Leipzig University
Faculty of Mathematics and Computer Science
Degree Programme Computer Science

Profiling the Gender Identity of Non-binary Authors

Bachelor's Thesis

Lukas Gehrke

1. Referee: Jun.-Prof. Dr. Martin Potthast

Submission date: January 20, 2020

Declaration

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann. Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, January 20, 2020

.....
Lukas Gehrke

Abstract

Gender is one of the social traits most frequently predicted from texts in author profiling. All research in author profiling treated gender as a binary trait, only distinguishing male and female. Therefore, these studies measure sex rather than gender. But texts as source of data reflect an authors gender identity. And gender identities are not exclusively male and female. We rethink these approaches by adding transgender and non-binary authors to classical author profiling tasks. Therefore, we access data from Reddit and show how to label authors of Reddit with gender. We present a dataset of Reddit comments from which we removed a topic bias. By applying features that previously performed well on binary gender prediction to non-binary gender prediction as a multiclass-classification task, we find a common classification model to achieve solid performance. We conclude with incentives to refine dataset development and prediction setups for better performance in future works.

Contents

1 Introduction **1**

2 Related Work **3**
2.1 Computer Science 3
2.2 Linguistics 5
2.3 Psychological Diagnostics 7
2.4 Personality Inventories Measuring Gender 9

3 Data **10**
3.1 Twitter Data 10
3.1.1 Preprocessing of the Twitter Dataset 11
3.2 Reddit Data 12
3.2.1 Generating a Reddit Dataset 13
3.2.2 Preprocessing of the Reddit Dataset 20

4 Method **27**
4.1 Classification Setup for the Twitter Dataset 27
4.2 Classification Setup for the Reddit Dataset 30

5 Results **31**
5.1 Predictive Performance with the Twitter Dataset 31
5.2 Predictive Performance with the Reddit Dataset 33

6 Discussion **36**
6.1 Future Work 37

7 Conclusions **39**

Bibliography **40**

Chapter 1

Introduction

Gender, refering to *socially constructed characteristics of women and men* [59], is a topic that causes great controversy. Stars, politicians and managers have fallen in the wake of abusive behavior of men against women, with the #MeToo movement being a recent example of fighting gender-related discrimination. Non-binarity and gender diversity are other controversial topics. Societies are rethinking their understanding of gender broadening it in official documents like passports [16]. Gender is being untied from being binary male and female, with huge scenes that do not identify with the standard gender identities or with those they have been assigned at birth. People connect on plattforms such as Twitter or Reddit in order to talk about gender-related topics.

The concept of gender has also made a huge impact on the scientific world. With gender studies, gender has formed its own interdisciplinary scientific field [51]. Social scientists, psychologists, linguists and biological scientists conduct studies about gender and work with an understanding of gender that uses a spectrum of identities rather than a binary between masculinity and femininity. Recent studies by biologists and clinical geneticists substantiate the concept of gender's variety by proposing a spectrum in sex characteristics, ranging from male to female [3].

Computer science differs from other scientific disciplines regarding its treatment of gender. Gender has been inferred from texts in *author profiling*, the field of predicting demographics of authors from texts with statistical learning models. Performance of prediction whether an author is a man or a woman has even become a benchmark for new methods. Most studies use only male and female, as these are easily accessible, but this does not meet a modern, diverse understanding of gender. To be precise, scientists do not really predict gender. They predict sex rather than gender, if they classify authors as male and female. By doing so, the tasks in author profiling confuse what is actually measurable: Language as data source does not reflect an authors sex. Language

does more likely reflect authors expectations of how to behave according to their gender [10]. But by following the idea that gender is not binary, it should be measured with more than two classes. This would be a multi-class classification task using for example male, female, non-binary and transgender as classes. Gender could even be measured as a spectrum where all individuals would have an associated degree of masculinity and femininity. In computer science, this would be a regression task. We claim that tasks like these, untied from gender binary, would truly match the concept of gender and thus match what is actually measurable in text: Gender, not sex.

This thesis introduces a more diverse perspective on gender in computer science. The contributions are as follows: (1) Firstly, we survey the previous study of gender and language in computer science, linguistics and psychology. (2) Afterwards, we process an already existing dataset consisting of data from the microblogging plattform Twitter and use texts from Reddit to create new datasets with a focus on gender identities different from male and female. (3) Lastly, we propose and evaluate models to classify authors in diverse genders and provide experimental evidence for the non-binarity of gender expression in language use by comparing prediction performances in different class-setups. Thereby this thesis aims to extend state-of-the-art methodologies in computer science with an up-to-date understanding of gender.

The Twitter dataset for this thesis contains 33,881 authors of mostly male and female gender. Only a small fraction of 77 authors are transgender or of non-binary gender identity though. Classification of these authors performed poorly, as they tended to be misclassified as female. Afterwards, we created a new dataset. The task of creating a relatively balanced dataset with authors of all the aforementioned gender identities is not trivial. There are few reliably labeled corpora containing high fractions of transgender or non-binary individuals. We used dumps of Reddit as source of data and collected labeled text with two methodologies of matching authors with gender: Distant supervision which searches texts written by authors for self-reportings of gender and labeling by author *flairs*, user tags on Reddit which sometimes contain gender information. From the self-created corpus, we got around 6.533 out of 84,619 authors that identify as non-binary or transgender. When classifying these in a five-class setup, we yielded respectable performance with an F1-score of `0.47` compared to a random baseline performance of `0.21`.

Chapter 2

Related Work

Many disciplines contributed to the modern understanding of gender, which is generally defined as the social component of sexual differences [59]. This definition is quite young in comparison to other traits of personality in social sciences. Its origins lie in psychological studies from the 1950s. Money et al. [28] untied the term gender from its previous purpose of marking grammatical categories. They defined gender as the behavioral side of differences between the sexes to oppose the biological side which includes the anatomy of individuals, their reproductive system and secondary sex characteristics [28]. Notably, Money did not limit his concept of gender to masculinity and femininity by describing gender as *'ones experience of individuality, behaviors (...) as clearly and unlimited male, female or ambivalent on a smaller or larger scale'* [28]. During the following decades, the concept of gender gained popularity when sciences [19], societies and individuals started to turn their attention to it. The World Health Organization defined gender as *'socially constructed characteristics of women and men - such as norms, roles and relationships of and between groups of women and men'* and also even stated that gender influences health, underlining its meaning today [59].

2.1 Computer Science

With the rise of automated text categorization during the 1990s, especially using machine learning, scientists started training learning models with different representations of texts, hoping that their model would learn real-world differences and generalize to new, previously unseen data [53]. The discipline of machine learning-supported author profiling emerged, focused on predicting social variables such as age, gender, native language or personality from text [40]. Early approaches that targeted gender classification of texts by Koppel et al. [22] used lexical and syntactic features including a list of function words

and part-of-speech n-grams to successfully classify unknown authors as male or female. Other pioneering works in author profiling used more different linguistic variables such as word-groups representing emotions [37]. Notably, Schler et al. [50] identified two main categories of textual features to use for author profiling: Content-related features and style-related features, reflecting that authors write about different topics (content) and that authors write about the same topics differently (style) [6]. Table 2.1 provides an overview of corpora, labeling strategies and features being used in these studies and similar ones. Rao et al. [45] used n-gram features weighted by term frequency and introduced usage of *socio-linguistic features* such as occurrences of *OMG* ('Oh my god'), *Honorifics* ('Dude', 'Bro') *Ellipses* ('...'), *Excitement* ('!!!'), *Repeated Alphabets* ('Whaaat') or *Laugh* ('LOL') in text, which are similar to style features. They found out that women use more ellipses, excitement and alphabetic character repetition than men while men use more honorifics. Burger et al. [12] used n-grams on texts of different languages such as English, Portuguese and Spanish. Marquardt et al. [27] worked with content-based features extracted with tools and collections: The MRC (Medical Research Council) machine usable dictionary contains a collection of words and 26 categories of linguistic and psycholinguistic attributes. These include for example *familiarity*, *concreteness* and *imagery* and are measured with frequencies of associated words [62]. LIWC (Linguistic Word Count and Inquiry) is another collection which counts words in psychologically meaningful categories such as motion, anger or certainty in order to deduce psychological correlates such as emotional stability or agression [36]. Sentiments, representing the degree of negativity and positivity in text, were used as well [27]. Style features used by Marquardt et al. [27] included readability scores, HTML tags, emoticon usage and rates of spelling- and grammatical errors. Fatima et al. [17] profiled facebook authors in English and Roman Urdu language and generated language independent style-based features such as average sentence lengths, average paragraph length, number of sentences or vocabulary richness. As content-based features, n-gram models with the help of feature selection methods such as Information Gain were used [17].

The largest shared task on author profiling takes place at PAN[1] with many submitted papers for author profiling tasks every year. Although many different models were tested there, best approaches were n-grams of characters and words [42].

Few studies deal with author profiling based on Reddit data so far. In most of them, corpus creation plays a major role. Burkhart [13] extracted around 100,000 labeled usernames by processing a Reddit BigQuery database [18]. He

[1] https://pan.webis.de

used regular expressions on self-assigned labels of Reddit authors, so-called *flairs*, to obtain gender labels. These were used to generate an overview of gender distribution over Subreddits. Vasilev [58] also focused on flairs and extracted labeled users and their Reddit comments from the same BigQuery database, yielding around 305,000 users with gender information. He also used submission titles from subreddits, topic-specific sub-sites of Reddit, that require demographic user information for labeling such as `/r/relationships`. Similar approaches yielded 8,592 [14] or 40,806 [32] gender labeled users. When it comes to gender prediction, the aforementioned approaches used classifiers such as logistic regression as well as neural networks. Gender was treated as male or female throughout all these studies.

Before computer scientists started working on texts and possibilities of extracting social variables like age or sex from it, research in fields like psychology, linguistics and social sciences had been conducted.

2.2 Linguistics

Special markers of gendered language were studied by Lakoff [23] who claimed that women's language and language towards women reflects powerlessness and marginality. In order to proof this hypothesis, he analysed usage of adjectives or particles and grammatical constructs in sentences. Findings were that women use more adjectives that indicate admiration or fewer swearing words than men. Mulac and Lundell [30] used linguistic variables to regress individuals genders. They found features like *Impersonals* ('it', 'there are'), *Fillers* ('okay', 'like', 'well') or *Elliptical Sentences* ('Just beautiful'; sentences missing either subject or predicate) to correlate with male authors and *Intensive Adverbs* ('really beautiful'), *Personal Pronouns* ('I', 'we') and *Negations* ('It doesn't look windy') to correlate with female authors. Other linguistic studies focus on the relation of gender and language in different contexts and also often include detailed explanations of the linguistic variables used [31].

More recent studies question the direct influence of gender on language. Herring and Paolillo [21] analyzed weblogs with style-based features that had previously proven to distinguish men and women in author profiling. The features were used to predict gender and blog genre, resulting in significant predictive performance for blog genre, but not for gender. Herring and Paolillo concluded that blog-genres and their requirements influence linguistic choices of authors regardless their gender. Bamman et al. [7] proposed an approach that goes beyond prediction of binary gender. They clustered texts to identify groups of authors that use similar words. Some of the resulting clusters were clearly gender-dominated, while others were homogenous regarding the topic

Table 2.1: This Table gives an overview of datasets and features used for gender prediction in related work.

Contributor	Dataset	Authors & Labels	Labeling Method	Features for Gender Prediction (Overview)
Misc				
Koppel et al. [22, 2002]	566 documents	566 male/female	pre-labeled data	Part-of-speech (POS) n-grams, function words
Schler et al. [50, 2006]	37,478 blogs, 295,526,889 words	37,478 m/f	pre-labeled data	style features, content features
Newman et al. [33, 2008]	14,324 text files, approx. 45,700,000 words	11,609 m/f	pre-labeled data	Linguistic Inquiry and Word Count (LIWC) words
Argamon et al. [6, 2009]	avg. 7,250 words/author	19,320 m/f	pre-labeled data	style features, content features
Rao et al. [45, 2010]	405,151 tweets	1,000 m/f	manual labeling	socio-linguistic and n-gram features
Mukherjee and Liu [29, 2010]	3,100 blog posts	3,100 m/f	pre-labeled data	stylistic features, gender-preferential features, POS n-grams
Burger et al. [12, 2011]	4,102,434 tweets	183,729 m/f	following links to labeled blog profiles	character and word n-grams
Schwartz et al. [52, 2013]	15,4 Mio facebook messages	approx. 75,000 m/f	asking users for labels	LIWC words
Bamman et al. [7, 2014]	9,212,118 tweets	14,464 m/f	labeling by most-likely gender of profile user name	10,000 most frequent lexical items in dataset
Fatima et al. [17, 2017]	facebook posts, avg 2156 words/author	479 m/f	asking users for labels	style features, content features
PAN				
Rangel et al. [40, 2013]	blog posts	346,100 m/f	pre-labeled data	various (mostly style features, content features; n-grams)
Stamatatos et al. [55, 2014]	hotel reviews, tweets, blog posts	490 twitter authors	pre-labeled (blogs), manual labeling (tweets)	various (see PAN 2013)
Rangel Pardo et al. [44, 2015]	tweets	1070 m/f	asking users for labels	various (see PAN 2013)
Rangel et al. [41, 2016]	tweets, up to 1,000/author	428 m/f	manual labeling	various (see PAN 2013)
Rangel et al. [42, 2017]	tweets, 100/author	19,000 m/f	manual labeling and automatic labeling with dictionary of proper nouns	various (see PAN 2013), including deep learning techniques
Rangel et al. [43, 2018]	tweets, 100/author	12,600 m/f (4,900 English)	see PAN 2017	various (see PAN 2013), including deep learning techniques
Wiegmann et al. [61, 2019]	156,4 Mio tweets	33,836 m/f/**non-binary**	linking wikidata profiles	n-grams range 1-4
Reddit				
Muller [32, 2018]	10,7 Mio comments	40,806 m/f	flairs and submission titles	n-grams range 1-3
Vasilev [58, 2018]	193 Mio comments	305,000 m/f	flairs and submission titles	characters and words
De Pril [14, 2019]	500 comments/author	8,592 m/f/nb	flairs and submission titles	words used by 5 different users, graphbaed features

the contained authors write about or some habiual attitudes rather than gender. They concluded that social-networks of authors influence selection of language markers, so that often *'gender emerges as individuals position themselves relative to audiences, topics, and mainstream gender norms'* [7].

2.3 Psychological Diagnostics

Psychology as the science of mind and behavior [4] has its own subfield dedicated to measuring social phenomena, one of these being gender. It is called psychological diagnostics and includes the creation of so-called personality inventories. We analyzed methodology for creating personality inventories in order to find out, which criteria psychologists use to select texts for gender measurement. The goal was to know what text would reflect gender differences best and how to obtain or generate this kind of text for our own dataset.

According to Aiken [2] a personality inventory consists of items that measure *'personal characteristics, thoughts, feelings, and behavior'*. In contrast to other means of determination in psychological diagnostics like rating scales or checklists, personality inventories are constructed with higher quality standards and measure a variety of personality variables at once [2]. Personality inventories are used for pathological purposes, where they detect mental illness, behavioral disorders or personally and socially destructive behaviors [2]. Additionally, there are inventories that focus on healthy individuals and measure positive features and coping behaviors which are used for scientific purposes, *'guidance, (...) personal development, and applicant selection'* [2].

The assessment instruments used by personality inventories are so-called items. Items can be questions or statements a respondent has to react to. For inventories, items mostly are in single or multiple choice form and require participants to respond accordingly with one out of two answers (true-false, yes-no, agree-disagree) or one out of mulitple answers, so-called Likert scales (completely false, mostly false, partly false, neutral, partly true, mostly true, completely true) [2]. This format makes items easily and quickly to answer to and enables inventories to cover a wide range of topics [2].

According to Aiken, items are constructed by three different approaches. The first one is called the *rational-theoretical strategy* and involves *'reasoning and theory'* [2]. In this approach, the design of an inventory begins with a definition of personality constructed from different sources, for example *'(...) common sense, research findings, professional judgments and theories of personality'* [2]. Definitions are also derived from particular concepts, such as *'depression, (...) and self-esteem'* [2].

When developing an inventory that measures gender identity, according to

Rentzsch and Schütz [48] one would start by defining personality and deduce definitions, first for gender identity in general and then different gender identities. In the next step, items that measure each gender identity would be created according to these definitions [48].

The second method according to Aiken is the *criterion-keying strategy*. Items are selected if their answers or scores distinguish between pre-defined criterion groups [2]. In relation to the target concept of an inventory, these criterion groups might be *dichotomous* - which means two-dimensional - groups such as patients in a psychiatries versus healthy people or musicians versus non-musicians in general [2].

To generate a gender identity inventory, the *criterion-keying strategy* would start with a collection of items that might be useful to differentiate gender identities according to Rentzsch and Schütz [48]. These items would be answered by test-participants which are clearly divided into the criterion groups *male*, *female*, *non-binary*, and so on, so that the same items would be answered by individuals with each gender identity. Then, the inventory would be generated with those items that have been answered most differently by the criterion groups [48].

A third approach for generating personality inventories according to Aiken is the *factor-analytic strategy*. By using this strategy a group of items with high intercorrelation is created [2]. This method helps finding social variables if these originally are not existent beforehand. If the correlations of different internally consistent scales with each other are low, these scales represent social variables. This strategy is based on the concept of factor analysis [2]. It is rather exploratory and therefore used if no theories or test participants for evaluation are at hand [48].

In order to develop an inventory with the *factor-analytic strategy*, one would have to collect items and data of test-participants of different genders that potentially work for measuring gender identity first. In a second step, factors or groups of items, each representing a gender identity, would have to be deduced. Therefore, factor analysis would be used to determine internally homogenous groups of items. These groups would reflect differences in-between the results from the test-participants.

When discussing how these approaches might help to generate a dataset for gender prediction, the *criterion-keying strategy* appeared most useful to us. In order to generate a dataset, we slightly modified its methodology: We identify as many authors with their respective gender identities as possible and collect texts written by them. The different gender identities of the authors are the criterion groups, their texts are a collection of representations of their behavior. Before measuring gender in these texts, we should remove other structures inside this data, which potentially differentiate the groups, such

as topics [7]. Afterwards, discriminative features are extracted, similarly to most-discriminative items for an inventory. Finally, the performance of a machine learning model is used to determine the ability to measure gender in the generated data.

2.4 Personality Inventories Measuring Gender

Sex and gender play a major role in some psychological inventories. The Minnesota Multiphasic Personality Inventory (MMPI) measures stereotypical masculine or feminine behavior among other traits . The degree of stereotypical masculine or feminine behavior is determined using questions that have to be answered on a five point Likert-scale [20]. For the MMPI's construction, the *criterion-keying strategy* was used.

Sandra Bem's Sex Role Inventory (BSRI) is an example for a test constructed with the *rational-theoretical strategy*. It measures identification with stereotypical male or female behavior as well as sexual androgyny. The BSRI uses items that were supposed to represent behavior typical for men and women in the United States at the time of its creation, the 1970s [9]. Bem based her work on the assumption that men have an instrumental orientation, a focus on *'getting the job done'*, while women would have an expressive behavior, *'an affective concern for the welfare of others'* [9]. A pre-selection of items that were constructed with those assumptions were judged by students of Bem and those items that were not rated as desirable for one sex by both male and female judges independently were used for the androgyny score [9]. Items, that both groups agreed on were used for the masculinity or femininity score. The Bem Sex Role Inventory contains 60 items, 20 for male, 20 for female and 20 for androgyny which have to be answered on a seven-point Likert scale. The evaluation of the BSRI assigns every participant a masculinity score, a femininity score and an androgyny score [9].

Building on top of Bem's Sex-Role Inventory is the Open Sex Role Inventory (OSRI) [34]. The OSRI was created with the *criterion-keying strategy*, using a list of items that show a huge gender difference in answers of test participants [34]. From an original list of 2,610 items a selection of 44 items remained in the final inventory [34]. The OSRI measures masculinity, femininity and androgyny with question items such as *'I have studied how to win at gambling'*, *'I save the letters I get'* or *'I like guns.'* [34].

Chapter 3

Data

In order to measure authors genders with text and form valid conclusions about the relation of gender and language, first a collection of texts, labeled with its authors genders, was required. The sources of data used for this thesis are two-fold. On the one hand, we used an existing dataset containing posts from the micro-blogging platform Twitter[1]. The social news aggregation and web content rating site Reddit[2] war our second source. In this chapter we explain the structure and processing of an already existing dataset of Twitter data and explain the creation of a Reddit dataset.

3.1 Twitter Data

The first part of our research is based on a corpus of tweets. Twitter allows users to post messages with a maximum length of 280 characters that will then be displayed to all followers of a user. Until November 2017, post length was restricted to 140 characters [49].

The Webis Celebrity Profiling Corpus

The Twitter corpus used throughout this thesis is the Webis Celebrity Profiling Corpus [61]. It was constructed from a crawl of all verified profiles on Twitter as of May 2018. These profiles have a blue checkmark assigned by Twitter, confirming that the profile is of public interest and authentic at the same time [57]. The originally crawled 297,878 Twitter accounts were linked to their respective Wikidata[3] entries which contain demographic data about the authors. Because many of those lack a profile on Wikidata, only 71,706 Twitter feeds

[1] `https://twitter.com`
[2] `https://www.reddit.com/`
[3] `https://www.wikidata.org/wiki/Wikidata:Main_Page`

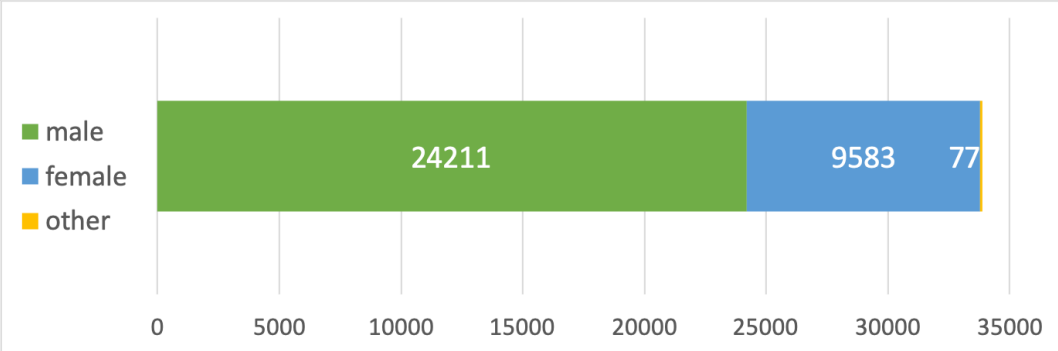

Figure 3.1: Distribution of genders in our version of the Webis Celebrity Profiling Corpus, *'other'* contains transgender authors as well as authors with non-binary gender identity.

remained as valid matches [61]. This thesis used another subset of the corpus, containing only 33,836 authors as dataset. The same subset was also used in the Celebrity Profiling Task at PAN @ Clef 2019[4]. The dataset contains 24,211 male, 9583 female and 32 non-binary authors and their Twitter feeds. We extended it with 45 more non-binary and transgender authors from the original corpus and yield an overall distribution as observable in Figure 3.1.

3.1.1 Preprocessing of the Twitter Dataset

Before using a machine learning model, the twitter dataset was preprocessed in order to remove information that misleads the learning process or that harms computational performance without having any informational value. Multiple steps were performed. Table 3.1 gives an overview of the preprocessing steps. Not-English feeds and single not-English tweets were removed by using the library langdetect [54]. Not-English feeds were defined as authors, whose language in all their tweets concatenated as detected by langdetect was not English. Additionally, single tweets of mostly English-writing authors that were not in English as detected by langdetect, were removed. Afterwards, retweets, tweets beginning with `RT`, were removed as these obviously do not represent an authors own text. We replaced mentions, links and hashtags with `<user>`, `<url>`, `<hashtag>` using regular expressions. These are elements of tweets that refer to different content or users and are not helpful as n-grams, as they are often unique. The occurrence count of these placeholders was used as a more stylistic n-gram. Finally, lower-casing was applied to all tweets. We did not remove stopwords since their usage differs between different genders [5]. Thus they are informational for the classification of tweets.

[4] `https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html`

Table 3.1: Overview of the numbers of tweets removed or affected during preprocessing.

Description	**Number of Tweets**
Feeds overall	33,881
Tweets overall	49,747,895
Avg. Tweets per user	1,468
Removal	
Not-English feeds	1,967
Not-English Tweets	161
Retweets	19,974,097
Replacement	
URLs	23,618,209
User mentions	41,774,076
Hashtags	19,414,743

3.2 Reddit Data

Reddit proclaims itself as *the frontpage of the internet*, has been launched in 2005, and ranks as the sixth of the most visited websites in the United States as of November 2019[5]. On Reddit, millions of users, also called *redditors*, discuss topics they are interested in and rate posts and comments of others. Posts can be links, texts or media content such as pictures or videos. Those are commented and up- or down-voted by other users with the most up-voted posts making it to Reddit's front-page temporarily, where all visitors of Reddit will eventually see them [60]. Users are able to create so-called *subreddits* around topics they are interested in. Subreddits mostly are named after their topic, beginning with *'/r/'* [60]. Some very active users are promoted to be moderators, which oversee the activity in their subreddits and ensure that its rules are obeyed by all redditors [56]. There are thousands of subreddits including ones about news, music, sports, science, painting, photography, technology, movies, video-games or social topics such as relationships or genders.

The fact that there is lots of activity in subreddits like */r/NonBinary*, */r/askTransgender* or */r/genderqueer* makes Reddit an excellent source to collect text-data of gender identities different from male and female. In many of those subreddits, users are allowed to assign themselves tags, so-called *flairs* [47]. Often, these contain gender information about users and thus can be used to reliably label redditors with gender [14]. Additionally, Reddit comments are not limited in length, in contrast to Tweets, and thus provide more text.

[5] `https://www.alexa.com/topsites/countries/US`, retrieved on 4th Nov. 2019

3.2.1 Generating a Reddit Dataset

In order to create a dataset of gender-labeled Reddit users and their texts, we had to find ways of accessing Reddit data first. Afterwards, the data had to be processed in order to label users with their respective gender identities. This thesis presents two major approaches of accessing texts from Reddit to generate a dataset: By using the official Reddit API and by using the Reddit dumps collected by the website *pushshift.io*.

Scraping with the Reddit API

The Reddit API[6] is supposed to be used by developers in order to build applications based on Reddit data [46]. It enables acquisition of for example redditors, comments and submissions via API endpoints. Libraries such as PRAW for Python work as API wrappers and generate instances from returned API data [38]. Keys, that are contained in every Reddit application, are used by the wrapper to authenticate with the API. The possible way of gender-labeling with the Reddit API we propose uses the `Comment` instance of PRAW. It contains a property called `author_flair_text` which includes the flair text the author of the comment has assigned himself in the subreddit it was posted in. This structure allows to obtain a set of labeled authors from subreddits using gender-revealing flairs. Therefore for labeling, two steps are required: First authors are labeled with gender by processing the `author_flair_text` property of `Comment` instances with regular expressions. In the second step, the `Author` instance is used to obtain all comments of the labeled authors, regardless their subreddits. Notably, Reddit's Api limits scraping to the 1,000 most recent instances, e.g. the 1,000 most recent comments of a user [25].

Reddit Dumps from pushshift.io

In contrast to the limitations of the Reddi API, Pushshift.io provides the complete Reddit activity of authors, only limited by time. Pushshift is a self-proclaimed place to *'Learn about Big Data and Social Media Ingest and Analysis'* [39]. On its homepage, different statistics about Reddit-activity are shown which are live updated every second. But most interestingly, the site provides access to Reddit-dumps[7] of comments, submissions or authors on a monthly basis. Similarly to the `Comment` instance of PRAW, comments from pushshift contain an `author_flair_text` property. We used the most recent 6 months of Reddit comments available on Pushshift as of November 2019 to obtain gender-labeled authors and their texts. Table 3.2 gives an overview

[6] `https://www.reddit.com/dev/api`
[7] `https://files.pushshift.io/reddit/`

Table 3.2: Overview about the Reddit dumps from pushshift we use with time, size and amount of comments.

Dataset	Size in GB	Comments
RC_2018-12 (Dec. 2018)	139.1	121,953,600
RC_2019-01 (Jan. 2019)	147.0	129,386,587
RC_2019-02 (Feb. 2019)	139.4	120,645,639
RC_2019-03 (Mar. 2019)	160.5	137,650,471
RC_2019-04 (Apr. 2019)	162.9	138,473,643
RC_2019-05 (May 2019)	168.1	142,463,421
Total	971.0	790,573,361

of sizes and amounts of comments in the dumps. Overall, the six offered dumps contain 790,573,361 Reddit comments. This work prefered the dumps from pushshift over data from the Reddit API, because the dumps contain all comments of users in their respective period of time.

Labeling Reddit Users

Using the six pushshift dumps, we labeled as many Reddit users with gender as possible and obtained all of their comments. The following gender groups were used: *male*, *female*, *transgender male-to-female* ('mtf'), *transgender female-to-male* ('ftm'), and *other*, which includes *non-binary*, *genderqueer*, *genderfluid*, and *agender*.

Gender labeling for our work was done by processing *flairs* of redditors with regular expressions [13] [14] [32] [58] and by applying *distant supervision* [15] to comment-texts.

Labeling by User Flair

To obtain a set of gender-labeled users by flair, we first collected subreddits that allow or engage their users to set flairs with gender information. Table 3.3 contains an overview of subreddits, example flairs and rules for setting flairs we have identified. The collection of subreddits was created by manually searching for subreddits about gender topics and exploiting references between related subreddits. Many subreddits link to other, associated ones in the info box on their main page. Afterwards, we retrieved all comments belonging to any of the identified subreddits from the the six dumps provided by pushshift. Finally, we extracted author gender from the comment-flairs by applying regular expressions which can be found in Table 3.4 to the `author_flair_text` property and created a collection of author names and respective gender identities.

Table 3.3: Subreddits that have been identified as using gender-revealing user flairs.

Subreddit	Flairing Rules	Example Flair
	Female subreddits	
/r/askwomen	*'flair for men, women, trans folks, and gender neutral people'*	♀
/r/askwomenover30	male, female with ages; transgender; flairing recommended	♀female 46-49
/r/askwomenadvice	♂, ♀, transgender, gender neutral	♀
	Male subreddits	
/r/askmen	male, memale, transgender, agendered, 'Bane'	♂
/r/askmenover30	male, female with ages; transgender; flairing recommend	♂male 30-34
	Other or transgender subreddits	
/r/asktransgender	Free text flair; Variety of Gender Identities with flags	22 MtF // HRT 7/27/17 // Dallas
/r/transgender	Free text flair; Transgender Question, Transgender, Genderqueer	trans mtf
/r/genderqueer	Free text flair, Flags with optional text	GQ bisexual
/r/nonbinary		nb trans guy
/r/nonbinarytalk	Free text flair	nonbinary
/r/traaaaaaannnnnnnnnns	Colors, free text	Closeted MtF
/r/lgbteens	Flags with free text inside	16 // F // girls
/r/ainbow	Flairing encouraged; Flags with free text	trans and bi
/r/ennnnnnnnnnnnbbbbbby	Different colors with free text	they/them nb
	No predominant gender	
/r/40something	Flairing of sex as male, female or neutral with age	♂48
/r/sexover30	Flairing of sex as male, female or non-binary and age encouraged	♂31 gentle giant
/r/datingoverthirty	Flairing of sex as male, female or neutral with age	♀35
/r/relationshipsover35	Flairing of sex as male, female or neutral with age	♀35
/r/wellnessover30	Free text flair	47, f, inherently lazy
/r/okcupid	Username, age, gender, profile name	M/24/Phillyyy
/r/keto	Free text flair	27/M/5'11 SW:350 CW:295 GW:200
/r/childfree	Free text flair	32 F AU Tubal
/r/xxketo	Free text flair	5'5" SW: 220 CW: 168 GW:125-135?
/r/loseit	Free text flair; alternatively colored flair with amount of weight lost	24F l 5'6" l 17LBS Lost
/r/fatlogic	Free text flair	F 19 SW: 205 CW:149 GW:125
/r/financialindependence	Free text flair	31M reformed spendypants 70% LeanFI
/r/infj	Personality Type and optional Age Gender	INFJ F17
/r/100daysofketo	Free text flair	F/33/5'7" UGW: 150 KETOCHOW
/r/tall	Size in foot, centimeters and gender as male or female through flair color	6' 5" | 195 cm and color
/r/short	same as /r/tall	5'4" | 164 cm and color
/r/relationship_advice	male, female with ages or only age	Early 20s male

The regular expressions checked flair texts for simple gender-revealing strings such as `'male'`, `'woman'`, `'non-binary'`, `'trans ftm'`. Additionally, the color of flairs from subreddits such as */r/tall* were checked and `A|S|L` (Age|Sex|Location) information from subreddits such as */r/loseit* or */r/okcupid* was checked to extract male and female users, similarly to Burkhart [13].

We excluded authors that are listed in a bot-list or which were already deleted, indicated by a `[deleted]` behind their usernames, from further processing.

The resulting set of labeled authors was evaluated manually by randomly picking authors and looking at their assigned gender and the flair text used for labeling. We can't guarantee to exclude false-positives completely with our labeling approach but found no mismatches when randomly picking 100 authors and comparing their flairs with assigned genders. The regular expression processing started with trying to match every comments flair as *transgender*, then *other* and then *female*. At this point, all matchable *transgender* individuals

Table 3.4: Regular expression patterns for gender labeling by flairs.

Target	Regular Expression
Male	
Directly	`([\S\s]*|)(?<!hu)(?<!ger)(?<!ro)[\s]?(♂|man|male')`
A/S/F	`^[^/]+/(m)/[^/]+`
Color	`flair_color == 'male'`
Female	
Directly	`([\S\s]*|)(♀|woman|girl|female)`
A/S/F	`^[^/]+/(F)/[^/]+`
Color	`flair_color == 'female'`
Transgender mtf	
Directly	`([\S\s]*|)(mtf|trans(|gender)[\s](woman|gal|girl|female))`
Transgender ftm	
Directly	`([\S\s]*|)(mtf|trans(|gender)[\s](man|boy|boi|male))`
Other	
Non-binary	`([\S\s]*|)((gender|)queer|non[-]?binary)`
Agender	`([\S\s]*|)(agender|genderfuck|neut(er|rois))`
Genderfluid	`([\S\s]*|)((|gender)(|-)fluid)`

Table 3.5: Overview for preceeding phrases excluded with negative lookbehind assertions for labeling with distant supervision.

Preceeding Phrases	Main Part
`feel(s) like, where, (as) if, hoping, assume (that), think(s), thought then, (that) means, imply, think(s), tell(s) me, guess, expect(s) (that)`	`i(\'|\sa)m\s(a)`*gender_pattern** *see Table 3.4 *Target Directly*

had already been excluded and did not have to be excluded in the regular expression pattern for *female*. Finally, we extracted *male*.

Labeling by Distant Supervision

The *distant supervision* approach retrieved all comments that matched self-labeling patterns as observable in Table 3.5 from the Reddit dumps. With these labels, a second collection of author names and genders was created. The key concept of the distant supervision method is to apply regular expression patterns to the comment texts. We looked for phrases such as `'i am a boy'` and used *negative lookbehind assertions* [24] with the basic form `'(?<!`*phrase*`)'` to exclude a variety of preceding phrases that flip the meaning of self-labeling phrases such as `'I am a <gender role>'` with `'I think I am a <gender role>'` [15]. Figure 3.2 gives an overview of how many authors with the different gender identities have been identified with the two labeling approaches, while Figure 3.3 compares the distributions of text lengths of authors in both approaches. Figure 3.4 shows the distribution of amounts of words retrieved for subreddits and Figure 3.5 presents a larger scale. For a handful of very active subreddits we received far more than 1,000,000 words while the average word count for a subreddit in the combined collection was 24,796.

Figure 3.2: Comparison of gender distributions of authors retrieved with the two labeling approaches.

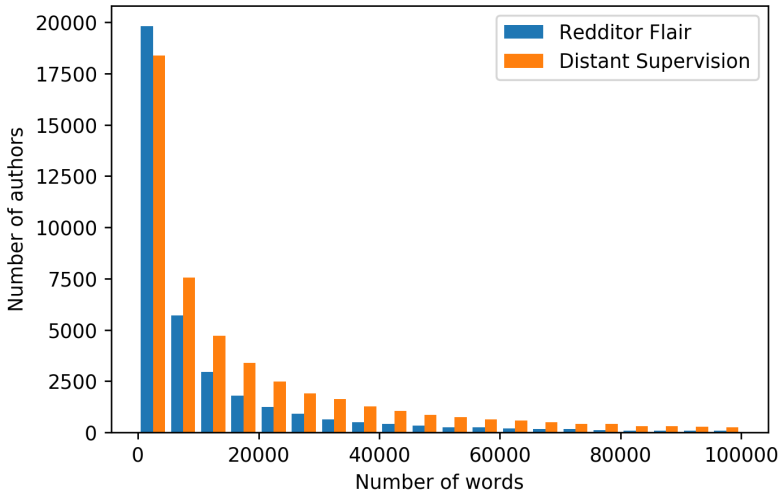

Figure 3.3: Comparison of the distributions of text lengths in words *across authors* retrieved with both labeling approaches.

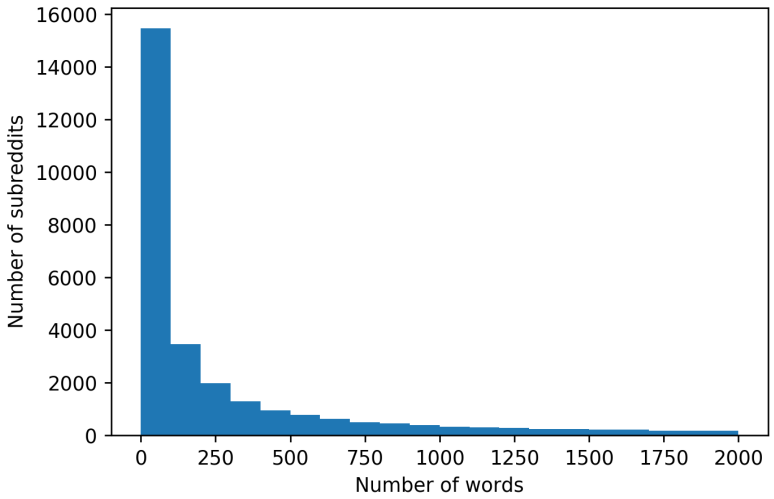

Figure 3.4: Distributions of text lengths in words *across subreddits* in the combined collection.

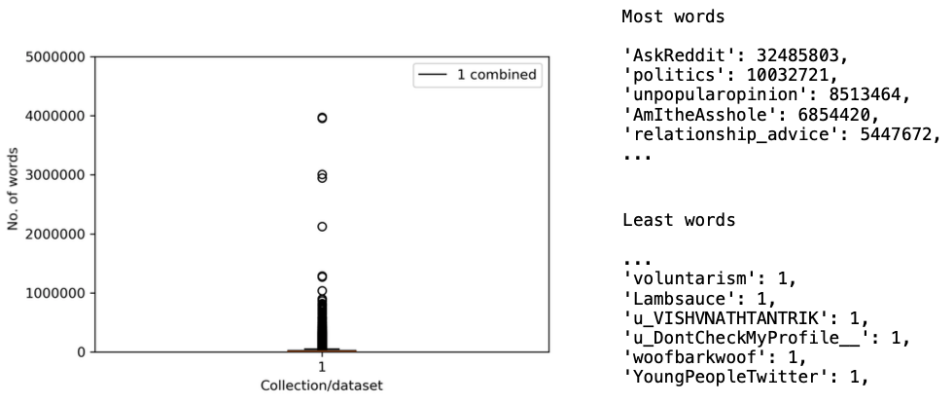

Figure 3.5: Distribution of Subreddit text lengths in the combined collection with median of 150 and average word count of 24.796 as well as most and least commented Subreddits. 5 Subreddits with more than 5,000,000 words have been excluded from the box plot.

3.2.2 Preprocessing of the Reddit Dataset

After merging the collections of authors generated with the two labeling approaches, they were checked for duplicates. All 548 authors which had been *labeled with different genders by both approaches* were removed. These authors often were labeled as *transgender male to female* or *female to male* by the flair approach, and got *female* respectively *male* assigned by the distant supervision approach. For example *transgender male to female* authors often assign themselves `mtf` as flair but self-report *female* as the gender they identify with or transitioned to in some of their comments. Other differing labels result from quotations included in the comments of authors, which mostly lead to ambiguous labeling of *male* and *female*. The 2,418 authors that had been *labeled with the same gender by both approaches*, remained in the corpus. Afterwards, the first raw version of the dataset was created by processing the pushshift dumps again and retrieving all comments written by the labeled authors, including their subreddits. Table 3.7 compares the different collections. The resulting dataset contained 84,619 authors with the following labels: 40,633 female, 37,453 male, 4,592 transgender male to female, 1,170 transgender female to male, and 771 other.

Next, the dataset was further processed for cleaning and normalization with steps as observable in Table 3.6: Removal of authors with very low and very high amounts of text, topic de-biasing, and replacement of user mentions and urls. The results are presented by Table 3.7 which compares statistics of the collections of comments retrieved with the labeling approaches with the processed dataset. We first removed authors with low amount of text. In order to define low amount of text in our collection, we used the ACL recommendation for length of Abstracts for scientific work as an orientation for a minimum number of words [1]. This lead to removal of authors with less than 200 overall words as counted with the nltk `word_tokenizer` [11]. Furthermore, authors with more than 5 times the average amount of words overall in the collection were removed, as these bare the risk of overfitting the classifier for single groups.

Additionally, we noticed a *topic bias* in the dataset, resulting from the subreddit structure of Reddit. Obviously, the subreddit of a comment influences its content. A set of comments from different subreddits bares the risk that a machine learning model learns the differences between subreddits and their topics rather than between texts with different gender identities behind them. Therefore, we attempted to balance the amounts of comments from different subreddits as much as possible. The goal was to have a relatively *even amount of text for each gender group for each subreddit.*

Table 3.6: Overview about the amount of data removed while cleaning the Reddit dataset.

Description	Amount
Raw Dataset	
Redditors overall	84,619
Comments overall	41,923,253
Removal	
Redditors with less than 200 words	3,226
Redditors with more than 106,030 words	3,124
De-Biasing	63,197
Replacement	
URLs	226,851 URLs in 16,417 comments
User Mentions	31,660 mentions in 3,652 comments
Processed Dataset	
Redditors overall	15,308
Comments overall	4,798,268

Table 3.7: Comparison of the different comment collections.

Description	Flair Collection	Distant Supervision Collection	Combined Collection	Processed Dataset
Author-level				
Redditors overall	36,412	50,626	84,619	15,308
Comments overall	12,395,733	31,697,700	41,923,253	4,798,268
Words overall	486,709,158	1,409,275,177	1,794,463,924	183,463,828
Avg. am. Comments/Redditor	340	626	495	313
Avg. am. Words/Redditor	13366	27837	21206	11984
Avg. am. Words/Comment	39	44	42	38
Subreddit-level				
Subreddits overall	44,766	63,873	73,298	34,920

Algorithm 3.1: Subreddit de-biasing

```
input: Dataset raw_dataset, list subreddits, list genders,
dict gender_counts
output: Dataset debiased_dataset
begin
  debiased_dataset = Dataset()
  for subreddit in subreddits:
    max_words = min(gender_counts[subreddit])
    for gender in genders:
      gender_words = 0
      while gender_words < max_words:
        author = raw_dataset.get_active_author(gender, subreddit)
        debiased_dataset.add(author)
        gender_words += author.get_subreddit_words(subreddit)
  return debiased_dataset
```

The basic idea to deal with the topic bias was to iterate all subreddits of the collection of authors and comments and build up a new dataset that contains relatively equal amounts of text for all gender groups for every subreddit. To do so, we applied the algorithm explained in Listing 3.1. Firstly, for each subreddit the amount of words for each gender-group in this subreddit was counted. Afterwards, the maximum amount of words for this subreddit was declared as minimum amount of words over all gender-groups for this subreddit. Then the de-biased dataset was created with the following iteration: For each subreddit, for each gender group, add as many redditors with *all their comments* to the dataset as possible, before the maximum amount of words for all groups for this subreddit is reached. Authors were pre-selected as having the subreddit in question as their most commented-in, to ensure that all authors in the resulting dataset are representative for their subreddits and not just 'one-time visitors'. The resulting dataset contains a gender distribution as observable in Figure 3.6.

The de-biasing approach used a heuristic: For simplicity, authors were added to the new dataset with *all their comments*. Therefore, in some cases the maximum amount of words for a subreddit and gender is reached or even far outreached by adding a single, very active user. A more fine-granular algorithm should add *as many comments as possible from as many different authors as possible* to the de-biased dataset until the maximum amount for a subreddit and a gender is reached. Figure 3.7 shows of the distribution of text lengths in the cleaned collection in more detail, revealing that most of the authors provide less than 5,000 words, while Figure 3.8 presents a broader overview of amounts of words per author in the different collections. Notably,

the collection of authors retrieved with distant supervision contained some authors with very high amounts of words, the top author having written more than 10 million words. Some of these authors turned out as bots that had not been filtered out. Figure 3.9 compares the distribution of comment authors for the most-commented-in subreddits of the combined collection with the processed dataset. It shows that the de-biasing resulted in higher fractions of the three smaller groups *trans mtf*, *trans ftm* and *other*, although no even distributions were achieved.

Lastly, strings that have no gender specific meaning were removed from the remaining comments. User mentions beginning with `u/` and hyperlinks were replaced with placeholders `<user>`, `<url>` by using regular expressions.

Table 3.8 contains the most frequently occurring words for each gender group in the processed dataset that only appear among the top 1,000 words for this particular groups. We noticed, that business-related words such as `market`, `total` or `000` which seems to be part of high digits and technology-related words like `speed` or `code` dominate the top-words for men, while family-related n-grams such as `husband`, `son`, `daughter` and `marriage` dominate women's top words. For transgender female-to-male people we noticed words such as `testosterone`, `penis` which `chest` indicate texts about being or becoming a man. Notably, for transgender male-to-female there were no such words among the top 1000. For non-binary authors, `non-binary` and `genders` indicated that texts were written about gender-related topics. All-in-all, these discriminating word-1-grams indicate an ongoing existence of topic-related differences, even in the processed dataset.

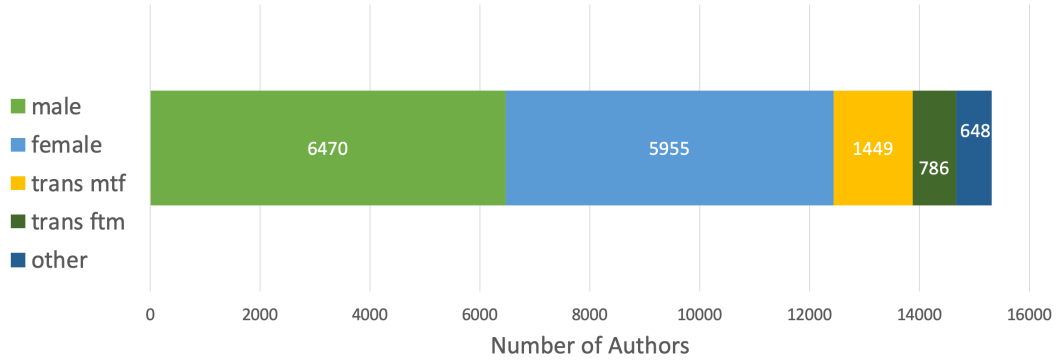

Figure 3.6: Distribution of gender labels in the processed dataset.

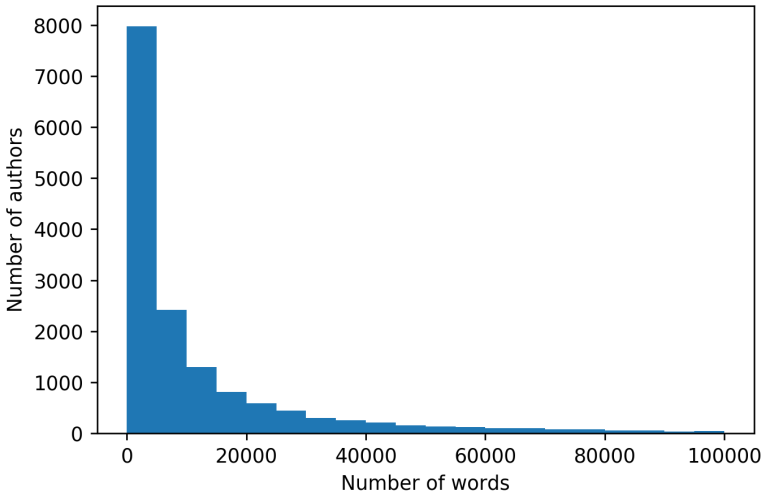

Figure 3.7: Distribution of text lengths in words *across authors* in the processed Reddit dataset.

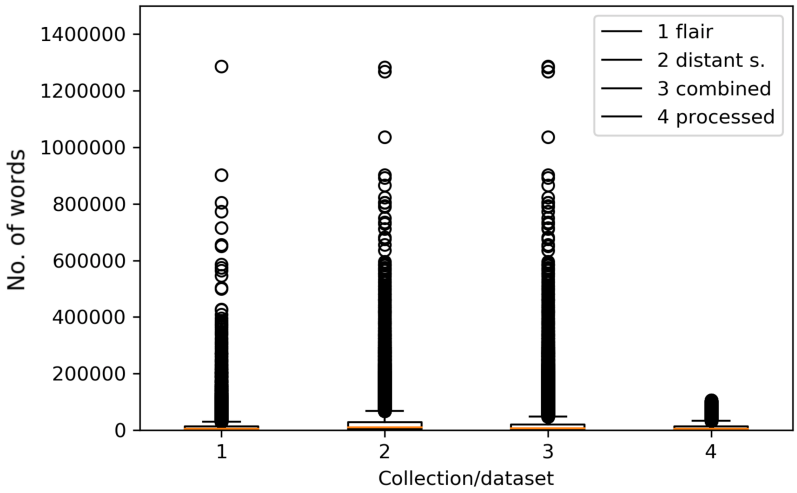

Figure 3.8: Comparison of distributions of text lengths in words *across authors* for the different labeling collections and the processed dataset.

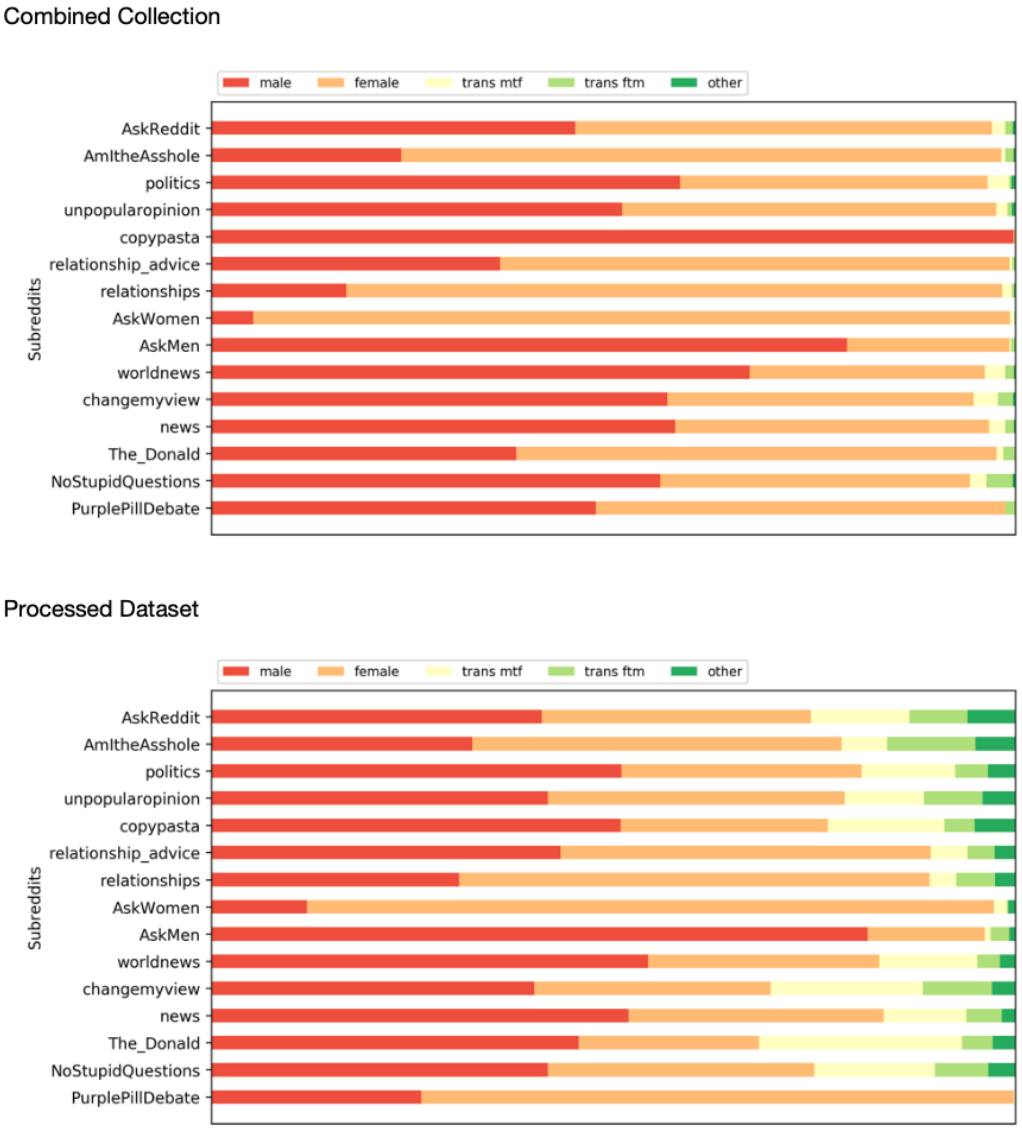

Figure 3.9: Comparison of gender distribution for comment counts over the 15 most-commented-in subreddits from the combined collection with the processed dataset.

Table 3.8: ngrams and occurrence counts of word 1-grams among the top 1,000 for each group that appear only for this group within the top 1,000 in the processed dataset.

Male	Female	Trans mtf	Trans ftm	Other
market	husband	page	ftm	nonbinary
8381	14241	1661	1298	1211
cheap	sweet	magic	testosterone	particular
8194	7323	1614	1080	506
air	son	claim	boys	dress
8160	6972	1589	950	506
building	omg	directly	chest	slightly
7870	6326	1470	941	504
speed	sit	basic	emotional	dream
7662	6263	1441	857	492
pull	info	anime	penis	meaning
7407	6262	1410	840	492
total	posting	disagree	upset	genders
7390	6261	1368	839	491
000	period	es	scared	four
7278	6220	1367	838	490
bet	moved	en	cats	present
7267	6005	1344	836	485
box	watched	private	study	beyond
7267	5712	1338	820	478
quick	marriage	definition	angry	yep
7231	5672	1336	796	478
code	daughter	result	girlfriend	individual
7058	5657	1314	793	478
tax	fake	forced	deserve	five
7044	5649	1308	792	475

Chapter 4

Method

This chapter contains the classification setups used for predicting multiple different gender identities from text in comparison to predicting only male and female. Additionally, it compares the two sources of data of this thesis: Twitter and Reddit. In brief, there are two classification setups being used: (1) *binary classification* of male and female, and (2) *multi-class classification* of male, female and other genders.

Both setups were used for a different purpose: Binary classification yields a baseline prediction performance and a first means of comparison for the predictability of sex from the two sources of text. Multi-class gender classification as male, female, other, transgender male-to-female ('mtf') and female-to-male ('ftm') was used to gain insights about the measurability of gender in a more differentiated way. Therefore, results in binary and in multi-class classification were also compared directly.

4.1 Classification Setup for the Twitter Dataset

The Twitter dataset helps to measure multidimensional gender predictability in a corpus that has already been used in author profiling. In the following section we explain our classifier and evaluation setup.

Related work shows feature-based machine learning still works best [42]. Thus, the classifier for this thesis was set-up based on features frequently used in author profiling and which performed well in the PAN competitions. The approach with the highest accuracy of Pan 2017 has been proposed by Basile et al. [8]. They used *Support Vector Machines* as classifier, character 3- to 5-grams and word 1- to 2-grams with minimal document frequency of 2 (`min_df=2`) which means that an n-gram has to appear at least in two authors feeds in the collection to be used for all authors. The n-grams were weighted with term-frequency, inverse document-frequency (`tf-idf`-weighting) which

means that every n-gram is counted with its occurrence count in an authors collections of text (`tf`) multiplied with the logarithmically scaled inverse fraction of the documents in which the n-gram appears (`idf`) [26]. Furthermore, they used five-fold cross-validation, dividing the dataset into different subsets for training and prediction testing for the classifier five times and generating an aggregated performance score.

Our setup ended up similarly, while being slightly modified because of performance differences in test-setups. *logistic regression* implemented by scikit-learn [35] was used as classifier, due to its multi-class classification support. Additionally, the char n-grams were cut down to 2- to 4-grams, as we didn't find any performance improvement in test runs. On the other hand, we added `min_df` up to 3, as better performance was gained in test runs. Furthermore, `idf` was used for word n-grams.

The classifier was not used on the whole dataset. When labeling Reddit users, we found many more *male* and *female* authors than *transgender* authors or authors with *non-binary* gender identity. Nonetheless, we wanted to check whether these amounts were enough to classify gender successfully. In order to find a reasonable amount of authors required for solid classification results in general, we tested gender prediction with an iteratively growing dataset, with 125 additional individuals of both, *male* and *female* gender, in each iteration. Results are observable in Figure 4.1. As setup, we used `logistic_regression` with 5-fold cross validation and character (2-4) and idf-weighted word (1-2) `n-grams`. As a result of these experiments, 750 was used as a reasonable amount of authors of each group used for classification. With an accuracy of `0.83`, it achieved similar accuracy to the aforementioned approach from PAN 2017 [8]. Although the results of our experiments revealed a classification accuracy constantly growing with more authors, we furthermore found 750 authors to be a turning point, which marked a tilt of the curve into a decreased but steady growing. We derived 750 as author group size in subsets for our classification setups. Subset structures we used for different classification setups are observable in Table 4.1.

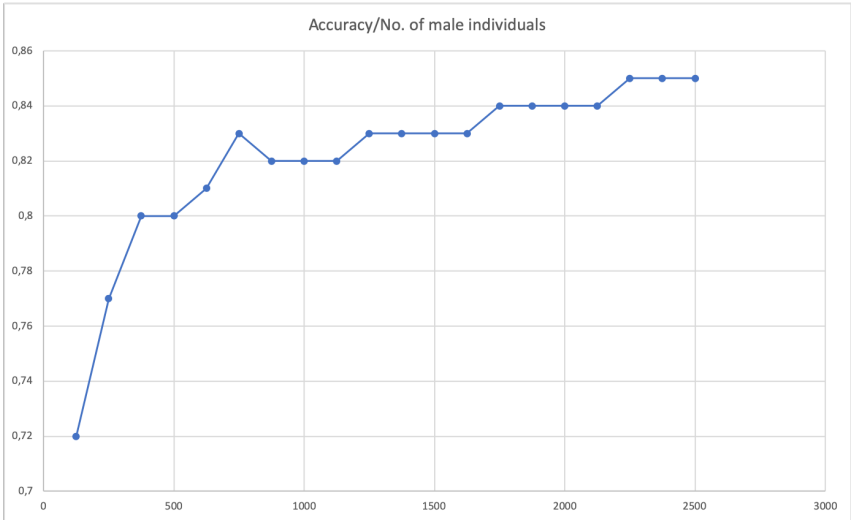

Figure 4.1: Results of the iterative gender prediction.

Table 4.1: Overview of the classification setups used for this thesis.

Setup Name	Setup Genders
Twitter Dataset	
Binary	750 male authors, 750 female authors
3-class	750 male, 750 female, 77 other
3-class balanced	77 male, 77 female, 77 other
Reddit Dataset	
Binary	648 male authors, 648 female authors
3-class	648 male, 648 female, 648 other
5-class	648 male, 648 female, 648 trans ftm, 648 trans mtf, 648 other

4.2 Classification Setup for the Reddit Dataset

In order to compare our Reddit dataset with the Twitter dataset, we used the same feature extraction and similar classification setups on the Reddit dataset. One major difference compared to the Twitter classification was the limitation of group size because of the smallest group in the processed and normalized Reddit dataset. There are only 648 individuals of other gender in the processed dataset. Therefore the group size for Reddit classification was limited to 648 authors to have a balanced dataset. Moreover, with the Reddit data were able to generate larger *balanced* subsets to test if more authors of all groups yielded better prediction performance.

All Reddit subsets were created with twice, with randomly chosen authors from both, the raw collection of authors and the de-biased dataset. The exception was, that all remaining 648 authors with *other* gender identity from the de-biased dataset were used. With these authors, subsets were created so that the same authors from each group were used in all setups. For example, the same 648 randomly chosen *female* authors were used in the *binary*, *3-class*, and *5-class* setup for Reddit. We furthermore used a dummy classifier provided by scikit-learn for a baseline performance in the multi-class Reddit setups [35].

Chapter 5

Results

5.1 Predictive Performance with the Twitter Dataset

Predicting *binary* gender from Twitter with 750 randomly chosen authors of both *male* and *female* gender identities yielded an accuracy of `0.81` when using 5-fold cross-validation and the aforementioned features. After adding the 77 authors with *other* genders for the *3-class setup*, accuracy dropped to `0.77`. We further noticed a strong confusion of the *other*-group towards *female*. Table 5.1 gives a more detailed overview about F1-scores and accuracy for the single groups. The normalized confusion matrices are observable in Figure 5.1 and Figure 5.2. The *3-class balanced* setup, which used the first 77 *male* and *female* authors from the multi-class setup to have an equal amount of authors from all groups yielded an accuracy of `0.61`. Notably, the *female* authors in this setup got the worst performance. *Female* achieved an F1 score of `0.51` compared to *male* with `0.62` and *other* with `0.68`.

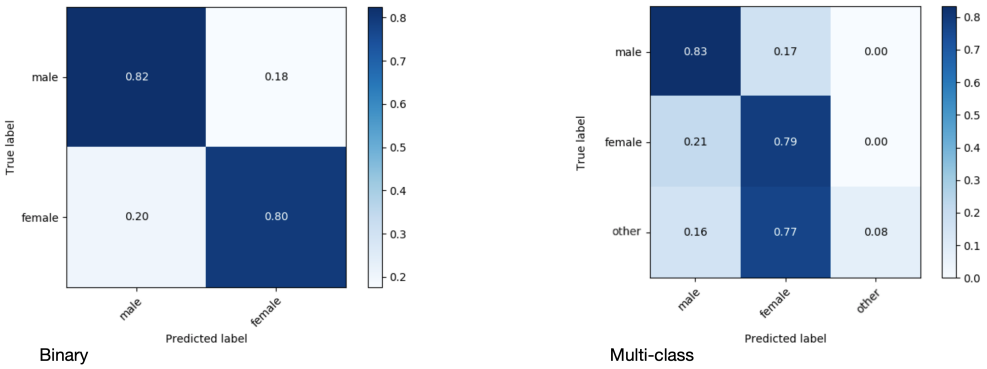

Figure 5.1: Confusion Matrices of *binary* and *3-class* Twitter classification.

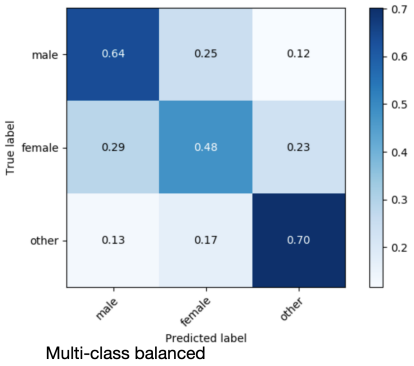

Figure 5.2: Confusion Matrix for *3-class balanced* Twitter classification.

Table 5.1: Results from binary and multi-class Twitter classification.

Description	**Precision**	**Recall**	**F1-Score**
Binary, 1,500 authors, Feature Count: 2,795,317			
Male	0.80	0.82	0.81
Female	0.82	0.80	0.81
Accuracy			**0.81**
Multi-class, 1,577 authors, Feature Count: 2,927,940			
Male	0.78	0.83	0.81
Female	0.76	0.79	0.77
Other	1.00	0.08	0.14
Accuracy			**0.77**
Multi-class balanced, 231 authors, Feature Count: 575,913			
Male	0.60	0.64	0.62
Female	0.54	0.48	0.51
Other	0.67	0.70	0.68
Accuracy			**0.61**

5.2 Predictive Performance with the Reddit Dataset

Table 5.2 presents the results of Reddit gender prediction. The setups for Reddit classification were similar to those for Twitter classification except that for Reddit we used a *5-class* setup with *male*, *female*, *trans ftm*, *trans mtf*, and *other*) alongside the three class setup without the transgender groups. Moreover, the group sizes for Reddit were limited to 648 instead of 750, as this was the smallest size throughout the groups in the Reddit dataset. Additionally, for Reddit we compared the raw comment collection with the processed and de-biased dataset and ran all setups with selections of authors from both datasets.

Gender classification in the *binary* setup yielded slightly worse performance when compared to Twitter, as observable in Figure 5.3: When classifying authors as *male* and *female*, accuracy on the raw dataset was `0.76`. The processed dataset resulted in `0.71`, compared to `0.81` for Twitter.

Figure 5.4 presents classification performance in the *3-class* setup. Here, the accuracy was `0.66` for the raw collection of comments and `0.63` for the de-biased dataset. We noticed that in this setup Reddit data did not perform much better than Twitter data which yielded accuracy of `0.61` in the *3-class balanced* Twitter setup, using only 77 authors instead of 648. Accuracy with the dummy classifier was `0.34` and `0.32`, indicating that the nearly doubled accuracy we achieved with logistic regression was far from a baseline performance. Similarly to the Twitter dataset, *female* authors had the highest misclassification tendencies in the three-class Reddit setup as their F1 on the de-biased dataset was only `0.59` compared to `0.64` for *male* and `0.66` for *other*.

Figure 5.5 shows the classification performance in the *5-class* setup, which resulted in an accuracy of `0.49` for the raw and `0.47` for the de-biased texts. These results seemed weak at first sight, with *trans mtf* appearing almost randomly distributed in the confusion matrices because of a recall of only 27%. When compared to the results of a dummy classifier, the performance appeared better, as the dummy classifier did not even perform half as good with accuracy of `0.22` and `0.21` for 5-class classification.

Table 5.2: Results from binary and multi-class Reddit classification with comparison of the raw and the processed collection.

	Precision de-biased	**Precision** raw	**Recall** de-biased	**Recall** raw	**F1-Score** de-biased	**F1-Score** raw
Binary, 1,296 authors, Feature Count: 1,362,333/2,047,746						
Male	0.70	0.75	0.72	0.77	0.71	0.76
Female	0.72	0.77	0.69	0.74	0.70	0.75
Accuracy					**0.71**	**0.76**
Multiclass I, 1,944 authors, Feature Count: 1,787,837/2,500,179						
Male	0.61	0.65	0.67	0.67	0.64	0.66
Female	0.61	0.65	0.56	0.58	0.59	0.61
Other	0.67	0.63	0.66	0.68	0.66	0.65
Accuracy, logistic regression					**0.63**	**0.66**
Accuracy, dummy classifier					0.32	0.34
Multiclass II, 3,240 authors, Feature Count: 2,707,233/3,588,645						
Male	0.49	0.54	0.63	0.67	0.55	0.60
Female	0.50	0.54	0.51	0.57	0.51	0.55
Trans mtf	0.39	0.34	0.27	0.31	0.32	0.33
Trans ftm	0.42	0.47	0.39	0.45	0.41	0.46
Other	0.51	0.53	0.56	0.46	0.53	0.49
Accuracy, logistic regression					**0.47**	**0.49**
Accuracy, dummy classifier					0.21	0.22

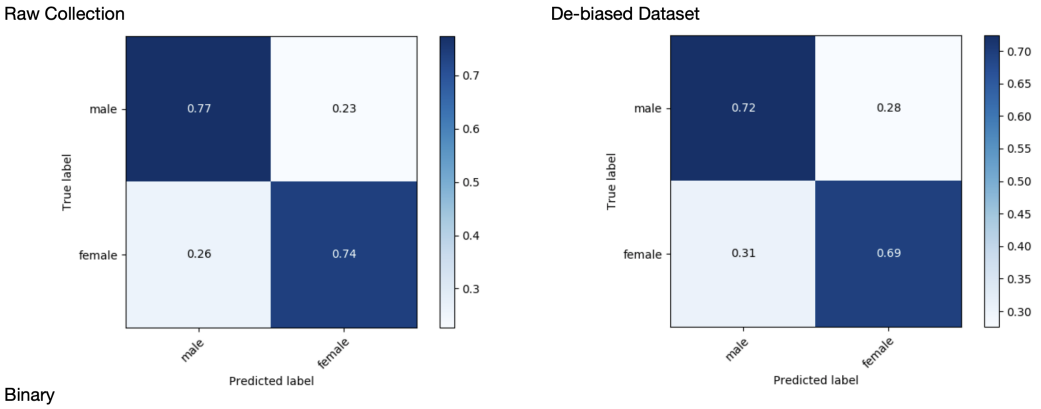

Figure 5.3: Confusion Matrices for binary Reddit classification.

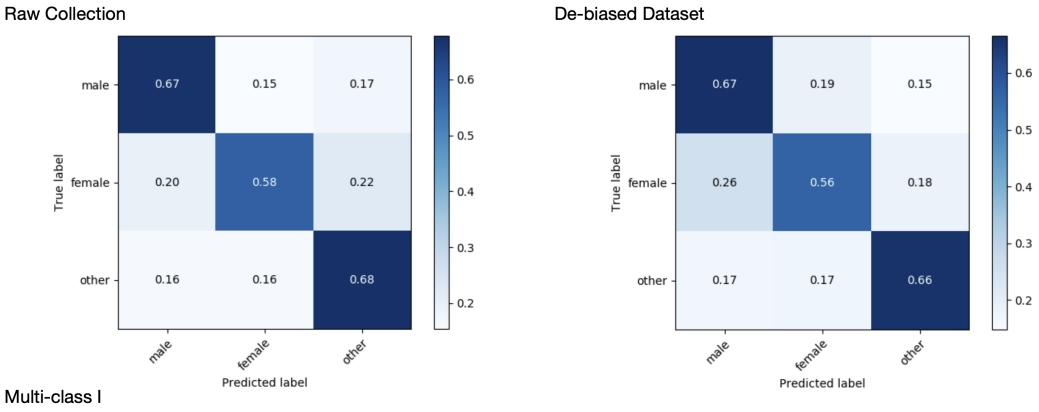

Figure 5.4: Confusion Matrices for 3-class Reddit classification.

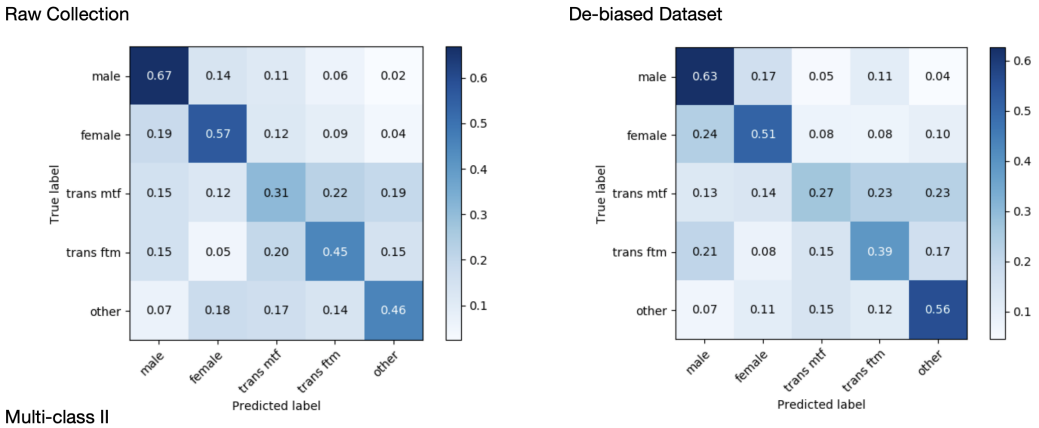

Figure 5.5: Confusion Matrices for 5-class Reddit classification.

Chapter 6

Discussion

This work starts with a survey of gender in author profiling, a subfield of computer science, and other scientific disciplines. We noticed that author profiling mostly treats gender as a binary of *male* and *female* when trying to predict gender from texts. But texts as samples of authors language reflect their beliefs of how to behave appropriately according to the gender they identify with rather then whether they are biological men or women. In other words: Texts are influenced by authors gender identities, and gender identity is not limited to the binary categories male and female. Therefore, we attempted to measure multiple gender identities instead of *male* and *female* from text, in order to find out whether those are measurable with similar performance. As datasets and corpora used for related work contained no authors or only small minorities of authors labeled with gender identities different from *male* and *female*, we furthermore generated a new dataset from Reddit data, which contained higher fractions of authors with other gender identities. Alongside the new dataset, we used an already existing dataset with Twitter data that contained only a very small fraction of authors that are not male or female.

Before creating the dataset, a survey of methodologies of psychological diagnostics, the field of psychology that measures personality traits such as gender with different tests, was done. It revealed that one way of creating inventories is usage of real-world data that discriminates groups of individuals with a certain trait a test is supposed to identify from others. These results substantiated our approach of generating a dataset that reflects differences between authors of different gender identities: We label authors with gender, collected as much text written by them as possible afterwards and finally try to generate a homogenous dataset, in which the gender groups preferably differ according to their gender identities and not according to other structures such as topics authors mainly write about.

Afterwards, we came up with a dataset created from Reddit data which

contains many *transgender* and *non-binary* authors. With our newly created dataset, we showed that it is possible to reliably label transgender and non-binary authors with gender and obtain samples of texts written by them. We attempted to balance our data when we defused a topic-bias which resulted from the topic-related site structure of the data source Reddit with its division into subreddits.

Finally, we checked whether the amount of users of other gender identities we retrieved was enough to classify authors according to gender identity in a machine learning setup previously used in author profiling and yield comparable performance. We derived our setup from a successful author profiling study and classified authors from both datasets according to their genders. Binary classification yielded best performance and was comparable to the submission we oriented after. *3-class* classification with *male*, *female* and *other* but differently sized groups - *other* was a minority - revealed poor performance for this third group, as authors with *other* gender identity were mostly mis-classified as *female*. With balanced groups of authors, meaning that same amounts of authors of every group were used, no strong misclassification tendencies were visible. The classification performance was worse than for binary classification, but comparison with a dummy classifier revealed that it is far from random performance. Comparison of small (77) and larger (648) balanced groups revealed only a small performance gain. *5-class* classification yielded similar results, with a random classifier being less than half as performative. When comparing the unprocessed Reddit dataset with the processed and de-biased datasets, the debiased dataset always performed worse, which might be due to the removed topic-bias but could also depend on different authors.

6.1 Future Work

Lastly, we propose some possible refinements for dataset creation and gender prediction in future works.

Generating Datasets with multiple Gender Identities

For this thesis, we used 6 Reddit dumps to create a dataset of gender-labeled authors and their texts. More Reddit dumps could be processed to find even more authors or to find more comments written by the authors that have already been labeled. Additionally, the labeling methods themselves bare some potential for refinements. The regular expressions used for distant supervision labeling could cover more self-labeling phrases such as *"for me as a ..."* and exclude even more preceding phrases. The flair labeling approach could be extended with more variations of flairs, such as more complex "A|S|L" Flairs.

Both labeling approaches in general could target even more different gender identities such as *pangender*.

The resulting collection of comments written by the labeled authors might also be processed more precisely, leading to a more even distribution of text written by the different gender groups across subreddits in the dataset. In order to achieve this, the algorithm described in Listing 3.1 could be refined and the heuristic removed. The fact, that the raw collection of authors performed better than the processed dataset in every Reddit classification setup, might indicate that the classifiers has learned differences between the gender groups from the subreddits and their topic related words rather than from differences in the general selection of words. Additionally, all comments that are matched by the distant supervision method should be removed. They bare the risk, that the machine learning model learns gender differences from these comments only.

Predicting multiple Gender Identities from Text

The classification setup with an ever increasing amount of authors we used, proofed that with more authors, the discriminative performance of a classifier gets better. It is to be verified, whether larger datasets also improve performance in classification setups such as the *3-class* or *5-class* setups used throughout this thesis.

Future Work might also make a step ack from using different classes for gender profiling: Instead of determining discrete classes, the sexes male and female could be used as poles to regress all authors on a spectrum, similarly to [9], where every author is assigned a continuous value.

Chapter 7

Conclusions

This thesis studies how author profiling could attempt prediction of gender from text differently, with a more diverse understanding of gender. We demonstrated ways of doing so by generating a collection of labeled authors and comments from Reddit and attempting multi-class gender prediction.

Our methodology of accessing and labeling Redditors with gender has proven its capability of identifying authors with gender identities such as *transgender* and *non-binary*. Thus, datasets with texts of authors with multiple gender identities can be built, laying the foundation for future work to look for discriminative features and find the best learning model to distinguish different gender identities of authors. Therefore, a more multifaceted treatment of gender in author profiling would be possible.

Finally, the multi-class classification scenarios we used on our data gained solid discriminative performance for multiple gender identities, as a comparison with a baseline performance revealed. It remains open, if larger and more precisely processed datasets as well as different learning models improve this performance.

Bibliography

[1] The 57th annual meeting of the Association for Computational Linguistics (ACL 2019) ACL. Call for papers, 2019. URL http://www.acl2019.org/EN/call-for-papers.xhtml. Accessed: 2019-12-05. 3.2.2

[2] Lewis R Aiken. *Personality assessment methods and practice, 3rd rev.* Hogrefe & Huber Publishers, 1999. 2.3

[3] Claire Ainsworth. Sex redefined. *Nature News*, 518(7539):288, 2015. URL http://theavarnagroup.com/wp-content/uploads/2015/11/Sex-Redefined.pdf. 1

[4] APA. Apa dictionary of psychology, 2019. URL https://dictionary.apa.org/psychology. Accessed: 2019-11-10. 2.3

[5] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text-The Hague Then Amsterdam Then Berlin-*, 23(3):321–346, 2003. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.1172&rep=rep1&type=pdf. 3.1.1

[6] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, 2009. 2.1, 2.1

[7] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2): 135–160, 2014. 2.2, 2.1, 2.3

[8] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*, 2017. URL N-gram: Newgroningenauthor-profilingmodel. 4.1

[9] Sandra L Bem. The measurement of psychological androgyny. *Journal of consulting and clinical psychology*, 42(2):155, 1974. 2.4, 6.1

[10] Cynthia L Berryman and James R Wilcox. Attitudes toward male and female speech: Experiments on the effects of sex-typical language. *Western Journal of Communication (includes Communication Reports)*, 44(1):50–59, 1980. 1

[11] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 3.2.2

[12] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics, 2011. 2.1, 2.1

[13] Blake Burkhart. Subreddit gender ratios, 2017. URL `https://nbviewer.jupyter.org/github/bburky/subredditgenderratios/blob/master/Subreddit%20Gender%20Ratios.ipynb`. Accessed: 2019-11-10. 2.1, 3.2.1, 3.2.1

[14] Robin De Pril. User classification based on public reddit data. 2019. URL `https://lib.ugent.be/fulltxt/RUG01/002/785/835/RUG01-002785835_2019_0001_AC.pdf`. 2.1, 2.1, 3.2, 3.2.1

[15] Chris Emmery, Grzegorz Chrupała, and Walter Daelemans. Simple queries as distant labels for predicting gender on twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 50–55, 2017. URL `https://www.aclweb.org/anthology/W17-4407.pdf`. 3.2.1, 3.2.1

[16] Transgender Europe. Germany introduces third gender - fails trans people, 2018. URL `https://tgeu.org/germany-introduces-third-gender-fails-trans-people/`. Accessed: 2020-01-20. 1

[17] Mehwish Fatima, Komal Hasan, Saba Anwar, and Rao Muhammad Adeel Nawab. Multilingual author profiling on facebook. *Information Processing & Management*, 53(4):886–904, 2017. 2.1, 2.1

[18] Google. 1.7 billion reddit comments loaded on bigquery, 2015. URL `https://www.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery/`. Accessed: 2019-11-25. 2.1

[19] David Haig. The inexorable rise of gender and the decline of sex: Social change in academic titles, 1945–2001. *Archives of sexual behavior*, 33(2): 87–96, 2004. 2

[20] Starke R Hathaway and John C McKinley. A multiphasic personality schedule (minnesota): I. construction of the schedule. *The Journal of Psychology*, 10(2):249–254, 1940. 2.4

[21] Susan C Herring and John C Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2006. 2.2

[22] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002. 2.1, 2.1

[23] Robin Lakoff. Language and woman's place. *Language in society*, 2(1): 45–79, 1973. 2.2

[24] The Python Standard Library. Regular expression operations, 2020. URL `https://docs.python.org/2/library/re.html`. Accessed: 2020-01-10. 3.2.1

[25] PRAW limitation. Praw listinggenerator, 2020. URL `https://praw.readthedocs.io/en/latest/code_overview/other/listinggenerator.html`. Accessed: 2020-01-10. 3.2.1

[26] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999. URL `https://nlp.stanford.edu/fsnlp/`. 4.1

[27] James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media. In *CLEF (Working Notes)*, pages 1129–1136, 2014. 2.1

[28] John Money, Joan G Hampson, and John L Hampson. An examination of some basic sexual concepts: the evidence of human hermaphroditism. *Bulletin of the Johns Hopkins Hospital*, 97(4):301–319, 1955. 2

[29] Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 207–217, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1870658.1870679`. 2.1

[30] Anthony Mulac and Torborg Louisa Lundell. Linguistic contributors to the gender-linked language effect. *Journal of Language and Social Psychology*, 5(2):81–101, 1986. 2.2

[31] Anthony Mulac, James J Bradac, and Pamela Gibbons. Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research*, 27 (1):121–152, 2001. 2.2

[32] Gytha Muller. *Distant Labelling and Author Profiling on Reddit*. PhD thesis, Tilburg University, 2018. URL `http://arno.uvt.nl/show.cgi?fid=147835`. 2.1, 2.1, 3.2.1

[33] Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236, 2008. 2.1

[34] Openpsychometrics. Open sex-role inventory, 2019. URL `https://openpsychometrics.org/tests/OSRI/`. Accessed: 2019-11-10. 2.4

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011. 4.1, 4.2

[36] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001. URL `http://www.depts.ttu.edu/psy/lusi/files/LIWCmanual.pdf`. 2.1

[37] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003. 2.1

[38] PRAW. Python reddit api wrapper, 2020. URL `https://praw.readthedocs.io/en/latest/index.html`. Accessed: 2020-01-10. 3.2.1

[39] Pushshift. Pushshift - learn about big data and social media ingest and analysis, 2019. URL `https://pushshift.io`. Accessed: 2019-11-10. 3.2.1

[40] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT, 2013. 2.1, 2.1

[41] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th author profiling task at pan

2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.*, pages 750–784, 2016. 2.1

[42] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*, 2017. 2.1, 2.1, 4.1

[43] Francisco Rangel, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*, 2018. 2.1

[44] Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8, 2015. 2.1

[45] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010. 2.1, 2.1

[46] Reddit. Reddit api access, 2016. URL `https://www.reddit.com/wiki/api`. Accessed: 2019-11-10. 3.2.1

[47] Reddit. How do i get flair (the text/image next to my username)?, 2019. URL `https://www.reddithelp.com/en/categories/using-reddit/your-reddit-account/how-do-i-get-flair-textimage-next-my-username`. Accessed: 2019-11-10. 3.2

[48] Katrin Rentzsch and Astrid Schütz. *Psychologische Diagnostik: Grundlagen und Anwendungsperspektiven*, volume 16. W. Kohlhammer Verlag, 2009. 2.3

[49] Aliza Rosen. Tweeting made easier, 2017. URL `https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html`. Accessed: 2019-11-10. 3.1

[50] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006. 2.1, 2.1

[51] Franziska Schößler. *Einführung in die Gender Studies*. Akad.-Verl., 2008. 1

[52] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013. 2.1

[53] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002. 2.1

[54] Nakatani Shuyo. Language detection library for java, 2010. URL `http://code.google.com/p/language-detection/`. 3.1.1

[55] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A Sanchez-Perez, and Alberto Barrón-Cedeño. Overview of the author identification task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–21, 2014. 2.1

[56] Kaitlyn Tiffany. Inside r/relationships, the unbearably human corner of reddit, 2019. URL `https://www.theatlantic.com/technology/archive/2019/10/reddit-moderation-relationships-subreddit-memes/600322/`. Accessed: 2019-11-10. 3.2

[57] Twitter. About verified accounts, 2019. URL `https://help.Twitter.com/en/managing-your-account/about-Twitter-verified-accounts`. Accessed: 2019-11-10. 3.1

[58] Evgenii Vasilev. Inferring gender of reddit users. 2018. URL `InferringgenderofRedditusers`. 2.1, 2.1, 3.2.1

[59] WHO. Gender, equity and human rights - glossary of terms and tools - gender, 2020. URL `https://www.who.int/gender-equity-rights/knowledge/glossary/en/`. Accessed: 2020-01-19. 1, 2

[60] Jake Widman. What is reddit? a sbeginner's guide to the front page of the internet, 2019. URL `https://www.digitaltrends.com/social-media/what-is-reddit/`. Accessed: 2019-11-10. 3.2

[61] Matti Wiegmann, Benno Stein, and Martin Potthast. Celebrity profiling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2611–2618, 2019. 2.1, 3.1

[62] Michael Wilson. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1): 6–10, 1988. URL `https://link.springer.com/content/pdf/10.3758/BF03202594.pdf`. 2.1