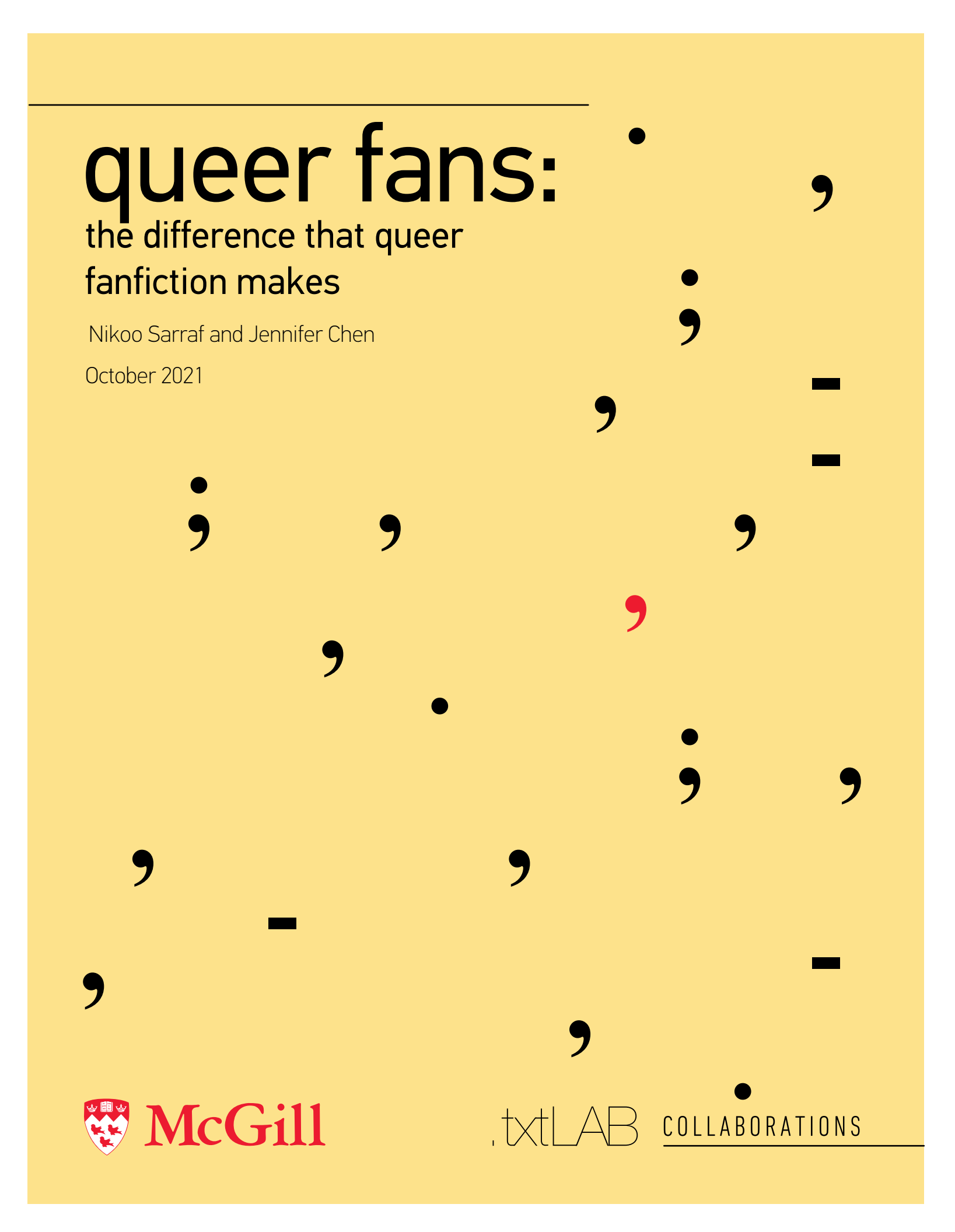
queer fans:

the difference that queer fanfiction makes

Nikoo Sarraf and Jennifer Chen

October 2021



McGill

.txtLAB COLLABORATIONS

Introduction

Fanfiction is a powerful and transformative form of writing that provides a safe space for youth to explore their gender and sexuality, knowing that they are part of a larger community. Recently, researchers have discovered that fanfiction is useful to sexual and gender minority youth (SGMYs) **by providing a sense of belonging,**[1] **helping young readers come to terms with their own sexuality,**[2] **and exposing them to a community where they are able to explore their identities.** Moreover, survey data has shown that fandom-participating youth are more active online daily and reach established identity milestones earlier than non-fandom participating peers, reporting sentiments such as, "Without fandom, I wouldn't have found who I was" and "By being exposed to this community I was able to figure out pretty early on who I was."[3]

Research on fanfiction has suggested that fandoms allow participating writers **to create queer information worlds that restructure heteronormative narratives that dominate mainstream publishing, reclaiming the power to define what is normative**[4]. According to Deborah Kaplan in "Construction of Fan Fiction Through Narrative": "fan analysis draws on textual, fannish, and extratextual interpretation, and contributes to the community's understanding of character...characters who were created and who exist outside the fan fiction texts therefore become available for complex play and re-creation."[5] Queer fanfiction provides a *community* where readers can form emotional connections to stories that they already feel familiar with and attached to.

Our aim in this collaboration is to better understand the stylistic qualities of queer fanfiction that might be contributing to these strong reader attachments. By comparing queer fanfic with books published in more 'mainstream' (i.e. corporate) venues, we want to surface some of the properties of fanfic that distinguish it from its corporate counterpart. We use data-driven methods in text analysis in order to provide empirical support for existing insights from the scholarly & popular literature on fanfiction's uniqueness. Accordingly, we analyze a collection of 13,827 fics posted to Archive of Our Own expressly tagged as queer in comparison to a collection of 426 novels reviewed in the *New York Times* over the past ten years. We divide our analysis into three realms of exploration:

1. Exploration of Taboo Topics

What potential effects does the greater emphasis on intimacy and sex have for sexual and gender minorities?

2. Character-Focused Storytelling

Is fanfiction more character-driven in comparison to the more plot-driven narrative style adopted by conventional literature?

3. Emotional Vulnerability

Do fanfic authors feel more comfortable exploring vulnerable topics? How does the range of emotion explored in fanfiction compare to that of mainstream books?

Summary

Overall, we found significant literary differences between fanfiction and mainstream fiction. Content-wise, fanfiction is much more open to exploring the topics of sex and intimacy. In fact, our analysis found that **a fanfiction narrative is about 6 times more likely to mention sex compared to a mainstream fiction narrative**. Additionally, the percentage of character-related vocabulary (e.g. "he," "she," "they", as well as proper names) was **9.9% in fanfiction but only 8.7% in the NYT dataset**, suggesting that fanfiction is noticeably more character-centric than mainstream fiction. Finally, we found that while events in mainstream fiction mimic those of everyday events in the real world, events in fanfiction hone in on introspective feelings. The prevalence of narrative events that encourage and celebrate expression of self indicate that **fanfiction normalizes intimacy and emotional vulnerability for sexual and gender minority youth**.

Data

Our research focuses on the comparison of two datasets, which will be described in this section. Please refer to the Appendix section for more detailed descriptions of the methodologies used to interpret the data. All of our Python scripts and data are available on GitHub.

Data Overview

The fanfic dataset (FanFic) consists of fanfictions posted to Archive Of Our Own (AO3), and includes 13,827 randomly selected texts that were equally drawn from the top 15 fandoms as of 2019 and explicitly tagged as queer. Texts were randomly selected from the entire dataset of over 750,000 fics and were restricted to having a length between 2,000 and 10,000 words and either an "M/M" or "F/F" tag in the metadata to indicate their explicit queer orientation. While we aimed to select 1,000 texts per fandom with an equal balance between M/M and F/F tags, in some cases we ended up with fewer because some fandoms did not have enough texts to meet these criteria. The second dataset (NYT) consists of 426 novels that were reviewed in the *New York Times* since 2002. The goal of this dataset is to approximate popular, mainstream literary fiction. Because the documents vary significantly in length, we employed data chunking methods to make them more commensurable (sizing the chunks in a similar distribution to the FanFic data). As Table 1 indicates, the overall word counts of the two data sets are reasonably similar.

Data Set	Texts	Average length	Total words
FanFic	13,827	3,574	49,413,594
NYT	426	100,399	42,769,877

Table 1: Summary of fanfiction and NYT source texts

Limitations

Before beginning our analysis, we want to acknowledge certain limitations to the retrieved data:

1. **Lack of Diversification in Fanfiction Dataset:** While hundreds of fandoms exist, our corpus only contains fanfictions from the top fifteen most popular fandoms. Certain patterns may exist within these 15 worlds that differ from other fandoms. Nevertheless, these fandoms capture the most popular fanfiction sources.
2. **Existence of Heterosexual Relations in Fanfictions:** Some of the fanfictions include a combination of M/M, F/F, and M/F tags, meaning not all the character relations were exclusively M/M or F/F, but at least one relationship has been tagged as such by the work's author.
3. **Difficulty in Labeling Topics:** One form of analysis we employed was a topic model. Although certain generated word groups could be classified easily with intuition, it was more difficult to label groups of words that seemed to bear minimal connection with one another.

Exploration of Taboo Topics

Given the audience for queer fanfic, we hypothesize that fanfiction authors will be more open to using more "taboo" language, such as profanity or sex-related vocabulary. To test this theory, one form of analysis we employed was a topic model that clustered together various words belonging to the same topic, then found the likelihood of that topic occurring in FanFic or NYT texts. Both corpuses were split into 500-word text segments to account for the length differences between the average fanfiction and average NYT novel.

The aim of the topic model is to investigate disproportionate interest in subject matter in one or the other dataset. We found that it was immediately noticeable that fanfiction writers are much more comfortable exploring certain topics which tend to be "off-limits" to mainstream fiction authors.

Fanfiction writers are much more comfortable exploring topics which tend to be "off-limits" to mainstream fiction authors.

Topic Frequency by Dataset

After comparing topic models with 20, 40, 60, and 80 topics, we selected the 40-topic model because it allowed for the most extensive coverage of subjects discussed in the corpus, without there being considerable overlap of topics. In order to get the topic frequency, we stored the topic distribution for all documents in our dataset and ran 1000 bootstrap samples per data set to get an average topic frequency for both the fanfiction and mainstream fiction datasets.

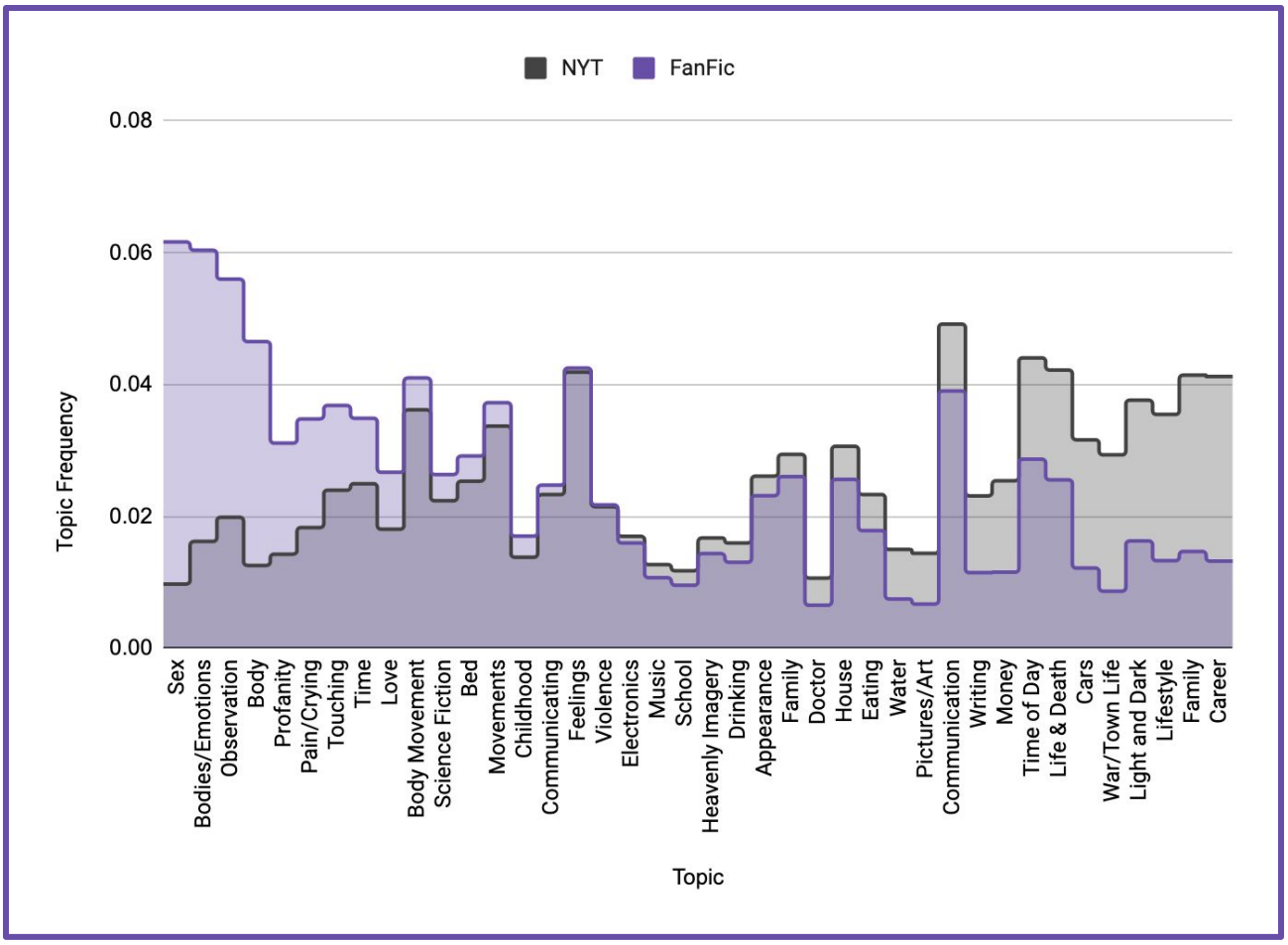


Figure 1: Topic frequencies sorted by most disproportionate use of topics per dataset.

As seen in Figure 1 above, the topics most disproportionately favoured by fanfiction are about sex, emotions and observation, while the topics more prevalent in mainstream fiction concern family and career. For example, FanFic is about 6 times more likely to write about sex, and 3 times more likely to write about intimate emotions than mainstream fiction. Other intimate topics such as observation, pain/crying, and the body are also far more frequently mentioned in the fanfiction corpus. **Specifically, there seems to be a heavy emphasis on intimacy and body language in the fanfictions that is absent from the mainstream fiction dataset, creating a space where queerness is normalized and expansively explored.** It is important to point out that such intimacy is not only based in sexual contact but crucially encompasses a broad spectrum of emotions as well.

A fanfiction text is about 6 times more likely to write about sex, and 3 times more likely to write about intimate emotions than a mainstream text.

We also noted that published mainstream fiction is distinguished by its attention to more public-facing activities (e.g. school, work, family). This attention to **social spaces** in mainstream fiction has the effect of maintaining social “taboos” around forms of human intimacy, a point worth further exploration.

Character-Focused Storytelling

Connoisseurs of fanfiction often emphasize that fanfiction is a character-driven medium.[6] Fanfiction is premised on an engagement with existing characters and redeploying them in novel narrative settings. A survey conducted in the Harry Potter fanfiction community shows that in the hierarchy of textual elements in fanfictions, characters were placed first, followed by world-building and plot.[7] Thus, we hypothesize that fanfiction would emphasize character relationships, resulting in stories that are more character-driven rather than plot-driven.

Character Centrism by Character Mentions

In order to test this theory, we used David Bamman's BookNLP library to obtain tables that provide data about every word in every corpus. To find character centric words, we filtered on two factors. First, we collected all words with ner tag 'PERSON.' Additionally, we collected any entry that had 'she', 'he', or 'they' in the lemma column. This would ensure that we collected all third person pronouns without including various forms of 'it' that would refer to an object rather than a character. Then, we calculated the two ratios by dividing character-centric words by the total number of non-punctuation tokens.

	Fanfiction	**NYT**
Percentage of Character-Centric Words	0.099137	0.086944
Standard Deviation	0.0391	0.0388

Table 2: Percentage of Character-Centric Words

Doing so, we found that LGBTQ+ fanfiction is more character-centric than conventional literature. According to our tests, about 8.7% of the words in NYT books are character-centric, and about 9.9% of the words in fanfic are character-centric. Although this may seem inconsequential, this equates to about 1,200 more character mentions in a novel-length work.

Character Centrism by Narrative Events

To further test the hypothesis that fanfiction is more character-centric, we examined the nature of "narrative events" in our data. For our purposes, we define an event as an action undertaken by an agent. Using David Bamman's BookNLP tool, we identify bigrams of the form ('Person,' 'verb') and trigrams of the form ('Person,' 'verb,' 'object'), where connections are derived by their dependency relationships. If agents of actions are people, then we resolve them to the single superordinate category 'Person.' This may include pronouns or proper names. We also categorize events by their BookNLP "supersense," which aggregates verbs and objects into broader categories, such as "communication" or "artifact." We then use a log-likelihood ratio test (G-test) to calculate the events that are significantly more likely to appear in one data set relative to the other. Figures 2 and 3 present the events that are most distinctive of each corpus.

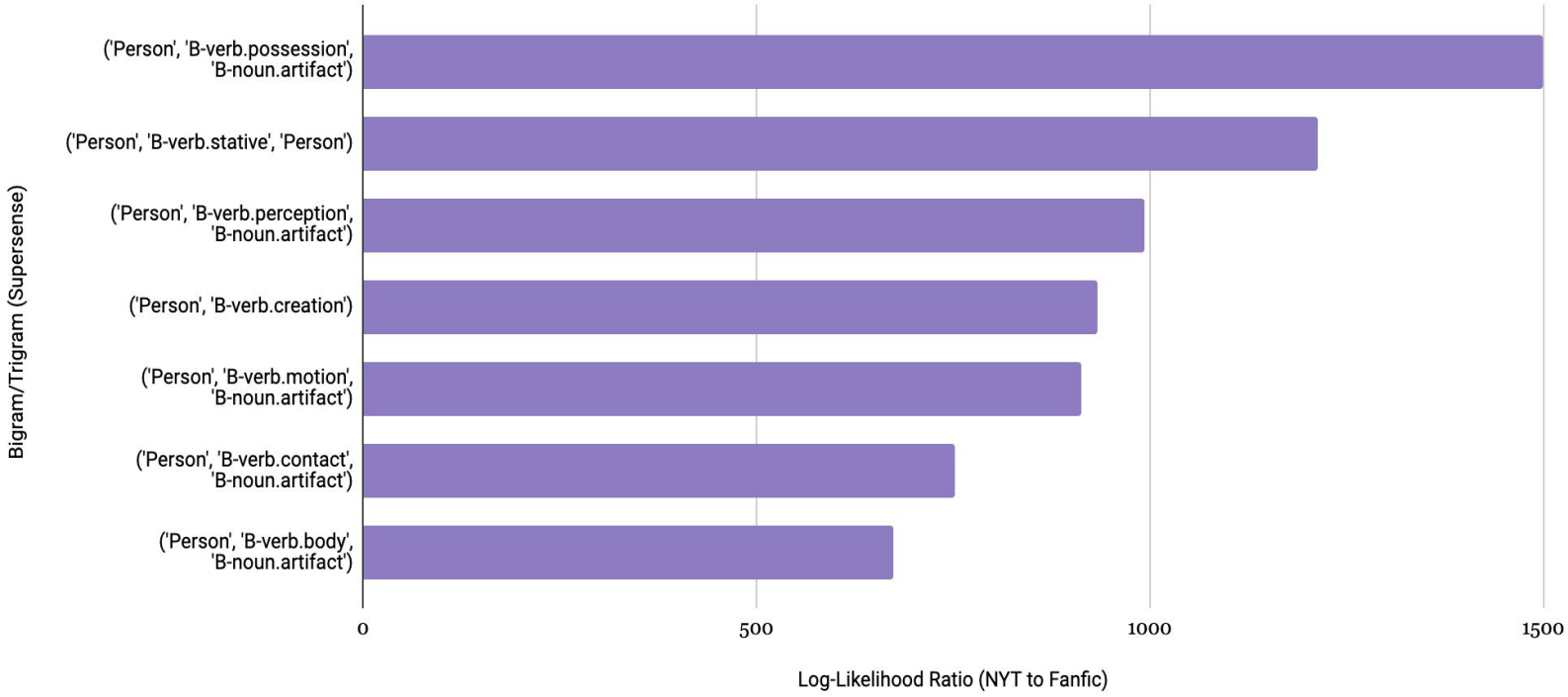


Figure 2: Supersense Events Most Skewed Towards NYT Dataset

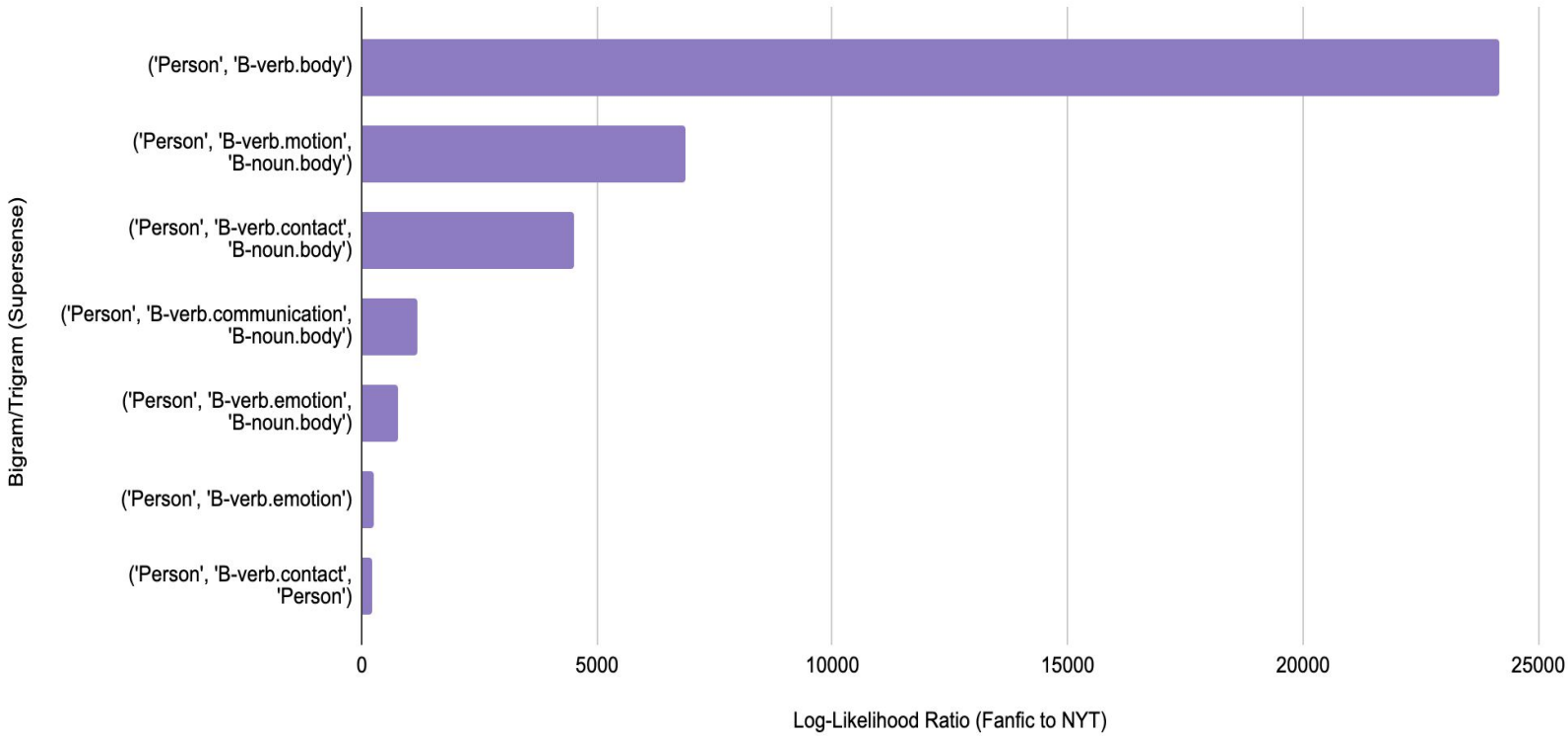


Figure 3: Supersense Events Most Skewed Towards Fanfiction Dataset

Bodies, Not Objects. We find that many of the NYT-skewed events are actions performed on objects, as indicated by the 'B-noun.artifact', while an abundance of fanfic-skewed events relate to the body ('Person', 'B-verb.contact', 'Person'). One way to interpret this data is to suggest that mainstream fiction centres more strongly around 'plotlines' (i.e. events in which characters are interacting with their immediate environments), while queer fanfic tends to revolve more around *interpersonal* experiences. Since fanfictions take place in worlds that have already been built by authors, there is less focus on world-building and more focus on character-building. What makes fanfiction unique is that it takes characters that readers already feel some degree of connection to, then adds more depth and variation to these pre-existing relationships to connect more deeply with readers seeking out less visible forms of human connection.

Emotional Vulnerability

The final style difference we investigate is the frequency of discussion of emotional vulnerability within the two datasets. Our preliminary research suggests that the connections formed between SGMYs and their fanfiction worlds are extremely personal and strengthened by feelings of belonging and acceptance. We seek to discover if sexual and gender minority youth are drawn to fanfiction because the subject matter is distinctively designed to connect with readers on a deeper, emotional level.

Events Analysis

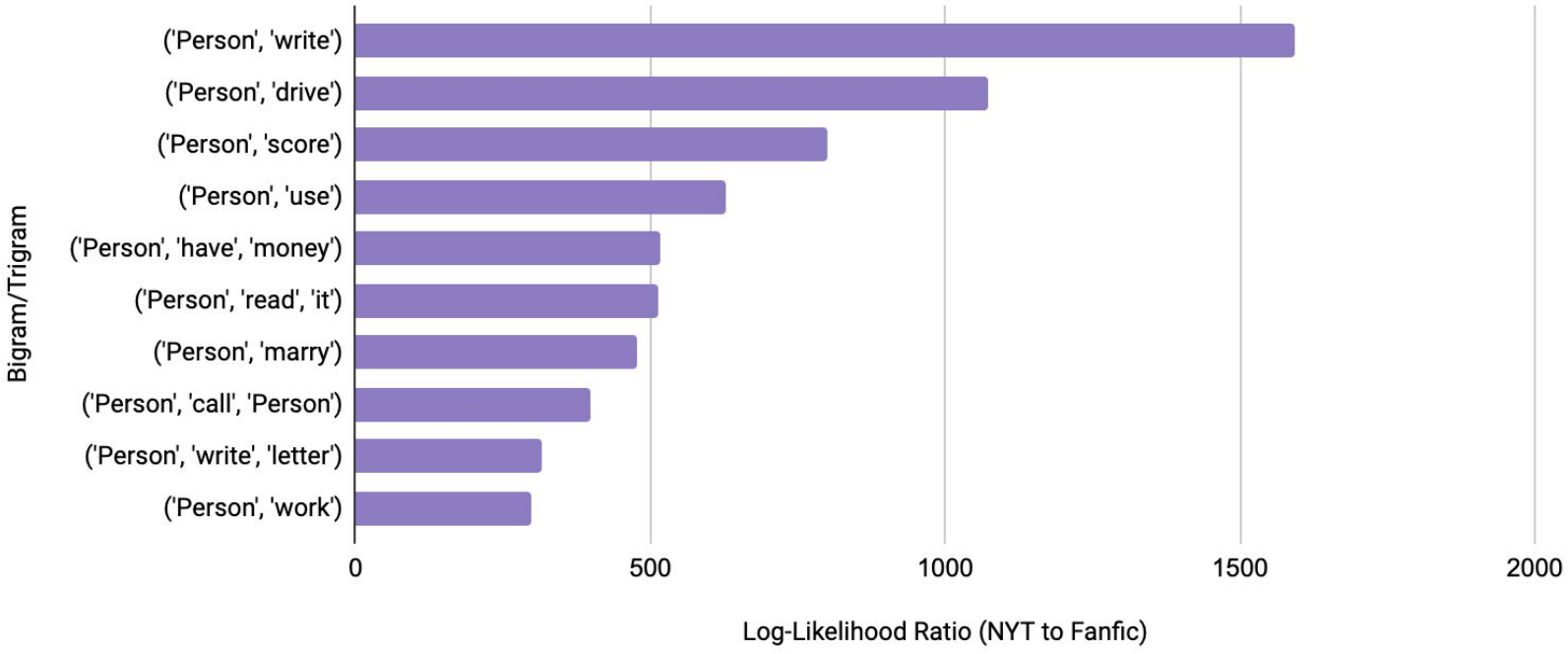


Figure 4: Events most skewed towards NYT data set

As illustrated in Figure 4, the narrative events commonly occurring in mainstream fiction but not fanfiction are absent of words that suggest emotional vulnerability, instead describing quotidian events commonly performed in everyday activities. Several are also career/family-oriented, such as working, marrying, and having money.

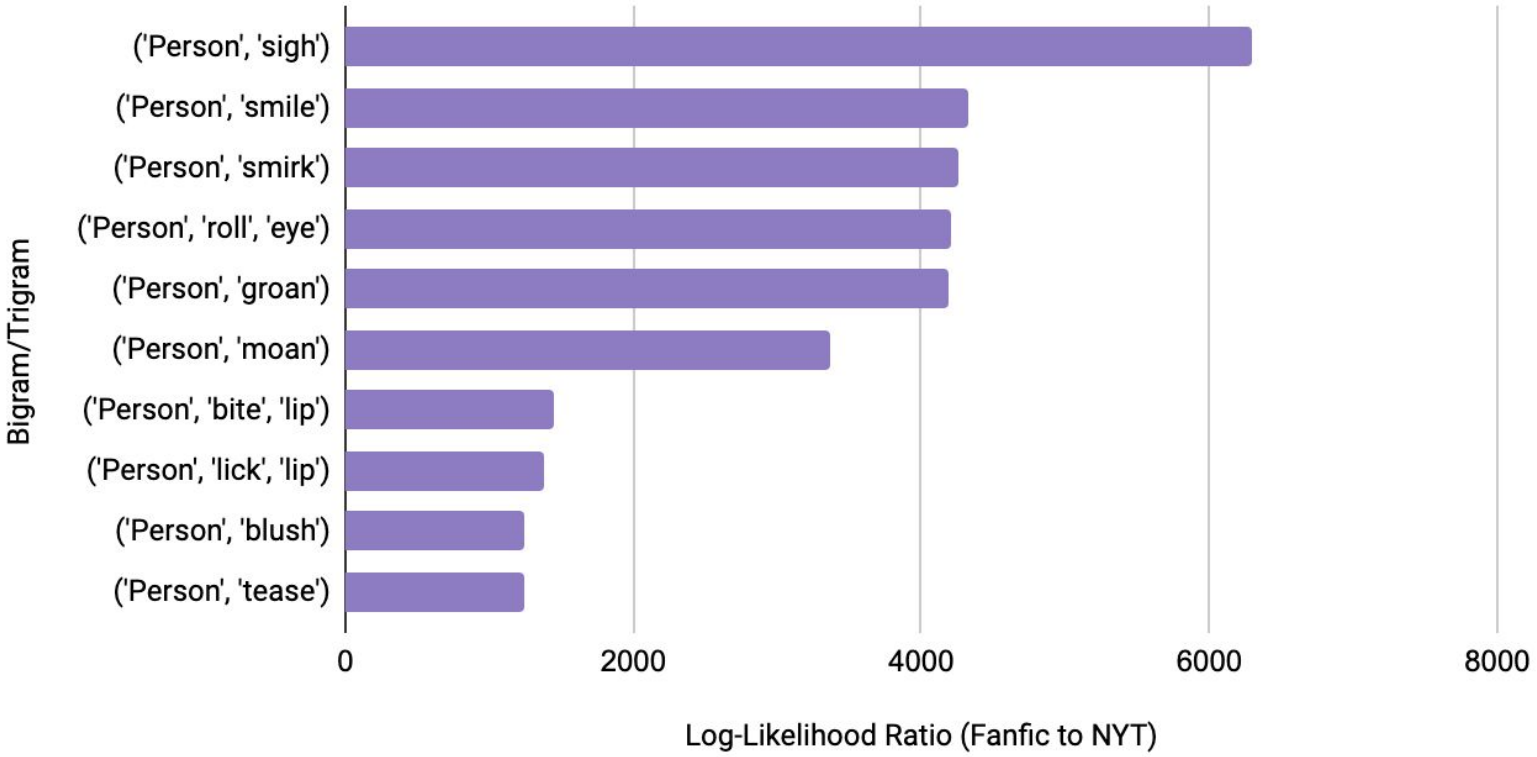


Figure 5: Events most skewed towards fanfic data set

While events in mainstream fiction mimic those of real life, events in fanfiction hone in on introspective feelings. The supersense bigram most heavily skewed towards fanfiction is ('Person', 'B-verb.body'), which would encompass events seen in Figure 5 such as ('Person', 'sigh'), ('Person', 'smile'), ('Person', 'smirk'), ('Person', 'groan'). These actions all communicate emotion from one character to another. Furthermore, some of the fanfiction-skewed events have sexual undertones, such as ('Person', 'lick', 'lip') and ('Person', 'moan'). This conveyance of intimacy is another demonstration of vulnerability.

While events in mainstream fiction mimic those of everyday life, events in fanfiction hone in on interpersonal feelings.

Word Frequency Analysis

In a sense, fanfiction explores the topics that are absent from other forms of contemporary culture, as it is a space where the discussion of sexuality is desired rather than stigmatized. The prevalence of narrative events that encourage and celebrate expression of self illustrates the manner in which fanfictions normalize intimacy and emotional vulnerability for SGMYs. This is also demonstrated through the patterns that emerge in the words from the topic model for the two datasets:

Category	Words
More Prevalent in Fanfic	body, feel, mouth, finger, moan, lip, sigh, expression, kiss, soft, face, smile...
Equal Prevalence	blood, kill, speak, question, thing, feel, happen, dance, play, music, day...
More Prevalent in NYT	work, mother, money, office, family, father, business, lawyer, police, soldier...

Table 3: Sample Words from Topic Model Classified Based On Prevalence in Datasets

Mainstream fiction indicates an interest in public-facing social institutions such as family, business, and careers (e.g. "lawyer", "police," "soldier"). Fanfiction draws greater attention instead to issues of personal interaction and feeling. Rather than reinforce public norms, its aim is to provide alternative spaces of feeling that are not well represented in mainstream fiction.

Our data suggest that fanfiction writers are creating a separate world for the LGBTQ+ community. The fanfiction medium itself is unique in the sense that it often builds off of worlds that are already fictitious, such as Harry Potter or Naruto. The fanfictions thus further explore topics that are absent from traditional media, such as more intimate forms of gay romance, or interactions where characters open up about their insecurities and vulnerabilities. Thus, rather than attempting to inject queernormativity into the existing world, fanfiction creates distinctively unique worlds where activities such as family life, career, and school are far less prevalent. Instead fanfiction creates a community that validates and normalizes the sexual and emotionally sensitive experiences that are unique to SGMYs.

Conclusion

In this paper, we analyze some of the key literary differences between fanfiction and mainstream fiction that enable sexual and gender minority youth to emotionally connect to fanfiction. Our aim has been to complement and validate existing research using computational text analysis. We find that fanfiction represents a distinct "queer information world" as hypothesized by Diana Floegel, one that consists of the exploration of sexuality, intimacy, and vulnerability, strongly differentiating it from mainstream fiction[8].

Our analysis tries to highlight the stylistic features that might inform the strong identification with fandoms that social science survey research has indicated. Future work will want to explore these specific identificatory mechanisms further to better understand why it is that sexual and gender minority youth find fanfiction so valuable in terms of self-formation. Our hope with this project is to help break down the stigmatization of fanfiction that comes from literary systems of value predicated on hierarchical social norms and instead highlight the social value that creative work like fanfiction offers to the world.

Acknowledgements

We would like to thank Professor Andrew Piper for the many hours he put into guiding our learning process and helping us navigate the natural language processing resources that brought this remarkable project to life. We would also like to thank the other members of txtLAB for providing us with an extraordinary community of data scientists within McGill University.

Works Cited

1. Tague A. M., Reysen, S., & Plante, C. (2020). Belongingness as a mediator of the relationship between felt stigma and identification in fans. *The Journal of Social Psychology*, *160*(3), 324–331. https://doi.org/10.1080/00224545.2019.1667748
2. Fielding, D. M. (2020). Queernormativity: norms, values, and practices in social justice fandom. *Sexualities*, *23*(7), 1135–1154. https://doi.org/10.1177/1363460719884021
3. Lauren B. McInroy & Shelley L. Craig (2018). Online fandom, identity milestones, and self-identification of sexual/gender minority youth. *Journal of Lgbt Youth*, *15*(3), 179–196. https://doi.org/10.1080/19361653.2018.1459220
4. Kraicer, E., & Piper, A. (2019). Social characters, the hierarchy of gender in contemporary english-language fiction. *Journal of Cultural Analytics*, 1(1). https://doi.org/10.22148/16.032
5. Hellekson, K., & Busse, K. (2006). Construction of Fan Fiction Character Through Narrative. *Fan fiction and fan communities in the age of the Internet: new essays* (pp. 134–170). McFarland & Company.
6. Peeples, D., Yen, J., & Weigle, P. (2018). Geeks, fandoms, and social engagement. *Child and Adolescent Psychiatric Clinics*, *27*(2), 247-267. https://doi.org/10.1016/j.chc.2017.11.008
7. Samutina, N. (2016). Fan fiction as world-building: transformative reception in crossover writing. *Continuum*, *30*(4), 433–450.
8. Floegel, D. (2020). "Write the story you want to read": world-queering through slash fanfiction creation. *Journal of Documentation*, *76*(4), 785–805. https://doi.org/10.1108/JD-11-2019-0217

Appendix

In this section, we will detail some specifics of the methodology that we employed for our analyses.

Topic Modelling Methodology

The first method of analysis involved using topic modeling to search for recurring themes that emerge distinctively in one dataset but not the other. Several software libraries were employed in this process. To begin, Python's Natural Language Toolkit filtered out the stop words of little significance (e.g. "the," "and," "but," "or"). Then, Spacy was used to lemmatize the corpus; this ensured that different forms of a word (e.g. "organize" and "organizing") were grouped together under one common base form. We then used a dictionary of words to create a bag of words (BoW) corpus. Finally, MALLET was used to create multiple term-topic matrices with each column containing 20 words belonging to a given topic. The process was repeated three times to create four different matrices with 20, 40, 60, 80 topic columns. It was determined that 40 topics allowed for extensive coverage of various topics, without there being considerable overlap of topics.

Character Centrism Methodology

In this section we executed a frequency analysis of the rate of characters in fanfiction v. mainstream fiction. To begin, we used David Bamman's BookNLP library to obtain tokens tables that provide data about every word in every corpus. We then counted the total number of words by counting each line in the tokens file and subtracting the ones with 'deprel' value 'punct' (i.e. we removed marks of punctuation). We discovered that entries labeled 'punct' would always outnumber any other 'deprel' tags, so we simply listed out the quantity of each deprel tag from greatest to least and added them all up, excluding the first one (which would be punctuation). Then, to get the character centric words, we filtered on two factors. First, we collected all words with ner tag 'PERSON', which means that the entry is the name of a character. Additionally, we collected any entry that had 'she', 'he', or 'they' in the lemma column. This would ensure that we collected all third person pronouns without including various forms of 'it that would refer to an object rather than a character. Then, we calculated the two ratios by dividing character-centric words by total words.

Narrative Events Methodology

The aim in finding "narrative events" is to use the BookNLP tokens tables to identify the types of actions that most commonly occur in the two datasets. In our definition, an event is an action undertaken by an agent, potentially on an object. To collect and analyze our narrative events, we began by parsing through our bookNLP files and identifying the anchor verbs of each sentence. Anchor verbs were words that had both -1 for headTokenID and a verb POS tag (ie: VB, VBD, VBG, VBN, VBP, or VBZ). From there, we backtracked to find the subject associated with our anchor verb, a noun with deprel 'nsubj'. Then, we verified that the nsubj was a person by running it through a list of strings for a match, checking its ner tag for 'PERSON' and checking its supersense for 'noun.person'. Then, we collected the objects by forward tracking from the anchor verb to find a word with deprel 'dobj'. For both the subject and object, we verified that they were associated by checking that its headTokenID pointed to the ID of the anchor verb. We built our bigrams with (subject, verb) and our trigrams with (subject, verb, object). In order to begin our analysis, we replaced all subjects and person-based objects with the word 'Person'. To focus on the more common narrative events, we only kept the 1000 most frequently-parsed narrative events and removed the rest. With the remaining events, we built contingency tables using NumPy, and then found the log-likelihood of a given event being in a fanfiction corpus compared to a NYT corpus. Log-likelihood ratios were used rather than regular ratios to account for the fact that there were different numbers of occurrences for different narrative events, and events with more occurrences are more important for analysis. To aggregate verbs and objects into more general categories, we also grouped bigrams and trigrams by their supersense tags. For example, ('Person', 'say') turned into ('Person', 'B-verb.communication'), and ('Person', 'roll', 'eye') became ('Person', 'B-verb.motion', 'B-noun.body'). Finally, once log-likelihood values and their corresponding p-values were found using NumPy and SciPy, we sorted the events by their log-likelihood ratio values.