, based on the analysis produced by the character coreference module.

Concurrently, we extract the quotes uttered by each character. Each quote is associated with a character, a paragraph index from the original fic, and the paragraph index of a quote it replied to (with a null option). Quotes are attributed by extracting features such as character name placement (Elson and McKeown, 2010), and providing these features to a pretrained SVM-Rank classifier (Joachims, 2006) which ranks likely characters for each quote. We follow the assumption made by He et al. (2013) that all quotes of a single paragraph are uttered by only one character.

We have not yet evaluated the coreference, assertion extraction, and quote attribution modules of this pipeline. A systematic evaluation of these modules and our adaptations from BookNLP for the fanfiction domain is part of the proposed work in this thesis (see Chapter **??**).

5.3.2 Text Representation

Our computational goal is to build vector representations of characters that reflect in a quantized form how characters are presented in specific fanfiction stories, as well as in canon. To do so, we borrow an approach from NLP for learning vectorized word representations that not only captures co-occurrence relationships, as in earlier approaches such as Latent Semantic Analysis (Landauer et al., 1998) and Latent Dirichlet Allocation (Blei et al., 2003), but also captures proximity and sequencing information within texts. More specifically, word vectors model shallow lexical semantics by positioning words that occur in similar contexts close within a vector space. One such method is the popular word2vec model (Mikolov et al., 2013a); we use an adaptation of this approach, FastText, which also trains vectors for subword information (Bojanowski et al., 2017).

In the remainder of this section, we review a range of embedding based approaches we have used within this work. We explore three types of techniques to generate embeddings that allow us to compare between canon texts and derivative fics: (1) training approaches that build up vector-based representations from corpora, (2) alignment approaches that construct a shared embedding space trained on different corpora, and (3) contextualization approaches that enable document-specific adaptations of word embeddings.

Embedding Space Training

Word embedding spaces offer dense vector representations of words that reflect word co-occurrence and lexical semantics in the corpora on which they were trained. As a base embedding space, we pretrained FastText embeddings (Bojanowski et al., 2017) on a collection of 5130 science fiction and fantasy e-books, the closest genre to the Harry Potter series on which we would apply our methods. Starting from the pretrained background embeddings, we trained a shared embedding space using the combined canon and fanfiction corpora, as well as separate embedding spaces for canon and fanfiction.

Embedding Space Alignment

In order to compare the separate fanfiction and canon embedding spaces, we want to map the embeddings from both spaces into a shared space so we can compare how word embeddings in canon and fanfiction have changed. We base our transformation paradigm on Mikolov et al. (2013b), who describe an approach for learning a linear projection from monolingual word vectors for a source language into a target language vector space for machine translation.

Given a set of n known word translation pairs $\mathcal{T} = \{x_i, y_i\}_1^n$ where $x_i \in \mathbb{R}^{d_1}$ is an embedding for word w_i in the source language and $y_i \in \mathbb{R}^{d_2}$ is an embedding for the translation of w_i in the target language, our goal is to find a linear translation $W \in \mathbb{R}^{d_2 \times d_1}$ such that Wx_i can approximate y_i. W can be learned using stochastic gradient descent to solve the following optimization problem:

$$\text{argmin}_W \, ||Wx_i - y_i||^2 \, , (x_i, y_i) \in \mathcal{T} \tag{5.1}$$

In our approach, rather than considering translation pairs (x_i, y_i), we want to consider pairs (f_i, c_i) where f_i is the embedding for w_i in the fanfiction space and c_i is the embedding for the same word in the canon space. Adapting this method for comparing fanfiction with canon texts, however, requires additional considerations. First, equation 5.1 aims to find a projection W that maintains similarity between seed pairs (x_i, y_i). For translation, it is easy to select seed translation pairs (x_i, y_i), as any known translation pair can provide grounding for learning the projection matrix W. Our goal, on the other hand, is to investigate both similarities and differences between the canon and fanfiction vector spaces. As a result, the choice of seed words has a large effect on what concepts will be similar in canon and fanfiction after the projection.

Traditionally, seed translation pairs are drawn from the most frequent words in the source language for translation. For our task, however, the most frequent words in the canon and fanfiction corpora are often important character names or concepts that could have different representations in canon texts versus fanfiction. We instead use the top 1500 verbs in the canon text that are also in the fanfiction corpus. The next 500 most frequent verbs are used as a validation set for learning the transformation. We expect the meaning of verbs to be relatively stable across fanfiction and canon in comparison to nouns and adjectives.

Contextualized Word Embeddings

The approaches for text representation described above 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 5.4.4). As an alternative, we also 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., 2013b).

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 fics, for each character within each fic we average contextualized representations of all of that character's mentions within the fic. This gives us a representation of the character that can be compared across fics.

5.4 Validation Experiments

As an illustration of some of the capabilities of the proposed set of tools, we present an investigation of how relationship types are portrayed in fanfiction, often in contrast to how they are framed in canon. We begin with a motivation for a specific task as well as its specifications, and then describe the data, the computational setup, and the experimental results.

5.4.1 Prediction Task Experimental Design

To investigate one form of canon divergence common in fanfiction, we set up a validation task predicting whether fics present character relationships similarly or differently than in canon.

Our assumption is that a computational system that can distinguish between fics that are presenting character relationships in the same way romantically as canon is capturing intertextual divergence from the original text. We hypothesize that focusing on specific portions of texts relevant to characterization provides a stronger signal for learning relationship differences than using the whole text. We also hypothesize that explicitly representing differences between canon and fanfiction in embedding spaces can act as a signal for relationship characterization. We use the embedding approaches described above to construct fic-specific representations of character portrayals in line with these hypotheses as we build and evaluate our predictive models.

For each pairing of main characters within each specific fic, we distinguish between three separate binary characterizations:

1. **Canon**: If a pairing in a fic is represented the same way romantically as in canon. That is, if the pairing is romantic in canon and romantic in the fic, then Canon is true. If the pairing is non-romantic in canon and non-romantic in the fic, then Canon is true. Otherwise it is false.

2. **Romantic**: If a pairing in a fic is romantic, regardless if the characters are romantic in canon, then Romantic is true, otherwise it is false.
3. **M/M**: If the two characters in the pairing, regardless of romance, are male, then M/M is true, otherwise it is false.

The first task is the one that is most of interest to us. However, the other two allow us to confirm that we are actually able to perform the first task, and not that we simply get what appears to be good performance just because the corpus is distributed in a way that provides useful proxies.

Predicting romance between characters does not necessarily capture anything about any shifts in framing from canon. Predicting gender could serve problematically as a proxy for canon divergence since most fanfiction relationships are male-male pairings and most in mainstream media are female-male. For gender, we predict whether the pairing is male-male or female-male. Fanfiction has a strong historical trend of pairing male characters together (Tosenberger, 2008; Fazekas, 2014) and none of the most popular pairings we examined were female-female.

Gold labels for all three tasks come from Archive of Our Own metadata, an extensive and well-used system where authors tag the characters and relationships present in each fic.[2] Many fics contain multiple relationships, including relationships with more than two characters, but we narrow our focus to two-character pairings.

The constructed embeddings are input to a logistic regression model. Though better performance could likely be obtained with models other than logistic regression, our goal is to evaluate the expressiveness of just the embedding spaces, not model architectures.

5.4.2 Data

To be able to compare fanfiction with canon, we selected J.K. Rowling's *Harry Potter* series due to its relatively large canon text source and active fandom, with over 100,000 works of fanfiction on Archive of Our Own alone. We collect all fics from the Harry Potter fandom posted by November 19, 2018.

To create our dataset, we first selected 5 main characters in the Harry Potter series who were often represented in a variety of canon and non-canon romantic pairings. We then narrowed consideration to pairings for which at least 1000 fics listed that pairing as romantic in their story. These selected 5 characters and 6 pairings are listed in Table 5.1.

For each pairing, we sampled 900 fics containing at least 5 paragraphs with both characters mentioned. Ninety percent of the fics in this set are taken as a training set, with the rest as a test set. We also sampled 332 fics per pairing for a validation set for tuning. Each prediction instance is a particular pairing within a fic. The same fic could be sampled for different pairings, but there was no overlap between fics in the training and test sets. Dataset statistics are presented in Table 5.2.

[2] `https://archiveofourown.org/faq/tags?language_id=en#whatisatag`. Accessed 30 August 2019.

	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 5.1: Selected characters and pairings in the Harry Potter fandom. Starred pairings are romantic in the canon text.

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

Table 5.2: Number of instances per data split.

5.4.3 Embedding Input Text

We argue that relationship framing is best captured by locating which parts of texts are most relevant to characterization. To test this hypothesis, we compare using the whole fic to using extracted assertions and quotes where both characters in the selected pairing are interacting.

We remove all Harry Potter character names[3] 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 fanfiction pipeline to our selected characters by matching character clusters that contain the first name of the selected character.

Text segments used in experiments are described below:

- **Entire fic**: 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.

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

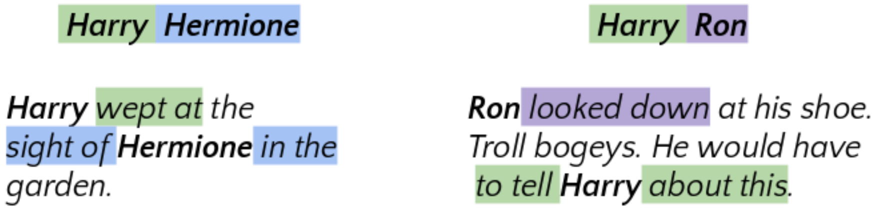

Harry Hermione
Harry Ron
Harry wept at the sight of Hermione in the garden.
Ron looked down at his shoe. Troll bogeys. He would have to tell Harry about this.

Figure 5.2: Example of fic-specific pairing representation construction. Context words around character mentions are averaged for each character, and then concatenated into a representation for the pairing.

5.4.4 Fic-specific Pairing Representations

To construct representations for pairings within specific fics, we take a TF-IDF weighted average of word embeddings in a window around the co-referenced mentions of each character in the quotes or assertions (see Figure 5.2).

To find the optimal window size, we try embeddings averaged over context windows of 5, 10, 25 and 50 words before and after character names. 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 10 words as our window for the following experiments:

- **Unigrams**: TF-IDF weighted bag-of-words unigram features over the context.
- **Embeddings**
  - **Background**: Pairing representations from word embeddings trained only on a background corpus of mainstream science fiction and fantasy (as described in Section 5.3.2).
  - **Shared**: Pairing representations from word embeddings pretrained on the background corpus and then trained on a corpus of Harry Potter fanfiction concatenated with the Harry Potter canon novels.
- **Embeddings + canon difference**: Like the pairing representations from fanfiction, we construct canon representations for each pairing. We average embeddings from 10-word windows before and after character names in paragraphs where both characters are mentioned. Coreference is less necessary in the case of canon since names are frequent enough in a book-length text. We then use this canon representation in each of the embedding spaces described above to compute cosine distance or vector difference (vector subtraction) from canon to the fic-specific representation of the pairing. This is our operationalization of fic-specific difference from canon.

  In the **Aligned** version of this experiment, the canon pairing representation uses word embeddings pretrained on the background corpus and then adapted to just the canon text, while the word embeddings for the fanfiction are taken from a space adapted just to fanfiction. Both embedding spaces are then aligned to the same background space (see Section 5.3.2), from which cosine distances and vector differences between the canon and fanfiction embedding space representations are computed.

	Prediction Task		
Feature Set	**Canon**	**Romantic**	**M/M**
Entire Fic (F)	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 5.3: Prediction accuracies across embedding input types.

5.4.5 Results and Discussion

Our first set of experiments test whether extracting texts specific to characterization gives a stronger signal than using the whole text, as well as if the inclusion of a distance value or difference vector between canon and fanfiction improves performance. Results are reported using the **Shared** embedding space.

Focusing on text that is relevant for characterization (quotes and assertions) performs substantially better at capturing divergence from canon than approaches using the entire fic text (Table 5.3). This supports our hypothesis that locating text that is relevant to specific social meanings can help clarify a signal that otherwise might be washed out amidst full texts.

Also evident in Table 5.3, difference vectors improve results across all input ($p < 0.05$ using McNemar's test) with assertions and assertions with quotes when compared to no difference vectors. This suggests that vector differences in lexical-semantic embedding spaces 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 fics in our experimental dataset, so we only report results in Table 5.3 on quotes in combination with assertions. On instances where quotes are present, they perform slightly better than assertions for the canon prediction task (66.7 accuracy over 66.0), and slightly worse for predicting M/M, 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 serve to complement each other.

Table 5.4 shows prediction results across different embedding construction methods. These results are with canon-fanfiction difference vectors concatenated with fic-specific pairing embeddings , as that was found to be the best setting (Table 5.3). The shared embedding space was found to give the highest performance on canon prediction, with a significant ($p < 0.05$) improvement in performance with the A+Q shared space compared with A+Q with the background

		Prediction Task		
Approach	**Feature Set**	**Canon**	**Romantic**	**M/M**
Majority	–	52.40	73.16	75.72
Unigrams	F ‘	55.30	65.08	65.56
	A	62.04	73.61	70.40
	A + Q	63.00	75.74	68.95
Background	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 5.4: Prediction accuracies across different embedding approaches and baselines. All the embedding-based feature sets (**Background**, **Shared**, and **Aligned**) include the difference between the embedding for each fanfiction pairing instance against the embedding for the pairing in canon.

space.

The embedding space alignment technique more formally 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 prediction task, when compared to simply training embeddings on a shared space of canon and fanfiction text. This suggests that there may be differences in how romantic relationships are portrayed in canon texts versus in fanfiction; aligning the embedding spaces may clarify that signal. However, further comparisons among canon, fanfiction, and shared embedding spaces would be necessary to confirm where the alignment method may provide more useful difference vectors.

Results on the two corollary tasks, predicting whether a pairing is romantic in a fic regardless of canonicity and whether both characters in the pairing are male, pattern differently from our main canon prediction task (Table 5.4), leading us to conclude that our representations are capturing an element of canon divergence in relationship framing beyond that of simply predicting romance or gender. For predicting whether the two characters are both male, the best performance comes from just assertions. This could be because narrative and exposition provide more gender cues than quotes.

5.5 Exploration of Character Framing Through Visualized Embeddings

In this section, we compare a multidimensional representation of character framing across multiple fics at a time using 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, we aimed to construct a representation that was not specific to the characters involved, or to any specific fic, 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 fic. 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. Thus for this exploration into character vector modeling and visualization, we apply the contextualized vector approach described in Section 5.3.2 to the names of characters.

For data, we again use fanfiction from Archive of Our Own (archiveofourown.org) 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.

Reducing the dimensionality of these vectors with an approach like t-SNE (van der Maaten and Hinton, 2008) allows 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 5.3). 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 5.4).

In Figure 5.5, 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

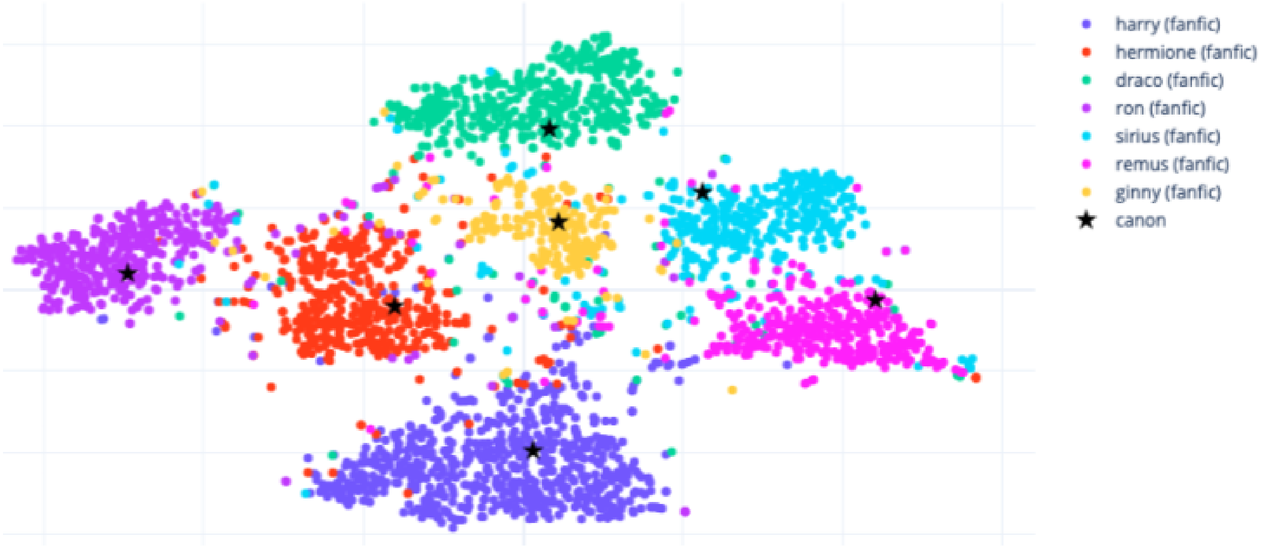


Figure 5.3: Character name vector visualizations. The closest star to each cluster is the vector for that character learned from the canon (original media).

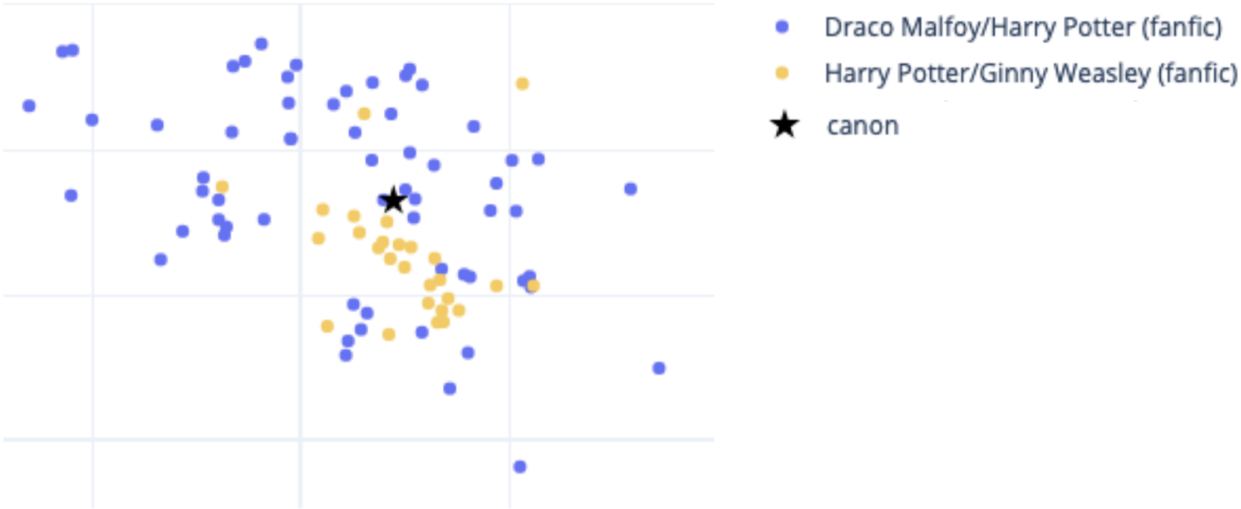


Figure 5.4: Representations for ‘Ginny’ in fanfiction and canon, colored by relationship. A few outliers are omitted.

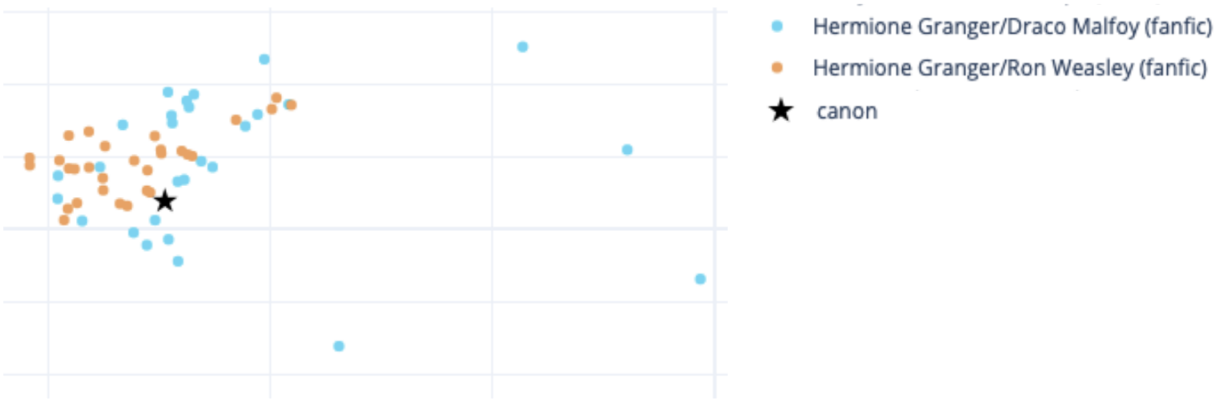


Figure 5.5: Representations for 'Ron' in fanfiction and canon, colored by relationship. A few outliers are omitted.

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.

5.6 Conclusion and Future Work

In this work, we take first steps toward computationally modeling how derivative texts transform the framing of characters and relationships. We find that vector differences in word embedding spaces between our canon and fanfiction representations capture differences in how relationships are portrayed. We also find that focusing on specific portions of texts that we believe to be more relevant to desired distinctions, in this case characterization, is effective. We use a computational pipeline for text extraction adapted from prior work.

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.

In future work we hope to validate these findings with further qualitative investigations on other fanfiction corpora and try other dimensionality reduction techniques for visualization. The types of word representations investigated here could also be used to identify other trends in fanfiction or other corpora, such as identifying semantic shifts across time. They could also be used to identify texts that use certain words in archetypical ways or in radically different contexts than most of the other texts in a corpus.

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

Part II

Proposed Work

Chapter 6

Fanfiction NLP Pipeline Evaluation

Our pipeline for extracting character features from fanfiction stories, described in Section 5.3.1, includes character coreference and the extraction of spans of text that encompass quotes, narrative, and evaluation of characters. Many of these modules are adapted from BookNLP (Bamman et al., 2014) or Stanford CoreNLP (Manning et al., 2014) for our specific domain of fanfiction, and some are added new.

Each part of this pipeline needs to be evaluated for effectiveness against other approaches on a set of diverse fanfiction stories. This is the work proposed in this chapter.

6.1 Coreference Evaluation

In our context, we wish to group mentions of characters (including pronouns) that co-refer to the same character, for example all mentions of *Captain Marvel* that corefer with *he* or other names for the same character grouped together. This requires entity recognition, pronoun resolution, and character name clustering.

We first tried BookNLP's character clustering module. However, the rules used to cluster character names were found to have errors when applied to our set. Frequently, mentions of the same character were grouped in different clusters due to tight substring requirements for combining clusters. For example, "Mon" and "Monica" were clustered into different clusters, as well as "Harry Potter" and "Mr. Potter". For this reason, we choose to use the trained statistical coreference model included in Stanford CoreNLP (Clark and Manning, 2015).

After iterations of qualitative checks on performance of the CoreNLP model, we chose to make the following modifications:

- Removing recognition of entities marked by NER as "Organizations" from possible character mentions, as this was one common source of error.
- Tracking a gender for each cluster of character mentions.

  This gender can have values of *female*, *male*, or *unknown/other*. Gender for the cluster is determined when pronouns are added to the cluster. We add a constraint that characters with genders of male and female cannot be merged (either binary gender can be combined with *unknown/other*).

Note that this would fail for characters undergoing gender transition in stories.

- Finding a "canonical" cluster name. This cluster name is the longest proper noun mention under 5 words after merging mentions that are substrings of each other. This step does not affect the clustering of mentions.

To evaluate the character coreference system, we will use the LEA link-based evaluation metric that avoids issues with precision and recall interpretability of earlier scoring systems (Moosavi and Strube, 2016). LEA considers links between entities in the *key* (gold, correct) partitions of character mentions against links in the *response* partitions produced by the system. Each link is between pairs of character mentions in the cluster, so there are for every entity e with n mentions there are $link(e) = \frac{n(n-1)}{2}$ unique coreference links. LEA evaluates a set of entities (clusters) with the following equation for each entity:

$$\frac{\sum_{e_i \in E} importance(e_i) \times resolution(e_i)}{\sum_{e_k \in E} importance(e_k)} \tag{6.1}$$

The "importance" of an entity cluster is simply

$$importance(e_i) = |e|$$

.

The "resolution" of a cluster is a measure of coreference link overlap from key to response clusters or vice versa. In the example of recall, we calculate resolution between keys K and response entities R as follows:

$$resolution(k_i) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)} \tag{6.2}$$

For precision, e_i is all $r_i \in R$ response clusters in Equation 6.1, and r and k are reversed in Equation 6.2.

We will evaluate the labeling of clusters with canonical names separately, reporting the accuracy of each cluster choosing the most fully formed (descriptive) name out of all mentions in its cluster.

6.1.1 Baselines

BookNLP For its character coreference, BookNLP (Bamman et al., 2014) uses a named entity recognition (NER) system to extract mentions of probable characters. It then uses a rule-based system to group these mentions into clusters, as well as a log-linear classifier for pronoun resolution from training on annotations on three classic British novels.

Stanford CoreNLP Statistical This model, proposed by Clark and Manning (2015), uses pairwise comparisons of mentions to determine mergers into clusters.

Model	#Mentions	Correct output	Incorrect output	Precision (%)
BookNLP	93	64	29	68.8
CoreNLP	148	101	47	68.2
CoreNLP + modifications	135	96	39	71.1

Table 6.1: Preliminary precision results from output of three coreference approaches on one fanfiction story. #Mentions is the number of mentions detected by each model; the true number of character mentions is unknown.

spaCy neural model The spaCy implementation of the neural mention-ranking model of Clark and Manning (2016). In a small test so far, this model has been found qualitatively to perform no better than the Stanford CoreNLP statistical approach on a small fanfiction sample.

6.1.2 Annotation

Our goal will be to annotate at least 500 coreference links in fanfiction stories of varying lengths from the 10 selected fandoms used in proposed analysis (Chapter 7). This is similar to the amount annotated in the development of BookNLP's pronoun resolution module. We plan to annotate these stories with a crowdsourcing task on Amazon Mechanical Turk, asking crowdworkers to identify whether a word in a sentence is a likely mention of a character, and further to identify links between entities: whether or not two character mentions refer to the same character in at most a paragraph.

6.1.3 Preliminary results

Character mention and cluster output from BookNLP, Stanford CoreNLP, and our adapted Stanford CoreNLP approach were evaluated on one fanfiction story from the Friends TV show fandom. Note that only the output was evaluated and no gold label annotation from the text, so recall is not possible to calculate. Precision values are given in Table 6.1. The number of mentions is the number of character mentions detected by each model, not the true number of mentions in the text. Before annotation for the true number of mentions is completed, the likely recall can be approximated by the number of mentions given as output from the models. So far, our modifications equal or improve precision percentages over BookNLP while increasing the number of mentions considered.

Though BookNLP uses the gender of pronouns as a feature in its log-linear classification for pronoun resolution, our constraint approach to gender appears to improve results.

6.2 Quote Attribution Evaluation

BookNLP uses a heuristic for quote attribution in which quotes are attributed to the nearest character mention. This can easily misattribute implicit quotes, in which the names of the character who said the quote is not attributed with a quotative tag such as *said*. This can be exacerbated

Model	Accuracy (%)
BookNLP	26.5
He et al. (2013)	40.2

Table 6.2: Preliminary accuracies on quote attribution averaged over three fanfiction stories N = 132 utterances.

in conversation chains, where a back-and-forth conversation between two characters is described without quotatives.

Based on the approach of He et al. (2013), we use a classifier to rank character mentions in close proximity to the quote based on features of that character mention, such as the distance from the quote to the mention. Quotes are extracted based on quotation marks. As described in Section 5.3.1, we extract features such as character name placement around the quote. These features are provided to a SVM-Rank classifier pretrained on the *Pride and Prejudice* novel to rank likely characters for each quote. These annotations are provided by He et al. (2013).

We evaluate quote attribution based on accuracy, following Elson and McKeown (2010) and He et al. (2013).

6.2.1 Annotation

Annotation of quotes in 100 fics of varying length will be accomplished in a crowdsourcing task on Amazon Mechanical Turk. Crowdworkers will be asked to match extracted quotes with the correct character mentions. Stories of varying lengths and fandoms will be selected, all with non-explicit ratings.

Note that this quote annotation will be for the purpose of building a test set; more annotations would have to be collected to train a new model. We plan on using the pretrained classifier of He et al. (2013) unless performance is particularly poor.

6.2.2 Baselines

BookNLP, which matches quotes to the closest character mention, is our baseline.

6.2.3 Preliminary results

One annotator (a colleague) annotated 132 utterances in 3 stories of fanfiction from the *Detroit: Become Human* videogame fandom. Results on these stories compared with BookNLP are reported in Table 6.2.

6.3 Assertion Extraction Evaluation

In FanfictionNLP, we also extract spans of text that relate to characters but that are not quotes from the character or from other characters. The goal of this extraction is to identify spans of text that relate to characters for the context of word embedding training. We are not aware of any other work that does this in the same way. To select the boundaries of these text spans, we use TextTiling (Hearst, 1997), which looks for segment boundaries based on word frequency changes.

6.3.1 Annotation

To evaluate our assertion extraction, we annotate all non-quote narrative and exposition that relates to a character. This is not an easily defined distinction. Since the goal of this pipeline extraction is to extract spans of text that are relevant to characterization, our guiding question for each sentence becomes: “Is this sentence relevant for portraying a character to a reader?” This encompasses narrative and evaluation that directly includes a character, but also could include more indirect characterization. For example, “Mary stood by the hill.” is narrative relevant to the character of *Mary*. Though a bit more indirect, “The hill towered over her” could also be also relevant. More unclear might be a following sentence such as, “The grass on the hill rippled pleasantly”, though we would include it as characterizing Mary since there is a reasonable chance that the subjective perspective taken in this statement is the character Mary. As in Chapter 5, the use case in mind for this text extraction is training word vectors, so in difficult annotation cases, we tend toward including more text as possible characterization.

We again annotate these sentences with a crowdsourced task on a set of 25 fics with careful specification of which sentences could include this perspective-taking relevant to characterization.

We evaluate our text spans with precision and recall on the sentence level.

6.3.2 Baselines

We compare our approach to selecting all non-quoted sentences in paragraphs containing mentions of the character (from the output from the coreference module), as well as only selecting sentences that contain mentions of the character.

Chapter 7

Positioning Gender and Sexuality in Fanfiction

7.1 Introduction

Authors of fanfiction often explore same-sex relationships and gender expressions outside societal norms, leading researchers to label fanfiction as a "queer space" (Lothian et al., 2007). Fanfiction writers, many of whom identify as LGBTQ (Dym et al., 2019), transform characters, plots, and relationships in original stories in virtually limitless directions. But what are the contours of this queer space? In their myriad re-writings and re-making of characters and relationships, how do fanfiction authors position particular genders and sexual identities? Do they use fictional characters to create worlds that transcend normative configurations of gender and sexuality (Barthes, 1977)? Or do they, as some critics of fanfiction culture posit, further existing heteronormative or cisnormative narratives operating underneath the representation of queer relationships (Walton, 2018)? Alternatively, does fanfiction redeploy traditional gender roles and stereotypes in queer relationships parodically, like the practice of drag (Butler, 1990)? Distinctions between the reproduction of norms and parodies of them likely require extensive qualitative analysis, but in this chapter we make the first step in using computational tools to quantify associations made with identity types across hundreds of thousands of fanfiction stories.

Specifically, we investigate how gender roles are portrayed in fanfiction. Though the same-sex male "slash" relationships are often the focus of fanfiction studies, gender in fanfiction is less often studied. In an early application of natural language processing to fanfiction, Milli and Bamman (2016) find a stronger focus on female characters than in original media in some fanfiction communities. However, little is known about the roles that female, male, and nonbinary characters are cast in on a large scale in fanfiction, and how this portrayal varies across other dimensions of identity. For example, characters in gay male relationships could be presented in a way that still fits normative heterosexual gender roles, one character being domestic and submissive and the other being dominant.

We present three research questions related to gender roles in fanfiction:

RQ1. How do roles characters take in fanfiction compare with stereotypical gender roles?

RQ2. Do characters with the same gender play similar roles in fanfiction that they do in canon?

RQ3. Do character with the same gender play similar roles when cast in queer relationships in fanfiction compared to being cast in straight relationships?

To explore each of these questions, we discover popular "roles" that characters play based on topic modeling of character descriptions, inspired by Bamman et al. (2013). We will also use a crowdsourced lexicon from the work of Fast et al. (2016) relating popular terms in online fiction writing to gender to interpret these character roles in relation to gender.

7.2 Methods

7.2.1 Discovering character roles

Instead of directly mapping character descriptions to conceptions of masculinity and femininity as Fast et al. (2016) do in their exploration of gender stereotypes, we wish to first discover configurations of character activities and descriptions, *roles*, that frequently co-occur. We have two reasons for doing this. First, fanfiction is known as a queer space where gender and sexuality are represented in ways that may be significantly different from mainstream gender and sexuality categorization organized around binaries such as *gay/straight* and *male/female*. Considering these possibilities is part of the goal of this exploration. We thus want to consider character roles that may relate to gender in ways other than simple male and female stereotypes, and we also want to consider non-binary characters. Second, sociolinguists argue for an *indirect* relationship between language and gender. Ochs (1992), for example, posits that identity types such as gender are not directly referenced by particular words, but rather that they reference particular stances in interaction, such as being violent, being friendly, being polite, being assertive, etc. These interactional stances are then in turn ideologically associated with categories such as gender. Following this structure, Fast et al. (2016) asks crowdworkers to relate terms to topics that are stereotypically male or female to construct their lexicon of gendered terms. We wish to incorporate this framework in our model of gender roles. To do so, we relate the language used in fanfiction to portray characters to gender through an intermediate level of roles that those characters play. For example, if we simply find that female characters in fanfiction are being described with submissive words such as "wilted" and "dainty", we might miss a trend in fanfiction that characters described in this way are commonly paired with action terms in casting female characters in superhero roles.

Note that we currently collapse distinctions between what roles are selected and how those roles are portrayed: the learned character roles are simply what are portrayed in the corpus of fanfiction. These roles do not directly correspond to character types in the original series or in general media, but are simply the patterns in the portrayal of characters in our corpus of fanfiction that are frequent enough to be discovered by a topic model.

The documents for our topic modeling consist of extracted features for every character in every fanfiction story. To extract character features, we start by running fanfiction stories through our FanfictionNLP pipeline (described in Section 5.3.1). Part of the output of this pipeline are assertions, i.e. narrative, evaluation, and description, of coreferenced mentions of characters in each story. Early topic modeling of these assertions led to topics that often related to a setting or situation rather than character roles, so we further extract terms more directly related to characterization. Following Bamman et al. (2013) and Fast et al. (2016), we extract verbs for which a character mention (name or pronoun) is the grammatical subject or object, and "attributes", adjectives and appositives, that are used to describe the character.

Specifically, we use spaCy (Honnibal and Montani, 2017) to parse assertions into syntactic dependencies. To extract character subjects, we select verbs for which the character mention has a *nsubj* or *agent* relation. For verbs where the character is an object, we extract *dobj, subjpass, dative* and *pobj* relations. To extract attributes that describe the character, we select dependents of the character mention who are adjectives with a *amod* or *appos* relation or nouns with a *nsubj* or *nmod* relation, as well as copulas with *attr* relations to "is" or "was" with a character mention as a dependent. This attribute extraction captures cases such as "he was *great*" or "Molly, a *powerful* singer, stepped on stage".

For topic modeling, each document is features for each character in each story. We use the Structural Topic Model (STM), which is based on Latent Dirichlet Allocation (Blei et al., 2003), to model co-occurring character features. STM, developed by political scientists Roberts et al. (2014, 2019), directly incorporates document covariables of interest so that topic distributions and word distributions within topics can be learned separately for separate values of covariates.

We consider these topics to be character roles, since they are learned over extracted character features. In our case, we allow separate parameter values to be learned for role (topic) distributions among distinct character genders and distinct relationship types that characters are portrayed in. As a first step, we only allow topic *prevalence*, that is, parameters determining the likelihood of a topic in a document, to vary according to this metadata. Thus topic proportions θ can be influenced by a set of covariates X through a regression model, $\theta \sim \text{LogisticNormal}(X\gamma, \Sigma)$, where γ are coefficients for topic prevalence and Σ is a covariance matrix between topics. STM estimates this model with variational expectation maximation (EM). We do not currently (in our pilot project) allow the learning of separate parameters for topic *content*, which words are selected within a topic, for characters and documents with different metadata. We may include this at a later point to see, for example, if male characters are depicted in a violent role with different terms than female characters are.

To choose the number of topics, we first consider qualitative interpretability of distinct, coherent roles judged from the words with the highest probabilities in each topic. We will balance this qualitative examination with quantitative coherence metrics such as the selection strategy of Lee and Mimno (2014), who find the number of topics for which convex hulls from a word cooccurrence matrix, reduced to two dimensions with t-SNE (van der Maaten and Hinton, 2008) can be computed. In preliminary experiments, 10 topics is found qualitatively to contain useful topics that can be viewed as corresponding to character roles.

7.2.2 Relating character roles to gender

We relate character roles to gender in two ways:

1. Comparing the roles with genders of characters frequently enacting those roles. In the pilot project, we only capture characters labeled as female or male, but we have plans to capture non-binary characters in this step as well.

2. Comparing the top-ranked terms in roles with the stereotypical gender association in a crowdsourced gender-term lexicon (Fast et al., 2016). Note that this method only relates character roles to stereotypically masculine or feminine terms.

7.2.3 Annotating character gender and sexuality

Character gender To annotate characters for gender, we use the Wikia FANDOM pages when available. These are community-created wikis that often list characters' gender in a page about the character [1]. Fanfiction authors, however, often transform the gender of characters; this sort of behavior is exactly what we wish to include in our models. For stories where this is the case, including the practice of genderswap (McClellan, 2014), authors often signal this change to readers in the "additional tags" added to the story. Thus to capture deviations from the gender of characters in canon, we will search for tags added by the author that include terms that indicate gender, such as *trans*, *nonbinary*, *male*, and *female*, as well as the name of a character. If the character whose features are in a document is marked with a gender tag, we override the wiki canon gender annotation and annotate that character in that story as having the gender indicated by the author.

Character sexuality To annotate sexual orientation for characters, we use author-provided values for relationships between characters present in each story, a field commonly provided in our dataset (see Section 7.3). For each document (character-story pairing), we search for the name of the character in these relationships and find the name of the other character in the relationship. For a pilot experiment we ignore pairings with more than 2 characters and only take the first pairing that a character is in, but we plan on considering these cases in the future. We find the gender of the other character in the relationship from the Wikia reference, allowing for modification based on the author tags. Possible values for this field include:

- **Character not in relationship**: the character is not in a listed relationship in the story.
- **Straight**: if the character is male, they are in a relationship with a female character, or vice versa.
- **Queer**: the character is in a relationship that is any other than a female-male straight pairing.
- **Unknown**: the gender for the other character in the relationship is unknown.

7.3 Data

We use fanfiction from Archive of Our Own (AO3), on which fanfiction authors provide rich metadata about their stories, including names of characters in the main relationships in stories. More details on this dataset can be found in Section 3.2.

We select data from the most popular ten fandoms as of September 2018. Statistics for the data for these fandoms can be found in Table 7.1. For the pilot project, we start with data from

[1] See for example, `https://harrypotter.fandom.com/wiki/Main_Page`

Fandom	**Media type/s**	**Number of stories**
Marvel	comics, movies	256,153
Supernatural	TV show	187,764
Harry Potter	books, movies	176,531
DC	comics, movies	119,703
Sherlock Holmes	books, TV show	115,647
Teen Wolf	TV show	95,696
Star Wars	movies	76,346
Doctor Who	TV show	62,240
The Lord of the Rings	books, movies	56,113
Dragon Age	videogames	55,505

Table 7.1: Dataset statistics.

the *Harry Potter* fandom. We filter stories to English-language, complete stories with 1,000-50,000 words to exclude overly large stories that may dominate our role discovery, or extremely short stories in which there may not be enough character features. After this filtering in our pilot project, as well as filtering out stories for which we no character features from our pipeline were extracted, we settle on a dataset of 66,061 stories and 262,634 character-story pairs, the instances for our machine learning analysis. In preprocessing, we remove stopwords, character names, lowercase, and lemmatize words.

7.4 Character roles

Here we present the discovered character roles from STM in a pilot project on Harry Potter data. Ten topics were found to be qualitatively interpretable, distinct, and coherent based on top word probabilities in each role.

Topics are summarized in Table 7.2. Top words for each topic are given with the FREX evaluation described in Roberts et al. (2019). FREX is calculated as:

$$\mathrm{FREX}_{k,v} = \left(\frac{\omega}{\mathrm{ECDF}(\beta_{k,v} / \sum_{j=1}^{K} \beta_{j,v})} + \frac{1-\omega}{\mathrm{ECDF}(\beta_{k,v})} \right)^{-1}$$

where v is the word within topic k, ECDF is the empirical CDF, ω is a weight hyperparameter (set to 0.7 in the pilot), and $\beta_{k,v}$ is the probability of word v occurring in topic k.

These topics will change in the full proposed work after including data from more fandoms and a quantitative analysis to tune the number of topics.

Topic	Name	Top words (FREX)
1	facial gestures	nod, grin, smile, sigh, chuckl, shrug, shook, laugh, spoke, smirk
2	quotatives	said, ask, repli, suggest, mean, agre, point, repeat, think, protest
3	cognition	knew, saw, felt, heard, found, rememb, realiz, woke, thought, seen
4	violence	kill, die, defeat, trust, save, betray, destroy, protect, attack, deserv
5	movement	say, shake, sit, goe, take, catch, stand, hear, begin, add
6	love, leave	without, love, gone, visit, son, man, taught, taken, born, disappear
7	sex	moan, thrust, lick, arch, whimper, stroke, slid, suck, pant, gasp
8	disgust	sneer, scowl, note, snarl, snap, regard, strode, stalk, glare, attempt
9	dating, marriage	told, date, mention, marri, introduc, invit, wrote, fanci, friend, talk
10	desire	want, care, imagin, feel, touch, wish, expect, need, meant, hope

Table 7.2: Learned topics for Harry Potter data. Topic names are manually assigned.

7.5 RQ1: Character roles and gender stereotypes

7.5.1 Analysis

We first explore whether character roles align with stereotypical male/female roles. STM is designed to analyze relationships between learned topics and document metadata. To examine this effect, we train linear regression models predicting topic proportions for each document from selected document metadata (the recommended approach from STM). Topic proportions are drawn 25 times from the variational posterior and regressions computed. Average coefficient values from these regressions are then used as estimates of the effect of different metadata values on topic proportions.

7.5.2 Results and interpretation

We first consider the effect of binary gender (nonbinary and possibly trans values will be considered in the proposed work). Average coefficient values for the effect of binary gender are shown in Figure 7.1.

We first note that learned coefficients for effects are relatively small, though they are significant ($p < 0.01$) for all roles except for 'cognition' and 'sex', indicating that binary gender does not have a large association with topic proportions.

What effect is visible largely fits stereotypes, often concurring with those found by Fast et al. (2016) in online original narratives from Wattpad. The topics with the highest learned coefficients for male characters were 'violence' and 'disgust'. The prevalence of the topic of 'disgust' may be due to a few notable Harry Potter characters, such as Severus Snape and Mad-Eye Moody, who are characterized as often exhibiting sneers and scowls in the original series, which fanfiction writers appear to have taken up. Though Fast et al. (2016) found that male characters were characterized as overly sexual, this effect is not seen in our analysis. This may be due to the inclusion of variables for sexual orientation of relationships, for which we see a larger effect on the 'sex' topic (see Figures 7.2 and 7.3).

Topics with highest learned coefficients for female characters were the 'dating/marriage',

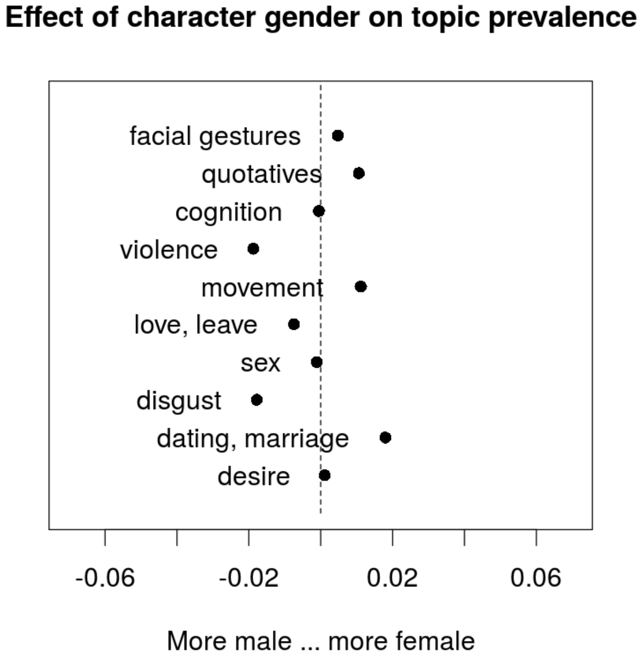


Figure 7.1: Effect of binary gender on topic proportions. Point estimates are labeled with topic names.

'quotatitives', and 'movement' topics. 'Dating/marriage' appears to be a character role that would match stereotypes of females as concerned with families, dating, marriage, the functions of "social reproduction". However, this trend is lessened for female characters in queer relationships (see Figure 7.3). Qualitative analysis of these results may nuance this finding, but it appears to be matching stereotypes of femininity. The topic labeled as 'movement' is currently dominated by the term *say*, and other terms which are in the present tense. Updates to lemmatizing that merge *say* and *said* may adjust this topic. However, the topic of 'quotatives', generally with terms such as *said* in the past tense, is also associated with female characters. More dialogue attributed to female characters (as opposed to other types of character action?) will be interesting to qualitatively examine. For each of these qualitative explorations, we will look at documents with particular character metadata (gender, for example) with high proportions of estimated topic prevalence.

We also plan on comparing the gender of characters listed by fanfiction authors with the gender stereotypes of terms used to describe these characters. Following the indirect indexicality model of Ochs (1992), we will first report mixtures of words in intermediate themes in the crowdsourced lexicon from Fast et al. (2016) and then links from those themes to masculinity and femininity. For each role, we will calculate the distribution of stereotypically male and female terms (as determined by the crowdsourced lexicon) in the top-ranked terms for each topic. This will give us an idea of the degree to which these topics align with stereotypes of binary genders. Statistics on how character roles fit gender stereotypes will be given overall and per fandom, as well as qualitative analysis of specific roles.

7.6 RQ2: Gender roles in canon vs. fanfiction

Fanfiction is based on original works of media, referred to as *canon*, which in the vast majority of cases do not represent same-sex relationships to the same degree as the fanfiction based on them do. But does this "queering" of characters and actions in fanfiction extend to gender roles? This exploration can also be seen as a more focused investigation of how gender roles differ in fanfiction compared to mainstream media, this time comparing directly to the original media on which particular fanfiction is based instead of indirectly through distributions of gendered characters or stereotypical associations with terms.

To examine gender roles in canon, we apply the same character feature extraction pipeline to the two fandoms in our dataset with book-based canons: Harry Potter and the Lord of the Rings. These series also have popular movie adaptations, and it is difficult to determine whether the majority of fanfiction stories in these fandoms are responding to the book or movie settings. We checked representative stories from a sample of stories and chose these two fandoms as ones where the book series is sufficiently important in the consciousness of fanfiction written on AO3. We excluded Sherlock Holmes, as the large majority of stories we saw were based on the recent BBC *Sherlock* adaptation running from 2010 to 2017. It is important to note that only including two fandoms, both from the fantasy genre, limits the generalizability of claims made about shifts in gender roles. It would be possible to extract character features from movie scripts or comic book texts, but narration and evaluation of characters is presented much differently in these media; quotes form a much larger part of characterization.

Proceeding with the analysis with Harry Potter and the Lord of the Rings, we will run STM with the same number of topics over a combined corpus of documents in canon in addition to fanfiction. Estimates of word length will be made such that each document in canon roughly corresponds with the average length of fanfiction stories (chapters in canon text will likely correspond better to fanfiction than full books). In this combined corpus, we will include a metadata field whether the document was 'canon' or 'fanfiction'.

We will then estimate the effects of this variable on topic distributions over characters of different genders (as in the analysis plotted in Figures 7.2 and 7.3). We then can qualitatively explore the changes in specific roles and the distribution of gendered language they contain. For example, this would reveal if topics that male characters in canon perform consistently shift to being more feminine, even more masculine, or with no clear pattern in fanfiction.

Note that these topics would be different from those learned over data from all 10 fandoms that we plan to include in other analyses.

7.7 RQ3: Gender roles in queer vs. straight relationships

Similarly to gender, we estimate the effect of a character's inclusion in certain types of relationships with regressions predicting the topic proportion of a document based on its metadata. Values for the sexual orientation of characters in relationships (if they are in one) are **character**

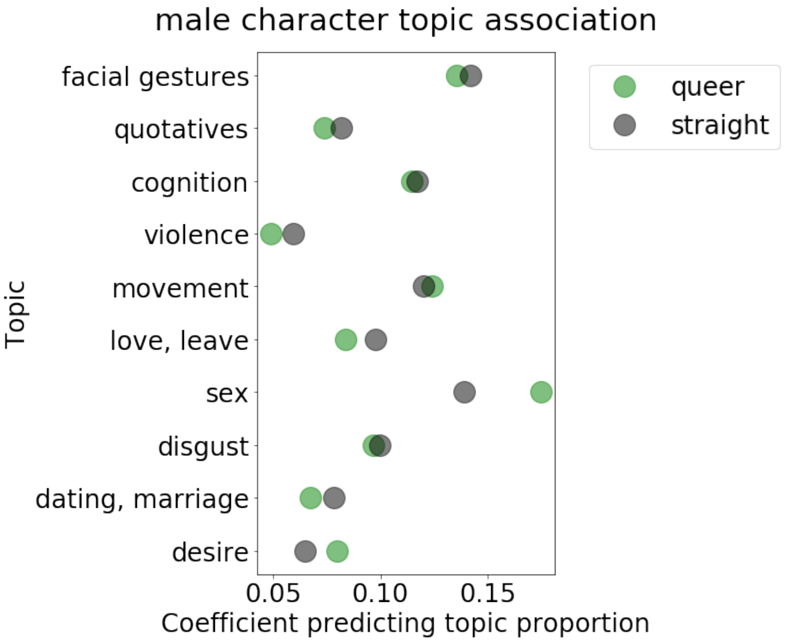


Figure 7.2: Associations between character sexuality and topic prevalence for male characters. Topic names are given on the y-axis.

not in relationship, **straight**, **queer**, and **unknown** (discussed in Section 7.2.3). These labels are annotated by comparing the the gender of the character whose features are extracted in the document with the gender of the other character listed in a relationship with that character. From the learned coefficients, these metadata appear to have a larger effect on topic prevalence than gender. See Figure 7.2 for average coefficients of regressions on 25 draws of document topic proportions from the variational posterior for male characters, and Figure 7.3 for the same estimates for female characters. Note that this only compares characters marked in **queer** and **straight** relationships, excluding those who are not in relationships in the story and those for whom the gender of their romantic partner is unknown (often the case for relationships that cross-over fandoms).

In straight relationships, does one character play roles that feature stereotypically feminine terms, while the other plays stereotypically masculine roles? For queer relationships, we could have a variety of different results:

1. Gender roles in queer relationships could match those in straight relationships, i.e. with one character tending toward masculine roles and the other toward feminine roles. This case would have to be closely examined qualitatively, as it could point to reproduction of heteronormative relationships (suggested in Walton (2018)), but could also be a parodic re-interpretation of heterosexual roles (Butler, 1990).
2. Roles of characters in queer relationships could match the genders of the characters they portray, such as both characters in **M/M** stories playing largely masculine gender roles. This could point to a more limited queering of gender along with the acknowledged queering of heterosexuality in fanfiction more generally, but qualitative exploration would be

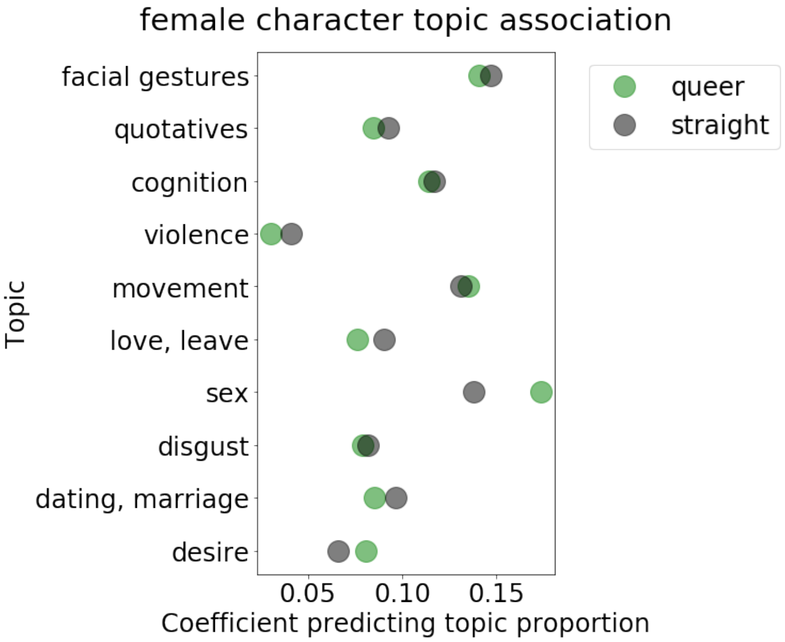


Figure 7.3: Associations between character sexuality and topic prevalence for female characters.

necessary to make any claims.

3. Roles of characters in queer relationships could exhibit a difference from characters in straight relationships that is more complicated than simply tending toward the stereotypical roles of the characters' genders or toward heterosexual gender roles. For example, characters in **F/F** stories could both tend toward masculine roles, or toward particular more masculine roles and particular more feminine roles. This could point to trends in "queering" gender roles as well as relationships in fanfiction.

From the pilot analysis, it appears that both men and women in queer relationships are portrayed as more sexual, less violent, less concerned with dating, marriage, and love, and more full of "desire". Qualitative analysis of documents with straight and queer characters that exhibit these patterns will have to be examined to see in what ways characters take up these topics. Furthermore, is it both characters in queer or straight relationships that exhibit these trends, or are there particularities between characters? Do these changes reflect the choice of which characters are commonly portrayed in queer or straight relationships? One important proposed step will be to directly compare the same characters that are portrayed in both types of relationships and chart changes in role distributions.

These findings would likely vary across types of queer relationships, which also must be taken into account.

7.8 Conclusion

In this project, we will investigate roles that characters in fanfiction in many popular fandoms are portrayed as playing. We relate these roles to the genders of characters playing them, as well as to stereotypical associations with binary characters. We also investigate how these roles change between fanfiction and original series, as well as across straight and queer relationships.

In a pilot project, we find that a character's gender has a relatively association with particular character roles, though where there is an effect, stereotypical trends such as violent men and relationship-focused women are found. There is a larger association with characters' sexual orientation; queer characters are cast as more sexual, with more longing and desire, and as less violent than straight characters. Qualitative analysis of these findings, as well as comparisons for characters cast in varying sexual orientations and genders in specific stories, will be next steps.

We also plan on comparing gender roles in original book-based stories with the fanfiction they inspire.

7.9 Future Work

We are interested in examining how different sorts of characters (gendered and sexed) are voiced in fanfiction. Our fanfiction pipeline includes quote extraction and attribution for characters. How do quotations from characters index their gender and sexuality, which may change from story to story and from canon to fanfiction? For example, do characters in gay male pairings speak differently from the same characters in straight (and likely canon) pairings? What language features (lexical, syntactic, or morphological) are most strongly associated with gender or sexuality changes? We would be strongly interested in investigating these questions in future work.

Which characters are "lead" or "supporting" characters may also play a role in which roles they are portrayed in. We will include this metadata value in proposed or future work.

Chapter 8

Identity Representations and Community in Social Media

8.1 Introduction

Similarities and differences between identity labels can occur on multiple possible dimensions. For example, *male/female, trans/cis, binary/nonbinary*, and *family-oriented/individual-oriented* are all possible axes of difference when comparing identity labels given for gender, such as 'male', 'transman', 'cishet', 'grandma', 'husband', 'bad boy', or 'fangirl'. Outside of social media analysis, sociolinguists Bucholtz and Hall (2005) argue that identity labels are always relational to other labels. They note that this relationality is along multiple axes, such as similarity, authenticity, and institutional legitimacy. In this chapter, our goal is to learn relational representations of identity labels that provide insight to which dimensions of similarity and difference are relevant with respect to content propagation.

As in Chapter 4, we use data from Tumblr, a social media site known for both identity talk and the strong presence of cultural subcommunities. However, in Chapter 4 we define feature extractors for identity labels based on observed popularity of those labels, and grouped these labels into categories based again on popularity and the researchers' own intuitive notions of similarity. The specific combinations of identity labels found to be relevant in the task, such as "anime" and "manga", seemed to indicate particular notions of similarity. However, the similarity we discovered did not always match our categories based on popularity. Here we remove those assumptions and learn a vector representation space for identity labels based on user behavior in content propagation.

To learn these flexible, relational vector representations, we use a multi-step training procedure where each step presents a training objective that incorporates important relationships between labels. We start by training representations that place identity labels occurring in similar textual contexts in similar spaces, and then further adapt these representations to adjust labels used by users who follow each other toward similar spaces (inspired by the concept of homophily from network analysis). Finally, we supervise these learned representations with the reblog prediction task to incorporate a direct signal of user interaction in our representations for identity labels. In our first evaluation of the learned identity label representations, we compare against the feature-rich baseline of identity label comparisons based on textual similarity and popular categories of identity developed without attention to community or network structure (described in Chapter 4). We hope to demonstrate that our neural architecture incorporating network structure outperforms this baseline on reblog prediction, and will present findings of relevant weighting of identity labels from content propagation.

But the similarities and differences in these representations, trained to reflect network structure and supervised with reblog prediction, do not necessarily correspond to the identity labels that are most representative of communities. Differences and similarities between symbols of self-positioning, such as the use of particular labels in blog descriptions, are only made meaningful through the interpretation of a particular community (Fong, 2004). If identity labels are symbols that "buy" a user recognition as a particular type of person, can we use the representations of these labels to understand the community economy in which those symbols are given meaning? For example, a label of 'anti reylo' only makes sense within a fandom framework where 'reylo' is recognized as being the proponent of a specific pairing of characters ('Rey' and 'Kylo'), and even further within a community-specific Star Wars framework where 'reylo' is known to be a major relationship community and so this user is positioning themselves as opposed to that majority and possibly "mainstream" following.

In a second evaluation, we compare the representations against communities learned from network structure and confirmed by experts on Tumblr subcommunities. We ask if the portions of blog descriptions that we find relevant for reblog prediction are also those that define communities from each other. Do our learned representations for identity labels in blog descriptions reflect notions of similarity and difference that are determined by specific communities? If so, this points to the utility of our assumption that identity and community are inherently wrapped up together on Tumblr, and the possibility of finding community-specific identity similarities and differences with this approach. If not, this basic assumption would be in question and point to users responding to identity signals in content propagation that differ from those relevant to community.

8.2 Related Work

8.2.1 Neural representations for social media users

Many researchers have used both self-descriptions and network information to learn neural representations for social media users (Yang et al., 2014).

[Fill in with other work on joint representations of text and behavior in online communities.]

This work differs in focus: our goal is to learn neural representations for identity labels first, and representations for social media users who use such labels second.

8.2.2 Community labeling

[Work outside of computer science on the use of labels within communities; especially how labels' "meaning" is relative to community understandings (Fong, 2004). Anything from communities of practice?]

8.3 Data

We use two sources of data from Tumblr for this project: text blog descriptions and network activity (including following and sharing content, *reblogging*). Similar to the work in Chapter 4, we restrict blogs to those which have added content to blog descriptions and which post or reblog a minimum of 10 reblogs from June 1 to November 30, 2018, a set we refer to as users U. However, we sample a larger number of users in this set than in Chapter 4: we aim for at least 10,000 compared with the previously studied 1000.

Blog descriptions. More information on the use and visibility of blog descriptions on Tumblr can be found in Section 3.1. In contrast to the approach in Chapter 4, we include all terms identified in short (25-character) spans between bootstrapped delimiters. We still exclude text in longer spans since we have found that these often indicate quotes or other content that are not identity labels. We do not limit our feature extraction from blog descriptions otherwise since we will be learning neural representations for these identity terms and do not need to think as much about label sparsity as in the feature-based approach in Chapter 4.

Network structure. We build a directed, weighted network among the users U from follow behavior. Reblog behavior could also be considered when building a network between Tumblr users, but we exclude this 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 to Blog A. Note that edges can be bidirectional between nodes if both blogs follow each other.

8.4 Methods

Our goal is to learn positions for identity labels in multi-dimensional vector space. We train these embeddings in a multi-step procedure, with separate optimizations to encode multiple aspects of relevant similarity among labels. Starting with embeddings of identity labels from use in Tumblr posts, we further train these representations so that identity labels that occur in similar contexts in blog descriptions are in similar space. We then incorporate network structure and adjust embeddings to be similar to those blogs who the user follows. Finally, we adjust embeddings to be similar to those who the user chooses to reblog in our reblog prediction task, incorporating an even more active level of interaction. Each of these training steps is described in more detail below.

8.4.1 Similar contexts training

We start with 100-dimensional embeddings trained on text found in a large corpus of Tumblr posts. For this pre-training, we select FastText (Bojanowski et al., 2017) as an algorithm that incorporates subword information, since that may be important for rare or novel identity terms on Tumblr. We further train these embeddings for identity labels on our training corpus of blog

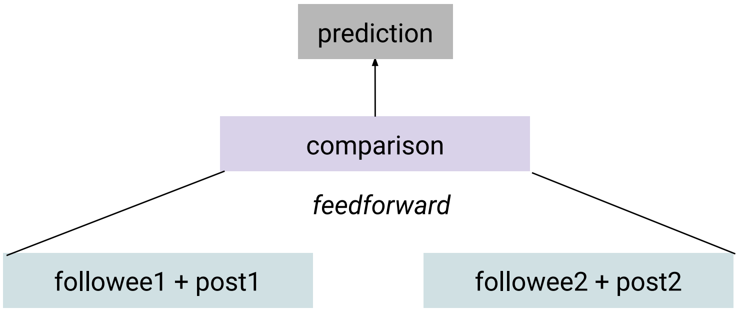


Figure 8.1: The comparison in the learning-to-rank reblog prediction task. Details on the representations for each followee+post are found in Figure 8.2

.

descriptions to encourage identity labels that occur in similar blog description contexts to be in similar vector space.

8.4.2 Follow network training

We then train identity label representations based on the follow network neighborhoods of users' blogs. We use node2vec (Grover and Leskovec, 2016) to learn these positions. Node2vec extends the skip-gram architecture of word2vec to graphs, learning to predict the context nodes in a random walk from a target node. We represent nodes (users) as the identity labels extracted from their blog descriptions. This can be thought of as an extension of word2vec (Mikolov et al., 2013a) in which documents are the identity labels used by users in localized network neighborhoods.

8.4.3 Reblog prediction training

We allow backpropagation through the identity label representations in our reblog prediction task in order to encode the relevance of actual user interaction (other than follows) in our identity label representations. This is a major goal of this work: to learn representations for identity labels based on machine learning over the visible social interaction that indicates community behavior. The details of the reblog prediction task are described in the next section.

Each instance consists of pair of posts and their associated "followee" user information. The follower of the two followee decides between these two instances in our reblog prediction task. In place of the vector subtraction in Chapter 4, we use a feedforward layer between vectors representing this comparison to produce a final prediction, as seen in Figure 8.1.

Each vector representing the followee information consists of a learned identity representation concatenated with post content information (Figure 8.2). To construct uniform-length representations from identity labels used by each user, we use a feedforward layer with a weighted average (attention) over the embeddings for the first n identity labels. We may also consider RNNs and CNNs, but ordering of these identity labels is less relevant than in typical sentence-

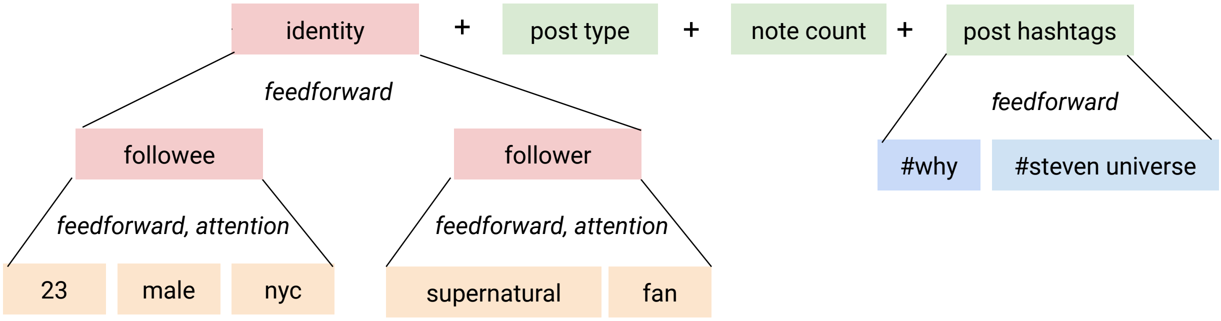


Figure 8.2: Details on the representations learned for users and post information
.

or phrase-based language tasks. These identity vector representations are concatenated with features representing post content: learned 100-dimensional vector representations of hashtags and count features for note (like and comment) counts. Both one-hot features and learned 20- and 50-dimensional vectors will be tried for representations for post type (image, text, chat, video, etc.).

In training, we will backpropagate through all layers of this model, since learning embeddings for identity labels that encode information about content propagation is part of our goal.

8.5 Reblog prediction evaluation

We first evaluate our neural approach with the reblog prediction task introduced in Section 4.4. This task is a pairwise prediction of which post a user will reblog from two followers who post in a similar timeframe (see Figure 4.2). In this evaluation we wish to demonstrate the capabilities of our flexible representations of identity labels that incorporate both blog description context and network structure against several baselines, including the feature-rich baseline developed in Chapter 4.

8.5.1 Baselines

We wish to test the utility of our neural identity label representations, as well as each separate training step in their development. We compare performance against the approach in Chapter 4, which uses count and binary features expressing word overlap between identity labels extracted from blog descriptions, as well as specific interactions between labels which were found to be informative (Section 4.5.2).

However, performance gains from our model could largely be the product of a more expressive, multi-dimensional vector representation. To test if this is the case, we include a baseline with the same architecture, but using identity label representations pretrained on Tumblr blog posts.

To test the effectiveness of each training step, we also include ablation baselines without the **context similarity** and **follow network similarity** training procedures.

8.5.2 Results

Forthcoming results will demonstrate the effectiveness of the learned embedding identity label representations compared with a feature-rich baseline. We will also identify which training steps are relevant for the reblog prediction task and thus which label-based similarities are informative for predicting an important user behavior. For example, are network-based similarities encoded in identity label representations meaningful above simply the co-occurrence patterns found in blog descriptions? Performance comparisons among baselines will allow us to formulate partial answer to these questions; the interpretations in Section 8.7 will illuminate more specific findings.

8.6 Community-based evaluation

We also evaluate what community information is being learned in these identity representations compared with network-based community detection methods. To do this, we compare with communities detected with the popular Louvain algorithm (Blondel et al., 2008) over a directed follow graph. A varation of this community detection, over a graph of both follows and reblogs, is done at the time of this proposal, and the results are confirmed to be a good approximation of known communities on Tumblr by community experts within the company. We do not include reblogs in this graph since we wish to compare to our approach, which is supervised by reblog behavior.

To compare our identity representations with detected communities, we start with the most popular n identity labels (= 5, 10 or 20) from blog descriptions in each of a few selected large and medium-sized Louvain communities. We use the nearest neighbors to these identity labels in our constructed identity label embedding space to expand the list of labels. We then compare the blogs of users in the Louvain community with the blogs of users who use labels in the corresponding expanded identity label space. Do these correspond well? We will measure the overlap between these communities. If the blogs of users who use the expanded labels are all in one Louvain community, our representations will score highly in matching these Louvain communities.

As in the previous evaluation, we will compare with baselines that exclude certain embedding training steps (particularly the network training step). For reference, we will use a baseline of word embeddings pretrained on Google News corpora.

Another evaluation we will consider is clustering identity label terms in our embedding space, and then grouping users that use those terms together. We then can compare the clusters of users based on our identity label embeddings with the clusters found by the Louvain community detection and use extrinsic clustering evaluation measures such as BCubed or V-measure.

In qualitative analysis, we are particularly interested in times when users have similar identity labels according to our learned representations but whose Louvain community differs. Where do similarities according to identity labels differ from those learned by network structure? Similarly, what kinds of similarities between identity label representations are found in users in the same Louvain communities? Our identity label representations, trained with the reblog prediction objective, can be seen as a way of weighting labels in communities by their relevance to user behavior. For example, it could be the case that ‘aesthetic’, ‘photos’, ‘girl’, and ‘love’ are all popular identity labels with users in the “vintage” community (the largest Louvain community), but that ‘aesthetic’ and ‘love’ are particularly close in our embedding space, indicating that this identity dimension has particular merit in this community.

8.6.1 Results

Results from this evaluation will indicate if our learned identity representations are aligning with the identity information that sets communities apart from one another. If there is a high degree of overlap, this is evidence that identity labeling that indicates a shared community may be associated with content propagation. Furthermore, community-specific weightings of identity terms may be found in our learned embedding space.

If there is not high overlap, then we must conclude that content propagation is associated with self-presentation that does not vary uniformly across communities. Our learned representations for identity labels would thus not be good resources for finding community-specific meanings of identity labels. Either way, we will investigate these findings with interpretation methods described in the next section.

8.7 Interpretation

The goal of this work is not simply to build effective representations for the reblog prediction evaluation, but to interpret what similarities and differences between identity labels are relevant for content propagation. To interpret these representations, we use two approaches: 1) interpretation of learned attention weights in follower/followee comparison, and 2) investigation of the learned identity label embedding space.

8.7.1 Learned attention weights

To investigate which identity labels the model has learned to attend to in its predictions, we examine the learned weighted averages (attention) used in building the user representations (“followee” and “follower” in Figure 8.2). For each instance, we can visualize which identity terms were weighted most highly and thus which were most attended to in reblog prediction.

8.7.2 Embedding space investigation

We also visualize similarities learned between identity labels by reducing the dimensions of our embeddings with t-SNE (van der Maaten and Hinton, 2008), a standard method of visualizing embedding spaces. From this visualization, we can see which labels are grouped together. We look for labels whose embeddings are unexpectedly close considering weak character overlap, and those with large character overlap but larger distance in embedding space. This investigation will give insight into what information is encoded in our learned embedding space. If we think then in terms of abstracted levels of labels, the “categories” of Chapter 4, what “categories” or groups of identity labels does our learned embedding space produce? Along with t-SNE, various clustering techniques may be considered.

8.8 Conclusion

In this project, we attempt to induce vector representations for identity labels that take into account similarities and differences relevant to user behavior, especially content propagation. To do so, we train neural embeddings for self-presented identity labels with multiple objectives, including appearing in similar textual and network contexts. We evaluate these embeddings with a machine learning task predicting content propagation and compare with the feature-rich baseline presented in Chapter 4. Interpreting the learned embedding space of identity labels, we will discuss findings about axes of similarity and difference our model found particularly relevant to identity labeling in specific communities on Tumblr. Generalizing away from the specific context of Tumblr, we demonstrate learning flexible, relational representations for markers of identity with training objectives that incorporate indications of community values. In our case this was through supervision with what user behavior that is particularly relevant to the platform. In other contexts, one could imagine applying similar approaches with supervision of which posts are liked or which hashtags are selected, which also reflect community-specific meanings and values.

8.9 Future Work

Hashtag use is also considered central to community definitions on Tumblr. This framework could be expanded to incorporate hashtag use as a signal for identity positioning. For example, doc2vec (Le and Mikolov, 2014) could be used to coax representations for identity labels into similar spaces depending on the hashtags used by users who give those labels. Hashtags could also be treated as the community symbols which users trade in frequent interaction to define communities; representations for these could be learned jointly with identity labels and even blog representations based on network structure.

Chapter 9

Proposed Timeline

- **Chapter 4, Tumblr self-presentation and interaction**
  February: Make small modifications in manuscripts, submit to WebSci 2020.

- **Chapters 5 and 6, Fanfiction pipeline and character portrayal**
  January: Do pipeline evaluations.
  March: Revise manuscript, submit to COLING 2020 or TACL.

- **Chapter 7, Fanfiction gender roles**
  March-April: Work on fanfiction gender roles project.
  May: Prepare manuscript, submit to ICWSM 2021 or possibly a digital humanities journal such as the *Journal of Cultural Analytics*.

- **Chapter 8, Tumblr identity label representation project**
  January-April: Continue talking to Tumblr about using data collected under data agreement with Yahoo Research for the proposed project, as well as for negotiating a new data agreement.
  May-July: Work on Tumblr identity label representation project, submit to TACL or ICWSM 2021.

- **Write thesis**
  July-August.

When chapters will be ready to review:

- **Chapter 5, with pipeline evaluations from Chapter 6 included**: March 2020.
- **Chapter 7**: May 2020.
- **Chapter 8**: July 2020.

A chart of this timeline can be seen in Figure 9.1.

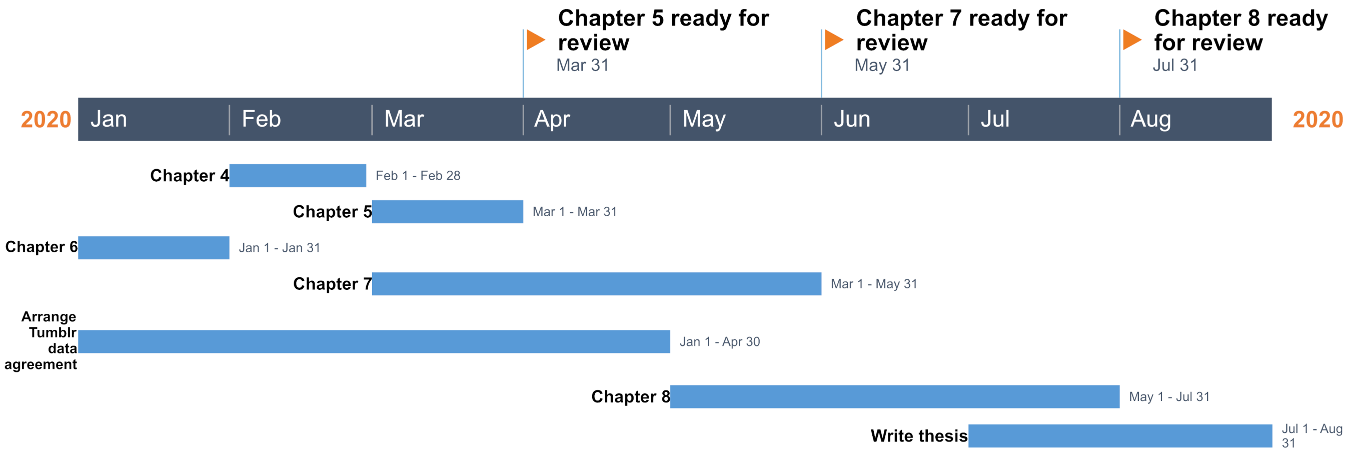


Figure 9.1: GANTT chart of proposed work.

Chapter 10

Conclusion

In this thesis, I attempt to bring a perspective of identity emerging in discourse and interaction into computational sociolinguistics. Doing so shifts focus away from building models that attempt to discover latent, personal identity attributes from linguistic and other social behaviors (user attribute inference, described in Chapter 2). In its place, I have built computational models that explore how social media users and writers of narrative position themselves and others with explicit identity labels and with implicit associations with those given those terms.

Importantly, this work takes into account users' *choice* of providing self-presentation labels (Chapter 4), which is otherwise often treated as missing data. This work also takes into account social media 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 our specific platform of interest, while in Chapter 8 we use a multitask machine learning approach to learn relational representations for these identity labels. In learning these representations, we evaluate how well machine learning tasks that predict relevant user interaction can be used to capture community interpretations of identity labels, the "economy" in which identity labels can act as symbols with particular value in positioning users.

Similarly, we demonstrate how machine learning approaches can be used to measure changes in representations of characters with particular identities in narrative. Here again, we treat identity as a flexible concept. Specifically, we examine the ways in which associations with straight and queer gender and sexual identities can be portrayed in ways that differ or align with mainstream societal narratives. We demonstrate this approach on fanfiction, well-known from qualitative work as a "queer space" in which writers transform the gendered and sexed behavior of characters in myriad ways. We first develop methods for the extraction of text segments and features about characters and relationships in fanfiction and test the ability of a variety of word embedding approaches to capture basic changes in how fanfiction stories change the representation of relationships from original media (Chapter 5). We then use unsupervised machine learning techniques to learn typical roles played by characters and compare these roles to stereotypically gendered terms (Chapter 7). We will compare the associations over these roles for gendered characters in both fanfiction and original series, as well as shifts in portrayal of gendered characters in queer and straight relationships.

Statistical and computational approaches to identity have often been used, and continue to be used, for the social classification and control (Chapter 2). Such methods often rely on discrete categorization, and designers frequently encode notions of identity that focus on popular and mainstream expressions and binaries, erasing or marginalizing the experiences of people who do not fit these mainstream categories. This does not have to be the case. Contemporary machine learning techniques often allow flexible, relational learned representations for the expression and construction of identity in language. Embedding-based word representations (Chapter 5) and representations for graph structure (Chapter 8) are examples of these. Furthermore, such techniques do not necessarily entail binary or categorical thinking about identity; these notions are supplied by the designers of such systems. We attempt to take a different approach to identity: one informed by qualitative approaches that note identity's active *construction* in language. In this thesis we hope to have demonstrated the applicability of computational models in studying how people position the identities of themselves and others in large datasets of linguistic and social interaction.

Bibliography

Graham Allen. 2011. *Intertextuality*. Routledge. 5.2.2

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. 5.2.1

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

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning Latent Personas of Film Characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. pages 352–361. 7.1, 7.2.1

David Bamman, Ted Underwood, and Noah A. Smith. 2014. 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, 5.3.1, 6, 6.1.1

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. 2.1.1

Roland Barthes. 1977. *Roland Barthes by Roland Barthes*. Hill and Wang. 7.1

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022. 5.1, 5.3.2, 7.2.1

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. 8.6

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. 5.3.2, 5.3.2, 8.4.1

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. 4.5.5

Mary Bucholtz and Kira Hall. 2005. Identity and Interaction: A Sociocultural Linguistic Approach. *Discourse Studies* 7(4-5):585–614. 1.1, 2.1.1, 4.2.2, 4.3.1, 4.3.1, 8.1

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. 1.1

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. 2.1.1

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge. 2.1.1, 7.1, 1

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, 4.2.1

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*. 2.1.1

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

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. 5.3.1, 6.1, 6.1.1

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings* pages 2256–2262. https://doi.org/10.18653/v1/d16-1245. 6.1.1

Nikolas Coupland. 2007. Sociolinguistic Resources for Styling; Styling Social Identities. In *Style: Language Variation and Identity*, Cambridge University Press. 1, 1.1, 2.1.1, 4.2.2

Kimberle 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. 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*. 4.1

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. 3.2.2, 7.1

Penelope Eckert. 2000. The Social Order of Belten High. In *Language Variation as Social Practice: The Linguistic Construction of Identity in Belten High*, Wiley. 2.1.1, 4.2.2

David K Elson and Kathleen R McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*. 5.3.1, 6.2

Norman Fairclough. 2003. *Analysing discourse: Textual analysis for social research*. Psychology Press. 5.2.2

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, 7.1, 7.2.1, 2, 7.5.2

Angela Fazekas. 2014. *Queer and Unusual Space: White Supremacy in Slash Fanfiction*. Master's thesis, Queen's University. 3.2.2, 5.4.1

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories. In *Proceedings of the Fourteenth Conference on Web and Social Media*. 5.2.1

Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4(1):1–14. 4.5.5

Marty Fink and Quinn Miller. 2014. Trans Media Moments. *Television & New Media* 15(7):611–626. 3.1.3, 4.1

Mary Fong. 2004. Identity and the Speech Community. In Mary Fong and Rueyling Chuang, editors, *Communicating Ethnic and Cultural Identity*, Rowman & Littlefield, chapter 1, pages 3–18. 8.1, 8.2.2

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences* 115(16):E3635–E3644. 5.1, 5.2.1

James Paul Gee. 2011. *An Introduction to Discourse Analysis: Theory and Method*. Routledge. 4.2.2

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday. 2.1.1

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. 4.2.2

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

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. 2017. Exploring the role of gender in 19th century fiction through the lens of word embeddings. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*. Springer International Publishing, Cham, pages 358–364. 5.1

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

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks .

8.4.2

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

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)*. volume 1, pages 1312–1320. 5.3.1, 6.2.3, 6.2, 6.2.1

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23(1):33–64. 5.3.1, 6.3

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, Thirteenth Conference of the European Chapter of the Association for Computational Linguistics*. April, pages 45–54. 1

Ryan James Heuser. 2017. Word Vectors in the Eighteenth Century. *Digital Humanities* . 5.1

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. 4.1, 4.2.2

Michael Holquist. 2003. *Dialogism: Bakhtin and his world*. Routledge. 5.2.2

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 7.2.1

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

Henry Jenkins. 2003. *Textual Poachers: Television Fans and Participatory Culture*. Studies in Culture and Communication. Taylor & Francis. 3.2.2

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. 4.1, 4.4.1

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of KDD*. pages 217–226. 5.3.1

Barbara Johnstone. 2010. Locating Language in Identity. In *Language and Identities*, Edinburgh University Press. 4.2.2

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

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, Wincing Leia, and a Seriously Great Friendship: Learning to Classify Emotional Relationships of Fictional Characters (2). 3.2.3

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the Relationship between Literary Genres and Emotional Plot Development pages 17–26. 3.2.3

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

Julia Kristeva. 1986. Word, dialogue and novel. In Toril Moi, editor, *The Kristeva Reader*, Basil Blackwell, Oxford. 5.1

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

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284. 5.1, 5.3.2

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*. pages 1188–1196. 8.9

Moontae Lee and David Mimno. 2014. Low-dimensional embeddings for interpretable anchor-based topic inference. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* pages 1319–1328. https://doi.org/10.3115/v1/d14-1138. 7.2.1

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. 5.4.3

Bill Yuchen Lin, Frank F Xu, Kenny Zhu, and Seung-won Hwang. 2018. Mining Cross-Cultural Differences and Similarities in Social Media. In *Proceedings of ACL*. pages 709–719. 5.2.1

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, 5.1, 7.1

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010. 5.3.1, 5.3.1, 6

Ann McClellan. 2014. Redefining genderswap fan fiction: A Sherlock case study. *Transformative Works & Cultures* 17. 3.2.2, 7.2.3

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119. https://doi.org/10.1162/jmlr.2003.3.4-5.951. 5.3.2, 8.4.2

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* . 5.3.2, 5.3.2

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, 5.1, 7.1

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

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers* 2:632–642. 6.1

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*. 4.2.1, 4.4.3

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. 5.1

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. 5.2.2

Abigail Oakley. 2016. Disturbing Hegemonic Discourse: Nonbinary Gender and Sexual Orientation Labeling on Tumblr. *Social Media + Society* 2(3):1–12. 2.1.1, 3.1.1, 3.1.2, 4.1

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. 7.2.1, 7.5.2

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. 5.3.2

Mario Piergallini, A Seza Doğruöz, Phani Gadde, David Adamson, and Carolyn Rose. 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. 4.2.2

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn A Walker. 2017. Modelling Protagonist Goals and Desires in First-Person Narrative. August, pages 360–369. 3.2.3

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

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. 7.2.1, 7.4

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. 7.2.1

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. 5.2.2

Yanchuan Sim, Bryan R Routledge, and Noah A Smith. 2016. Friends with Motives : Using Text to Infer Influence on SCOTUS. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1724–1733. 5.2.2

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

Catherine Tosenberger. 2008. Homosexuality at the Online Hogwarts: Harry Potter Slash Fanfiction. *Children's Literature* 36(1):185–207. 5.1, 5.4.1

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. 5.5, 7.2.1, 8.7.2

Jacky Visser, Rory Duthie, John Lawrence, and Chris Reed. 2018. Intertextual Correspondence for Integrating Corpora. *Conference on Language Resources and Evaluation (LREC)* pages 3511–3517. 5.2.2

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359(6380):1146–1151. 4.2.1

Sarah Schaefer Walton. 2018. The leaky canon: Constructing and policing heteronormativity in the Harry Potter fandom. *Participations: Journal of Audience & Reception Studies* 15(1):21. 7.1, 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. 2.1.1

Michael Worton and Judith Still. 1991. *Intertextuality: Theories and practices*. Manchester University Press. 5.1

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. 4.2.1, 4.4.3

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, 4.1

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

Diyi Yang, Miaomiao Wen, Abhimanu Kumar, Eric P. Xing, and Carolyn Penstein Rosé. 2014. Towards an integration of text and graph clustering methods as a lens for studying social interaction in MOOCs. *International Review of Research in Open and Distance Learning* 15(5):215–234. 8.2.1

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. 4.2.1

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. 4.2.1

Changtao Zhong, Hau Wen Chang, Dmytro Karamshuk, Dongwon Lee, and Nishanth Sastry. 2017. Wearing Many (Social) Hats: How Different Are Your Different Social Network Personae? In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. pages 397–406. 3.1.2