Universität Regensburg

Acquisition of a German Fan Fiction Corpus and Analysis in the Context of Gender Representation

Master's Thesis in Medieninformatik
at Institute for Information and Media, Language and Culture (I:IMSK)

Handed in by:	Jonathan Sasse
Address:	Am alten Schlachthof 20, 93055 Regensburg
E-Mail (University):	jonathan.sasse@stud.uni-regensburg.de
E-Mail (private):	jonathan.sasse@outlook.de
Student Number:	1872869
Primary Corrector:	Prof. Dr. Christian Wolff
Secondary Corrector:	Prof. Dr. Jürgen Reischer
Supervisor:	Thomas Schmidt
Current Semester:	9. Semester M.Sc. Medieninformatik
Date handed in:	Abgabetermin der Arbeit

Contents

List of Figures

List of Tables

In the course of this thesis, an extensive corpus of German fan fictions was acquired. Fan fictions are fan-made stories that use existing characters and plot elements from media such as literature, movies, or games and alter them as they see fit. Multiple sources for this kind of stories were evaluated and the most suitable ones, *FanFiktion.de* and *Archive of Our Own*, were selected. This corpus consists of 412,923 stories, their chapters and reviews, as well their respective authors and metadata. To analyze the corpus, we used the state-of-the-art named-entity recognition model *FLAIR* to extract story characters with their number of occurrences, trained an LSTM model using *TensorFlow* and *Keras* to predict the genders of these characters, and in the process counted all personal pronouns used. We can confirm previous findings regarding a dominance of female authors, male character portrayals and erotic narratives across all genres. The observed use of gender-specific personal pronouns further supports these claims. In addition, we find that younger authors feature more female characters and write less about otherwise overrepresentative all-male relationships.

1. Introduction

Scrolling through fan fiction archives, one can find a wide variety of stories, ranging from the most innocent to the most explicit. Archive warnings, age restrictions, or many stories that deviate from heteronormativity have the potential to scare off readers. For many, this is precisely the refuge where they can express their artistic abilities, their desires and inclinations in the form of stories, whether deviating from social norms or not. But fan fiction sites are not just a platform for sharing stories, it's a vibrant fan community.

While fan fiction is a place for many to flourish, it represents a great opportunity for researchers in the area of natural language processing (NLP), digital social science and digital humanities as well. There is no shortage of web-based, freely accessible large bodies of narrative texts with rich metadata. For researchers, this offers the potential for large-scale analysis utilizing the wide variety of writing styles, social science observations such as the evolution of social norms in the anonymity of the Internet, or the development and evaluation of new methods (Yoder et al., 2021; Liu et al., 2019; Muttenthaler et al., 2019; Vilares & Gómez-Rodríguez, 2019; Zhang et al., 2019). Consequently, there are already many studies in that field.

While most of these focus on English-language texts and corpora, in this paper we would like to examine specifically German fan fiction. The objective of this work is therefore to acquire a corpus of German fan fiction from suitable sources that is as comprehensive as possible. In addition to the stories, users and reviews, this should incorporate all available metadata. We then examine the elicited texts with regard to the representation of the characters' genders.

First, we will highlight previously conducted studies that have addressed a similar research question 3. Then, potential sources for fanfictional texts will be located and a suitable tool for scraping the data will be evaluated and implemented. After

the data acquisition is completed, the data is preprocessed by extracting appearing story characters and gender-specific personal pronouns with their frequencies. For determining the gender of characters, a suitable neural network model is trained. Finally, the results will be presented and discussed.

The entire source code including all web crawlers, programs and scripts, as well as all figures presented in this paper and beyond, and a detailed *README* can be found on *GitHub*[1]. Due to legal restrictions, the corpus is currently only available on request[2].

[1] https://github.com/Cele3x/fanfiction
[2] jonathan.sasse@ur.de

2. Objectives of this Thesis

Missing data for possible data analysis of fan-fictional texts for the German language in digital humanities has already been determined.

We therefore want to generate a comprehensive corpus of German fan fiction texts and accompanying metadata. Specifically, this corpus should contain all stories as well as their chapters, reviews, associated genres, and fandoms. It should also include the authors of stories and reviews alike, as well as their profiles. In order to subsequently analyze a distribution of gender portrayals, some preparatory steps must first be taken.

Usually, a story consists of a collection of characters. These must first be extracted by a suitable named-entity recognition model. Since the name of a character as such does not tell us anything about its gender, we need to train a prediction model. In addition to counting and predicting the character names for each story, we also want to count the number of personal pronouns that refer to a particular gender.

The information obtained in this way can then be used to study gender representation, taking into account suitable metadata.

Opportunities offered by fan fiction have already been addressed and will be explored in the following chapter based on previous research in the area.

3. Previous Research in the Field of Fan Fiction

Fan fiction is an extraordinary resource for a huge variety of freely available texts by as many authors. Especially in the field of natural language processing (NLP), fan fiction is a great source of material, which many publications take advantage of.

3.1. History of German Fan Fiction

Since this paper's research is about German fan fiction, it is worthwhile to look at it in terms of its history. Cuntz-Leng & Meintzinger (2015) state in their "A Brief history of fan fiction in Germany" that many fan activities, objects and phenomena in Germany are unique to German culture. In the early 19th century Karl-May (1842-1912) was (and still is) one of the most widely read, translated and adapted writers (Petzel & Wehnert, 2002), and the closest thing to a pop-cultural phenomenon. The growing fan base led to a growth of the German fan fiction community in general, but was abruptly halted by two world wars and the repressive politics of National Socialism. According to Odin (2008), Germans are still struggling with this trauma and loss of identity as of today. The introduction of the internet and the success of anime and manga (Malone, 2010) are said to have been responsible for the resurgence around the year 2000 when with *Animexx*[3] and *FanFiktion.de*[4] two fan fiction platforms were founded. While there were relatively few stories from the genre of anime and manga on Anglo-American fan fiction platforms like *FanFiction.Net*[5] (25.3%) and *Archive of Our Own*[6] (12%), there were significantly more on *FanFiktion.de* (29%) and *Animexx* (49.5%), which only reinforces the importance of this genre for German fan fiction culture. Table 1 from Cuntz-Leng and Mentzinger

[3]https://www.animexx.de/fanfiction/
[4]https://www.fanfiktion.de/
[5]https://www.fanfiction.net/
[6]https://archiveofourown.org/

Fandom	FanFiktion.de	Archive Of Our Own
Harry Potter	36,877 (12.2%)	69,072 (4.8%)
Naruto	26,404 (8.7%)	9,987 (0.7%)
Twilight	13,954 (4.6%)	4,397 (0.3%)
One Piece	8,781 (2.9%)	3,175 (0.2%)
One Direction	6,308 (2.1%)	33,217 (2.3%)
Yu-Gi-Oh!	4,522 (1.5%)	2,339 (0.16%)
Tokio Hotel	4,453 (1.46%)	725 (0.05%)
Supernatural	3,449 (1.14%)	91,848 (6.3%)
Sherlock Holmes	2,867 (1%)	72,637 (5%)
The Avengers	1,074 (0.35%)	53,888 (3.7%)
Doctor Who	677 (0.2%)	36,896 (2.5%)
Total	303,316	1,452,704

Table 1.: Top 7 fandoms on *FanFiktion.de* and top 5 on *Archive of Our Own*. Adapted from Cuntz-Leng & Meintzinger (2015, Table 1).

shows that German fan fiction was much less diverse than Anglo-American was. The *Harry Potter*, *Naruto* and *Twilight* fandoms on FF.de accounted for a very large proportion of German fanfiction in 2015, with over 25 percent of all published stories. Because each fan fiction community differs in its "sets of rules, the socialization and education of their members, and the popularity of certain characters, pairings, tropes, or genres", Cuntz-Leng and Mentzinger contend that generalizing assumptions about fan fiction communities is highly questionable. They also assumed that German fan fiction writers and readers are more likely to migrate to English-speaking areas of fan culture with their greater variety, larger readership, and less strict youth-protection regulations.

Just as important as the historical and cultural background of fan fiction is analyzing who writes fan fiction.

3.2. Analyzing the Authorship and Readership of Fan Fiction

Since this paper intends to analyze gender roles, and we assume that they are directly dependent on the author and the reader, it is of particularly high importance to find out how they define themselves.

In the media industry men are observed as the dominating gender in divisions

like television shows, news coverage, social media like, like *Twitter*[7] conversations, and text authors in general (Milli & Bamman, 2016; Bergstrom et al., 2012; Bretthauer et al., 2007; Jia et al., 2015; Garcia et al., 2014). The question now arises as to whether differences can be identified in the area of fan fiction.

Duggan (2020) provides a detailed analysis of people writing *Harry Potter* fan fiction on the online fan fiction repository *Archive Of Our Own* (*AO3*). Target of research were the contents of almost 2,000 story paratexts as well as their authors user profiles regarding age respectively life stage, gender, location and sexuality. This wide-scale qualitative study is based on the assumption that fan fiction is no longer the domain of female, white, straight, middle-class, adult and Anglophone people (Busse & Lothian, 2017; Hellekson & Busse, 2006; Scott, 2013; Lothian et al., 2007; Stanfill, 2011). This is due to the fact that fan fiction has become more mainstream (Barnes, 2015, 74) and more accessible to "all fans despite their ages, financial means, ethnicities, nationalities, locations, linguistic knowledge, sexualities and genders" (Duggan, 2022). Duggan categorizes genders in female, male, trans, genderqueer respectively genderfluid (transgender), non-binary and genderless. For being able to make statements regarding the text authors gender the pronouns used in story paratexts and profiles were examined (see Table 3). Despite the fact that the majority of fans on AO3 did not provide any extractable demographic information in their notes, the paper comes to the conclusion that "gender identity in the *Harry Potter* fandom is much more diverse than previously acknowledged" based on the data that could be obtained. While Table 2 shows that 50.39% stated themselves to be females, 74.02% preferred she/her(s) pronouns (see Table 3) why it can be inferred that the actually number lies somewhere between those values. With 36.22% non-cis individuals like genderless, non-binary and transgender outline a much larger group than previously anticipated. Though one should assume that heteronormativity is regarded as default and is therefore not being particularly emphasized by the authors. Lastly only 13.39% reported themselves to be male. There were no distinctions made between various types of genres and fandoms and rather viewed

[7] https://twitter.com

Gender	Frequency	Percentage
Female	64	50.39%
Non-binary	27	21.26%
Male	17	13.39%
Transgender	14	11.02%
Genderless	5	3.94%

Table 2.: Gender distribution in *AO3* paratexts. Adapted from Duggan (2020, Table 1).

Preferred pronoun(s)	Frequency	Percentage
She	94	74.02%
He	17	13.39%
They	6	4.72%
She/They	4	3.15%
She/He/They	2	1.57%
"Pronounless"	2	1.57%
He/They	1	0.79%
She/He	1	0.79%

Table 3.: Preferred pronoun distribution in *AO3* paratexts. Adapted from Duggan (2020, Table 2).

from a more general perspective.

From this it can be deduced that female individuals make up the larger proportion of readers and authors of fan fiction compared to male individuals. Assuming that the *Harry Potter* fandom can be generalized and mapped to the entire sector of fiction.

3.3. Stereotypes and Gender Bias in Fan Fiction

Based on the assumption that the majority of fan fiction is written by women, is it possible to draw conclusions about differences in stereotypes and gender bias?

In analyzing films, books and music lyrics, women tend to be portrayed as younger, domestic, rather emotional than rational, and with a greater focus on beauty than on their intellect (Towbin et al., 2004; Gooden & Gooden, 2001). In comparison, men are portrayed as physically stronger and more active, more violent, more economically successful, less in control of their sexuality, and emotionally cold (Emons et al., 2010; Bretthauer et al., 2007; Lauzen et al., 2008; Soulliere, 2006).

In a comprehensive study on a dataset of more than 1.8 billion words from the *Wattpad*[8] writing community Fast et al. (2016) investigated gender bias and stereotypical questions. High-level gender statistics provided by *Wattpad* on request could not confirm Duggan (2020)'s thesis regarding an overrepresentation of female authors (male/female ratio was 1.16 for 655,295 stories). To address the question of how the gender of the characters affects the actions, they examined the verbs that referred to each character. Both male stereotypes with common verbs like *thrush*, *abuse* or *roar*, and female stereotypes with verbs like *squeal*, *giggle* or *shriek* could be verified with this method. It was noteworthy, however, that the *angry man* stereotype failed to show up and verbs in this category tended to have darker connotations (like *beat*, *rip* or *snarl*).

Female	**Odds**	**Sample verbs and adjectives (female odds)**
weak	1.73	fragile (6.3), faint (3.2), sick (1.8), tired (1.4)
submissive	1.66	helpless (3.5), shy (2.9), timid (2.8), whimper (1.7)
childish	1.54	squeal (11.1), naive (7.8), giggle (4.9), silly (1.7)
afraid	1.46	shriek (4.8), frightened (2.3), shiver (1.8)
dependent	1.43	clingy (3.2), vulnerable (2.5), desperate (1.8)
hysterical	1.25	bitchy (11.4), dramatic (3.2), suicidal (3.1)
domestic	1.16	cook (2.3), wash (1.8), marry (1.7), clean (1.5)
emotional	1.04	meanest (7.9), gush (5.1), sob (37), fiery (2.8)
angry	1.05	bellow (3.1), growl (2.7), curse (1.4), snarl (1.3)

Table 4.: Dominant female stereotype categories. Adapted from Fast et al. (2016, Table2).

Male	**Odds**	**Sample verbs and adjectives (male odds)**
strong	2.02	intense (3.1), smash (2.6), intimidating (2.1)
arrogant	1.30	cocky (7.1), smirk (2.8), smug (2.6), rude (1.4)
sexual	1.22	sexiest (3.1), kiss (2.4), hot (2.1), flirt (1.5)
active	1.17	jog (2.5), lift (2.4), dodge (1.7), spin (1.4)
dominant	1.15	rich (28), protective (2.7), royal (2.0), command (1.4)
violent	1.10	abuse (4.4), hurt (2.3), beat (2.0), kill (1.5)
beautiful	1.06	dreamy (8.14), attractive (4.09), cute (3.3), hot (2.14)
angry	1.05	bellow (3.1), growl (2.7), curse (1.4), snarl (1.3)

Table 5.: Dominant male stereotype categories. Adapted from Fast et al. (2016, Table 2).

[8]https://www.wattpad.com

They further investigated which adjectives were used to describe characters and categorized these as well as depicted in 4 and 5. In general, stereotypical characteristics were confirmed here as well, but a relatively large number of male characters were described with the rather feminine adjective *beautiful*. As a third hypothesis, they explored the question of how stereotypes in stories affect their user ratings. They find that many common stereotypes, such as sexual, arrogant and violent men, have a positive impact on the ratings, while others, such as strong women or domestic men, have a negative impact (see 6).

Positive with rating	**Coefficient**	**Negative with rating**	**Coefficient**
sexual (male)	+2.03	strong (female)	-0.96
arrogant (male)	+1.45	domestic (male)	-0.66
sexual (female)	+1.24	afraid (female)	-0.66
violence (male)	+0.92	weak (male)	-0.63
active (male)	+0.90	domestic (female)	-0.57
hysterical (male)	+0.59	strong (male)	-0.51
hysterical (female)	+0.57	dominant (female)	-0.44
anger (male)	+0.56	emotional (female)	-0.44
violence (female)	+0.46	beautiful (female)	-0.39
childish (male)	+0.42	weak (female)	-0.34
angry (female)	+0.20	dependent (female)	-0.25
emotional (male)	+0.12	childish (female)	-0.21
submissive (female)	+0.02	active (female)	-0.03

Table 6.: Categories of stereotype with positive respectively negative effects on story ratings in a logistic regression. Adapted from Fast et al. (2016, Table 4).

Finally, they investigated whether female authors stereotyped men in their stories and vice versa. To this end, they trained a logistic regression model with frequency counts of words captured by their stereotype categories as feature inputs for male and female characters. The goal here was to predict the authors gender of a story. However, due to stereotypical features, the author could not be specified by the model and achieved only 53% accuracy, so the authors concluded that both men and women write indistinguishably stereotypical genders.

3.4. Possibilities for Natural Language Processing

One of the early works focusing on similar aspects of source of material and research topic as this paper is a proceedings paper by Milli & Bamman (2016). In their work they analyzed the possibilities provided by fan fictional texts for NLP, computational social science and the digital humanities in general. Large-scale literary data as well as a vibrant social network provide a huge opportunity for research in these areas. A large fan fiction corpus as depicted in Table 7 was collected from *FanFiction.Net* and served as study object.

Type	**Count**
Canons	9,246
Stories	5,983,038
Tokens	55,264,185,653
Reviews	159,914,877
Users	2,093,601
– Authors	1,364,729
– Reviewers	1,438,721
Languages	44

Table 7.: Summary of the *FanFiction.Net* corpus. Adapted from Milli & Bamman (2016, p. 2049).

They stated that fan fiction with its mainly female authorship (Duggan, 2020; Barnes, 2015) deprioritizes the main protagonists, which are mostly male, and consists in return of more and stronger female characters (Handley, 2012; Scodari & Felder, 2000; Leow, 2011; Busse, 2009). For extracting and comparing characters they utilized the Natural Language Processing pipeline *BookNLP* (Bamman et al., 2014). While freely available canonical texts from *Project Gutenberg*[9] serve as object of comparison, Milli & Bamman can proof their hypothesis regarding deprioritization with a statistical significance of $p < 0.001$ (40.1% female characters in canonical texts to 42.1% in fan-written texts).

As an example they presented in Figure 1 the difference of character appearances in texts about the *Sherlock Holmes* fandom. A secondary character as in *Watson* receives substantially more mentions in fan fiction than in the canonical text, while

[9] https://www.gutenberg.org

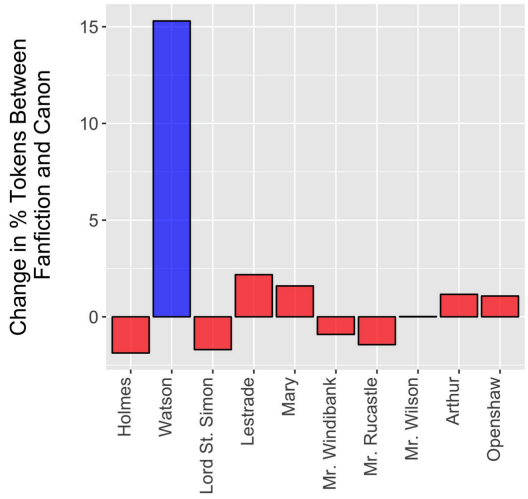

Figure 1: Difference in percent character mentions between fan fiction and canon for *Sherlock Holmes*. Reprinted from (Milli & Bamman, 2016, p. 2049).

the main character (*Sherlock Holmes*) loses them slightly.

Furthermore, they identified that more than half of the fan-authors (52%) are in return reviewers of other works. In an exploratory analysis they ran a Latent Dirichlet Allocation model (Blei et al., 2003) on the review texts and determined that most reviews are about positive encouragement, pleas for updates, requests to the author about the progression of the story and emotional reactions. For training a sentiment classifier predicting the users reactions in reviews they conducted a study asking the participants to judge the sentiment towards the character expressed in the response (positive, negative, neutral or not applicable).

While there are other works highlighting the possibilities that these large quantities of diversely written texts provide, they pursue other approaches that are impractical or beyond the scope of this research paper.

One of a few popular mentions is Liu et al. (2019) who fine-tuned a pre-trained BERT model for a multi-class emotion analysis using texts from previously mentioned *Project Gutenberg* and *Wattpad*, another platform intended for users to read and write stories.

Muttenthaler et al. (2019) developed and compared three n-gram models to identify authors of fan-fictional texts. They concluded that their standard n-gram model (2 - 5 gram) performed best and a combination of different text representations best

reflects the author's writing style.

The task of predicting upcoming actions from textual descriptions of scenes using *Harry Potter* fan fiction was examined by Vilares & Gómez-Rodríguez (2019). A model based on an LSTM, a recurrent neural network with feedback connections, performed best for frequent actions and extensive scene descriptions, while logistic regression performed well for infrequent actions.

By collecting a million stories and their summaries on *Wattpad*, Zhang et al. (2019) identified common components to describe the stories' characters. The objective of the work was to automatically generate these character summaries using inferring salient attributes. They developed two models of which one extracted and used attributes from the source story by ranking them, while the other classified abstractly using a list of attributes drawn from the entire corpus of stories which performed better.

Kleindienst & Schmidt (2020) took a more specific approach and studied only one fandom, the quite popular TV show *Supernatural* (see also 8). In doing so, they examined how the fan fiction community adapted to the source material over the time the show ran using, while using the show's scripts and 7,000 texts from the AO3 platform. They can confirm the thesis of Milli & Bamman (2016) concerning the overrepresentation of secondary characters and observed an overwhelming amount of male-male relationships in fan fiction (91.99%; Tosenberger (2008); Hellekson & Busse (2006); Duggan (2017)) compared to the source material.

Not just one specific fandom, but rather the combination of individual fandoms, the so-called crossovers, were investigated by Schmidt et al. (2022). They used a tool named *Gephi*[10] for a computational social network analysis (SNA) depicting unique character relationships visually. They found that original characters are important for crossovers, and that popular fandoms, stories of the same genre (e.g. Sci-Fi) and nationality (e.g. British as in *Sherlock Holmes* and *James Bond*) are often linked.

These are just a few examples that show the possibilities offered by the analysis and study of fan-fictional texts.

[10]https://gephi.org/

3.5. Analyzing German Fan Fiction

Because most fan fiction gets published in the english language and though an increasing number of german authors tend to switch to it as well (Cuntz-Leng & Meintzinger, 2015), research focuses primarily on english-written texts.

Cuntz-Leng & Meintzinger (2015) stated in their paper about the history of german fan fiction that style, content and progression of fan fiction differ on a cultural background. Which is why Schmidt et al. (2021) gathered a corpus of 9,640 german writings from the previously mentioned *AO3* platform for analysis. While metadata and general text statistics like the most frequent words were the subject of the investigation, they identified "attributes that are very specific and unique to German culture". Next to popular fan fiction fandoms like *Harry Potter* and *Supernatural*,

Fandom	**Frequency**	**Percentage**
Tatort	986	10.2%
Harry Potter	800	8.3%
Supernatural	413	4.3%
Sherlock (TV)	405	4.2%
Original Work	349	3.6%
Football RPF	295	3.1%
Stargate Atlantis	220	2.3%
Stargate SG	191	2.0%
Historical RPF	151	1.6%
Glee	141	1.5%
Rest (1,603 Fandoms)	5,830	58.9%

Table 8.: Top 10 fandoms of the *AO3* corpus. Adapted from Schmidt et al. (2021, p. 4).

Tatort, a german crime television show, real person fan fiction (*RPF*; e.g. *Football*) as well as stories about *Goethe* and *Schiller* (*Historical RPF*) were often used as source material as seen in 8. The assumption that *Anime* is strongly related to the rise of fan fiction in Germany (Cuntz-Leng & Meintzinger, 2015) could not be reflected in the AO3 corpus consisting only of 97 stories. It was also noticeable that all-male romances (56%), male characters in general (see also 3.4 respectively Kleindienst & Schmidt (2020)), and an excess of stories of an erotic nature dominated the corpus. The fascination that fan fiction has with same-sex relationships has already been

identified and analyzed in the humanities by Hellekson & Busse (2006), Tosenberger (2008) and Duggan (2017). Furthermore, after analyzing the most frequently occurring words in the corpus, they were able to determine that they are strikingly often words that describe physical attributes.

4. Data Acquisition

In previous research multilingual archives such as *Archive of Our Own* (Duggan, 2020; Cuntz-Leng & Meintzinger, 2015; Kleindienst & Schmidt, 2020; Schmidt et al., 2022), *Fanfiction.net* (Milli & Bamman, 2016; Vilares & Gómez-Rodríguez, 2019), *Wattpad* (Fast et al., 2016; Liu et al., 2019; Zhang et al., 2019) were the objects of study, while only Cuntz-Leng & Meintzinger (2015) gathered some statistical data from *FanFiktion.de*, a german fan fiction website. As can be seen, little to no research has been done in the field of German fanfiction, despite Cuntz-Leng & Meintzinger's (2015) emphasis on the importance and opportunities afforded by the study of culturally diverse texts. For this reason, we dedicated ourselves to the task of assembling and then analyzing a comprehensive corpus consisting exclusively of German fan fiction.

4.1. Sources Evaluation

First, we had to locate and evaluate all potential sources of fan fiction texts. We looked at the languages provided, the number of stories (at the time of collection), the metadata provided, and whether scraping was allowed and feasible under the terms of service of each source. Table 9 shows an excerpt of the evaluation results (see attachment for a comprehensive table).

Sources were examined according to their provided languages, number of German stories, structure, genres, story metadata, review existence, user information and finally whether scraping is allowed due to the terms of service (TOS). As seen in Table 9 *FanFiction.NET*, *wattpad* and *tumblr* were excluded from the list of possible candidates for our corpus. Although *FanFiction.NET* would have been an ideal candidate for inclusion in our corpus with its substantial amount of German text (estimated only), story and user information, and structure in general, scraping in

Name	Language/s	German Stories	Scraping Permitted	Usable
FanFiktion.de[11]	German	412.033	Yes	Yes
FanFiction.Net[12]	Multilingual	N/A*	No	No
AO3[13]	Multilingual	N/A*	Yes	Yes
wattpad[14]	Multilingual	1.600	No	No
tumblr[15]	Multilingual	N/A*	No	No

Table 9.: Results of the fan fiction source evaluation. (*not available; Acquisition date: January 8, 2022)

any form is prohibited. *wattpad* and *Tumblr* were poorly structured, could not be filtered by language or had too few stories in German, had little to no metadata on individual stories, and again, where not allowed to scrape. Both *Archive of Our Own* and *FanFiktion.de* offered a solid site structure, a large amount of German texts and rich metadata. While *FanFiktion.de* allowed scraping for non-commercial purposes, *Archive of Our Own* allowed it as long as it did not interfere or disrupt with their services.

Accordingly, *Archive of Our Own* and *FanFiktion.de* are getting used as source material for our corpus and will be referred to as *AO3* and *FF.de*, respectively, in the following paragraphs.

Now that the sources could be determined, the next step was to decide which tool to use to obtain the data from the websites.

4.2. Web Scraper

First, the difference between the terms *web crawling* and *web scraping* should be clarified. While *web crawling* is about finding and discovering URLs or links from websites, *web scraping* is about extracting specified data from websites (Kenny, 2022). Usually both procedures are needed: first to crawl the URLs of a website and second to scrape the content.

There is a wide range of tools for crawling and scraping the internet. Factors such as customizability, scalability, the type of input data to be processed, the desired output format, the crawling interval, and the amount of data transferred in

the process all play a role. Since we planned to scrape continuously and free of charge, Software as a Service (SaaS) platforms like *ScrapingBee*[16] or *Diffbot*[17] were not suitable. Desktop scraper applications such as *ScrapeBox*[18] were due to the lack of customizability impractical. In contrast to SaaS and desktop scraper applications, frameworks like *Scrapy*[19], *Beautiful Soup*[20], *pyspider*[21], *Goutte*[22], *Cheerio.js*[23] and *Puppeteer*[24] were all open source, highly adaptable and scalable, and therefore applicable to our purposes.

4.2.1. Scrapy

We have chosen *Scrapy* as scraping framework because of aspects like those requirements mentioned in the previous section. *Scrapy* uses the *Python*[25] language, is open source, platform-independent, actively maintained, well documented, highly scalable, and got in general lots of features that solve the most common web scraping problems. Furthermore, one can easily add extensions like proxies and fake user-agents for the obfuscation of queries, or auto-throttle capabilities for adjusting request delays based on the sites current download latency. Auto-throttle is a particularly valuable feature for not disrupting the sites services while also being able to automatically increase the work load at low usage times, as this was for example requested by *AO3* (see Section 4.1).

Figure 2 describes the architecture and work flow of the *Scrapy* framework. Each fan fiction archive, *FF.de* and *AO3*, differs in its page structure and data presentation. It requires an individual definition of where to find the corresponding data and how to search the respective archive. These definitions are implemented in so-called *Spiders*. The *Spider* sends a request with retrieval information via the *Engine* to the *Scheduler*, which stores these requests and forwards them to the *Downloader* with a

[16] https://www.scrapingbee.com/
[17] https://www.diffbot.com/
[18] https://www.scrapebox.com
[19] https://scrapy.org/
[20] https://www.crummy.com/software/BeautifulSoup/
[21] http://docs.pyspider.org/en/latest/
[22] https://github.com/FriendsOfPhp/Goutte
[23] https://cheerio.js.org/
[24] https://pptr.dev/
[25] https://www.python.org/

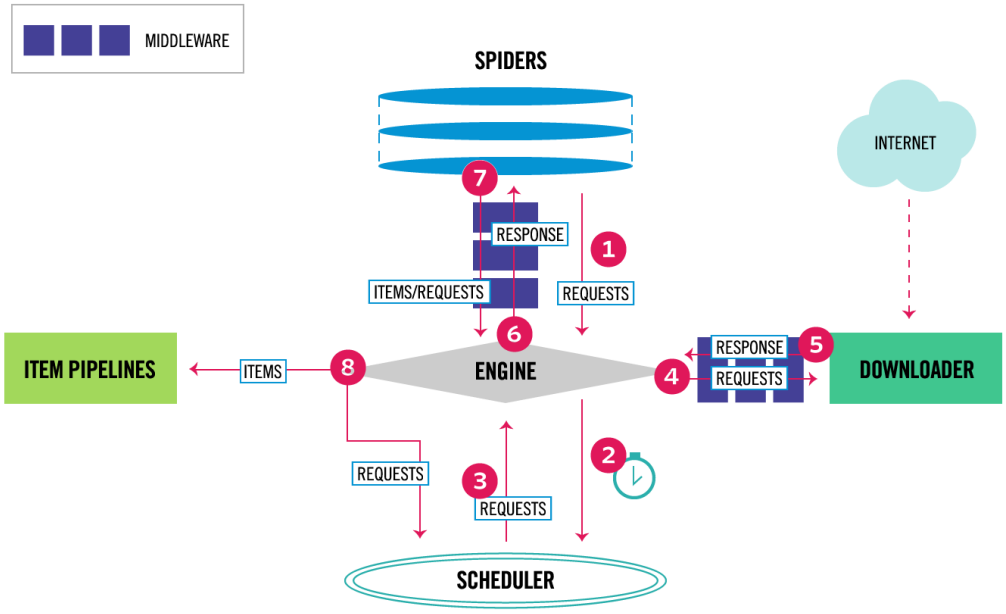

Figure 2: The architecture of the *Scrapy* framework. Reprinted from Architecture overview, In Scrapy, n.d., Retrieved October 29, 2022, from https://docs.scrapy.org/en/latest/topics/architecture.html. Copyright 2008–2022 by Scrapy developers.

defined delay (fixed or by using auto-throttle). The *Downloader* returns the found *Items* to the *Item Pipeline* if successful, where some previously programmed cleanup operations are performed and the objects are stored in a database.

The crawl process runs in a way that the *Spider* seeks and follows a URL for the next page at defined positions. To filter on a language, headers with appropriate filters must be passed to the request.

4.2.2. Concurrent Crawling

For time efficiency, we developed and ran two *Spiders*, each with its own public IP, to operate on different hosts at the same time. Each host was assigned a different fan fiction genre to crawl so that a web page did not have to be processed multiple times. Since the database for storing the scraped data was only local, we needed a different approach for the remote host.

While the local host's *Spider* stored the data directly in the database, the remote

host downloaded the full HTML page, compressed it into an archive (in stacks of 1000 files), and kept a CSV list of the archive's contents. As per previous definition in Section 4.2, the local host crawled and subsequently scraped the data, while the remote host merely crawled the URLs and downloaded the web page.

The CSV manifest in each archive contained a list of the names of the downloaded HTML files, the source URLs, the internal IDs of the items (stories, chapters, users or reviews; see 4.3.2), and the chapter number (if applicable). This list could then be used to skip URLs that had already been processed by searching for the URL found by the crawler. The HTML files in combination with the CSV files could then be scraped by another spider without any time delay, since the data did not have to be downloaded from a website.

4.3. Data Storage

Just like with the web scraper tools, we had to decide on a suitable data storage format. We needed a storage format that would allow us to store large amounts of data in a well-structured, performant and analyzable way. Data sets had to be easily supplemented, while avoiding duplicates and data loss.

4.3.1. MongoDB

We used *MongoDB*[26] as database for storing the scraped data.

MongoDB is a *NoSQL*, non-relational database and has several advantages over few disadvantages for this project. Data is stored in documents in a *BSON* format which is a *JSON*-like format (see Figure 3).

These documents are extremely flexible and can store any data without the risk of losing data because the wrong format was used (Bruce, 2021). Unlike with relational database management systems (*RDBMS*), which usually store the data in the 3rd normal form, *NoSQL* databases generally do not use normalized forms (Chapple, 2022). *MongoDB's* design philosophy states that data should be embedded rather than referenced, resulting in extremely efficient and performant queries since there

[26]https://www.mongodb.com/

```
{
  "_id": { "$oid": "62dee5f34cbb9245c00daec6" },
  "title": "Concerning Atlantis",
  "source": "ArchiveOfOurOwn",
  "category": "Crossover",
  "url": "https://archiveofourown.org/works/9518678",
  "fandoms": [
    { "tier1": "Temeraire", "tier2": "Naomi Novik", "name": "Temeraire - Naomi Novik" },
    { "tier1": "Harry Potter", "tier2": "Harry Potter", "name": "Harry Potter - Harry Potter" }
  ],
  "publishedOn": { "$date": "2017-01-30T00:00:00Z" }
}
```


Figure 3: Excerpt of an exemplary story document in the *BSON* format.

are no expensive join queries (MongoDB, n.d.-a).

These embeddings lead to redundancies and thus to a higher storage requirement, which is, however, negligible due to the constantly decreasing storage prices (McCallum, 2022). The schemaless structure can also cause data clutter and loss of data quality.

In addition, *MongoDB* with their *BSON*-documents features an easily accessible format for any programming language without the need for object-relational mappers (*ORMs*) (Bruce, 2021). The aggregation pipeline can be utilized to build high-performance queries where data is queried and processed in multiple stages (MongoDB, n.d.-b). Each stage processes its input documents and forwards them to the next stage. For example, documents can be first filtered, then grouped, and finally sorted, with changes to the output data possible at each of these stages.

4.3.2. MongoDB Schema Design

For data acquisition, we first designed a different database schema than what was later used for the data analysis.

Data Acquisition Schema

Although not in compliance with the design philosophy of mongodb (MongoDB, n.d.-a), this schema rather aligned to the principles of the 3rd normal form common in relational database models. An excerpt of this can bee seen in Figure 4 while detailed database model schema can be viewed in Chapter A.

This relational structure had the advantage that, on the one hand, duplicates

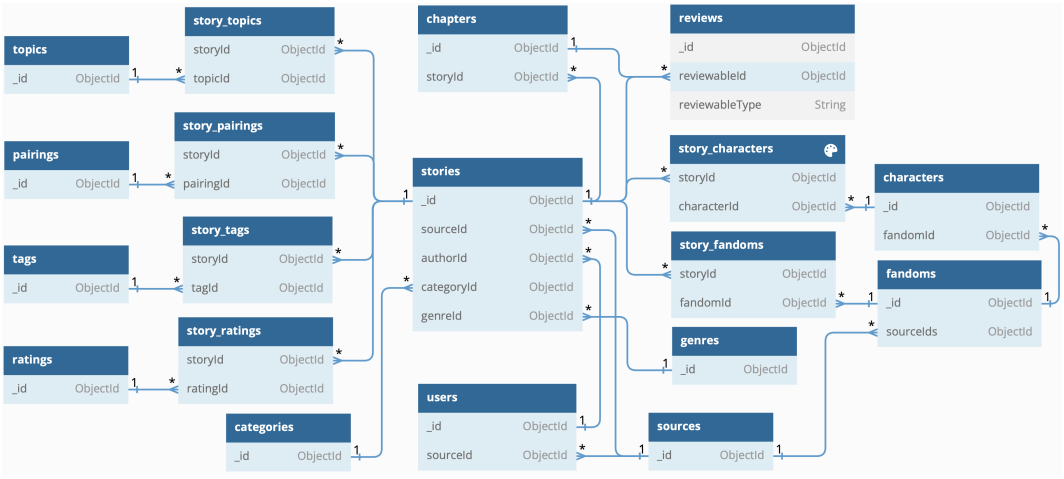

Figure 4: Excerpt of the database schema for the data acquisition.

could be avoided and, on the other hand, data could be cleansed and merged more easily later. For example, this allowed defining alternative names that are specific to each source archive. These could be translations (e.g. the genre could be *Bücher* for *FF.de* and *Books & Literature* for *AO3*) or differentiating designations (e.g. the stories pairing could be *MaleSlash* for *FF.de* and *M/M* for *AO3*). In this way, it was possible to check during the acquisition process whether an item already existed. Furthermore, if a fandom was to be renamed, this only had to be done in one place. Both data quality and a consistent presentation and structure could thus be ensured.

Data Analysis Schema

To benefit from the capabilities of MongoDB to the full extent, we subsequently restructured the schema for analyzing the data.

Fields that were specifically needed for the scraping process were removed. These were, for example, fields that compared the total number of chapters in the story, as indicated in the story overview, with the number of chapters already scraped, without having to query the database each time. When a story was created in the database, an author was required. Therefore, we initialized an empty author that contained only a user url and ID for referencing, and set a field for that user indicating that it was provisional and needed to be scraped later in the process.

All the information were then merged into four tables: *stories*, *chapters*, *users* and *reviews*. An overview of this database schema is illustrated in Figure 5. The chapter

table contained the largest documents, since they stored the entire narrative of a story and were due to *BSON*'s size limitation of 16 megabytes not merged into their story document.

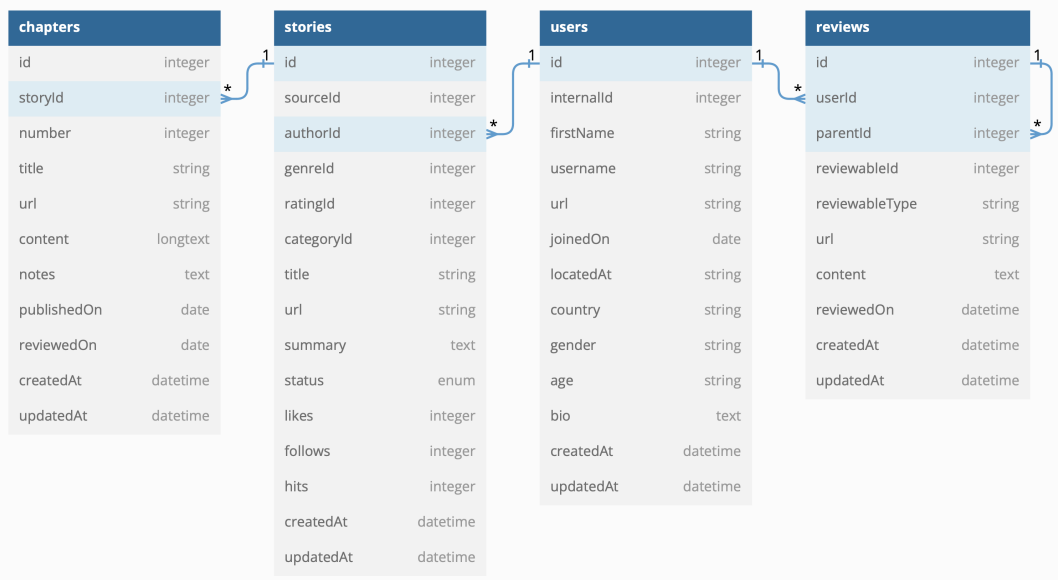

Figure 5: Database schema for the data analysis.

As can be seen, this schema is much leaner than the previous one. Fields from other tables were joined into data structures such as arrays and dictionaries, disregarding the concept of avoiding duplicates.

Characters from a story were consolidated into an array containing the character name as well as the fandom the character originates from. We structured the fandoms referenced in a story into tiers to allow easier analysis of popular fandoms with many subgenres as illustrated in Figure 6.

4.3.3. Cleansing and Plugging Data Holes

In the next step of the data acquisition phase we had to fill in missing data. This was done by implementing another set of *Spiders*, one for the *AO3* archive and one for the *FF.de* archive. These *Spiders* were used for data that was missing from the initial scraping process by querying the database for stories and chapters that contained no content or missed other crucial elements like fandom or genre and then scraping it from the archive's website. After scraping the missing information we still had some

stories that contained no text which is why we implemented another, very simple scraper using the *BeautifulSoup* library. For that we used the URLs referencing the stories from the database and stored the obtained texts.

Numerous stories are published that contain merely an image (e.g. a cartoon), an embedded video or audio player, a link to a story on another platform or a link to a *YouTube*[27] video. Story chapters containing those or did not reach a specified character threshold of 15 characters were removed from the database. Additionally, low character chapters were manually checked for their content since sometimes rather short but totally valid phrases like "nicht schwanger" (german for "not pregnant") were used while others were just placeholders. Stories that subsequently did not include any chapters were also removed from the database.


```
[
  {
    "tier1": "J.R.R. Tolkien",
    "tier2": "Mittelerde",
    "tier3": "Der Herr der Ringe",
    "name": "J.R.R. Tolkien - Mittelerde - Der Herr der Ringe"
  },
  {
    "tier1": "Harry Potter",
    "tier2": "Harry Potter",
    "name": "Harry Potter - Harry Potter"
  }
]
```


Figure 6: Example of refactored story fandoms.

To improve the quality of the data and increase analyzability, we have consolidated fandoms from both archives under a common name. For the comparison of fandom names, any punctuation, whitespaces, tabs, line breaks or carriage returns were temporarily removed. We then used the Levenshtein distance (Levenshtein, 1966), a metric that measures the difference between two strings based on the minimum number of single character edits such as insertions, deletions or substitutions, to determine the best match from the other archive. For overview and manual correction, all fandoms of *AO3* were then listed in a table and the most likely counterpart for each based on previous calculations was suggested. Finally, these values were used for the consolidation in the database.

[27] https://www.youtube.com/

4.4. Difficulties during Data Acquisition

We had several problems when collecting the data.

The first problem was that the number of requests per minute to the archives was limited, and we needed to probe the sites for the least time delay between requests possible without being blocked by the host. Because of the time delay and the huge amount of data available, the scraping process took its time. To improve this, we ran a second web crawler on another machine, as described earlier in Subsection 4.2.2. While this reduced the overall time required for scraping, it created another problem in that the fanfiction data was stored in a local database and access from this second computer was therefore not possible, so we used the CSV file as a temporary storage.

Another issue was the age restrictions for certain stories. While this was handled at *AO3* by confirming the age of majority in a popup window, at *FF.de* we had to either restrict scraping of restricted stories to between 11pm and 4am, or prove our user's age with a valid ID. This could be handled after authentication via *Scrapy*'s cookie middleware, where we attached our user's session to the requests. Sessions, on the other hand, expired after a certain time, and the stories had to be marked as age-restricted for later scraping.

AO3 distinguishes between fandoms from different media (e.g. *Sherlock Holmes* ***TV*** and *Sherlock Holmes* ***Books***), while *FF.de* is much stricter in this regard and only allows one type of media (e.g. just *Sherlock Holmes*). Consolidating the fandoms as outlined in the previous Subsection 4.3.3 solved this issue.

Another difficulty arose from the unreliability of the users. For example, sometimes authors used the title input box to enter the text of the story, which resulted in documents having empty chapters and overfilled titles. These stories then had to be scraped anew after they were identified.

5. Data Supplements

For an improved analyzability of the data we had to append additional information to our documents. This already covered the first steps of analysis at document level.

5.1. Counting Text Tokens

First, we implemented an asynchronous script that traversed through the corpus and collected general text statistics for both chapters and stories in batches. This included the number of sentences, words, letters and text characters.

For sentence segmentation we used the open-source library *spaCy*[28], written in *Python*. It provides several pre-trained natural language models for a number of different languages, including German. Besides its speed and ease of use, no recognition failures were detected for the sentences we tested.

After these more general supplements for the statistical analysis, we required additional information for analyzing the gender representation in the corpus.

5.2. Pronouns Detection

More and easier-to-analyze information was needed to verify gender distribution. Utilizing the *spaCy* library again, we implemented a script that detected all pronouns in the corpus and counted their occurrences. This was based on the approach of Duggan (2020), who analyzed pronouns in story paratexts and user profiles (see Section 3.2). We restricted ourselves to the personal pronouns in the third person singular, since other forms as well as the plural do not allow a clear assignment to one gender. Consequently, these were *er*, *sie*, *ihn*, *ihm*, *ihr*, *seiner* and *ihrer*.

[28]https://spacy.io/

5.3. Extracting Story Character Names

Another very time-consuming step was the filtering of all person names from the corpus on a chapter-by-chapter basis. For this purpose, a named entity recognition (NER) model was used, which was already pre-trained for the German language and did not require any further adaptations. We initially tested this with the fast *spaCy* library, but then decided to use the state-of-the-art *FLAIR* (Akbik et al., 2019) natural language processing (NLP) framework instead for much better results. While Milli & Bamman (2016) used the *BookNLP*(Bamman et al., 2014) library for this task, we were unable to utilize it, even though it was more tailored to our problem, as it was only available for the English language. *FLAIR* was trained on the *CoNLL-2003* (Tjong Kim Sang & De Meulder, 2003) dataset, which consists of news articles from *Reuters*[29] and *Frankfurter Rundschau*[30]. The NER models provided by *FLAIR* detect named entities such as persons, locations, organizations, and miscellaneous. For this study, we limited the model to the recognition of persons, in our case characters, who appear in the stories. Although we did not use the *spaCy* library for entity recognition, we still used its previously tested sentence segmentation capabilities. We processed each chapter at the sentence level and then appended the found person tags with the number of their appearances to the respective database document.

When using the *FLAIR* models, processing was associated with accurate results, but very slow. We have therefore tried and implemented several approaches to speed up the process.

Initially, we used the model *ner-german-large*, but then decided to use model *ner-multi-fast* because it was noticeably faster. Even though it was trained on English, German, Dutch and Spanish language corpora, one could also pass German as a parameter to the sentence tagger and got equally good results. For comparison: the German tagger (*ner-german-large*) required about 1 minute and 52 seconds for a text with a length of 752 sentences (the average text length of a chapter), while the

[29]https://www.reuters.com/
[30]https://www.fr.de/

multi-language tagger *ner-multi-fast* required about 1 minute and 44 seconds. While the difference of about 8 seconds may not seem like much, it makes a significant difference in a corpora of almost two million chapters like this one.

Another approach was to use the fast *spaCy* library and cleansing the results of any unwanted tokens afterwards, but this was not applicable for usable results.

For counting text tokens (see Section 5.1), we used the concept of asynchronous batch processing for acceleration. We tested this with various process counts and batch sizes, but the processing times increased instead of decreasing. The likely reason was the amount of memory necessary to run multiple taggers simultaneously.

Due to one machines limitation to memory and CPU, we set up a cluster of 18 cloud computing machines to process the data in parallel. Cloud computing space using the free quota provided by *Microsoft Azure*[31], *Digital Ocean*[32] and *Vultr*[33] was used for this purpose. We also configured an Ubuntu 20.04 LTS server and paired it with a static hostname for remote access so that any of these machines can access the shared database. Each of these computers ran a single token classifier.

Due to time constraints we reduced the number of stories to be tagged to 280.000. This was roughly equivalent to 68% of the total number of stories. Stories were randomly selected, locked and processed.

5.4. Determine Story Character Name Genders

After the names of the characters were extracted from the stories, the gender of each individual had to be determined. Along with all the other code, the script for training the gender classifier can be found on *GitHub*[34].

Acquiring a Training Dataset

The first step was to acquire a training dataset. In addition to the *NLTK* name corpus (Steven Bird et al., 2009), we obtained a large dataset from *Google BigQuery*[35] containing all names from applications for *Social Security* cards for births in the *United*

[31]https://azure.microsoft.com/
[32]https://www.digitalocean.com/
[33]https://www.vultr.com/
[34]https://github.com/Cele3x/fanfiction/
[35]https://cloud.google.com/bigquery/public-data

States after 1879 provided by the *United States Social Security Administration*[36].

To increase diversity, we have also included names from the *babynames.com*[37] website in our dataset. For example, names for categories such as biblical names or names of fictional characters with corresponding gender information are listed there, but also names of various geographic origins. The *babynames.com* website names were hosted on a dynamic website, so we had to scrape them with *Selenium*. *Selenium* is usually used to automate web browser interactions for testing purposes, though we used it to scrape the website and store the names in a file.

All datasets were then merged, removing all duplicates, gender-neutral names, and names that were not clearly assigned to a gender. The result was a list of 106,000 rows, each consisting of the name and a gender classifier.

Preprocessing the Data

Names had to be encoded in *ASCII* format because the *scikit-learn*[38] library only accepts *ASCII* characters, therefore we used the *unidecode*[39] library to convert all names (e.g. á became a). This entailed a further duplicate check and subsequent cleanup.

Furthermore, we padded all names with spaces and converted them to lowercase. To be able to use the data in a neural network, we also encoded the story characters in numbers. The gender for the training set was one-hot encoded as 0 for 'F' (female) and 1 for 'M' (male).

Training the Model

For training the model we use the *Keras*[40] library with the *TensorFlow*[41] backend. The model consists of three layers: an embedding layer, a bidirectional LSTM layer and a dense layer. We then trained the model using the Adam optimizer and the binary cross-entropy loss function. Splitting the data into a usual training and test set size of 80% and 20%, respectively, we trained for 50 epochs with a batch size of 64. For

[36] https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data
[37] https://babynames.com/
[38] https://scikit-learn.org/
[39] https://github.com/takluyver/Unidecode/
[40] https://keras.io/
[41] https://www.tensorflow.org/

not overfitting the model, we used early stopping with a patience of 5.

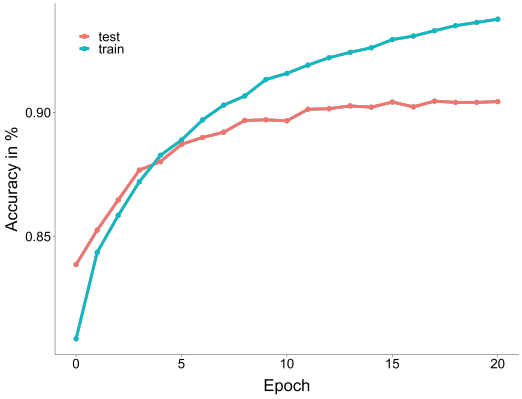

Figure 7: Model accuracies on train and validation datasets.

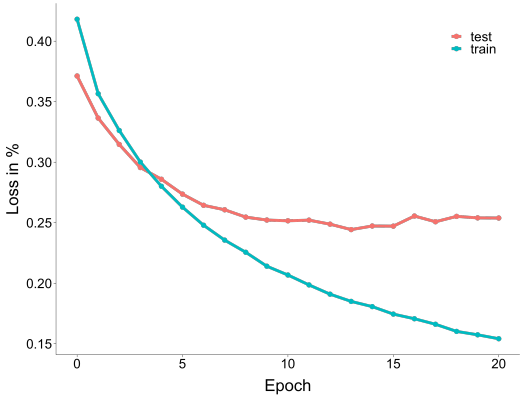

Figure 8: Model losses on train and validation datasets.

Accuracies (see Figure 7) were not rising significantly and losses (see Figure 8) for train and validation dataset departed after 21 epochs, so we stopped the training at this point. We restored the weights from the best epoch (16) with an accuracy of 0.93 on the train and 0.90 on the test dataset.

Predicting Story Character Genders

The trained model was then used to predict the gender of each fictional character, their occurrences in the story and chapter were counted, and this information was then attached to the respective document. In Section 5.3, we have already described how the story character names were extracted from the story texts.

To begin, we had to cleanse all the previously extracted story character names. After some general clean-up operations, like removing special characters, numbers and whitespaces, we compared the names with the names from the training dataset. If a match was found, the process was stopped immediately for that name, if not, we filtered words from the German *DE-LIWC* dictionary (Meier et al., 2019). For example, the name "Glinda die Gute" would become "Glinda". Words were extracted from the *LIWC 2015* poster in *PDF* format using *PyPDF2*[42] and *PDFMiner*[43]. Cleansed names with their cumulative occurrences were stored in the respective story document.

[42]https://github.com/py-pdf/PyPDF2
[43]https://github.com/euske/pdfminer

Given that iterating over each story in the corpus and predicting the genders of the story character names would have taken a significant amount of time, we split this process into several steps.

First, we extracted all story character names from the stories, removed any duplicates, and saved them to a separate *CSV* file. This newly created list was subsequently merged with the names from the training dataset and again duplicates were removed. We then prepopulated additional information to the list, such as the gender and a probability value of 1.00 for all names from the training dataset. This probability value was used to evaluate the confidence of the prediction, with 1.00 being the highest confidence and 0.00 being the lowest.

In the next step, this *CSV* list was divided into 100 chunks with 2,800 stories each for the total 280,000 tagged stories. Each chunk containing all the names from the 2,800 stories was then passed to the prediction model and the results with the predicted gender and probability were stored in a temporary data structure. After all chunks were processed, names with a low confidence of less than 0.80 and having at least one whitespace in the name were reprocessed. The whitespaces were used to split the name into two parts, and the first part, a potential first name, was resent to the prediction model. Finally, all prediction results were merged with the original list and saved in another *CSV* file. Figure 9 shows a fragment of this list to give an impression of what this looks like.

	name	gender	probability
23783	alec	M	1.00000
237773	doofus	M	0.88126
255015	edward cullen	M	0.99916
343333	glinda	F	1.00000
475928	joe black	M	0.86474
909514	sasori	F	0.60051
1035552	tracy	F	1.00000
1065130	violett	F	1.00000
1112360	zeus	M	1.00000

Figure 9: Fragment of the *CSV* list with the calculated gender of the names.

Utilizing this list, we iterated over all stories, comparing the names with the names from the prediction list. Statistics showing the sum of all female, male, and indecisive names, along with their percentages and a female-to-male ratio, were then appended to the story document. Whereas a ratio of 0.0 means that all decisive

persons are female and 1.0 means that all are male, adapted from our one-hot coding for gender that we performed earlier for model training. For example, a ratio of 0.7 means that the gender representation is rather male-dominated.

6. Results and Discussion

Utilizing all previously described and implemented methods, we analyzed the data and present the results in this chapter. The results are divided into two sections: the first part is about general statistical analysis of the corpus, the second part is about the analysis of the gender representation.

6.1. Corpus Analysis

Fan Fiction archives were scraped over a time period of 7 months, from January 28 to August 23, 2022. Updates for *FF.de* database documents range over the full duration, while *AO3* was processed over a period of 2 weeks, from July 25 to August 8, 2022. This was due to the fact that German-language fan fiction is less common on *AO3* in comparison and took therefore less time to scrape. The corpus acquired consists of 412,923 stories, their chapters (1,955,923) and reviews (4,887,367), as well their respective authors (149,975) as depicted in Table 10.

Archive	Stories	Chapters	Users	Reviews
FF.de	394,848	1,885,066	135,726	4,849,646
AO3	18,075	70,857	14,249	37,721
Total	412,923	1,955,923	149,975	4,887,367

Table 10.: Corpus overview.

The lack of diversity in german fan fiction mentioned by Cuntz-Leng & Meintzinger (2015) in Section 3.1 can be confirmed. Table 11 demonstrates that after 7 years, Harry Potter is still by a large margin the most popular fandom on *FF.de*. While in 2015 the top 7 fandom distributions have shifted only slightly, stories about musicians as *One Piece*, *One Direction* and *Tokio Hotel* lost some popularity, likely due

Fandom	FF.de	AO3
Tatort	644 (0.16%)	2,978 (16.48%)
Harry Potter	54,405 (13.78%)	1,793 (9.92%)
Music	38,568 (9.77%)	654 (3.62%)
Naruto	27,303 (6.91%)	200 (1.11%)
Marvel	5,311 (1.35%)	1,234 (6.83%)
Internet-Stars	17,830 (4.52%)	50 (0.28%)
Canon	N/A	671 (3.71%)
Twilight	13,734 (3.48%)	45 (0.25%)
One Piece	11,443 (2.90%)	89 (0.49%)
Sports	11,440 (2.90%)	663 (3.67%)
The Three Investigators	1,251 (0.32%)	509 (2.82%)

Table 11.: The top 7 fandoms on FF.de and AO3 respectively. Individual appearances in the top fandoms are color-coded, while mutual appearances are not.

to the high fluctuation of overall approval ratings in the music industry. With an increasing number of publications about celebrities from the Internet (YouTuber [44]) and sports (mainly soccer), these fandoms have been overtaken.

An interesting observation is the very high number and overall share of stories about *Tatort*, a German crime series, on *AO3* compared to this fandom on *FF.de*. This is likely due to the fact that the fandom community for *Tatort* established itself on *AO3* and authors active in this rather publish their stories there.

The same assumption can safely be applied to fan fiction distribution as a whole as depicted in Table 10. Users publish their stories where their language specific fan base and community is located. For German fanfiction, this is generally the *FF.de* archive, but in rare cases on other platforms, such as in the instance of *Tatort*. This can explain the scarcity of German-language fan fiction on *AO3*. Another attempted explanation is the tendency, mentioned in Section 3.1 (Cuntz-Leng & Meintzinger, 2015), to migrate to fan fiction written in English, even as a non-native speaker.

[44] https://www.youtube.com

Like Schmidt et al. (2021) for AO3, we also observe regarding the supposedly strongly intertwined anime fan fiction in Germany that the most popular anime fandom in our corpus is *Naruto* with only 27,503 stories and a share of 6.66%. Even in commonly male-dominated genres, there is a preponderance of female writers.

Also worth mentioning in this context are the works about stories by Karl May, who, according to Cuntz-Leng & Meintzinger (2015), is said to have been directly responsible for the emergence and development of the German fan fiction scene in the 19th century. This so-called phenomenon seems to have died before the establishment of online archives, because the number of publications about Karl May stories on *FF.de* is only 204 (0.05%) and on *AO3* 46 (0.25%).

Genre	Sentences	Words	Letters	Characters
Musicals	19,911	243,903	1,221,731	1,523,534
Books & Literature	1,777	21,389	107,618	133,041
Crossover	1,037	9,897	46,485	58,682
Other Media	334	3,263	16,298	20,096
Cartoons & Comics	318	3,230	16,111	20,588
Movies	197	3,155	15,256	19,193
Celebrities	110	1,351	6,510	8,210
Anime & Manga	101	812	4,248	5,460
Video Games	98	1,423	7,672	9,363
Tabletop & RPGs	70	788	3,746	4,719
TV-Shows & Podcasts	33	224	1,016	1,440
Total	157	1,763	8,685	10,881

Table 12.: Medians for sentences, words, letters and characters for each genre.

Table 12 shows the median values for the number of sentences, words, letters and characters for each genre. While publications about musicals are rather unpopular, with a share of only 0.68% (2,795 stories), they are the longest, with the highest number of sentences, words, letters and characters. In this regard, the statistics about *Books & Literature* and crossovers can be considered more significant, which in the case of the latter often includes works about written texts as well, with a total share of 28.21% (116,481) and the second longest with a median of 31,286 words. The

original works on which these stories are based are also typically written in a fairly lengthy and detailed manner. Consequently, the fan fictions that refer to them are often written in the same style and reflect this in their length as well. Works about games and *TV-Shows & Podcasts* are the shortest published stories on average.

Pairing	Frequency	Mean
Generic	276,188	66.50%
M/M	117,051	28.18%
F/M	14,784	3.56%
F/F	2,826	0.68%
Multi	2,702	0.65%
N/A	1,059	0.25%
Diverse	693	0.17%

Table 13.: Comparison of story pairings sorted by frequency.

Pairings in fan fiction are classifications of stories referring to the romantic relationship between characters. According to Table 13 the most common pairing is the generic one, which is used for stories where relationships do not exist at all or are not addressed. The second most common pairing is *M/M* or *MaleSlash*. It implies that romantic relationships are present in the story and are also thematized. Primarily, it is about romantic relationships between men. The opposing classification of this is *F/F* or *FemSlash*, which is used for stories about romantic relationships between women. While *F/M* describes heterosexual relationships, *Diverse* relationships that can not be classified and *Multi* relationships that do involve multiple previously defined pairings without a predominant one.

Potential reasons for the share of *MaleSlash* stories being so high are discussed later in this chapter.

After this general statistical analysis of the corpus, the focus in the following section shifts towards the representation of gender in fan fiction.

6.2. Gender Representation in Fan Fiction

Previous research has shown that women are underrepresented in a variety of media forms (Collins, 2011). In popular films, for example, men provide more than two-thirds of the speaking roles (Neville & Anastasio, 2019). This section examines whether this can also be observed in the media form of fan fiction. The subject of the investigation is the previously acquired corpus on German fan fiction.

6.2.1. Analyzing Character Genders

To analyze gender representation in German fan fiction, we used with *Flair*(Akbik et al., 2019) a named entity recognition model to extract the story characters' names from the stories and trained an LSTM model to predict their respective gender, as outlined in Section 5.4. The general results of this approach are presented in Table 14, which compares the number of male, female, and indecisive story characters, that is, story characters whose prediction did not reach a confidence of 0.80% or higher.

Archive	Ratio	Males	Females	Indecisives
FF.de	0.63	36,000,856 (60.85%)	19,471,225 (32.91%)	3,695,846 (6.25%)
AO3	0.74	579,925 (71.63%)	189,421 (23.40%)	40,212 (4.97%)
Total	0.63	36,580,781	19,660,646	3,736,058

Table 14.: Gender representation of *FF.de* and *AO3*. Ratio is using the arithmetic mean and depicts only Male and Female proportions with 1 = all male and 0 = all female. Shown percentages are on a per-archive basis.

It can be observed that most characters could be predicted sufficiently with only about 5% of all characters being indecisive. The majority of characters appearing in fictional stories are male at 60.85% on *FF.de* and even more so on *AO3* at 71.63%. Since the number of stories on *AO3* is significantly lower and tends to be less diverse with fandoms such as *Tatort* accounting for a large proportion of published texts, the percentage shown for *FF.de* is likely to be the relevant one, as the unchanged overall ratio suggests.

Consequently, a distinction between the two archives in this context is rather unnecessary and both archives can be considered as a whole. Statistics on characteristics that do not achieve a confidence of at least 0.80% (indecisives) are omitted in the following analysis. Furthermore, the ratio will always represent the distribution of male and female units with 1 = all male and 0 = all female.

Fandom w/ Genre	Ratio	Males	Females
Harry Potter *Books & Literature*	0.69	13,296,037	720,302
Musik *Celebrities*	0.66	978,215	379,515
Naruto *Anime & Manga*	0.57	571,593	393,172
Supernatural *TV-Shows & Podcasts*	0.90	294,194	52,332
Marvel *Movies*	0.77	318,954	107,142
Crossover *Crossover*	0.64	1,068,512	506,279
Online Games *Video Games*	0.58	365,981	266,543
Marvel *Cartoons & Comics*	0.86	45,670	5,157
Tanz der Vampire *Musicals*	0.65	207,777	94,481
Canon *Other Media*	0.58	12,146	3,308
Das Schwarze Auge *Tabletop & RPGs*	0.63	19,672	16,806

Table 15.: Gender representation of characters in each top fandom per genre. Sorted by their amount of stories in descending order. Ratio depicts only Male and Female proportions with 1 = all male and 0 = all female.

Table 15 illustrates this by listing the top fandoms for each genre, sorted by the popularity of the genre.

Three fandoms stand out among these: Supernatural, the Marvel cinematic universe, as well as their carton and comic counterpart. They all have an even larger amount of male story characters compared to the rest.

In the case of the television series Supernatural, this can be explained by the study

conducted by Kleindienst & Schmidt (2020) on AO3. They found that the texts in this fandom contained male-male relationships at an overwhelming rate of 91.99%. This trend also seems to be applicable to the *FF.de* corpus, though most pairings for *Supernatural* are declared as *Generic* (58.79%) rather than male-male (38.92%). This will be discussed in more detail later on.

Although there has been an increase in the number of female characters in the *Marvel Universe*, the overall number is still quite low, with a total percentage of 19.88%, according to Ray (2020)'s research on the *Marvel Cinematic Universe*. This scarcity of female characters in the canon, as well as an already strong tendency towards male characters in fan fiction in general, has likely led to this heavily male biased ratio.

6.2.2. Analyzing Gender-Specific Pronouns

Seen across all fan fiction genres, the distribution of used feminine and masculine personal pronouns is fairly even as shown in Table 16. This confirms the tendency that has already been identified with regard to the predominance of male characters in the stories of this community. A correlation was to be expected, since the characters introduced in the story are addressed with personal pronouns to the same degree they are occurring.

Genre	Feminine	Masculine	Total
Books & Literature	38.05%	61.95%	64,144,827
TV-Shows & Podcasts	35.96%	64.04%	27,431,159
Anime & Manga	33.55%	66.45%	18,979,393
Movies	37.65%	62.35%	8,948,841
Video Games	37.66%	62.34%	8,765,844
Celebrities	20.71%	79.29%	6,469,078
Cartoons & Comics	39.08%	60.92%	4,500,338
Crossover	36.86%	63.14%	2,640,782
Musicals	38.20%	61.80%	1,564,273
Tabletop & RPGs	41.39%	58.61%	347,633
Other Media	25.94%	74.06%	282,224

Table 16.: Distribution of feminine and masculine personal pronouns per genre.

Moreover, it can be assumed that this ratio would presumably have to shift even further in the direction of the masculine pronouns. This is due to the fact that we previously defined the German "sie" (she) as a feminine personal pronoun, although it can also refer to the third-person plural or "you" in the polite form. However, this does not apply analogously to the German "er" (he), which is used exclusively for the masculine third-person singular.

The genres *Celebrities* and *Other Media* stand out in comparison with an even greater discrepancy between the use of feminine and masculine pronouns. To this extent, this discrepancy could not be observed before when comparing the character genders (see Table 15). *Other Media* is a genre exclusive to *AO3*, much of which consists of canons or original works. In this case, the discrepancy was previously seen in the gender ratio of the characters. This gives the impression, and also confirms the thesis of Milli & Bamman (2016), that canonical works are leaning even more towards a male-dominated narrative, in spite of the fact that only a quite small sample can be referred to for comparison.

6.2.3. Adding the User's Sex to the Ratio

Duggan (2020) previously stated that the sex of authors was directly dependent on the portrayal of gender roles in their stories. While they analyzed story paratexts and profile biographies on *AO3* for the *Harry Potter* fandom to extract their writer's sex, we were able to use the information provided by the users themselves, since users can submit it to their profiles. Due to their approach Duggan had a very small sample size of 1,800 users of which only 265 provided extractable information regarding their gender. As a consequence, we limit our analysis to the sex of *FF.de* users, but this is negligible in respect to the rather small number of german stories on *AO3* in any case. Users have the option to specify in their profile whether they consider themselves "weiblich" (female), "männlich" (male) or "divers" (diverse). While Duggan extracted this information, we had to trust the users to provide it correctly which could lead to deviations.

In Table 17, we can observe that with about 70.94% the majority of users state

User's Sex	Frequency	Authors	Reviewers	Age
Female	87,784 (64.68%)	72,959 (68.04%)	51,084 (67.89%)	26.89
Male	7,834 (5.77%)	6,291 (5.87%)	4,521 (6.01%)	27.98
Diverse	671 (0.49%)	511 (0.48%)	450 (0.60%)	23.12
N/A	39,437 (29.06%)	27,463 (25.61%)	19,185 (25.50%)	27.10
Total	135,726	107,224	75,240	26.97

Table 17.: Frequencies of FF.de users regarding their sex. The distinction between Authors and Reviewers ist not mutual exclusive, but users are unique for each category. Age is the arithmetic mean of all users for the respective sex.

their sex. For all sexes, users tend to author stories more likely than review any.

We therefore cannot confirm Duggans assessment towards the diversity of users, with only 0.49% stating "diverse", but the one towards an overwhelming majority of female writers and readers (about 50%) and even surpass this with about 65% female users. This is contrary to Fast et al. (2016) findings on high-level gender statistics from *Wattpad*[45], which state that the majority of fan fiction creators tend to be male, at around 54%.

Table 18 illustrates the distribution of authors sex regarding the genre of the stories. We can observe that any genre are female-dominated by a large margin, even *Anime & Manga* which is expected to be not (see Malone (2010)). The only exception seem to be tabletop and role-playing games, where male authors are even in the majority. It is also particularly interesting that this genre has by far the largest gap in terms of gender information among authors. If the trend in this genre's distribution continues, it can be assumed that men are more secretive about these declarations. In addition, conclusions can be drawn about people who write works about musicals. People in this genre identify themselves more frequently as diverse compared to the other genres.

In Table 19 we illustrate how the sex of the author affects the used character's genders and personal pronouns. Since no clear correlation can be established, we can conclude that the author's sex does not influence how the characters are defined

[45]https://www.wattpad.com

Genre	Frequency	Females	Males	Diverse	N/A
Anime & Manga	107,045	76.20%	6.09%	0.81%	16.90%
Books & Literature	106,007	75.30%	4.54%	0.77%	19.39%
Celebrities	75,854	76.34%	2.65%	0.84%	20.17%
TV-Shows & Podcasts	51,942	75.67%	3.07%	0.85%	20.41%
Movies	19,093	70.03%	8.00%	1.30%	20.67%
Video Games	16,923	65.24%	14.31%	1.09%	19.36%
Cartoons & Comics	9,064	65.72%	12.41%	1.33%	20.53%
Crossover	5,414	63.96%	13.32%	1.00%	21.72%
Musicals	2,738	76.52%	3.10%	3.03%	17.35%
Tabletop & RPGs	768	30.21%	40.76%	0.26%	28.78%
Total	394,848	294,765	21,126	3,455	75,502

Table 18.: Distribution of authors sex regarding the genre of the stories. Sorted by the number of distinct authors.

in their gender nor how often they are referenced. Both men and women alike use equivalent gender ratios in their works.

In the following, we will therefore no longer differentiate between the sexes of the authors.

When examining Table 20, a significant tendency can be identified. The figure shows that the older the author, the more likely they are to use male characters and masculine personal pronouns.

Author's Sex	Female Characters	Male Characters	Feminine Pers. Pron.	Masculine Pers. Pron.
Female	35.82%	64.18%	37.66%	62.34%
Male	34.52%	65.48%	35.64%	64.36%
Other	31.03%	68.97%	31.69%	68.31%
N/A	31.85%	68.15%	31.82%	68.18%

Table 19.: Male and female characters and personal pronouns usage in relation to author's sex.

Author's Age	Frequency	Characters Ratio	Pers. Pron. Ratio
1-20	15,380	0.61	0.58
21-25	70,171	0.61	0.61
26-30	64,502	0.63	0.61
31+	56,684	0.67	0.65

Table 20.: Male and female characters and personal pronouns usage in relation to authors age. Ratio depicts male and female proportions with 1 = all male and 0 = all female.

Author's Age	Generic	M/M	F/M	F/F	Multi	Other
1-20	73.52%	19.07%	5.20%	0.62%	1.41%	0.18%
21-25	71.95%	24.12%	2.95%	0.28%	0.62%	0.08%
26-30	67.96%	28.43%	2.56%	0.57%	0.44%	0.04%
31+	64.25%	31.52%	3.25%	0.57%	0.35%	0.07%

Table 21.: Pairings usage in relation to authors age

This trend is also reflected in the distribution of used pairings per author's age (Table 21). Older generations in the fan fiction community seem more inclined to write about romantic relationships between two men (*male-slash* or *M/M*). As Russ (1985) describes in her work "Pornography by women for women, with love" women in particular are avid writers of often very explicit male-slash stories. This can be explained by the fact that fan fiction is used to express a disruption of heteronormativity, which was and still is to an extent suppressed by social norms. The dominance of writings about pure-male relationships can be understood as "a communal and grass roots critique not only of popular culture but also of heterosexual hegemonic notions of gender and sexuality" (Jung, 2002). While movements and groups such as the *LGBTQ+* community have been fighting for the acceptance of non-heterosexual relationships for decades, the younger generations have grown up with this acceptance. Therefore, it is not surprising that the younger generations are less likely to write about homosexual relationships.

7. Conclusion

In the course of this work, an extensive corpus of German fan fiction was acquired. Multiple sources were evaluated and with *FF.de* and *AO3* the most suitable ones were chosen. This corpus consists of 412,923 stories, their chapters and reviews, as well their respective authors and metadata.

While we have already analyzed it with regard to gender representation, it can be used for further research in the field of fan fiction or natural language processing in general.

For analyzing the corpus, we used the pre-trained named entity recognition model *FLAIR* for extracting persons with their occurrences, trained a LSTM model for predicting those persons genders and counted all used pronouns in the process. The data obtained in this way was subsequently used for further analysis.

Previously, we have stated that men are overrepresented in media such as canons, television shows, and social media conversations and raised the question of whether this also applies to fan fiction. Although the majority of readers and writers of fan fiction is female, the dominance of male characters and masculine pronoun usages in the corpus is not shifting.

Younger generations have a lower tendency to write stories about male-slash...

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, 6). FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 conference of the north* (pp. 54–59). Minneapolis, Minnesota, USA: Association for Computational Linguistics. doi: 10.18653/v1/N19-4010

Bamman, D., Unterwood, T., & Smith, N. A. (2014). A bayesian mixed effects model of literary character. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, *1: Long Papers*, 370–379.

Barnes, J. L. (2015, 2). Fanfiction as imaginary play: What fan-written stories can tell us about the cognitive science of fiction. *Poetics*, *48*, 69–82. doi: 10.1016/j.poetic.2014.12.004

Bergstrom, K., Jenson, J., & de Castell, S. (2012). What's 'choice' got to do with it? In *Proceedings of the international conference on the foundations of digital games - fdg '12* (p. 97). New York, New York, USA: ACM Press. doi: 10.1145/2282338.2282360

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bretthauer, B., Zimmerman, T. S., & Banning, J. H. (2007, 2). A Feminist Analysis of Popular Music. *Journal of Feminist Family Therapy*, *18*(4), 29–51. doi: 10.1300/J086v18n04_02

Bruce, D. (2021, 4). *Understanding the Pros and Cons of MongoDB.* Retrieved 01.11.2022, from `https://www.knowledgenile.com/blogs/pros-and-cons-of-mongodb/`

Busse, K. (2009). In focus: Fandom and feminism: Gender and the politics of fan production: Introduction. *Cinema Journal*, *48*(4), 104–108.

Busse, K., & Lothian, A. (2017, 8). A history of slash sexualities: Debating queer sex, gay politics and media fan cultures. In *The routledge companion to media, sex and sexuality* (pp. 117–129). Routledge. doi: 10.4324/9781315168302-12

Chapple, M. (2022, 2). *The Basics of Database Normalization.* Retrieved 01.11.2022, from `https://www.lifewire.com/database-normalization-basics-1019735`

Collins, R. L. (2011, 2). Content Analysis of Gender Roles in Media: Where Are We Now and Where Should We Go? *Sex Roles*, *64*(3-4), 290–298. doi: 10.1007/s11199-010-9929-5

Cuntz-Leng, V., & Meintzinger, J. (2015, 6). A brief history of fan fiction in Germany. *Transformative Works and Cultures*, *19*. doi: 10.3983/twc.2015.0630

Duggan, J. (2017). Revising Hegemonic Masculinity: Homosexuality, Masculinity, and Youth-Authored Harry Potter Fanfiction. *Bookbird: A Journal of International Children's Literature*, *55*(2), 38–45. doi: 10.1353/bkb.2017.0022

Duggan, J. (2020, 9). Who writes Harry Potter fan fiction? Passionate detachment, "zooming out," and fan fiction paratexts on AO3. *Transformative Works and Cultures*, *34*. doi: 10.3983/twc.2020.1863

Duggan, J. (2022, 11). "Worlds...[of] Contingent Possibilities": Genderqueer and Trans Adolescents Reading Fan Fiction. *Television & New Media*, *23*(7), 703–720. doi: 10.1177/15274764211016305

Emons, P., Wester, F., & Scheepers, P. (2010, 3). "He Works Outside the Home; She Drinks Coffee and Does the Dishes" Gender Roles in Fiction Programs on Dutch Television. *Journal of Broadcasting & Electronic Media*, *54*(1), 40–53. doi: 10.1080/08838150903550386

Fast, E., Vachovsky, T., & Bernstein, M. S. (2016, 3). Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. In *Tenth international aaai conference on web and social media.* Retrieved from `www.aaai.org`

Garcia, D., Weber, I., & Garimella, V. R. K. (2014). Gender asymmetries in reality and fiction: The bechdel test of social media. In *Eighth international aaai conference on weblogs and social media.*

Gooden, A. M., & Gooden, M. A. (2001). Gender Representation in Notable Children's Picture Books: 1995–1999. *Sex Roles*, *45*(1/2), 89–101. doi: 10.1023/A:1013064418674

Handley, C. (2012). "Distressing damsels": Narrative critique and reinterpretation in Star Wars fanfiction. In *Fan culture: Theory/practice* (pp. 97–118). Newcastle upon Tyne: Cambridge Scholars Publishing.

Hellekson, K., & Busse, K. (2006). Introduction: Work in Progress. In K. Hellekson & K. Busse (Eds.), *Fan fiction and fan communities in the age of the internet* (pp. 5–32). Jefferson: McFarland & Company, Inc.

Jia, S., Lansdall-Welfare, T., & Cristianini, N. (2015, 5). Measuring Gender Bias in News Images. In *Proceedings of the 24th international conference on world wide web* (pp. 893–898). New York, NY, USA: ACM. doi: 10.1145/2740908.2742007

Jung, S. (2002). Queering popular culture: Female spectators and the appeal of writing slash fan fiction. In *Gender forum* (Vol. 2, pp. 30–50).

Kenny, C. (2022, 7 26). *Web Scraping vs Web Crawling.* Retrieved 01.11.2022, from `https://www.zyte.com/learn/difference-between-web-scraping-and-web-crawling/`

Kleindienst, N., & Schmidt, T. (2020, 11). Investigating the Transformation of Original Work by the Online Fan Fiction Community: A Case Study for Supernatural. In *Digital practices. reading, writing and evaluation on the web.* Basel, Switzerland. doi: 10.5283/epub.50828

Lauzen, M. M., Dozier, D. M., & Horan, N. (2008, 5). Constructing Gender Stereotypes Through Social Roles in Prime-Time Television. *Journal of Broadcasting & Electronic Media*, *52*(2), 200–214. doi: 10.1080/08838150801991971

Leow, H. M. A. (2011, 9). Subverting the canon in feminist fan fiction. *Transformative Works and Cultures*, *7*. doi: 10.3983/twc.2011.0286

Levenshtein, V. I. (1966, 2). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Liu, C., Osama, M., & de Andrade, A. (2019, 11). DENS: A Dataset for Multi-class Emotion Analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 6293–6298). Hong Kong, China. doi: 10.48550/arXiv.1910.11769

Lothian, A., Busse, K., & Reid, R. A. (2007, 9). "Yearning Void and Infinite Potential": Online Slash Fandom as Queer Female Space. *English Language Notes*, *45*(2), 103–111. doi: 10.1215/00138282-45.2.103

Malone, P. (2010, 4). From BRAVO to Animexx.de to Export: Capitalizing on German Boys' Love Fandom, Culturally, Socially and Economically. In A. Levi, M. McHarry, & D. Pagliassotti (Eds.), *Boys' love manga: Essays on the sexual ambiguity and cross-cultural fandom of the genre* (pp. 23–43). Jefferson: McFarland & Company, Inc.

McCallum, J. C. (2022). *Historical Cost of Computer Memory and Storage.* Retrieved 01.11.2022, from `https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage?country=~OWID_WRL`

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). "LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015. *PsyArXiv*(a).

Milli, S., & Bamman, D. (2016). Beyond Canonical Texts: A Computational Analysis of Fanfiction. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2048–2053). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/D16-1218

MongoDB. (n.d.-a). *Advantages of MongoDB.* Retrieved 01.11.2022, from `https://www.mongodb.com/advantages-of-mongodb`

MongoDB. (n.d.-b). *Aggregation Pipeline.* Retrieved 01.11.2022, from `https://www.mongodb.com/docs/manual/core/aggregation-pipeline/`

Muttenthaler, L., Lucas, G., & Amann, J. (2019). Authorship Attribution in Fanfictional Texts given Variable Length Character and Word n-grams. In *Clef*.

Neville, C., & Anastasio, P. (2019, 4). Fewer, Younger, but Increasingly Powerful: How Portrayals of Women, Age, and Power Have Changed from 2002 to 2016 in the 50 Top-Grossing U.S. Films. *Sex Roles*, *80*(7-8), 503–514. doi: 10.1007/s11199-018-0945-1

Odin, R. (2008). Towns in Search of Identity. In *We europeans?: Media, representations, identities* (1st Edition ed., pp. 43–57). Bristol: Intellect Books.

Petzel, M., & Wehnert, J. (2002). *Das neue Lexikon rund um Karl May - Leben, Bücher, Filme, Fans. Von der Wüste zum Silbersee: Der große deutsche Abenteuer-Mythos*. Berlin: Schwarzkopf & Schwarzkopf.

Ray, K. (2020). *Gender Portrayal in Marvel Cinematic Universe Films: Gender Representation, Moral Alignment, and Rewards for Violence* (Unpublished doctoral dissertation). Brigham Young University.

Russ, J. (1985). Pornography by women for women, with love. *Magic mommas, trembling sisters, Puritans and perverts: Feminist essays*, 79–99.

Schmidt, T., Grünler, J., Schönwerth, N., & Wolff, C. (2021, 9). Towards the Analysis of Fan Fictions in German Language: Exploration of a Corpus from the Platform Archive of Our Own. In *2nd international conference of the european association for digital humanities (eadh 2021).* doi: 10.5283/epub.50829

Schmidt, T., Hoffmann, J., & Wolff, C. (2022, 6). Analyzing Character Networks in Crossover Fan Fictions of Archive of Our Own. In *Workshop on computational methods in the humanities 2022 (comhum 2022).* Lausanne, Switzerland.

Scodari, C., & Felder, J. L. (2000, 9). Creating a pocket universe: "Shippers," fan fiction, and the X-Files online. *Communication Studies*, *51*(3), 238–257. doi: 10 .1080/10510970009388522

Scott, S. (2013). Textual Poachers, Twenty Years Later: A Conversation between Henry Jenkins and Suzanne Scott. In H. Jenkins (Ed.), *Textual poachers* (2nd ed., p. vii-vii). London: Routledge. doi: 10.4324/9780203114339

Soulliere, D. M. (2006, 7). Wrestling with Masculinity: Messages about Manhood in the WWE. *Sex Roles*, *55*(1-2), 1–11. doi: 10.1007/s11199-006-9055-6

Stanfill, M. (2011, 11). Doing fandom, (mis)doing whiteness: Heteronormativity, racialization, and the discursive construction of fandom. *Transformative Works and Cultures*, *8*. doi: 10.3983/twc.2011.0256

Steven Bird, Ewan Klein, & Edward Loper. (2009). *Natural Language Processing with Python*. Melbourne & Edinburgh: O'Reilly Media, Inc. doi: 10.1162/coli_r_00022

Tjong Kim Sang, E. F., & De Meulder, F. (2003, 6). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the seventh conference on natural language learning at hlt-naacl 2003* (pp. 142–147). Retrieved from `https://aclanthology.org/W03-0419`

Tosenberger, C. (2008). Homosexuality at the Online Hogwarts: Harry Potter Slash Fanfiction. *Children's Literature*, *36*, 185–207. Retrieved from `http://resolver.scholarsportal.info/resolve/00928208/v36inone/185_hatohhpsf`

Towbin, M. A., Haddock, S. A., Zimmerman, T. S., Lund, L. K., & Tanner, L. R. (2004, 6). Images of Gender, Race, Age, and Sexual Orientation in Disney Feature-Length AnimatedFilms. *Journal of Feminist Family Therapy*, *15*(4), 19–44. doi: 10.1300/J086v15n04_02

Vilares, D., & Gómez-Rodríguez, C. (2019, 6). Harry Potter and the Action Prediction Challenge from Natural Language. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 2124–2130). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1218

Yoder, M., Khosla, S., Shen, Q., Naik, A., Jin, H., Muralidharan, H., & Rosé, C. (2021). FanfictionNLP: A Text Processing Pipeline for Fanfiction. In *Proceedings of the third workshop on narrative understanding* (pp. 13–23). Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.18653/v1/2021.nuse-1.2

Zhang, W., Kit Cheung, J. C., & Oren, J. (2019, 7). Generating Character Descriptions for Automatic Summarization of Fiction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 7476–7483. doi: 10.1609/aaai.v33i01.33017476

A. Appendices

Data-Source-Comparison

Erklärung zur Urheberschaft

Ich habe die Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie alle Zitate und Übernahmen von fremden Aussagen kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt.

Die vorgelegten Druckexemplare und die vorgelegte digitale Version sind identisch.

Regensburg, Abgabetermin der Arbeit

Signature

Erklärung zur Lizenzierung und Publikation dieser Arbeit

Name: Jonathan Sasse

Titel der Arbeit: *Acquisition of a German Fan Fiction Corpus and Analysis in the Context of Gender Representation*

Hiermit gestatte ich die Verwendung der schriftlichen Ausarbeitung zeitlich unbegrenzt und nicht-exklusiv unter folgenden Bedingungen:

- ❒ Nur zur Bewertung dieser Arbeit
- ❒ Nur innerhalb des Lehrstuhls im Rahmen von Forschung und Lehre
- ☒ Unter einer Creative-Commons-Lizenz mit den folgenden Einschränkungen:
  - ☒ BY – Namensnennung des Autors
  - ❒ NC – Nichtkommerziell
  - ❒ SA – Share-Alike, d.h. alle Änderungen müssen unter die gleiche Lizenz gestellt werden.

(An Zitaten und Abbildungen aus fremden Quellen werden keine weiteren Rechte eingeräumt.)

Außerdem gestatte ich die Verwendung des im Rahmen dieser Arbeit erstellten Quellcodes unter folgender Lizenz:

- ❒ Nur zur Bewertung dieser Arbeit
- ❒ Nur innerhalb des Lehrstuhls im Rahmen von Forschung und Lehre
- ❒ Unter der CC-0-Lizenz (= beliebige Nutzung)
- ☒ Unter der MIT-Lizenz (= Namensnennung)
- ❒ Unter der GPLv3-Lizenz (oder neuere Versionen)

(An explizit mit einer anderen Lizenz gekennzeichneten Bibliotheken und Daten werden keine weiteren Rechte eingeräumt.)

Ich willige ein, dass der Lehrstuhl für Medieninformatik diese Arbeit – falls sie besonders gut ausfällt - auf dem Publikationsserver der Universität Regensburg veröffentlichen lässt.

Ich übertrage deshalb der Universität Regensburg das Recht, die Arbeit elektronisch zu speichern und in Datennetzen öffentlich zugänglich zu machen. Ich übertrage der Universität Regensburg ferner das Recht zur Konvertierung zum Zwecke der Langzeitarchivierung unter Beachtung der Bewahrung des Inhalts (die Originalarchivierung bleibt erhalten).

Ich erkläre außerdem, dass von mir die urheber- und lizenzrechtliche Seite (Copyright) geklärt wurde und Rechte Dritter der Publikation nicht entgegenstehen.

- ☒ Ja, für die komplette Arbeit inklusive Anhang
- ☐ Ja, für eine um vertrauliche Informationen gekürzte Variante (auf dem Datenträger beigefügt)
- ☐ Nein
- ☐ Sperrvermerk bis (Datum):

Regensburg, Abgabetermin der Arbeit

Signature