Analysis of gender bias in popular Subreddits

Bachelor thesis to obtain the bachelor degree
Bachelor of Science in Data & Information Science
at the faculty of F03 - Information Science and Communication Studies
of the TH Köln - University of Applied Sciences

submitted by:	Andreas Konstantin Kruff

submitted to:	Prof. Dr. Philipp Schaer
second supervisor:	M.Sc. Fabian Haak

Bergisch Gladbach, 09.11.2022

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer oder der Verfasserin/des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Ort, Datum

Rechtsverbindliche Unterschrift

1 Abstract - German

Das Ziel dieser Arbeit ist "Gender Biases" in der Kommunikation der Nutzer von Subreddits der Plattform Reddit zu detektieren. Die Analyse wird hierbei exemplarisch für elf ausgewählte Subreddits durchgeführt. Darüber hinaus wird versucht verschiedene Nutzertypen mit Hilfe von einem k-means Clustering zu identifizieren und ebenfalls "Gender Biases" in deren Kommunikation zu analysieren. Auf Basis der aggregierten Datensätze werden fasttext Word Embedding Models trainiert, um Terme zu identifizieren, die eine hohe semantische Verwandtschaft in Bezug auf die Kosinusähnlichkeit ihrer Wortvektoren mit ausgewählten weiblichen und männlichen Termen aufweisen. Die Terme werden dazu auf ihr Sentiment mit Hilfe des NRC-VAD Lexicons analysiert und auf statistische signifikante Unterschiede überprüft. Darüber hinaus werden der Word Embedding Association Test (WEAT) durchgeführt, um unterschwellige Assoziationen zu überprüfen. In Bezug auf den betrachteten Textkorpus wird im wesentlichen beobachtet, dass Frauen häufig mit Adjektiven in Verbindung gebracht werden, die sie mit Äußerlichkeiten, Gebärfähigkeiten oder Anpassungsfähigkeiten auch in Bezug auf die Familie assoziieren. Im Gegensatz dazu werden Männer mit Adjektiven assoziiert und daran gemessen, welche sich auf ihr Ansehen, ihre Stärken und Schwächen, ihre Karriere oder physische Eigenschaften beziehen.

Schlagwörter: NLP, Reddit, Gender bias, Word Embeddings, Sentiment, WEAT

2 Abstract - English

The goal of this work is to detect "gender biases" in the communication of users of Subreddits on the platform Reddit. The analysis is carried out for eleven selected Subreddits. Furthermore, an attempt is made to identify different user types with the help of a k-means clustering and also to analyze "gender biases" in their communication. Based on the aggregated datasets, fasttext Word Embedding models are trained to identify terms that show high semantic relatedness in terms of cosine similarity of their word vectors with selected feminine and masculine terms. To this end, the terms are analyzed for sentiment using the NRC-VAD Lexicon and tested for statistically significant differences. In addition, the Word Embedding Association Test (WEAT) is performed to check for subliminal associations. In relation to the considered text corpus, it is essentially observed that women are frequently associated with adjectives that associate them with appearances, childbearing abilities or adaptability also in relation to the family. In contrast, men are associated with and measured by adjectives that refer to their prestige, strengths and weaknesses, career or physical characteristics.

Keywords: NLP, Reddit, Gender bias, Word Embeddings, Sentiment, WEAT

Contents

List of Figures

List of Tables

Acronyms

ACL Association for Computational Linguistics. 2, VI

AMA "Ask me anything". 4

CBOW Continuous Bag of Word. 14

ECT Embedding Coherence Test. 3

EDA exploratory data analysis. 32

GloVe global vectors for word representation. 1

HAL Hyperspace Analogue to Language. 3

IAmA "I Am A". 4

IAT Implicit Association Test. 3

LGBTQ Lesbian, Gay, Bisexual, Transgender, Queer. 3

LSA Latent semantic analysis. 3

MAC Mean Average Cosine Similarity. 3

NLP Natural Language Processing. 4

NLTK Unicode Transformation Format – 8 Bits. 11

NSFW "Not safe for work". 4

POS Part of Speech. 25

RIPA Relational Inner Product Association Test. 3

RND Relative Norm Distance. 3

RNSB Relative Negative Sentiment Bias. 3

USAS UCREL Semantic Analysis System. 25

UTF-8 Unicode Transformation Format – 8 Bits. 11

WEAT Word Embedding Association Test. 3

3 Introduction

Recent studies have shown that social media platforms "such as Reddit not merely reflect a distinct offline world, but increasingly serve as constitutive spaces for contemporary ideological groups and processes." (Aran, 2020, p.1). Another experiment from Germany analyzed the affect of Classmates' Gender Stereotypes on Student Math Self-Concepts and found out that "gender stereotypes shared by students' classmates can have a substantial impact on students' math self-concepts, beyond their individual gender stereotypes. This finding emphasizes the significance of classmates as important socializing peers in the process of students' self-concept formation." (Wolff, 2021, p.1). According to that, applying association tests to the state-of-the-art GloVe Word Embeddings, a study found out that within this pretrained embeddings female words like "woman" or "girl" were more associated with arts than mathematics compared to male words. Furthermore the same applied for arts and sciences.[4]

3.1 Motivation

As assumed above the platform Reddit can be regarded as a reflection of a distinct offline world. It should be considered that social phenomenons can also be transfered in the online world. Based on demographic statistics from Statista, it can be observed that Reddit covers a wide age range, but the largest group is between 10 and 39. These studies show that Reddit has a relatively young male community with a not insignificant percentage of women, who regularly participate on the platform.[5][6][7] This shows that the main user base of reddit might still be in a state of self-concept formation, that can be influenced by the gender stereotypes that are shared within the before mentioned "ideological groups". Although the studies in the research field of gender bias are increasing over the years (see Table 1) a study from 2018 found out, that there might be "the possibility of an underappreciation of the phenomenon of gender bias and related research within the academic community" (Cislak, 2018, p.1). They found out that articles on gender bias are funded less often and published in journals with a lower Impact Factor than articles on comparable instances of social discrimination like race bias. Within this article they also underline the importance of this research field: "Addressing this meta-bias is crucial for the further examination of gender inequality, which severely affects many women across the world" (Cislak, 2018, p.1).[8] Within the scope of this thesis eleven of the biggest Subreddits with atleast 1,000,000 subscribers will be analyzed in terms of gender bias and stereotypes, assuming that these Subreddits have a great diversity with regard to the users, because of the general interest in the topics discussed within them. Furthermore spokesperson and other user groups will be identified, to specifically analyze their influence on the gender bias within the Reddit community.

For this purpose the content of these Subreddits will be preprocessed, to be able to train word embedding models based on the observed Subreddits and for different kind of user groups. For identifying the user groups a kmeans clustering will be used. The resulting models will then be analyzed with the Lexicon and the Word Embedding Association Test to detect possible gender biases or stereotypical role models in this specific social media communities. Additionally the most similar adjectives for woman and man will be manually evaluated.

YEAR	ACL #Publications	ARXIV #Publications
2015	1	6
2016	2	8
2017	4	8
2018	8	31
2019	38	67

Table 1 Number of publications with "gender bias" as main topic in the ACL anthology and arXiv from 2015 to 2019 [2]

As described above, however, it is not obvious insults or misogynistic comments that will be analyzed within the scope of this work, but rather the social image in relation to the gender anchored in linguistic usage. Although the author is aware of the fact that the binary gender model is outdated, within this thesis the focus will be on detecting stereotypes related to the binary gender model. The implementations related to this thesis are publicly available at Github. [1]

3.2 Research Question

1. To what extent are gender biases or stereotypes measurable in eleven of the largest Subreddits and what are their qualitative and quantitative characteristics?
   (a) In what ways do the words that are more associated with one gender or the other differ?
   (b) Is a statistically significant difference for the associated words measurable between users with different user behavior? If yes, how do the user groups differ from each other?
   (c) Additionally, how do the various Subreddits differ in terms of the observed biases?

4 Related Work

While gender bias can occur in various forms a very popular form in the research field of Natural language is the **language bias**, that occurs in multiple forms and

[1] `https://github.com/AndyKruff/BA_Reddit_Gender_Bias`

contexts as shown in [9]. While the bias in language is a known problem in human interactions, not much attention was paid to the language biases that occured in text corpora for the training of machine learning models. Bolukbasi et al was the first, who identified that common pretrained word embedding models like w2vNEWS, that were trained on Google News articles, inherit these biases within the underlying corpora to a great extent. In addition he served a first approach for debiasing such word embeddings.[10] In 2008 Sahlgreen et al evaluated models like HAL and LSA, that were the state of the art models at that time, in terms of the distributional hypothesis. He stated that distributional models are models of word meaning and the underlying meaning of every word is based on the text it was trained on. Therefore "they embody a thoroughly descriptive perspective" (Sahlgreen, 2008, p.15) of the underlying corpus. When changing the data also the model will change accordingly. [11]. According to that, the more recent findings of Mendelsohn et al showed with the help of word embedding models, that were trained on New York Times articles for every year from 1986 to 2015, how the meaning and the related words of LGBTQ terms changed over time.[12] Therefore distributional models cannot just inherit biases in meaning from the text corpora influenced by the present society, but also from former societies, based on the underlying corpus. This finding was also made by Caliskan et al, who created one of the first and most common approaches to identify Association Biases within word embeddings, that was inspired by Greenwald et al, a socialpsychologist, who invented the IAT, that measured the time difference that people needed to combine attribute words with target words, that they associate with each other compared to the combination of this words, when they do not associate the terms with each other.[13] Within this work it will be tried to take advantage out of this finding, by training word embeddings on text corpora to uncover gender biases within the text. Besides the WEAT metric from Caliskan et al in the following years multiple more approaches like RND, RNSB, MAC, ECT or RIPA were developed for identifying biases within word embedding representations.

5 Declarations

5.1 Reddit - Research object

Reddit.com is a social media platform, that currently offers around 3,500,000 [2] different Subreddits, in which topic-specific content can be shared in the form of links, videos, images, surveys and text content. It was founded in 2005 by Alexis Ohanian and Steve Huffman. It is often referenced as the so called "front page of the Internet" (Baumgartner, 2020, p.1). Unlike other social media platforms Reddit decided in June 2008 to make the code open source and invited the public to submit code to improve the side.[14] The webside got a lot of attention in 2012, when

[2] `https://frontpagemetrics.com/`, [last accessed: 06.07.2022]

it hosted President Obama's AMA thread. Until today this thread is still by far the most popular thread in the "r/IAmA" Subreddit in terms of upvotes compared to other public personalities and regular users, with almost double the amount of upvotes compared to the second most popular thread from 2017. [15]

5.2 Demographic Information

A statistic from January 2022 shows that while most of the users are male with 63.8 %, there is also a big female group with 36.2 % [16]. Another study from February 2021 with focus on the percentage of U.S. adults who uses the platform, shows that the main audience of Reddit with 36 % is between 18 - 29 years old, while the second biggest group with 22 % is between 30 - 49. The aggregated older user groups are much smaller with 10 % for users between 50 - 64 and 3 % for users above 65 years old [7]. With 47.13 % of the users being from the US, 7.48 % being from the UK, 7.36 % from Canada and 4.15 % from Australia being the four biggest countries in terms of desktop traffic to Reddit.com in May 2022 it can be assumed that the predominant language is english [17].

5.3 Structure of Reddit

Alike in other social media platforms Reddit also offers the opportunity to rate the content that a user posted by up- or downvoting it, which reflects the users popularity in his karma score. Additionally a user can be awarded with "Reddit gold", a premium currency within Reddit, that allows to use premium features like ad-free browsing or access to the "r/lounge" Subreddit. The karma score consists of the "post karma", "comment karma", "awardee karma" and the "awarder karma". Subreddits are topic-specific communities, that can be created by every user and are moderated by Reddit users of this community or by implemented bots. For every Subreddit the so called "reddiquettes" can be defined, which describe the communities etiquette guidelines.[14] By subscribing to a Subreddit, content of this Subreddit will be added to the personal user's front page. From the 3,500,000 different Subreddits over 100,000 Subreddits are classified as active, which means that it contains at least five posts or comments on their About Page. At the time of the cited article (2015) the amount of active Subreddits was about 9379. [14] Several studies have already shown that NLP methods for the analysis of different kinds of biases are well suited for Reddit data sets, especially due to the thematic structuring of Reddit's Subreddits. Among others, the data has been used to classify discourses about mental health or narratives about domestic violence. Furthermore, other studies have used NLP methods and discourse analysis to analyze discriminatory language in the form of sexism, racism, and "toxic technoculture." [18] If a Subreddit contains more critical or offensive content, it is either marked with a "NSFW" tag,

placed in quarantine or removed from the platform. Since it can be assumed that Subreddits tagged with the "NSFW" tag or quarantined often contain sexual or sexualized content, it might be concluded that women are probably sexualized there in linguistic form and are therefore not taken into consideration for the selection of the Subreddits in order to avoid the distortion of the results and to make the communities of the Subreddits more comparable.

5.4 Gender bias - General Definition

As stated above this thesis will cover gender bias with a focus on gender stereotypes based on the traditional binary gender model. To clarify the bias observed within this work, the definition of the Council of Europe Gender Equality Strategy will be used [19]:

> Gender stereotypes are generalised views or preconceived ideas, according to which individuals are categorised into particular gender groups, typically defined as “women” and “men”, and are arbitrarily assigned characteristics and roles determined and limited by their sex. Stereotypes are both descriptive, in that members of a certain group are perceived to have the same attributes regardless of individual differences, and prescriptive as they set the parameters for what societies deem to be acceptable behaviour. Stereotyping becomes problematic when it is used as a vehicle to degrade and discriminate women. Abolishing negative gender stereotypes is essential to achieving gender equality, and the media are central to prompting this change. (Halonen, 2016, p.2)

5.5 Gender bias in language

According to Bolukbasi et al there are basically five major language biases and stereotypes, that are highlighted within his work, that should be taken into consideration when analyzing biases in Machine Learning Tasks like the training of word embeddings. The biases, he had chosen from, were extracted from [9] a book from 2008 that covers the state of research for languages biases from multiple decades.[10] The following four forms of language bias were considered to be most relevant within this thesis.

- **Association Bias:** At first he mentions the IAT, a methodology to measure human biases resisting self-presentational forces, which was first introduced by Greenwald et al. Common assumptions are for example that female terms are more often associated with arts, while male terms are associated with science or career.
- **Bias in Meaning:** In this case the finding is that there is an imbalance in words for males and females in terms of their underlying meaning. Due to the findings of Stanley et. al "Women insult men by reference to unpleasantness

in their personalities, but men insult women by reference to their availability for sexual use."[20].

- **Benevolent sexism:** Another gender bias lies in benevolent sexism.The inflationary use of subjectively positive words such as "attractive" can yield to the problem, that the association of females towards terms like "career" or "professional" weakens.
- **Complementary stereotypes:** These kind of stereotypes are awarding females or males with strengths, that should compensate their weaknesses and complement the strengths of the opposite gender. In [21] the example was made that men are agentic and not communal, while women are communal and not agentic, so every gender gets a strength associated that complements the other gender and justifies the status quo.

6 Data Preparation

Using unstructured social media data comes with certain challenges regarding the preprocessing, to make the data useable for the further analysis. In the following the preprocessing steps, that were performed, are described and explained.

- Acquisition of the data
- Removing non-relevant reddit related content
- Removing special characters and non-alphanumeric characters
- language filtering
- Tokenization and Stopword removal
- Lemmatization

´

6.1 Data acquisition

6.1.1 Pushshift

Pushshift is "a social media data collection, analysis and archiving platform" (Baumgartner, 2020, p.1) that is collecting Reddit data since 2015 and updates them in real-time. It covers the whole Reddit Corpus since the beginning of Reddit itself and provides the data to researchers. In addition to monthly dumps of the reddit dataset, it also offers an API infrastructure to "aid in searching, aggregating and performing exploratory analysis" (Baumgartner, 2020, p.1). [22]

6.1.2 Dataset

For the acquisition of the dataset the psaw library was used to download the data through the API of pushshift. The observation window was set from the 01.05.2022 at 00:00:00 (CEST) to the 01.08.2022 at 00:00:00 (CEST) to be able to achieve sufficiently large datasets to train word embeddings for single Subreddits. Unfortunately comments and submissions for threads, that started before or ended after the time window, could not be considered. Therefore some egocentric reply-graph, could not be fully reconstructed. For the analysis eleven Subreddits were manually picked from the domain redditlist.com, that contains statistics about the 5000 biggest Subreddits. For the selection only Subreddits with at least 1,000,000 subscribers and enough comments were taken into consideration. For example the biggest Subreddit "announcements" has by far the most subscribers with 174,979,089, but within June 2021 it just had 3594 comments leading to a corpus that is too small for training word embeddings. Additionally Subreddits were selected by their content to see if further hypothesis can be drawn from the results. In the end selection the following Subreddits were selected for the analysis.

Subreddit	Active users	comments	submissions
AskWomen	254,096	473,817	38,923
AskMen	109,132	1,627,851	66,008
teenagers	325,034	5,162,969	303,030
Conservative	73,833	822,440	34,829
funny	456,611	1,279,830	77,372
technology	282,149	1,080,251	18,017
science	159,620	545,452	7,772
unpopularopinion	236,404	1,126,488	63,970
Parenting	55,548	348,802	15,374
gardening	73,598	308,039	46,498
TwoXChromosomes	122,181	738,102	23,595

Table 2 Manually selected Subreddits for the analysis of underlying gender biases (bots and empty comments were excluded beforehand). Key figures are describing the Subreddits for the observation window

The Subreddits in table 2 were not just chosen regarding their amount of subscribers or the size of the resulting corpus, but also to be able to categorize potential findings more systematically. For example the Subreddits "AskWomen" and "AskMen" tend to encourage people to ask questions, that are relying on certain stereotypes and should somehow be gender related, but not mandatory insulting.
Questions like in figure 1, which a men probably would not be asked about, results in comments that might link women to terms like "family" and "children". The "r/teenagers" Subreddit might give valuable insights about the language usage of teenagers themselves, that therefore directly also influences all the teenagers within

Figure 1 Examples for submissions in the r/AskWomen Subreddit

this Subreddit, who might be affected in terms of their self-concept formation. As described in the Declarations part about the IAT, it is a common assumption in these tests, that women are more associated with children, parenting or gardening, while men are more associated with mathematics, science or technology. Therefore the Subreddits "r/science", "r/technology", "r/Parenting" and "r/gardening" where taken into consideration, to check if they inherit this associations and if the associations are also convered within this Subreddits.
The "r/TwoXChromosomes" Subreddits description states, that it is "a Subreddit for both serious and silly content, and intended for women's perspectives"[3] and might therefore be relevant to analyze, if there are differences in terms of sentiment or the associations towards women. At last there are the "r/Conservative", the "r/unpopularopinion" and the "r/funny" Subreddits, where the first two both suggest the presumption of having a potential of showing a more biased attitude towards more "traditional" gender role models, while the jokes within the "r/funny" Subreddit might contain stereotypes, because jokes are often made about certain aggregated social groups like women, men, blondes etc..

6.2 Filtering bots participation's from dataset

Reddit has a wide variety of topic-specific Subreddits, that are mostly open available for everyone and being moderated by their community or bots like the "AutoModerator". That being said, controversial comments and posts violating the "reddiquettes" of a Subreddit leading to a warning or removing of the content. Although the chosen Subreddits are not dealing with controversial topics directly, the fact that all these belong to the biggest Subreddits by subscribers lead to the assumption that this Subreddits have a very heterogeneous community. This might explain why a total of 754,439 comments were removed within the dataset. In addition to this the comments that the users deleted themselves are represented as "[deleted]". Both were excluded from the dataset. Another problem according the removed comments is, that they can be removed for various reasons regarding the mentioned "reddiquettes". For example the post "If I hit a window with a full glass bottle, which would break" from "u/Ch1cken_Nugget_eater" was removed because

[3] `https://www.reddit.com/r/TwoXChromosomes/`, [last accessed: 09.11.2022]

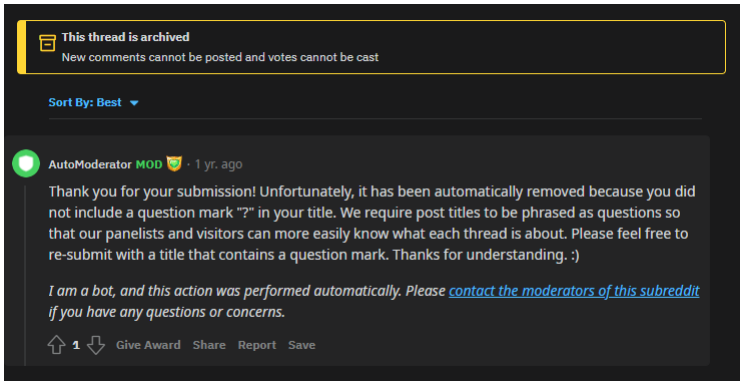

Figure 2 Example for the moderation of threads by the u/AutoModerator

it violated the first rule "Ask questions clearly and concisely in the title itself; questions should end with a question mark" from the "r/askscience" Subreddit. The "u/AutoModerator" automatically comments on why the post was removed (see figure 2).

These generic comments created by the "r/AutoModerator"-bot can be setup for the various Subreddits and the messages displayed by the bot are preformulated by the user setting it up, which might leave the possibility, that it inherits biases from the user. However in the dataset 206,056 comments are submitted from the "u/AutoModerator" artificially inflating the corpus with generic comments, which is distorting the dataset. That being said multiple bots are active in reddit for different purposes, which is why reddit formulated the so-called "bottiquette". Regarding the "bottiquette" you should "make sure your bot is actually adding something to the conversation it's posting in. A bot which says "Good post!" is pointless."[4] leads to bots like "u/WikiSummarizerBot". If a user includes a wikipedia link within his post, this bot is extracting the first paragraph of the linked wikipedia article and posting it as a reply comment. Like the "r/AutoModerator" this is also an example for bot comments inflating the corpus with comments that are not directly part of the conversation and that are not intended to be analyzed in terms of gender bias. For excluding these comments from the corpus as many as possible bots needed to be identified during research, to delete their comments based on the name. During research two web pages were identified, that ranked reddit bots based on their effectiveness or there popularity by measures like "Good bot votes", "Bad bot votes", "Comment karma" and "link karma" resulting in a score from 0 to 1. While the domain "https://botranks.com/" from Brandon McFarlin has a total of around 4,800 bots in their ranking, the domain "https://botrank.pastimes.eu/" has around 75,000 with a total amount of around 76,000 unique reddit bots.

[4] `https://www.reddit.com/wiki/bottiquette/`, [last accessed: 09.11.2022]

Regarding this medium article [5] the "https://botrank.pastimes.eu/" is implemented and provided by Branko Blagojevic, but the original idea of identifying and ranking bots in reddit regarding to him goes back to the user u/goodbot_badbot. While the user and the site do not exist anymore since 2018 (last successful snapshot 26.07.2022), the side can still be found in the waybackmachine. There the goal of the project is stated: "GoodBot_BadBot keeps track of the public's opinion on bots. Each user may vote one time per bot by replying to the bot with "good bot" or "bad bot". The aim is to identify which bots are the most and least popular. Please message GoodBot_BadBot if a user is not a bot, or if a real bot is not getting through the filtering."[6] The last successfull snapshot of the ranking of all bots covered by this project was taken at the 26.09.2017 and included a total of 1,306 bots [7]. The underlying script for u/goodbot_badbot always grabbed the 100 latest comments in r/all and scanned it for users replying to a bot's comment with either "good bot" or "bad bot", to add the name of the bot into the database and count the evaluation of the user.[8]. Based on the work of the team behind u/goodbot_badbot Branko Blagojevic used the existing rankings to further maintain a botranking webpage. All three projects follow a similar approach, by letting the users rate and indirectly manually annotate all the active bots. To be able to exclude the mentioned bots from the dataset the lists from were scraped from the webpages. It turned out, that the dataset included 16,137 bots from the list, that participated with 572,431 comments.

6.3 Pattern matching and deletion

The "u/WikiSummarizerBot" mentioned before uncovers another problem, because many users are posting links to other webpages, videos, gifs or photos. The markdown syntax can also be applied within a reddit comment/submission, therefore links are often embedded in a word like this:

$$[Linktext](https://dummy-link.org)$$

The link text is therefore often a part of a sentence and needs to be kept, while deleting the brackets and the embedded link. The link pattern for webpages, videos etc and links embedded in the markdown syntax where defined and deleted with regular expressions, before deleting the punctuations, to keep the traceable patterns. In addition some users even posted file names like "shockedpikachu.jpg". All of this are common phenomenons in social media platforms, but are leading to pointless

[5] Medium article, [last accessed: 15.10.2022]

[6] `https://web.archive.org/web/20180726121354/https://goodbot-badbot.herokuapp.com/`, [last accessed: 09.11.2022]

[7] ibid., [last accessed: 09.11.2022]

[8] `https://github.com/woodske/GoodBot_BadBot`, [last accessed: 09.11.2022]

tokens for the NLP Pipeline, which is why they needed to be extracted with the help of additional regular expression patterns. Furthermore another social phenomenon is tagging of people, Subreddits and hashtags which also leads to the same issue and that were faced the same.

6.4 Removing special characters and non-alphabetic characters

During preprocessing some digital typesetting artifacts were identified that were not resolved, like Non-breaking space (** **), the ampersand (**&**) or the "greater than" sign (**>**). These tokens needed to be removed before removing all the non-alphanumeric characters, to be able to make use of the characteristically pattern. In addition emojis needed to be extracted by using the database of emojis from the python library emojis. By building up a complete regex pattern based on the UTF-8 representation, emojis were also excluded. In addition all non-alphabetic characters like numbers, punctuation marks or line breaks were filtered with regular expression. Like mentioned in section "Reddit - Research object" the vast majority of reddit users are from countrys where english is the national language and in addition in international Subreddits the language used is english mostly. However during the analysis non-latin characters were found, partially within english comments. Because librarys like langdetect are pretty computationally expensive, regular expressions were used to exclude all characters, that are not covered within the Unicode range of **u0020** and **u007F**.

6.5 Tokenization and Stopword removal

For the training of the word embeddings, the comments and submissions needed to be separated into sentences and being splitted into tokens. For this purpose spacy's **en_core_web_sm** model was used. The documentation of spacy states that this model "is a small English pipeline trained on written web text (blogs, news, comments), that includes vocabulary, syntax and entities" [9], making it a good fit for the Reddit corpus. During the tokenization process every extracted token is compared to the english stopword list provided by NLTK and excluded when the token matches one of the stopwords. For the task the stopword list was adapted in that sense that the pronouns

$$['he', 'him', 'his', 'himself', 'she', 'shes', 'her', 'hers', 'herself']$$

were excluded from that list, because they can also carry valuable similarities with other female or male terms and are also included within the attribute sets of the WEAT Association tests and the centroid vectors, that are explained later on. Furthermore punctuations were deleted from stopwords like "don't" so that these

[9] `https://spacy.io/models`, [last accessed: 09.11.2022]

tokens still match the tokenized terms. However these tokens were already handled by the spacy tokenizer by splitting them up into "do" and "not".

6.6 Lemmatization

As a last step the generated tokens getting reduced to their root word. For this purpose the lemmatization was chosen over stemming, to be able to ensure that words are getting grouped up correctly, when the word embeddings are generated and to match the tokens in the NRC-VAD Lexicon.

7 Word Embeddings

7.1 Choose of model

Taking a look at the most recent publications in the field of word embeddings, the most widely used models in terms of pretrained models and selftrained models are still GloVe, word2vec and fasttext. For that reason these are also the models that were taken in consideration for this thesis. In addition, as stated in [12] based on the findings of (Devlin et al., 2018; Peters et al., 2018) contextualized embedding-based methods like the before mentioned, are suited very well for a more nuanced analysis of the language in terms of dehumanization.

7.2 Short model comparison

In general all three models are producing word-based embeddings that are aiming for semantic or syntactic similarity in order to optimize the model for a given task like text classification, part of speech tagging, named entity recognition or sentiment analysis. However structural differences in the implementation of these models making them more or less suitable for certain tasks. All three of them are dense static embedding models. Static means that the underlying method learns one fixed embedding representation for each word in the vocabulary of the training corpus. While sparse embeddings are capturing all co-occurrences of all words inside the vocabulary of the corpus leading to a matrix representation with a lot of zeros and huge dimensions, these models train embeddings that commonly have 50 - 600 dimension, with weights that are trained on the corpus. Unlike the co-occurence matrix, the dense vector representation is not really human understandable or having a clear interpretation. However dense vectors may do a better job in capturing synonyms, because the sparse vector representations for synonyms like car and automobile are distinct and unrelated. [23]
While fasttext is more like an extension to word2vec, the major difference lies between word2vec/fasttext and GloVe. While the GloVe model tries to capture semantic similarity based on global word-word co-occurences within the whole

corpus, to define similar terms, the models word2vec and fasttext try to identify similar words based on the surrounded words within a sentence. Therefore they are more focused on capturing local similarities based on the positive context words.[23] Unlike transformer based models like BERT it is not capable of differentiating multiple words occurrences with different meanings within the same sentence, because as stated above every term is just described by one word vector. However unlike fasttext they both have in common that they are trained on a word level, while fasttext uses n-grams of the words to capture semantic similarity. Making use of this might solve three problems that were faced, which is why the fasttext model was chosen for this task like in [24].

In general the use of ngrams instead of words has two advantages. At first the method is better in capturing semantic similarity for combound words. In general this is more useful in languages like german, so that the word "Torwart" and "Torwarthandschuhe" are semantically more connected by comparing multiple ngrams instead of just the two words. Another benefit is also related to the first benefit, because it solves the Out-Of-Vocabulary issue, meaning that the model is able to create a vector representation for a word that was not originally in the training corpus, by using words that share multiple ngrams with this word. For the WEAT analysis target and attribute words from already existing IAT tests will be used. Depending on the usecase a common preprocessing step for the training of word embeddings is the lemmatization of the corpus vocabulary. This is helpful to reduce the complexity of the model and probably also enhance the semantic similarity between words because of the aggregation. For this case it has the downside that attribute words like "children" or "mathmatics" occur, that are somehow related to their root words like "child" or "mathematic", although mathematic can also be seen as an adjective and child just refers to the singular form. For this reason the fasttext model was chosen, to be able to represent these words based on their root form and enhancing their meaning by words with similar ngrams. This allows to analyze rare words, but also words that are non-existent in the given corpus, without mixing up the meaning of words like "child" and "children" completely. Although the word "child" would affect the meaning of "children" it still stays as a standalone word with it's own representation.

Another advantage of the fasttext embeddings is, that typos in words still contribute to the semantical meaning of the original word. For example the teenagers Subreddit contained multiple typos for the word "woman" like "womam", "womann" or "womxn". It is just problematic, if it concerns a domain specific term, that should capture a different meaning. However these terms do not occur very often in the corpus. Another small disadvantage is that these terms occur with a very high rank, when measuring the cosine similarity of the word "woman". The top ranks tend to be occupied with this typos, but for the WEAT analysis it is not relevant. To fix

this issue libraries like autocorrect could be used, to correct these words, but this could also lead to a major distortion or manipulation of the data. For example the abbreviation "transfem" was autocorrected to "transfer" leading to a complete different meaning of the word within the sentence.

7.3 Hyperparameter

Like word2vec the fasttext embeddings also have the same two kinds of architectures for the optimization of the models: The CBOW model and the Continous Skip-gram model. While in the CBOW architecture the model tries to predict the next word based on the surrounding n-words, the Skip-gram model works the other way around by predicting the context words based on the input word. The comparison of CBOW and Skip-gram from [25] for the word2vec model shows, that the Skip-gram model significantly outperforms the CBOW method, when having a small text corpus and therefore low dimensionality and few training words. Especially for semantic evaluation it exceeds the CBOW accuracy by a lot. While the CBOW model worsens in semantic accuracy, when increasing the dimensionality and keeping the corpus the same, the Skip-gram model profits a lot from it and gains accuracy. This indicates that the Skip-gram model can handle rare terms better than the CBOW method. Overall the Skip-gram model typically outperforms the CBOW method in semantic tasks, while the CBOW method is slightly better in syntactic tasks. The good results for small text corpora, the good representation of rare words and the overall better performance in terms of semantic accuracy led to the conclusion that the Skip-gram model is better suited for the experiments, that were applied within this work.

Vector size 50		Vector size 300		Vector size 600	
Term	Score	Term	Score	Term	Score
manly	0.769	manly	0.521	manly	0.491
woman	0.768	woman	0.478	woman	0.428
boogie	0.758	dude	0.448	manga	0.414
male	0.757	guy	0.431	madman	0.395
boy	0.736	males	0.425	batman	0.387
female	0.716	men	0.425	dude	0.376
dude	0.689	manga	0.420	mans	0.370
men	0.679	soyboy	0.416	guys	0.361
masculine	0.476	feminine	0.413	masculine	0.360
slut	0.669	mans	0.411	men	0.359

Table 3 Comparison of the top 10 nearest terms to "man" based on the "r/Conservative" Subreddit, when varying the vector size

In table 3 the top 10 nearest terms for "man" are displayed based on three trained models on the "r/Conservative" Subreddit with different vector size dimensions.

The other parameters were identical for the training. While many terms occur when observing the top 20 terms, their position slightly varies. Some terms are also unique for one model like "batman" and "madman" for vector size 300. The cosine similarity however decreases by a lot when increasing the vector dimensions. Although "manly" is the most similar word to "man" for all the models it has a score of 0.869 for vector size and just 0.521 for the model with vector size 300. While the results are very similar, the computing time could also be a factor in terms of cost and benefit. However for the relatively small datasets the computing time is ≈ 7 seconds slower for a vector size of 300 compared to a vector size of 50. In general there is just a rule of thump, that with increasing size of the dataset, the vector size can be increased too. Due to the issue that the underlying datasets and therefore the amount of words within the corpus are very small, compared to pretrained word embeddings offered by word2vec or GloVe, a small vector size of 50 was chosen.
The findings of Levy and Goldberg et al show, that increasing the window size for the target word shift the embedding model from a model that represents similarity between words with similar meaning to a model that captures a broader topical content, that is also able to capture semantic relatedness. This can be seen in the results, where the five most similar words for target words like "hogwarts" are shown. While the word2vec Skipgram model with a window size of two suggests school names from other franchises like "evernight", "sunnydale" or "collinwood", the model with window size five suggests terms like "dumbledore", "hallows", "half-blood", "malfoy" and "snape" that are directly referring to the Harry Potter franchise in terms of semantical relatedness.[26] While within the thesis stereotypes or gender bias should be analyzed due to the semantic relatedness this is a desired effect, which will be made use of like in [12]. The window size was therefore set to 10, to make sure that every word of a sentence can be taken into account to achieve better semantic relatedness. According to that the "shrink_windows" parameter was set to "False", to ensure that all context words are equally taken into account.
When choosing the range of n-grams, that should be covered within the model, the findings of [27] show that for english texts the range of 3 - 6 is always a decent choice. It is stated that especially for analogy tasks longer n-grams increase the performance for semantic analogies. Furthermore it is stated that using ngrams ≥ 3 always improves the results, while ngram ≥ 2 "will not be enough to properly capture suffixes that correspond to conjugations or declensions, since they are composed of a single proper character and a positional one" (Bojanowski, 2016, p. 8).
Due to the stopword removal within the preprocessing step the sample parameter was set to zero, because the remaining frequently used tokens might be of relevance like the pronouns, that were excluded from the stopword removal in the preprocessing. This parameter is used as a threshold to define which high-frequency words should

be downsampled based on the amount of there appearance [10].
To reduce processing time the negative sampling was used instead of hierarchical softmax. This approach just calculates the new weights for every target word within one iteration for the terms within the window size and amount of k negative words, instead of rearranging the weights of the target word for every word in the corpus. When having a smaller corpus the amount of k is usually chosen higher between 5 and 20. Because all corpora are relatively small k was chosen at 20. Additionally the parameter ns_exponent, that controls which words are considered as negative words was chosen at 0.75 as it has been suggested by [28] for word2vec models. While choosing one would mean that only the most frequent words would be considered as negative words and zero would mean the frequency has no influence, 0.75 mainly prefers frequent words, but also leaves the opportunity for rare words.
For the sentiment analysis and the WEAT analysis two models were trained, having the same hyperparameter except for the min_count variable. For the sentiment analysis it was set to 15 to mainly tackle the issue of typos from words like woman in the most similar words. For the WEAT analysis however it was set to zero, to also consider rare terms for the training.

7.4 Evaluation

To test the quality of a word embedding model there are several datasets to evaluate the model in terms of semantic and syntactic quality. However having a good performance at one of this evaluation metrics does not automatically mean that the model is good for the task in mind as stated in the documentation of gensim. [11] However the results within these tests can still give valuable insights into the model quality. The gensim library already includes two datasets: One for the evaluation of each task and both were applied exemplary to one of the generated models. For the semantic evaluation task the **WordSim353 Test Collection**, that was introduced in [29], was used to evaluate the model in terms of similarity and relatedness. The dataset contains 353 word pairs that are human labeled with a score between zero and ten in terms of semantic similarity or relatedness. The underlying procedure for the evaluation process is described here [30]. The evaluation metrics are the Pearson's R and the Spearman's rho. When evaluating one of the trained models (model for "AskWomen" Subreddit), the results were surprisingly good without great efforts of optimizing the hyperparameters for this Subreddit. Pearson's R and Spearman's Rho were both resulting in ≈ 0.541 . The Association for Computational Linguistics provides some kind of benchmark from approaches of others literature's

[10] `https://radimrehurek.com/gensim/models/fasttext.html`, [last accessed: 09.11.2022]
[11] `https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html`, [last accessed: 09.11.2022]

on there domain [12], that shows that the model would still perform in average within this list of models. Although it should be kept in mind, that it mostly outperforms algorithms from the late 90s. The second evaluation was done on the analogy task introduced by [25]. The test contains multiple categories of evaluation tasks, that can be seen in Table 4. While the overall accuracy of the model is $\approx 44\%$, and therefore it has bad results in for example the geographical part, it does perform decent in the **family task**, that is very gender related with tasks like the famous man is to woman like king is to queen. For this evaluation there is also a benchmark provided by ACL, that shows that this model can compete with for example the skip gram model of Lai et al, with a way bigger corpus of 2.5 billion terms. [13] Applying these tests to the full corpus with all the 11 Subreddits included the metrics are nearly the same with similar observations. The used model for the Subreddit of "r/AskWomen" was trained basically on default parameters and window_size = 6, min_count=25,workers=12,sg=1, vector_size=50. When evaluating the model for this Subreddit with the final hyperparameters described in the word embeddings section, the quality in terms of the analogy results worsens a lot. It probably worsens mainly, because the threshold for the minimum frequency of occurrence for a word was set to 15 instead of 25 and therefore more questions in the evaluation tasks could be evaluated (2,947 instead of 2,329). But the analogies were therefore trained on just a few words due to the low frequency. However the hyperparameters could improve the similarity scores. (Pearson's R = 0.565, Spearman's rho = 0.578). Considering the small corpus, the results are decent compared to the model of Dobó et al from 2019.

8 Methodology

8.1 User classification - Reddit

In the conference paper of Morrison et al from 2013 he contributed an approach for identifying different kinds of user behaviour within the platform Reddit. This was done by recreating the egocentric reply graphs from all the users that participated in one of the 20 observed Subreddits within August 2011 and October 2011, to identify the amount of different user groups and their characteristics. The user centric measures were therefore inspired by the work of Chan et al [31] and Rowe et al [32]. The resulting vector representation covered nine different metrics of a user and were used for a k-means clustering, which yielded an optimal amount of four clusters. Afterwards the characteristics of the four clusters were interpreted and fitting user roles were assigned.

[12] `https://aclweb.org/aclwiki/WordSimilarity-353_Test_Collection_(State_of_the_art)`, [last accessed: 09.11.2022]

[13] `https://aclweb.org/aclwiki/Google_analogy_test_set_(State_of_the_art)`, [last accessed: 09.11.2022]

Task Category	Correct	Incorrect
capital-common-countries	0	12
capital-world	0	5
currency	1	17
city-in-state	0	25
family	**202**	**104**
gram1-adjective-to-adverb	356	294
gram2-opposite	194	148
gram3-comparative	23	67
gram4-superlative	2	4
gram5-present-participle	87	153
gram6-nationality-adjective	154	431
gram7-past-tense	0	30
gram8-plural	7	13
gram9-plural-verbs	0	0
Total:	1026	1303

Table 4 Results for the Google Analogy task for the default model trained on the "AskWomen" Subreddit

- **Contributors:** high initiation, medium engagement, high reciprocity and high popularity
- **Ignored:** high engagement, but low reciprocity
- **Lurkers:** very low engagement
- **Casual Commentators:** medium engagement but high reciprocity

These results were extracted from the clustering and then used for a decision rule set.

Group	Features
Initiation	th
Engagement	mpth, spth
Reciprocity	pr, bin , thbi
focus or breadth of interest	ent
popularity	ind, outdeg

Table 5 Groups for the different signals that classify the user role [3]

Unfortunately the description of the features, that were used for the clustering, often lacks the details of their definition, so in the following section the features, that were used, will be presented and clearly defined in the way they were interpreted and used within this work.

th describes the amount of new submissions that the user initiated. While Morrison et al. just defined it as " of submitted posts (new threads)", the article, that the measure was inspired by, named it "initiated %". The name clarifies that the

measure describes the percentage of threads that are initiated by the user. Dividing it by the total amount of submission lead to results in the same order of magnitude. The feature should help to distinguish users who initiated many threads from users who just reply. However during analysis it turned out that dividing it by the total amount of submissions, the th value was extremely small and therefore hard to differentiate. For this reason the total amount of submissions were taken and normalized for better interpretability.

mpth describes the average amount of comments, that the user contributes with in the submissions he is engaged in.

spth describes the standard deviation of the amount of comments, that the user contributes in the submissions he is engaged in. Both, the mpth and the spth, should provide information of the persistence of the user.

pr is described as "# of comments submitted by the user that received at least one reply" (Morrison, 2013, p. 2261) and also needs further clarification by the source article. In the original paper the feature is called "% Posts replied" indicating that it defines a ratio and gets divided by the total amount of comments from the user. A single comment or submission of a user can have many replies, but the average comment of the user just has a low percentage of replies, so this feature should serve as outlier adjusted measure about the acceptance of the user.[31] If nobody is replying to his comments, he might belong to the group "Ignored", that was identified in the paper. Therefore the amount of comments that achieved at least one reply were divided by the total amount of comments the user contributed.

bin is probably the most ambiguous feature, when following the descriptions of the papers. In Morrison et al it is described as followed: "proportion of a user's peers where bi-directional communication exists" (Morrison, 2013, p.2261) , while in Chan et al. it is described as "percentage of the neighbours of a user where there is both in and out edges (i.e. they have replied to each other)" (Chan, 2010, p.216). It leaves space for interpretation, when there is actually bidirectional communication. The additional example of Chan et al, mentioned above, leaves open, if they both have to reply to each other at least once, so you need the four comments like in figure 3. Otherwise, User B indicated by the red comments, already has one bidirectional communication, when he replies to someone and gets a reply by the same user.

Within this thesis it was assumed, that bidirectional communication is defined like case two in figure 3, so a user needs in and out edges. Another open question rises in terms of the calculation of the ratio. You could assume two ways. First is dividing it by the total amount of bidirectional communications, to compare the users with the other users. Alternatively it could be divided by the total amount of first replies of the user to analyze if the user tends to be involved in discussions. Looking at the

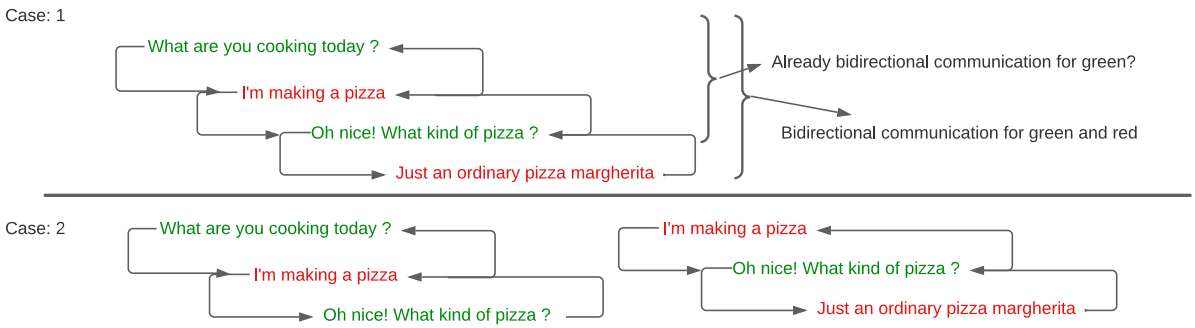

Figure 3 Illustration of bidirectional communication. Case 1: Visualization of the two possible options for bidirectional communication. Case 2: Final interpretation for bidirectional communication for User A and User B. [own representation]

average numbers of the bin features for the four user groups, it was decided that the second case was used. It would make sense, because overall the features are user-centered.

thbi describes the percentage of posts the user participates in where bidirectional communication exists at least once and divided by the amount of submissions where the user participated in.

ent describes the forum focus dispersion, so in the case of Reddit the amount of Subreddits the user engages in. In this specific usecase the percentage will be measured in how many of the 11 analyzed Subreddits the user participated in. This feature was excluded later on, because most of the users did not participate in two or more of the chosen Subreddits.

ind describes the amount of incoming edges to a comment of the user, but the last two features ind and outdeg also led to a lot of space for interpretation. These two features are as well defined as a ratio and therefore it must be clarified what the amount of indegrees should be compared to. You could argue that every user has the opportunity to answer to every comment in the given Subreddits, because all of them are public. You could also argue that it is unlikely that a user is not aware of new threads in Subreddits he did not join. Instead you could say everyone within the same Subreddit has the potential to answer the comment, but this would also take many users into account that are inactive, overlooked the thread or haven't joined the Subreddit at the time of the thread. So in the end it was decided to take only unique users into account that participated in this thread. The definition of Rowe et al raises another question. It is defined as "Number of followers of U" (Rowe, 2011, p. 409) within this paper, while the other two papers define it as the amount of incoming edges. So the question is if all unique users that replied to the user are divided by all unique users that had the opportunity. One assumption is that the users that replied are followers of the user, which is the main reason why they entered the communication with him. Alternatively all incoming edges from

the user will be divided by all comments that followed. The first assumption was used within this thesis.

outdeg describes the amount of outgoing edges from the users comment. Here the same questions arised like for the feature ind and answered with the same assumptions.

In contrast to the four clusters/user groups found in [3], the findings of Chan et al regarding the validation techniques and further manual inspection led to an optimal amount of eight or 15 clusters and the amount of different user roles respectively.

8.2 Sentiment Analysis

Like in [12] the NRC-VAD lexicon will be used within the Analysis part for comparing the sentiment of male and female related terms. It is suitable for this task, because it allows not just the comparison of female and male related words in terms of the valence, but also in terms of arousal and dominance. The lexicon contains around 20,000 english terms, rated by valence, arousal and dominance within a scope between zero and one. The resulting scores were manually annotated based on three crowdsourcing campaigns, one for each score, where people should rank 4-tuples of words.[33] The separation of the scores has the benefit that the hypothesis of [20] that "Women insult men by reference to unpleasantness in their personalities, but men insult women by reference to their availability for sexual use" (Stanley, 1977, p. 325) or in general in terms of the bias in meaning of gender related words can be analyzed not just on a valence level, but also in terms of sexism/arousal and dominance. Furthermore the lexicon approach works pretty well with single terms as output from the model, as long as they are covered within the lexicon. The Vader Sentiment library on the other hand could just assign neutral combound scores of zero for most of the words. Within this experiment the valence of the top 500 nearest words by cosine similarity to female or male terms were analyzed. The word vectors for female and male terms were weighted by frequency in the underlying Subreddit to gain their centroid vectors and were built based on the following terms like in [34]:

$$\overrightarrow{woman} = [\text{"}sister\text{"}, \text{"}female\text{"}, \text{"}woman\text{"}, \text{"}girl\text{"}, \text{"}daughter\text{"}, \text{"}she\text{"}, \text{"}hers\text{"}, \text{"}her\text{"}]$$

$$\overrightarrow{man} = [\text{"}brother\text{"}, \text{"}male\text{"}, \text{"}man\text{"}, \text{"}boy\text{"}, \text{"}son\text{"}, \text{"}he\text{"}, \text{"}his\text{"}, \text{"}him\text{"}]$$

The effect can be seen in figure 5 were the green centroid vector is influenced by the word vector for the different terms shown above. While in [12] this was done to capture all morphological forms of the label, in this case words with similar meaning

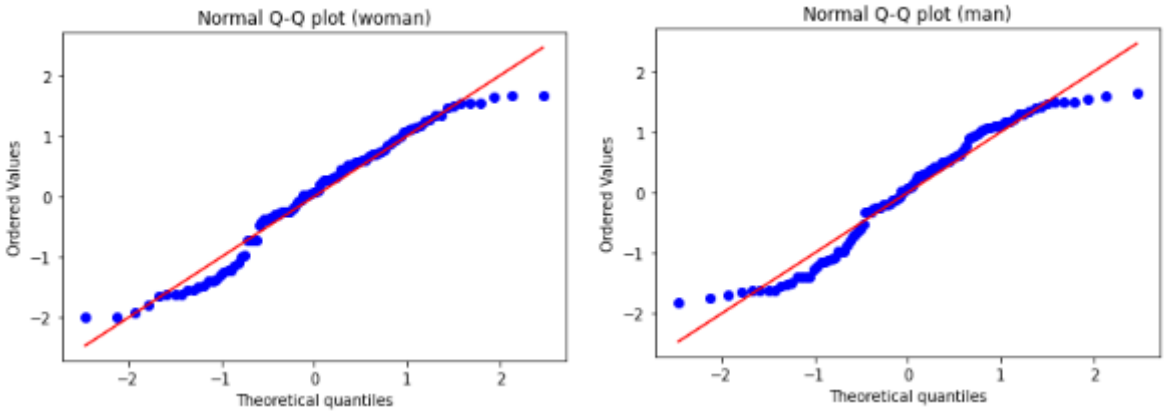

Figure 4 Q-Q plot for comparing the distributions of the valence of the top n terms related to woman and man to the normal distribution [own representation]

should be taken into account. The most similar words to the resulting centroid vectors were then evaluated in terms of their **valence**, **arousal** and **dominance** with the NRC-VAD Lexicon. To identify if there is a statistical significant difference of the female centroid vector and the male centroid vector, it will be tested if the measures follow a normal distribution.

If the measurements are normal distributed, the Student's t-test will be applied to gain information about statistical significance, and if not the Mann-Whitney U test will be used instead. In contrast to parametric tests like the Student's t-test, the Mann-Whitney u test does not require normal distribution.

Testing the distribution of the data can be done statistical and graphical. For this reason in the following the distribution of the "valence" for Subreddit "TwoXChromosomes" will be used exemplary [14] to identify the distribution graphically and statistically. While the statistical method can provide a statistical accurate result, it has some downsides as well. The main issue is that the test statistic and the p-value is very sensitive to higher sample sizes and can vary a lot, when the sample size is changed. For the statistical evaluation the method of Kolmogorov-Smirnov was used. In the first step the two samples were checked individually, if they are following a normal distribution. With p-values of $\approx 6.887e-25$ for both samples the null hypothesis is clearly rejected, because for p-values applies that: (Assumption: $\alpha = 5\%$)

$$6.887e-25 \ll 0.05$$

.

[14] All distribution figures regarding sentiment in : `https://github.com/AndyKruff/BA_Reddit_Gender_Bias/tree/main/experiments`

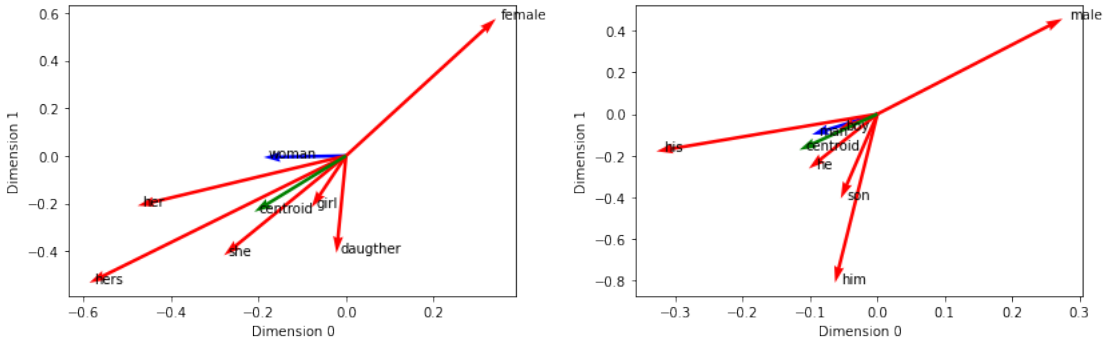

Figure 5 Building weighted averaged centroid vectors for woman and man for the fasttext embeddings based on the teenagers Subreddit. For the visualization the dimensions of the word vectors were reduced from 50 to two dimensions. It was created with the whatlies library. [own representation] [1]

When comparing the statistical test results to the graphical test in figure 4, the middle part of the graphs from -0.5 to 1.5 clearly follows the normal distribution or at least fits quite well to the normal distribution, the tails were however slightly deviated from it. Knowing that a sentiment analysis should normally be normal distributed, the graphical representation might be more trustworthy. However the findings of [35] indicate that pre-testing assumptions of the samples lead to unknown risks in terms of type-I and type-II errors, when the pre-testing is applied for the same set of observations, that should be tested. The authors therefore suggest to always use the Welch's t-test, a variation of the Student t-test that assumes unequal variances, instead of Student t-test or Mann-Whitney-U test, because it yields similar results as the t-test for homogenous variances. When the variances and skewness values are unequal it still has a robustness of 20 %, while the other two tests can not be recommended. In [36] the authors also suggest to prefer the Welch's t-test over Mann-Whitney-U and the Student T-test, therefore in the following sentiment analysis the Welch's t-test was used.

The following Welch's t-test hypothesis tests are defined as:

- **H0:** The two samples of (valence/arousal/dominance) regarding woman and man share identical average values.
- **H1:** The two samples of (valence/arousal/dominance) regarding woman and man do not share identical average values.

For further analyzing the strength of the significance of the Welch-Test results Cohen's d was calculated as:

$$d = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2 + s_2^2}{2}}} \tag{1}$$

8.3 Word Embedding Association Test

Word Embedding Association Test is the attempt to apply the IAT on word embeddings to identify biases within the model and the underlying text corpus. While in the original IAT it is the time difference that is measured to identify possible association biases, the WEAT uses the cosine similarity between the word vectors of the target and attribute sets. Within the study of Caliskan et al the same association biases that were found in the studies with the IAT could be identified within the word embeddings with the WEAT method. For the WEAT the null hypothesis, that will be tested, is defined as:

- **H0:** There is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words. [4]

The WEAT Score is therefore defined as

$$WEAT(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{2}$$

where $s(x, A, B)$ and $s(y, A, B)$ are defined as

$$s(w, A, B) = \frac{1}{n} \sum_{a \in A} cos(\overrightarrow{w}, \overrightarrow{a}) - \frac{1}{n} \sum_{b \in B} cos(\overrightarrow{w}, \overrightarrow{b}) \tag{3}$$

The bias measure can have a value between -2 meaning that the association is the other way around and +2 meaning there is a total association between Target Set 1 - Attribute Set 1 and Target Set 2 - Attribute Set 2. Zero would indicate that there is no difference in the association of the Target and Attribute sets.
The cosine similarity of every target word gets therefore calculated for the attribute words in A and B. It has to be considered that other than the IAT, this method cannot measure the association of a single pair of target words to one attribute, but of the complete target sets and the attribute. Therefore the association for single pairs of target words cannot be traced back. To evaluate the relevance of the association score a permutation test is applied and the degree of the significance will also be measured by Cohen's d as the effect size similar to the sentiment analysis.[4] As stated in [37] the choice of terms for the attribute word sets can manipulate the results, because any term could be considered gender biased, when the cosine similarity is higher for one of the target groups. So by creating new attribute sets, subconscious stereotypes or trial and error might have manipulated the results. To avoid such flaws predefined queries were taken from the WEFE library (see Jupyter Notebook [15]), that were aggregated as "Gender Queries" and were based on the

[15] `https://github.com/dccuchile/wefe/blob/master/examples/WEFE_rankings.ipynb`, [last accessed: 09.11.2022]

queries used in [4] and [38].

8.4 Analysis of most similar adjectives

The sentiment analysis only can give inside about the average scores of valence, arousal and dominance for the words that are covered within NRC-VAD lexicon. For the WEAT analysis just already existing tests were reproduced to avoid the problem that choosing the words manually might have led to subconsciously choosing them to produce more decisive results as described in section Word Embedding Association Test. For this reason this additional analysis was performed. So to be able to gain insight into the most associated adjectives to the female and male centroid vectors including slang words like "dateable" or "fugly" parts of the approach of [39] were used to identify the words that are most associated with the two word vectors. In the paper it was suggested to pay special attention to adjectives (POS: [JJ, JJS, JJR]), because they "are particularly interesting since they are modify nouns by limiting, qualifying, or specifying their properties, and are often normatively charged. Adjectives carry polarity, and this often yield more interesting insights about the type of discourses" (Aran, 2021, p.3). This was done by filtering the top-n most similar words by the identified POS tags with the help of the NLTK library. The bias measure was calculated very similar to WEAT and was defined as:

$$Bias(w, c1, c2) = cos(\overrightarrow{w}, \overrightarrow{c1}) - cos(\overrightarrow{w}, \overrightarrow{c2}) \tag{4}$$

In the analysis the top-30 most similar adjectives were taken into consideration for both centroids respectively. The centroid vectors were defined the same like in the underlying paper or in the sentiment analysis. Similar to the sentiment analysis the centroids were calculated by the weighted average of the words, to capture the meaning of the most frequent words, assuming that the related words are most expressive. The adjectives were analyzed half manually by identifying the terms that share a high intersection between the research objects (e.g. Subreddits, user groups) and were labeled with the help of the USAS that has an official implementation in python called pymusas, to gain a first insight about the characteristics, that the adjectives share. Furthermore the adjectives can be aggregated to a kind of topic group like "Anatomy and physiology" or "Health and disease" defined in USAS. The USAS is based on the Longman Lexicon of Contemporary English from 1981 by Tom McArthur and was developed by Rayson et al in [40].

9 Results

The analysis part will be divided into two separate experiments. In the first experiment the datasets were analyzed regarding a potential gender bias on a

Subreddit level. In the second experiment it was attempted to identify different kinds of user roles at first [3] within the extracted corpus, to analyze the potential gender bias on a user group level. For identifying these user roles at first the optimal amount of centroids for a k-means clustering must be found to compare the results with the results of Morrison et al. The procedure for identifying a potential gender bias will be identical for both experiments. At first the Sentiment analysis was performed as described in "Sentiment Analysis". In a second step the WEAT method was applied to the fasttext models to analyze if for example certain family related terms are more associated with female terms as with male terms. The used attribute and target sets can be found here [16]. At last the top 30 most similar adjectives regarding the centroids, that were also used for the sentiment analysis, were analyzed as described in "Analysis of the most similar adjectives".

9.1 Results - Subreddit Level

9.1.1 Sentiment Analysis - Subreddit level

Welch's t-test statistics (p-values)			
	valence	arousal	dominance
r/TwoXChromosomes	**0.001**	**0.015**	**0.030**
r/gardening	**0.003**	**0.009**	0.588
r/teenagers	0.342	0.109	0.558
r/funny	0.855	**0.003**	0.246
r/AskWomen	0.427	0.383	0.321
r/AskMen	0.796	0.054	0.512
r/Parenting	0.447	0.697	0.798
r/science	0.247	0.099	0.339
r/technology	**0.041**	0.220	0.085
r/unpopularopinion	0.982	**0.014**	0.064
r/Conservative	0.238	**0.012**	0.060

Table 6 P-values for two-sided Welch's t-test. P-value > 0.05 means H0 is not rejected

In the first analysis the two-sided Welch's t-test was applied to NRC-VAD Scores for the most similar words to the centroids of woman and men. Like in [12] the top 500 most similar words according to the cosine similarity were taken into account, although unfortunately a lot of words were not covered in the NRC lexicon and therefore the words to be analyzed were limited by the coverage of the lexicon. The amount of words varied and depended on the Subreddit. The lowest amount of intersections between the two lists was 124, while the highest amount was 381. This might be the case, because some communities might be using a lot of slang words, that are common within the Subreddit or the internet, but very likely not covered in

[16] `https://github.com/AndyKruff/BA_Reddit_Gender_Bias/blob/main/scripts/WEAT%20Querys.ipynb`

such a lexicon. The fact that the amount of intersection for the most similar words for man and woman was quite similar within the same Subreddit also supports this thesis.
That being said Table 6 showed the resulting p-values for valence, arousal and dominance in all the Subreddits. A p-values below $\alpha = 0.05$ indicated, that H0 was rejected and that there was a significant difference in the mean of this two samples. Therefore it could have been identified which sample had a significantly higher sentiment in terms of the mean. There were eight p-values within six Subreddits, where the sentiment of the related terms was significantly higher for one sample. The valence significantly differed for the Subreddits "r/TwoXChromosomes", "r/gardening" and "r/technology". While for "r/technology" the average valence was higher than for woman with a small effect size of ≈ 0.172, for the other two Subreddits the terms for woman had a higher mean and were therefore associated with more positive words. But with an effect size of ≈ 0.221 for "r/gardening" and ≈ 0.291 for "r/TwoXChromoses" the effect was also considered small. For the arousal the Subreddits "r/gardening" and "r/TwoXChromosomes" had significantly higher averages for the male terms, while the other two Subreddits "r/Conservative" and "r/funny" had higher means for the woman terms, but for the arousal the effect size for all Subreddits was also pretty low between ≈ 0.196 and ≈ 0.290. A significant difference for the dominance could just have been identified in "r/TwoXChromosomes". Here the mean dominance related to the female terms was higher, but the effect size was also low with ≈ 0.188, indicating that some significant differences were found, none of them were strongly significant. However it was interesting to see that the female terms in the "r/TwoXChromosomes" were all significantly different to the male terms in favor of the woman, while the other Subreddits which call themselves conservative or probably having a broader male community showed significant differences in favor to the male terms.

9.1.2 WEAT Analysis - Subreddit level

From the eight queries taken from [4] and [38] two needed to be excluded due to a lack of word representations within at least one or more target or attribute sets and therefore none of the Subreddits had any result for the queries "Male terms and Female terms wrt Science and Arts" and "Male terms and Female terms wrt Positive words and Negative words". The same applied for all missing barplots in figure 6. These were not zero but no results could be calculated. The reason for this was the transformation of the models for the WEFE library, that led to a high loss rate of words, that exceeded the threshold of 0.2. While most of the biases measured followed the associations that were expected, the association bias was mostly very small. An interesting observation was, to see that for some queries, especially regarding query 5, the association was reversed and therefore target set 1 was more

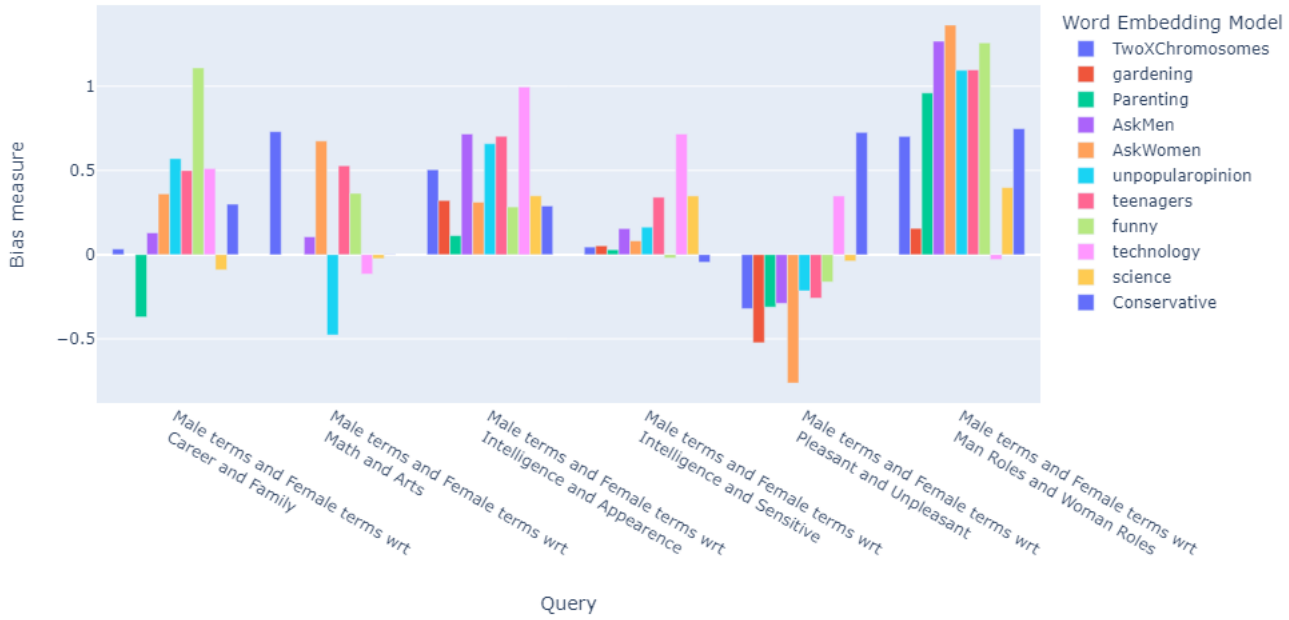

Figure 6 WEAT Test results on Subreddit level

associated with attribute set 2 and vice versa. Although the Subreddit "r/Parenting" had negative association of ≈ -0.371 for the query "Career and Family" and the Subreddit "r/unpopularopinion" had an even stronger negative association with the query "Math and Arts" with ≈ 0.477, these results could both be considered as not significant regarding the p-value. Although "r/AskWomen" had a quite high association bias regarding query 5 with -0.763, the null hypothesis could not be rejected. In contrast the "r/Conservative" Subreddit with 0.726 showed a significant association bias towards query 5 with a strong effect size of 0.977. On the other hand there were certain Subreddits that had significant positive association scores. While there were many Subreddits with scores in the range of 0.5 regarding the association, they were all not considered significant. But also higher scores for the bias measure could not automatically be considered significant, for example the bias score of 1.109 for query 1 and regarding the Subreddit "r/funny" still got rejected by the permutation test (p-value = 0.059). Besides query 5 just the Subreddit "r/technology" had significant results for other queries, namely "Intelligence and Appearance" and "Intelligence and Sensitive". The graphs showed that most of the Subreddits (8/11), all except for "r/gardening", "r/technology" and "r/science", yielded very high and significant bias measures for the association test concerning stereotypical occupations regarding the gender.

9.1.3 Analysis of the most similar adjectives - Subreddit level

Figure 7 shows the most similar adjectives from the observed Subreddits for the centroid vectors of woman and man that occured in the top 30 of most similar adjectives in three or more Subreddits. The size of the word is related to the amount of occurences in multiple Subreddits. The underlying models were the models used for the Sentiment analysis to exclude typos. Although the terms were filtered by their POS tag with NLTK, some nouns like "ovary" were still occuring in the results. However these terms still followed the same observations as the adjectives.

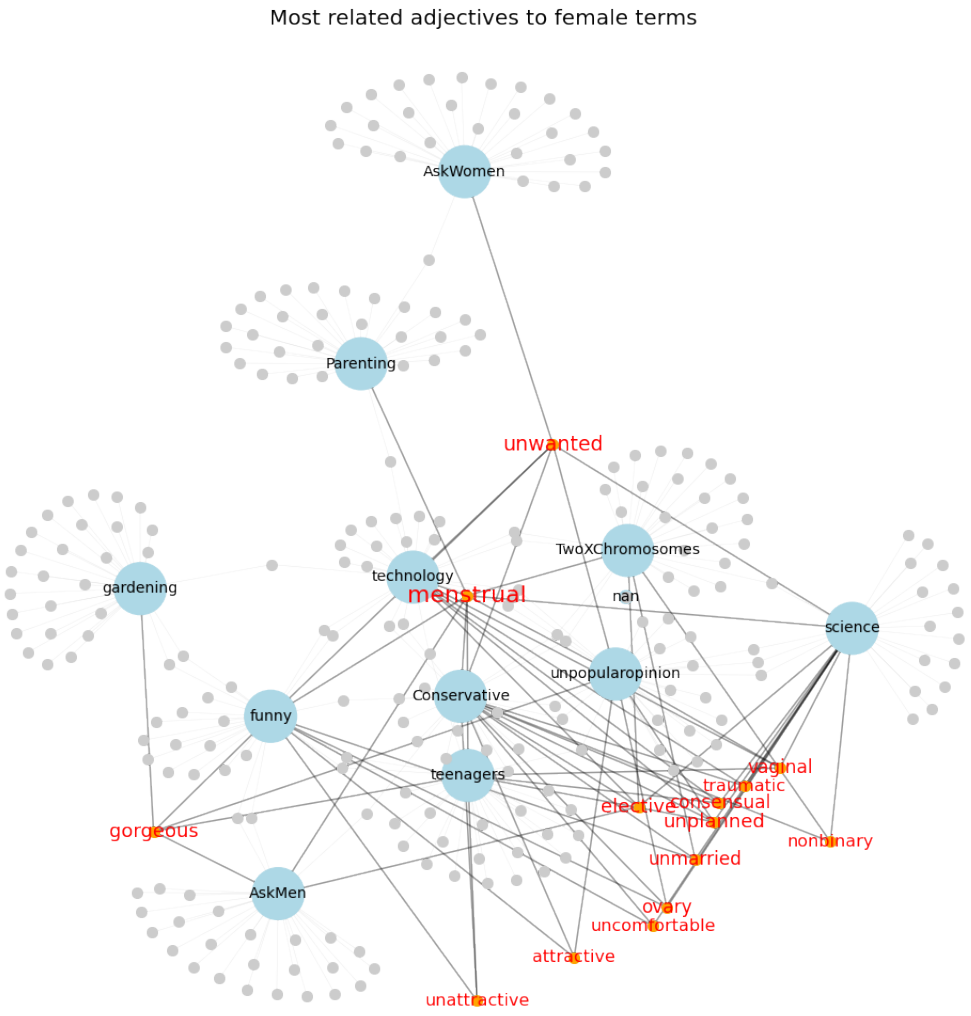

Figure 7 Most frequent adjectives related to female terms. Just words with occurrences in the top 30 in more than **two Subreddits (degree $>$ 2)** are displayed [own representation]

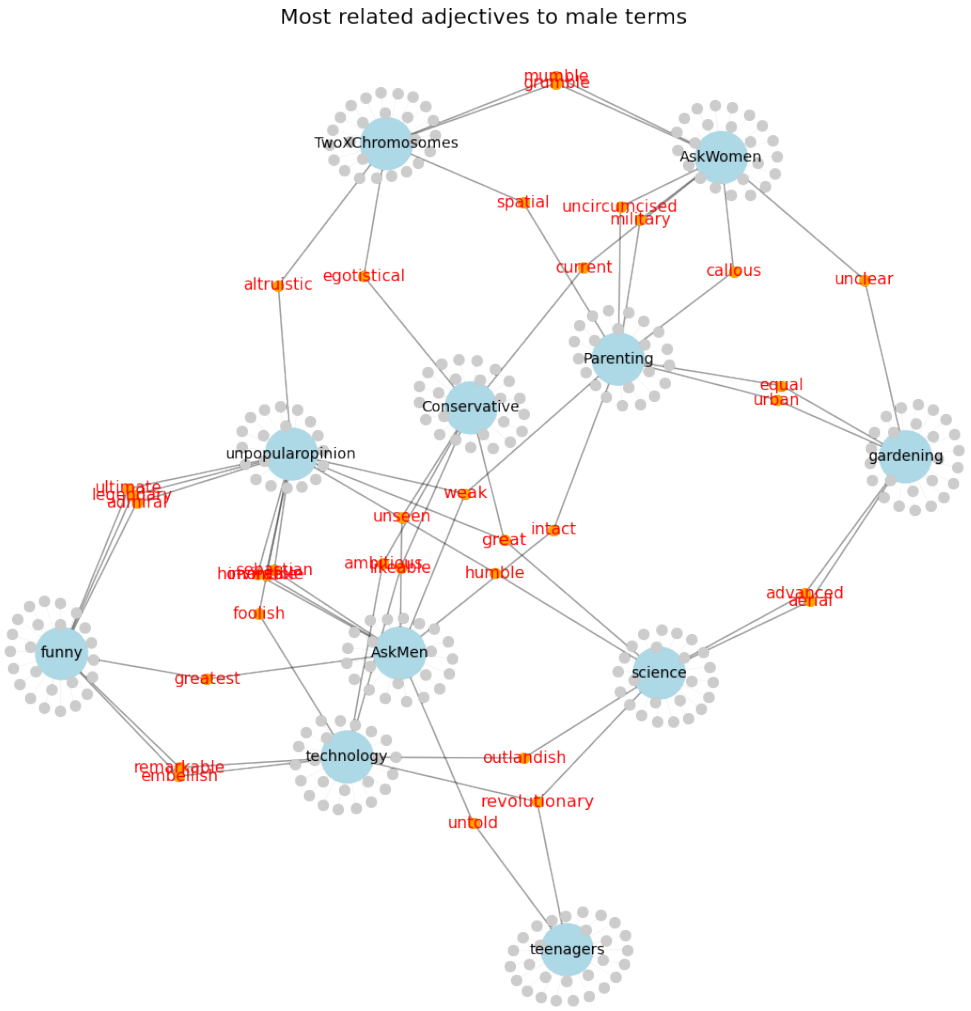

Figure 8 Most frequent adjectives related to male terms. Just words with occurrences in the top 30 in more than one Subreddit (degree > 1) are displayed [own representation]

When having a look at the biggest topic set for both vectors it was "Z99", which covered all unmatched words. However the next bigger groups did yield an insight about differences for man and woman. While for woman the biggest fitting sets were "B1", "T1.3", "X7-", "Z1mf/Z3c", "O4.2+", "B3", "A6.1-" , "S4" and "A6.1+". "B1" is therefore defined as "Anatomy and physiology" and was assigned to words like "menstrual", "ovary", "unborn" and "cervical". These words were not just general anatomy words, but directly linked to the childbearing abilities. On the other hand "B3" is defined as "Medicines and medical treatment" and was assigned

to terms like "surgical", "aborted" and "epidural", where at least the last two might have also referred to the female reproduction. "O4.2+" and "A6.1-" are defined as "Judgement of appearance (pretty etc.)" and "Comparing: Similar/different". While words like attractive, gorgeous, adorable, desirable and pleasurable rate the woman on their appearance, words like incompatible and contrary might have rated women in terms of their adaptedness in society. "S4" is defined as "kin" and was assigned to words that referred to the relationships within the family. When having a look at figure 7 the intersection of related words in the top 30 varied. While "r/AskWomen", "r/Parenting" and "r/gardening" just shared one adjective with other Subreddits, Subreddits like "r/funny" shared up to eight adjectives with other Subreddits. The Subreddits in the middle to the right are therefore the once, that shared the most adjectives as it can be seen by the amount of the intersecting edges. It should be noted that figure 7 only displays adjectives with a degree > 2, to ensure clarity and readability, while in the other networks it was filtered by degree > 1, because the amount of intersecting terms was a lot lower.
The adjectives regarding man on the other side were mostly part of the sets "A5.1+", "O4.1", "A11.1+", "A6.2-", "S7.1+", "A13.3", "A1.1.1" and "S1.2.6-" (Definitions can be found here [17]) and were all referring to the social status and the character traits a man gained in terms of importance, physical properties or power. What figure 8 additionally shows is, that although there were certain topic groups for the adjectives for men, the associated adjectives were very different. The term "weak" in the center below "Conservative" only occured in three different Subreddits in relation to men, while "menstrual" occured in nine out of eleven Subreddits for women. Most of the related adjectives for men that are displayed in figure 8 were shared between two Subreddits. It was interesting to see that in the figure Subreddit "r/AskMen" and "r/Parenting" were emerging to some kind of cluster centroids, by having multiple intersections between the Subreddits around them. Furthermore it showed some kind of similarity between the Subreddits. While some like "r/funny" and "r/unpopularopinion" shared three adjectives with each other, other Subreddits did not share any intersection. Therefore it seemed like there was a shared common sense between the users regarding the adjectives associated with women, while men were getting evaluated within the same topics, the associated adjectives differed between the Subreddits. While these terms might have been expectable for contextual Subreddits like "r/TwoXChromosomes", "r/Parenting", "r/"AskWomen" or "r/science", similar associations were drawn in Subreddits like "r/technology", "r/unpopularopinion" or "r/Conservative".

[17] `https://ucrel.lancs.ac.uk/usas/usas_guide.pdf`, [last accessed: 09.11.2022]

9.2 Results - Group Role Level

9.2.1 Identifying User Groups

Before beginning with the clustering the metrics, that were used, were evaluated in terms of correlation, to see if all metrics were uncorrelated and therefore meaningful.

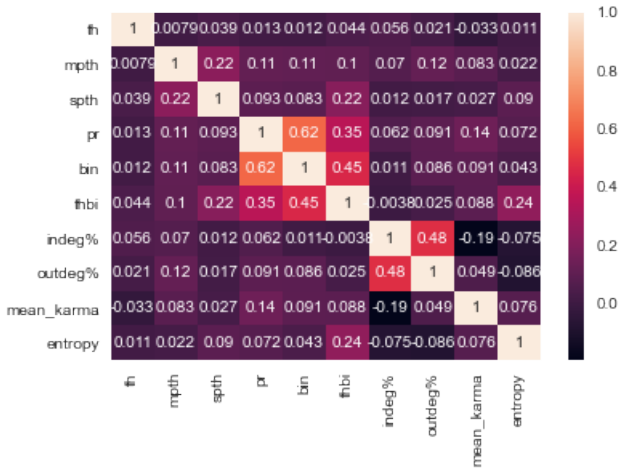

Figure 9 Correlation matrix for the metrics used in the clustering
[own representation]

When having a look at table 5 the features that were aggregated by Morrison et al were more correlated than others. So the engagement features were slightly more correlated with a score of 0.22. The reciprocity and the popularity features had a lot higher correlation, which was explainable due to their similar nature. However when looking at the findings of Morrison et al, the role of the Contributor was characterized by high pr, the bin was not that relevant for the decision rules in that case. For the reason that the unique features still carried unique valuable information and with the exception of the correlation between bin and pr the other correlation scores were not too high and it was decided to not exclude features due to their correlation.

Trying to reproduce the results of Morrison et al. the k-means algorithm was used for the clustering process of the users and their potentially different behaviours. While the k-means clustering has the benefit that it just needed the amount of centroids, the stability between independent clusterings is often problematic with k-means as stated in [3]. To address this problem the n_init parameter of sklearn [18] was used, to run the algorithm for n amount of times with different centroid seeds and the best result of the consecutive runs in terms of the summed up squared distance in relation to the nearest cluster was taken, as suggested in [3]. While in EDA Clustering finding the right amount of centroids is often problematic, in this case

[18] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`, [last accessed: 09.11.2022]

the analysis started with the assumption that four centroids might be a good fit for the data. After filtering out the bot accounts and the accounts that were deleted or removed from the dataset, it contained a total of 1,599,026 users for the clustering procedure. As stated above, to reproduce the results of the original article, four centroids were assumed as fitting for the first run.

Cluster	Cluster Size
0	1,589,735
1	2
2	1
3	9,288

Table 7 Clustering size for four centroids

The results of Table 7 showed very imbalanced cluster centroids, with ≈ 99% of the users being assigned to cluster 0. While in [3] it was already stated, that the user groups Ignored and Lurker could take up to 50 % of a Subreddit, these results did not represent the findings of the original article.
To identify the optimal amount of clusters for the underlying dataset the elbow method was used.

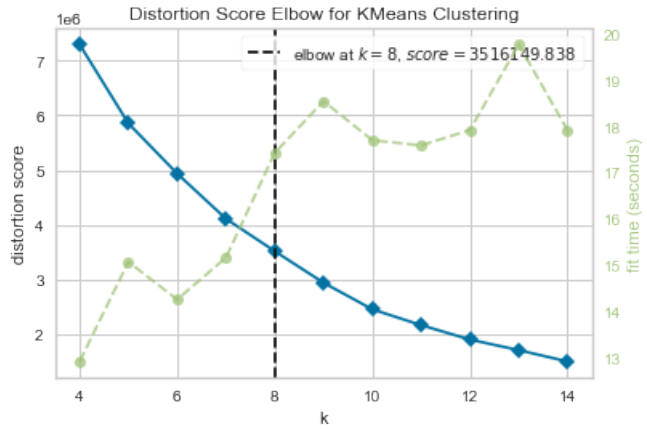

Figure 10 Identifying optimal amount of cluster by yellowbricks elbow visualizer [own representation]

Figure 10 shows the Distortion Score for the clustering of all users with centroids from 4 up to 15. According to these results the optimal amount of centroids was 8, because of the optimal ratio of the distortion score in relation to the computation time needed to fit the cluster.

When applying the kmeans algorithm with eight centroids on the dataset, there were still imbalances between the cluster sizes, but there were three to five clusters that might be of further interest. The two biggest clusters were characterized by overall low to medium values in all features, which made them potential candidates for the user groups Lurker and Ignored. While users in cluster 0 had a slightly higher

Cluster	Cluster Size
0	83355
1	1
2	1
3	1824
4	1
5	1512872
6	935
7	37

Table 8 Cluster sizes for 8 centroids as suggested by the findings of the elbow method

engagement than users in cluster 5, it was not high compared to other clusters. The two clusters both would be potential clusters for Lurkers or Ignored, but did not fit perfectly, especially for Ignored according to the low engagement metrics. Clusters 6 and 7 were very small, but they followed certain characteristics of the Contributor and even the size of the clusters indicated, that this clusters might represent the contributor, because in Morrison et al it was said that contributors are just a small core population. Although the two clusters just contained a total of 972 users, these users contributed 77,089 comments and submissions. Cluster 0 and 3 both had medium engagement and a medium to high reciprocity and would therefore fit to the description of the Casual Commentators. Although not all characteristics of the user clusters aligned with the findings of Morrison et al., the findings could still be identified within the clusters. Probably the amount of users in the clustering led to a lot noise and harmed the quality of the clusters. Within Morrision et al 20 of the most active were observed, where the most active Subreddit had 24,450 active users in the observation window. While in [3] the clustering was done for every Subreddit, within this work the user roles should be identified within the whole corpus with a total of 1,599,026 to be clustered. However when applying a role labelling decision rule set like in Morrion et al the clusters might be able to be further refined. Unfortunately the role labelling decision rule set was also not unequivocally explained within Morrison et al, therefore it was applied as described in the following. To do so for all features the average was calculated and the findings for the different user groups, as shown in listing of section "User classification - Reddit", of Morrison et al were applied. The entropy feature was excluded because nearly every user just contributed in just one of the observed Subreddits. The same applies for the average karma score of ≈ 1, that is shared within every resulting cluster. At first Contributors were extracted from the dataset, when all the characteristic metrics described in the listing, were above average. Therefore 21,166 were identified. After that the Ignored were extracted by filtering for mpth above average and bin and thbi below average. These group contained 99,247 users. For Lurkers all users with

a mpth below average were taken and resulting in a group of 1,141,730. At last the Casual Commentators were extracted by medium values for mpth and the spth value in the range of the standard deviation from the mean and high reciprocity values above average. The last cluster contained 225,028 users.

9.2.2 Sentiment Analysis - User group level

When analyzing the most similar words of the different user roles just two null hypothesis were rejected. For the Contributor and for the Ignored the average dominance score significantly differed between the terms for men and women. With a mean of 0.514 for the female terms and a mean of 0.609 for the male terms, men were significantly more associated with dominance within the user group of Contributors. Unlike the results of the sentiment analysis of the Subreddits, this finding had a Cohen's d of 0.548 and it therefore indicated a medium effect size. With an average dominance score of 0.553 for male terms and 0.510 for female terms, the same applied for the user group Ignored. In comparison to Contributors the effect size was low and similar to the effect sizes of the Subreddits with a Cohen's d of 0.252.

Welch's t-test statistics (p-values)			
	valence	arousal	dominance
Contributor	0.743	0.614	**0.000**
Ignored	0.169	0.162	**0.002**
Lurker	0.425	0.361	0.534
Casual Commentator	0.337	0.479	0.190

Table 9 P-values for two-sided Welch's t-test. P-value > 0.05 means H0 is not rejected

9.2.3 WEAT Analysis - User group level

When analyzing the different user groups regarding the association bias, the same two queries did not yield any results or just pretty small scores similar to the Subreddits. This was expectable because the text corpus stayed the same, just aggregated into four groups instead of eleven. However overall the results for the user groups had a lot stronger characteristics, meaning that the Bias Measure are a lot higher than for the Subreddits. Similar to the results in the Subreddit analysis query 5 yielded negative bias scores for all user groups except for the Casual Commentator. Additionally query 2 for the Ignored and the query 4 for the Lurker showed reversed association biases. However non of these observations were considered significant regarding the permutation test. Like query 5, query 1 and query 2 do not yield any significant associations biases for any of the user groups. Although query 4 had very low WEAT Scores below 0.511, the association biases for the Contributor and

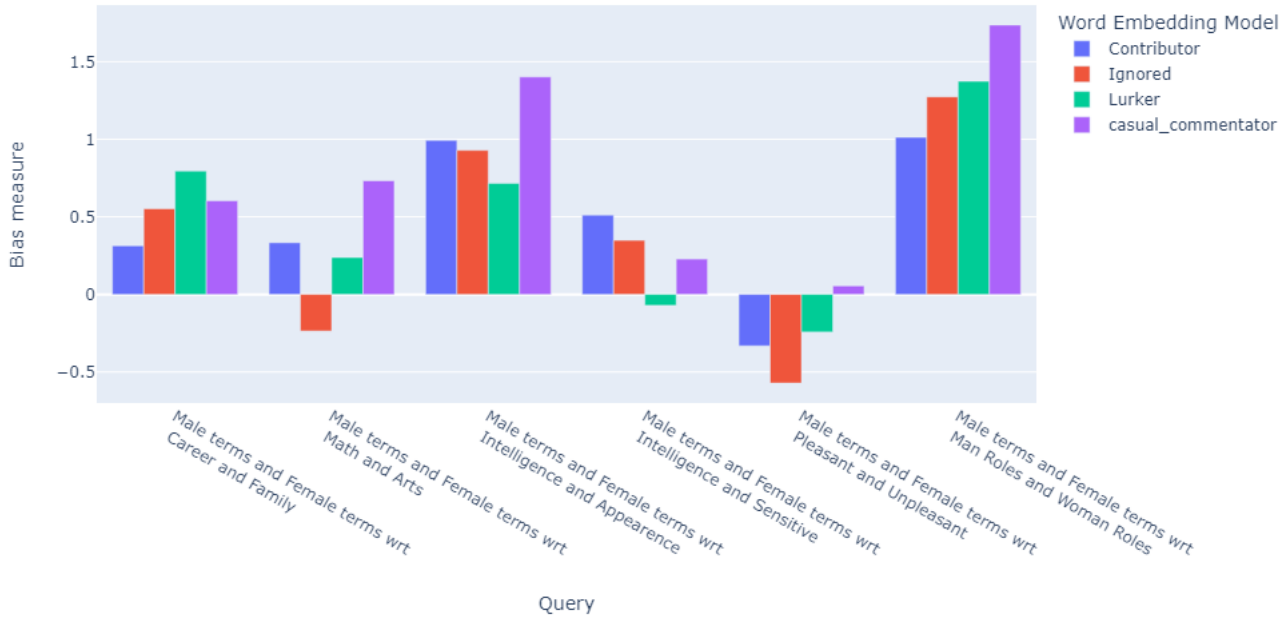

Figure 11 WEAT Test results for different user groups

the Ignored with 0.511 and 0.349 were considered significant by the statistical test. However this finding is one of the flaws underlying the mathematics behind WEAT described in [41] and should not be considered meaningful. Unlike the just analyzed queries and query results of query 1, 2, 4 and 5 the bias scores for the remaining queries were very high for the observed user groups. With WEAT scores of 0.993, 0.928 and 1.402 for Contributor, Ignored and Casual Commentator respectively, all of the biases were considered significant, with high effect sizes of 0.80, 0.71 and 0.97. The association biases for query 5 regarding stereotypic occupations were shared between all the identified user groups and showed even higher results. With 1.012, 1.273, 1.373 and 1.736 all scores were above 1 and considered significant due to the rejected null hypothesis. With high effect sizes of around 1 or higher, the significance for association bias were considered as highly relevant.

9.2.4 Analysis of the most similar adjectives - User group level

When analysing the most similar adjectives for the different user groups, the results were pretty much following the same common sense as described for the Subreddits. While "lesbian", "gorgeous" and "ovary" occured in the top 30 of every user group as it can be seen in the center of figure 12, the words "unwanted", "sensual", "spontaneous", "unplanned", "menstrual", "aborted" and "vaginal" occured in three out of four user groups. The figure additionally shows that the user groups Ignored and Contributor had a more similar vocabulary and the user groups Lurker and Casual Commentator were more similar respectively. On the other hand the

adjectives associated with men did not share any words within 3 or more user groups. Most of the terms were just present in the top 30 of one or two user groups. Like in the analysis for the Subreddits, the user groups barely shared any adjectives in the top 30. However words like "strategic", "powerful" or "presidential" followed the same observation of describing the man's status in society and his character traits that makes him successful, while many of the female related terms were associating woman with sexuality, female reproduction or the womans appearance.

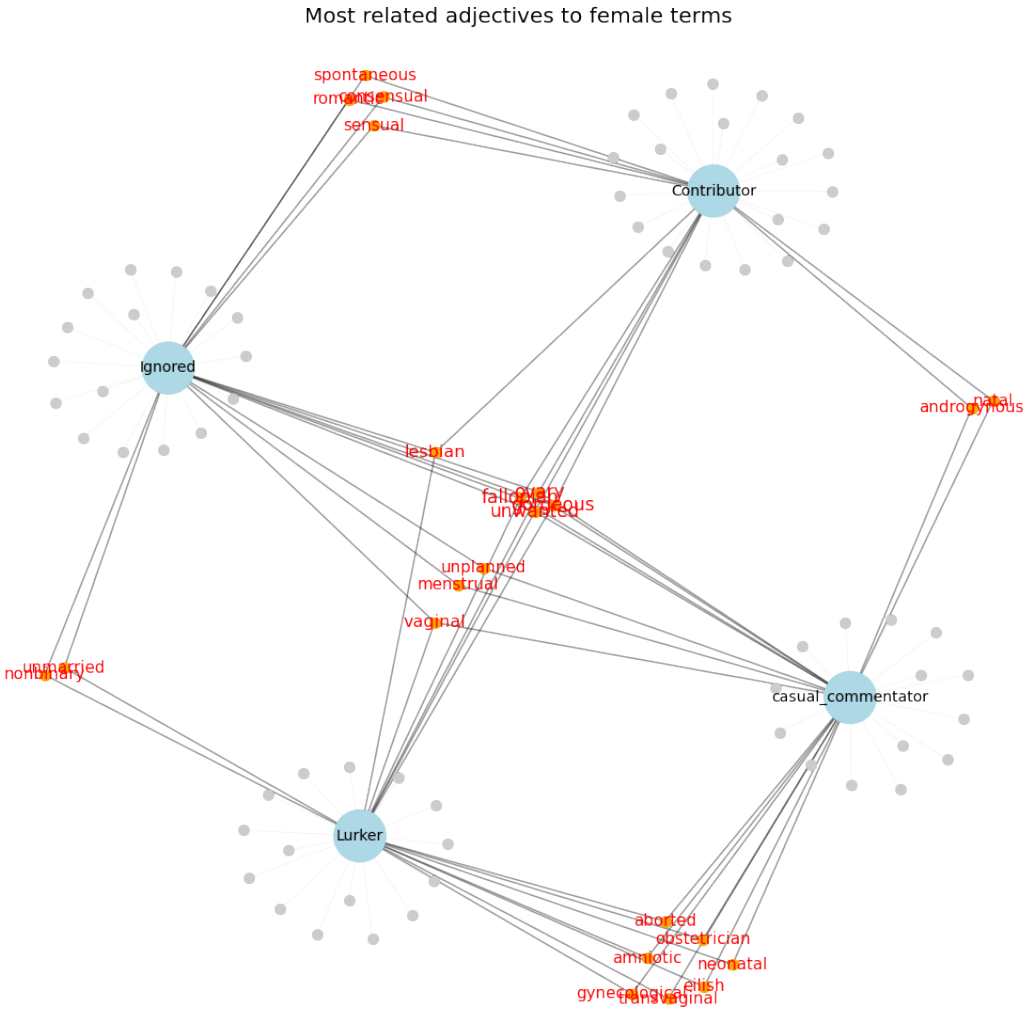

Figure 12 Most frequent adjectives related to female terms. Just words with occurrences in the top 30 in more than one Subreddits (degree > 1) are displayed [own representation]

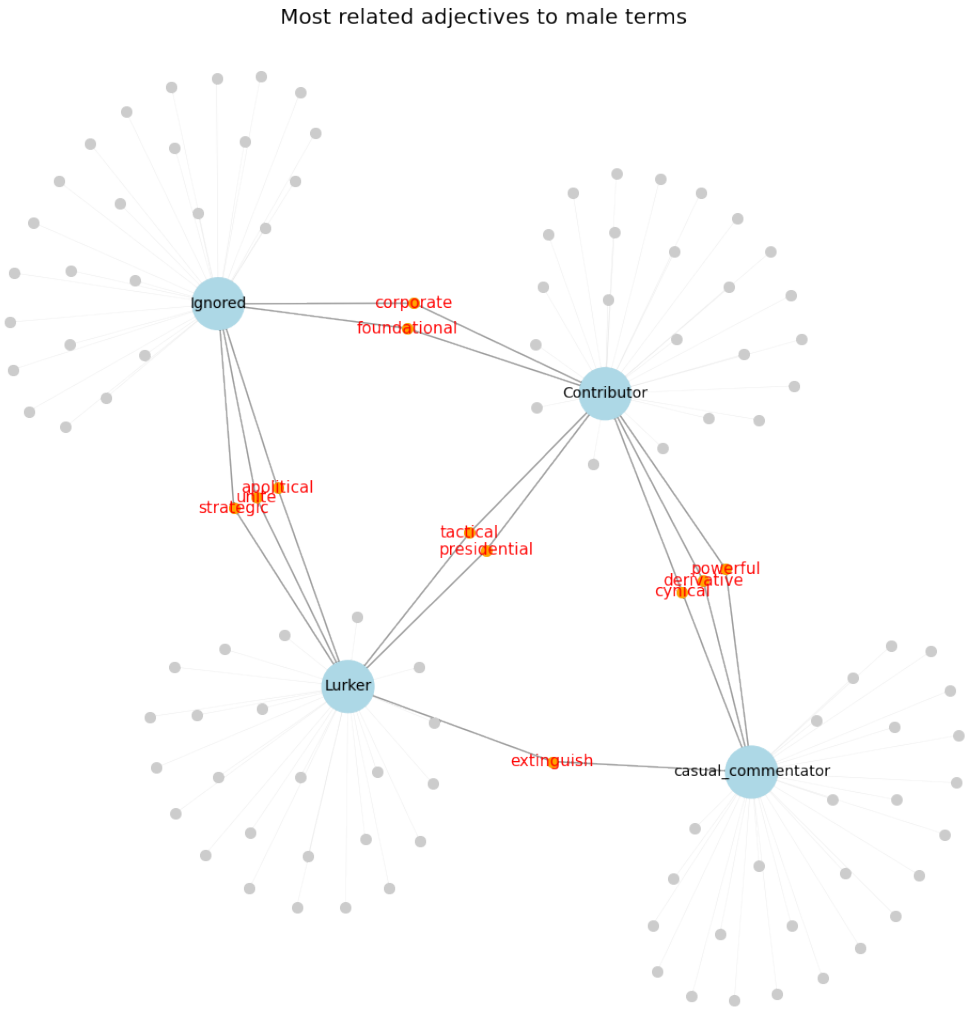

Figure 13 Most frequent adjectives related to male terms. Just words with occurrences in the top 30 in more than one Subreddits (degree > 1) are displayed [own representation]

10 Discussion

The aim of this thesis is to analyze the social media platform Reddit in terms of potential gender biases regarding different Subreddits and different user groups.
For the determination of the potential user groups, the findings of Morrison et al could partly be reproduced. However it turns out that the different user groups show very similar results for the sentiment analysis with few exceptions regarding the Contributor and the Ignored. While for the Ignored the effect size was still low, for

the Contributor the associated dominance for words related to men were significantly higher than for women and even got a medium effect size. However all user groups were sharing the same significant bias scores, with high effect sizes, for query 5 within the WEAT analysis. Furthermore, except for the Lurker, they all shared significant association biases for query 3 regarding "Intelligence and Appearance". For the other queries the identified user groups mostly followed the same associations. This leads to the assumption that the aggregation of the users led to very homogeneous groups, that mostly shared the same characteristics like the analysis of the single Subreddits. The significant results found in the sentiment analysis for the Subreddits all had very low effect sizes and were mainly not reflected in any of the resulting user groups. The differences for the Subreddits were either too small to be captured for one user group or could not be traced back to a single user group.

Although the sentiment analysis allows to gain general insight about the sentiment of the common related words, the high amount of slang words like "soyboy", that are not covered within the NRC lexicon nor covered in the USAS Semantic tagger, limited the analysis to mainly non domain specific words. For the analysis of the adjectives the NLTK POS Tagger still partly recognized slang words like "fugly" or "dateable", therefore some domain specific words could have been taken into consideration and the words still could be grouped manually. For the sentiment analysis many information got lost for the evaluation of the average valence, arousal or dominance score. Therefore it would be helpful to adapt existing sentiment lexicons for modern words or slang words in online communities.

Considering that the word embeddings were trained with the idea to enhance specific words like "mathematics" by using the n-grams of the fasttext approach and the WEAT scores tend to often be overestimating the biases in word embeddings [37], the amount of significant results for the Subreddits and the user groups turn out to be lower than expected. The most remarkable results within the thesis are the findings for the WEAT analysis of query 5, where it was shown that both, user groups and most of the Subreddits, share equal association biases regarding gender stereotypic occupations.

During the semi-manual analysis of the most associated adjectives for women and men, it turns out that the Subreddits respectively share a common sense in terms of reducing women to their childbearing abilities and their appearance. From the top 30 most similar adjectives most of the terms dealt with the female sexuality, her ability for reproduction with related terms like "menstrual" and the evaluation of her appearance. On the other hand adjectives most associated with men can be categorized as different adjectives describing the social status,the acknowledgement and character traits a man could achieve in society. While there is a high intersection between adjectives describing women, there is a lot variation for the adjectives used for men between the Subreddits. Nevertheless the adjectives for

men are topically related. Furthermore it can be observed, that two clusters were emerging in 8, showing that for example "r/TwoXChromosomes", "r/AskWomen" and "r/Parenting" seem to be relatively similar, just like the connection between "r/funny", "r/AskMen" and "r/technology". The biases found here fit very well to the "Bias in Meaning" and "Benevolent sexism" described in section "Gender bias in language" Although these results are not quantified, it leads to the conclusion that men are still more associated with their career, while women are associated with their family. In contrast the WEAT analysis for query 1 "Career vs family" taken from [4] followed the same association (positive bias scores), but did not show any significant results. These results were found for the Subreddits and user groups respectively. Regarding query 3 ("Intelligence and Appearance") of the WEAT analysis, the user groups Ignored, Casual Commentator, Contributor and the Subreddit "r/technology" shared an association bias regarding the covered adjectives and therefore support these findings. This might also lead to the assumption, that the Subreddit and the user groups are linked to each other, by sharing many of the same users. Additionally the Subreddit "r/technology" also shared a significant association bias regarding query 4 (Intelligence and Sensitive) also supporting these findings. For query 6 regarding the association with pleasant and unpleasant words, most of the Subreddits except for "r/technology" and "r/Conservative" showed reversed associations and therefore tend to associate female terms with more pleasant words. However it is remarkable to see that in contrast as the only Subreddit the "r/Conservative" achieved a significant bias score and therefore associated the female terms with unpleasant words. Although the sentiment analysis for the Subreddits just yielded significant results with low effect sizes, it is remarkable to see that Subreddits with potentially higher percentage of female users tend to associate words with better sentiment words for the female centroid vector. On the other hand the broad majority of Subreddits, that tend to have a higher percentage of men in the community, because of the underlying domain, associated words are favoring the male centroids in terms of the sentiment scores.

11 Conclusion

Within this thesis the following research question, defined in section "Research Question", will be answered based on the findings of this work. The main research question is subdivided into three different aspects.

1. To what extent are gender biases or stereotypes measurable in eleven of the largest Subreddits and what are their qualitative and quantitative characteristics?
    (a) In what ways do the words that are more associated with one gender or

the other differ?

(b) Is a statistically significant difference for the associated words measurable between users with different user behavior? If yes, how do the user groups differ from each other?

(c) Additionally, how do the various Subreddits differ in terms of the observed biases?

Regarding question a) the finding is that the female terms are more associated with the woman's appearance and her sexuality, with focus on the female reproduction. This findings can be concluded by the analysis of the most associated adjectives and are partly confirmed by the WEAT analysis (Query 3 and 4) for some user groups and Subreddits. Furthermore within all identified user groups and most of the Subreddits female terms are significantly more associated with stereotypical female occupations and vice versa.
Concerning question b) the findings of the sentiment analysis show almost no difference for the user groups. From three research objects per user group, just two were found significant, with low effect sizes, for the user group Contributor and Ignored. When taking a look at the WEAT analysis, all user groups shared the same association bias regarding stereotypic occuptions, with significant results and high effect sizes. However for query 3, regarding the association between intelligence and appearance, differences can be found for the user groups Ignored and Casual Commentators, both carrying strong association biases in regard to this query.
With regard to question c) the findings of the sentiment analysis allows the assumptions, that for the sentiment analysis Subreddits with percentage of female users above average, tend to slightly favor female terms in terms of the associated sentiments, while the other Subreddits tend to favor male terms instead. Similar to the user groups WEAT analysis most of the Subreddits share the same association bias regarding the occupations. Apart from that, only the Subreddit "r/technology" differs by showing significant association biases for query 3 and 4, similar to user groups Ignored and Casual Commentator.

12 Further Work

While gender biases were analyzed on Subreddit level and for different user groups it turns out that in Subreddits like "r/AskWomen" and "r/AskMen" it is common that users often put ♂ or ♀ at the beginning of their message, so that people are able to identify if the comment is written from male or female perspective. Within the dataset this symbols occured over 20,000 times each and a user that once used one of this symbols can be labeled for all his comments. Therefore for the potential female users in the datasets 1,134,239 comments could have been labeled

Figure 14 Example for Subreddits, that might enhance the model quality in terms of geography analogy tasks

and for male users 1,286,923 comments. This might open the opportunity to use this comments to train a supervised classifier and be able to analyze and differentiate gender biases regarding the gender of the author. Besides the gender also the awardings, that a user achieved by other users for his comment or submission, may be taken into account as a feature for an extended user clustering to get a deeper understanding of the user roles. Due to the thematic structure of the Subreddits another interesting analysis might be to analyze which Subreddits are most suitable to improve the model quality regarding analogy task, to create a robust and balanced dataset that might be used to train models, which is versatile in use. It has been shown that the "r/AskWomen" Subreddit could perform pretty decent in analogy tasks regarding family terms that are associated with the gender without further optimization. However the model could not find analogies for the nationality terms or for geographic analogies like capitals and the corresponding countries. Here Subreddits like "r/EarthPorn" might be able to enhance the model quality. Within this Subreddit people post pictures of beautiful landscapes and often refer to where the picture was taken or people are discussing about the location in the comments like in figure 14. Furthermore some abnormal behaviour was found in the dataset. For example the user "u/FutureGohanFan" participated a total of 20,002 times within one thread in a Subreddit. While this is an extrem outlier there are certain users that show this kind of behaviour. It might be interesting how the comments of these users develop in terms of biases or the use of language through the ongoing thread.

References

[1] V. Warmerdam, T. Kober, and R. Tatman, "Going beyond T-SNE: Exposing whatlies in text embeddings," in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 52–60. [Online]. Available: https://www.aclweb.org/anthology/2020.nlposs-1.8

[2] M. R. Costa-jussà, "An analysis of gender bias studies in natural language processing," 2019.

[3] D. Morrison and C. Hayes, "Here, have an upvote: communication behaviour and karma on reddit," in *GI-Jahrestagung*, 2013.

[4] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, apr 2017. [Online]. Available: https://doi.org/10.1126%2Fscience.aal4230

[5] S. Dixon, "Distribution of reddit app users in the united states as of march 2021, by age group," 2021, [last accessed: 17.06.2022]. [Online]. Available: https://www.statista.com/statistics/1125159/reddit-us-app-users-age/

[6] S. R. Department, "Distribution of reddit users in the united states as of february 2016, by gender," 2016, [last accessed: 17.06.2022]. [Online]. Available: https://www.statista.com/statistics/517155/reddit-user-distribution-usa-gender/

[7] S. Dixon, "Percentage of u.s. adults who use reddit as of february 2021, by gender," 2021, [last accessed: 26.08.2022]. [Online]. Available: https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/

[8] A. Cislak, M. Formanowicz, and T. Saguy, "Bias against research on gender bias," 2018.

[9] J. Holmes and M. Meyerhoff, *The Handbook of Language and Gender*, ser. Blackwell Handbooks in Linguistics. Wiley, 2008. [Online]. Available: https://books.google.de/books?id=tQqAbxMAmfQC

[10] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *CoRR*, vol. abs/1607.06520, 2016. [Online]. Available: http://arxiv.org/abs/1607.06520

[11] M. Sahlgren, "The distributional hypothesis," *Italian Journal of Linguistics*, vol. 20, 01 2008.

[12] J. Mendelsohn, Y. Tsvetkov, and D. Jurafsky, "A framework for the computational linguistic analysis of dehumanization," *Frontiers in Artificial Intelligence*, vol. 3, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frai.2020.00055

[13] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, "Measuring individual differences in implicit cognition: the implicit association test." *Journal of personality and social psychology*, vol. 74 6, pp. 1464–80, 1998.

[14] K. E. Anderson, "Ask me anything: what is reddit?" [Online]. Available: https://doi.org/10.1108/LHTN-03-2015-0018

[15] S. Dixon, "Most popular posts from public personalities and regular users in the r/iama ("ask me anything") community on reddit from 2009 to ytd 2022, by number of upvotes," 2022, [last accessed: 31.10.2022]. [Online]. Available: https://www.statista.com/statistics/1332905/top-iama-posts-reddit-by-upvotes/

[16] ——, "Reddit: distribution of global audiences 2022, by gender," 2022, [last accessed: 26.08.2022]. [Online]. Available: https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/

[17] J. Clement, "Regional distribution of desktop traffic to reddit.com as of may 2022 by country," 2022, [last accessed: 26.08.2022]. [Online]. Available: https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

[18] X. Ferrer, T. van Nuenen, J. M. Such, and N. Criado, "Discovering and categorising language biases in reddit," 2020. [Online]. Available: https://arxiv.org/abs/2008.02754

[19] T. Halonen, "Combating gender stereotyping and sexism in the media," 2016. [Online]. Available: https://rm.coe.int/168064379b

[20] J. P. Stanley, *Paradigmatic woman: The prostitute*, ser. David L. Shores and Caitlin P. Hines (eds) Papers in Language Variation, Montgomery, 1977.

[21] J. T. Jost and A. C. Kay, "Exposure to benevolent sexism and complementary gender stereotypes: Consequences for specific and diffuse forms of system justification," *Journal of Personality and Social Psychology*, vol. 88(3), p. 498–509.

[22] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," 2020. [Online]. Available: https://arxiv.org/abs/2001.08435

[23] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed., January 2022.

[24] M. Kurpicz-Briki and T. Leoni, "A world full of stereotypes? further investigation on origin and gender bias in multi-lingual word embeddings," *Frontiers in Big Data*, vol. 4, p. 625290, 06 2021.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[26] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 302–308. [Online]. Available: https://aclanthology.org/P14-2050

[27] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2016. [Online]. Available: https://arxiv.org/abs/1607.04606

[28] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013. [Online]. Available: https://arxiv.org/abs/1310.4546

[29] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," vol. 20, 01 2001, pp. 406–414.

[30] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics, Jun. 2009, pp. 19–27. [Online]. Available: https://aclanthology.org/N09-1003

[31] J. Chan, C. Hayes, and E. M. Daly, "Decomposing discussion forums and boards using user roles," in *ICWSM*, 2010.

[32] M. Rowe, S. Angeletou, and H. Alani, "Predicting discussions on the social semantic web," in *The Semanic Web: Research and Applications*, G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and

J. Pan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 405–420.

[33] S. M. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.

[34] X. Ferrer, T. van Nuenen, N. Criado, and J. Such, "Discovering and interpreting biased concepts in online communities," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 12 2021.

[35] D. Rasch, K. Kubinger, and K. Moder, "The two-sample t test: Pre-testing its assumptions does not pay off," *Stat. Pap.*, vol. 52, pp. 219–231, 02 2011.

[36] G. D. Ruxton, "The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test," *Behavioral Ecology*, vol. 17, no. 4, pp. 688–690, 05 2006. [Online]. Available: https://doi.org/10.1093/beheco/ark016

[37] K. Ethayarajh, D. Duvenaud, and G. Hirst, "Understanding undesirable word embedding associations," 2019. [Online]. Available: https://arxiv.org/abs/1908.06361

[38] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, apr 2018. [Online]. Available: https://doi.org/10.1073%2Fpnas.1720347115

[39] X. F. Aran, T. van Nuenen, J. M. Such, and N. C. Pacheco, "Discovering and categorising language biases in reddit," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, no. 1, 2021, pp. 140–151.

[40] P. Rayson, D. Archer, S. S. Piao, and A. M. McEnery, "The ucrel semantic analysis system," 2004.

[41] S. Schröder, A. Schulz, P. Kenneweg, R. Feldhans, F. Hinder, and B. Hammer, "Evaluating metrics for bias in word embeddings," 2021. [Online]. Available: https://arxiv.org/abs/2111.07864