Discovering and Categorising Language Biases in Reddit

Conference Paper · August 2020

4 authors, including:

Xavier Ferrer
King's College London
16 PUBLICATIONS   **27** CITATIONS

SEE PROFILE

Tom van Nuenen
King's College London
19 PUBLICATIONS   **58** CITATIONS

SEE PROFILE

Natalia Criado
King's College London
73 PUBLICATIONS   **625** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Discovering and Attesting Digital Discrimination (DADD) View project

Privacy in Social Applications View project

Discovering and Categorising Language Biases in Reddit*

Xavier Ferrer[+], Tom van Nuenen[+], Jose M. Such[+] and Natalia Criado[+]
[+] Department of Informatics, King's College London
{xavier.ferrer_aran, tom.van_nuenen, jose.such, natalia.criado}@kcl.ac.uk

Abstract

We present a data-driven approach using word embeddings to discover and categorise language biases on the discussion platform Reddit. As spaces for isolated user communities, platforms such as Reddit are increasingly connected to issues of racism, sexism and other forms of discrimination. Hence, there is a need to monitor the language of these groups. One of the most promising AI approaches to trace linguistic biases in large textual datasets involves word embeddings, which transform text into high-dimensional dense vectors and capture semantic relations between words. Yet, previous studies require predefined sets of potential biases to study, e.g., whether gender is more or less associated with particular types of jobs. This makes these approaches unfit to deal with smaller and community-centric datasets such as those on Reddit, which contain smaller vocabularies and slang, as well as biases that may be particular to that community. This paper proposes a data-driven approach to automatically discover language biases encoded in the vocabulary of online discourse communities on Reddit. In our approach, protected attributes are connected to evaluative words found in the data, which are then categorised through a semantic analysis system. We verify the effectiveness of our method by comparing the biases we discover in the Google News dataset with those found in previous literature. We then successfully discover gender bias, religion bias, and ethnic bias in different Reddit communities. We conclude by discussing potential application scenarios and limitations of this data-driven bias discovery method.

1 Introduction

This paper proposes a general and data-driven approach to discovering linguistic biases towards protected attributes, such as gender, in online communities. Through the use of word embeddings and the ranking and clustering of biased words, we discover and categorise biases in several English-speaking communities on Reddit, using these communities' own forms of expression.

Reddit is a web platform for social news aggregation, web content rating, and discussion. It serves as a platform for multiple, linked topical discussion forums, as well as a network for shared identity-making (Papacharissi 2015). Members can submit content such as text posts, pictures, or direct links, which is organised in distinct message boards curated by interest communities. These 'subreddits' are distinct message boards curated around particular topics, such as /r/pics for sharing pictures or /r/funny for posting jokes[1]. Contributions are submitted to one specific subreddit, where they are aggregated with others.

Not least because of its topical infrastructure, Reddit has been a popular site for Natural Language Processing studies – for instance, to successfully classify mental health discourses (Balani and De Choudhury 2015), and domestic abuse stories (Schrading et al. 2015). LaViolette and Hogan have recently augmented traditional NLP and machine learning techniques with platform metadata, allowing them to interpret misogynistic discourses in different subreddits (LaViolette and Hogan 2019). Their focus on discriminatory language is mirrored in other studies, which have pointed out the propagation of sexism, racism, and 'toxic technocultures' on Reddit using a combination of NLP and discourse analysis (Mountford 2018). What these studies show is 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.

Such ideologies and biases become especially pernicious when they concern vulnerable groups of people that share certain *protected attributes* – including ethnicity, gender, and religion (Grgić-Hlača et al. 2018). Identifying language biases towards these protected attributes can offer important cues to tracing harmful beliefs fostered in online spaces. Recently, NLP research using word embeddings has been able to do just that (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018). However, due to the reliance on predefined concepts to formalise bias, these studies generally make use of larger textual corpora, such as the widely used Google News dataset (Mikolov et al. 2013). This makes these methods less applicable to social media platforms such as Reddit, as communities on the platform tend to use language that operates within conventions defined by the social group

---

[1]Subreddits are commonly spelled with the prefix '/r/'.

itself. Due to their topical organisation, subreddits can be thought of as 'discourse communities' (Kehus, Walters, and Shaw 2010), which generally have a broadly agreed set of common public goals and functioning mechanisms of inter-communication among its members. They also share discursive expectations, as well as a specific lexis (Swales 2011). As such, they may carry biases and stereotypes that do not necessarily match those of society at large. At worst, they may constitute cases of hate speech, 'language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group' (Davidson et al. 2017). The question, then, is how to discover the biases and stereotypes associated with protected attributes that manifest in particular subreddits – and, crucially, which linguistic form they take.

This paper aims to bridge NLP research in social media, which thus far has not connected discriminatory language to protected attributes, and research tracing language biases using word embeddings. Our contribution consists of developing a general approach to *discover* and *categorise* biased language towards protected attributes in Reddit communities. We use word embeddings to determine the most biased words towards protected attributes, apply k-means clustering combined with a semantic analysis system to label the clusters, and use sentiment polarity to further specify biased words. We validate our approach with the widely used Google News dataset before applying it to several Reddit communities. In particular, we identified and categorised gender biases in /r/TheRedPill and /r/dating_advice, religion biases in /r/atheism and ethnicity biases in /r/The_Donald.

2 Related work

Linguistic biases have been the focus of language analysis for quite some time (Wetherell and Potter 1992; Holmes and Meyerhoff 2008; Garg et al. 2018; Bhatia 2017). Language, it is often pointed out, functions as both a reflection and perpetuation of stereotypes that people carry with them. Stereotypes can be understood as ideas about how (groups of) people commonly behave (van Miltenburg 2016). As cognitive constructs, they are closely related to essentialist beliefs: the idea that members of some social category share a deep, underlying, inherent nature or 'essence', causing them to be fundamentally similar to one another and across situations (Carnaghi et al. 2008). One form of linguistic behaviour that results from these mental processes is that of linguistic bias: 'a systematic asymmetry in word choice as a function of the social category to which the target belongs.' (Beukeboom 2014, p.313).

The task of tracing linguistic bias is accommodated by recent advances in AI (Aran, Such, and Criado 2019). One of the most promising approaches to trace biases is through a focus on the distribution of words and their similarities in word embedding modelling. The encoding of language in word embeddings answers to the distributional hypothesis in linguistics, which holds that the statistical contexts of words capture much of what we mean by meaning (Sahlgren 2008). In word embedding models, each word in a given dataset is assigned to a high-dimensional vector such that the geometry of the vectors captures semantic relations between the words – e.g. vectors being closer together correspond to distributionally similar words (Collobert et al. 2011). In order to capture accurate semantic relations between words, these models are typically trained on large corpora of text. One example is the Google News word2vec model, a word embeddings model trained on the Google News dataset (Mikolov et al. 2013).

Recently, several studies have shown that word embeddings are strikingly good at capturing human biases in large corpora of texts found both online and offline (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; van Miltenburg 2016). In particular, word embeddings approaches have proved successful in creating analogies (Bolukbasi et al. 2016), and quantifying well-known societal biases and stereotypes (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018). These approaches test for predefined biases and stereotypes related to protected attributes, e.g., for gender, that males are more associated with a professional career and females with family. In order to define sets of words capturing potential biases, which we call 'evaluation sets', previous studies have taken word sets from Implicit Association Tests (IAT) used in social psychology. This test detects the strength of a person's automatic association between mental representations of objects in memory, in order to assess bias in general societal attitudes. (Greenwald, McGhee, and Schwartz 1998). The evaluation sets yielded from IATs are then related to ontological concepts representing protected attributes, formalised as a 'target set'. This means two supervised word lists are required; e.g., the protected attribute 'gender' is defined by target words related to men (such as {'he', 'son', 'brother', ... }) and women ({'she', 'daughter', 'sister', ...}), and potential biased concepts are defined in terms of sets of evaluative terms largely composed of adjectives, such 'weak' or 'strong'. Bias is then tested through the positive relationship between these two word lists. Using this approach, Caliskan et al. were able to replicate IAT findings by introducing their Word-Embedding Association Test (WEAT). The cosine similarity between a pair of vectors in a word embeddings model proved analogous to reaction time in IATs, allowing the authors to determine biases between target and evaluative sets. The authors consider such bias to be 'stereotyped' when it relates to aspects of human culture known to lead to harmful behaviour (Caliskan, Bryson, and Narayanan 2017).

Caliskan et al. further demonstrate that word embeddings can capture imprints of historic biases, ranging from morally neutral ones (e.g. towards insects) to problematic ones (e.g. towards race or gender) (Caliskan, Bryson, and Narayanan 2017). For example, in a gender-biased dataset, the vector for adjective 'honourable' would be closer to the vector for the 'male' gender, whereas the vector for 'submissive' would be closer to the 'female' gender. Building on this insight, Garg et.al. have recently built a framework for a diachronic analysis of word embeddings, which they show incorporate changing 'cultural stereotypes' (Garg et al. 2018). The authors demonstrate, for instance, that during the second US feminist wave in the 1960, the perspectives on women as portrayed in the Google News dataset fundamentally changed. More recently, WEAT was also adapted

to BERT embeddings (Kurita et al. 2019).

What these previous approaches have in common is a reliance on predefined evaluative word sets, which are then tested on target concepts that refer to protected attributes such as gender. This makes it difficult to transfer these approaches to other – and especially smaller – linguistic datasets, which do not necessarily include the same vocabulary as the evaluation sets. Moreover, these tests are only useful to determine predefined biases for predefined concepts. Both of these issues are relevant for the subreddits we are interested in analysing here: they are relatively small, are populated by specific groups of users, revolve around very particular topics and social goals, and often involve specialised vocabularies. The biases they carry, further, are not necessarily representative of broad 'cultural stereotypes'; in fact, they can be antithetical even to common beliefs. An example in /r/TheRedPill, one of our datasets, is that men in contemporary society are oppressed by women (Marwick and Lewis 2017). Within the transitory format of the online forum, these biases can be linguistically negotiated – and potentially transformed – in unexpected ways.

Hence, while we have certain ideas about which kinds of protected attributes to expect biases against in a community (e.g. gender biases in /r/TheRedPill), it is hard to tell in advance which concepts will be associated to these protected attributes, or what linguistic form biases will take. The approach we propose extracts and aggregates the words relevant within each subreddit in order to identify biases regarding protected attributes as they are encoded in the linguistic forms chosen by the community itself.

3 Discovering language biases

In this section we present our approach to discover linguistic biases.

3.1 Most biased words

Given a word embeddings model of a corpus (for instance, trained with textual comments from a Reddit community) and two sets of target words representing two concepts we want to compare and discover biases from, we identify the most biased words towards these concepts in the community. Let $S_1 = \{w_i, w_{i+1}, ..., w_{i+n}\}$ be a set of target words w related to a concept (e.g $\{he, son, his, him, father$, and $male\}$ for concept $male$), and $\vec{c_1}$ the centroid of S_1 estimated by averaging the embedding vectors of word $w \in S_1$. Similarly, let $S_2 = \{w_j, w_{j+1}, ..., w_{j+m}\}$ be a second set of words with centroid $\vec{c_2}$ (e.g. $\{she, daughter, her, mother$, and $female\}$ for concept $female$). A word w is biased towards S_1 with respect to S_2 when the cosine similarity[2] between the embedding of $\vec{w}$, is higher for $\vec{c_1}$ than for $\vec{c_2}$.

$$Bias(w, c_1, c_2) = cos(\vec{w}, \vec{c_1}) - cos(\vec{w}, \vec{c_2}) \qquad (1)$$

where $cos(u, v) = \frac{u \cdot v}{\|u\|_2 \|v\|_2}$. Positive values of $Bias$ mean a word w is more biased towards S_1, while negative values of $Bias$ means w is more biased towards S_2.

Let V be the vocabulary of a word embeddings model. We identify the k most biased words towards S_1 with respect to S_2 by ranking the words in the vocabulary V using $Bias$ function from Equation 2:

$$MostBiased(V, c_1, c_2) = \arg\max_{w \in V} Bias(w, c_1, c_2) \quad (2)$$

Researchers typically focus on discovering biases and stereotypes by exploring the most biased adjectives and nouns towards two sets of target words (e.g. *female* and *male*). Adjectives are particularly interesting since they modify nouns by limiting, qualifying, or specifying their properties, and are often normatively charged. Adjectives carry polarity, and thus often yield more interesting insights about the type of discourses. In order to determine the part-of-speech (POS) of a word, we use the `nltk`[3] python library. POS filtering helps us removing non-interesting words in some communities such as acronyms, articles and proper names (cf. Appendix A for a performance evaluation of POS using the nltk library in the datasets used in this paper).

Given a vocabulary and two sets of target words (such as those for *women* and *men*), we rank the words from least to most biased using Equation 2. As such, we obtain two ordered lists of the most biased words towards each target set, obtaining an overall view of the bias distribution in that particular community with respect to those two target sets. For instance, Figure 1 shows the bias distribution of words towards women (top) and men (bottom) target sets in /r/TheRedPill. Based on the distribution of biases towards each target set in each subreddit, we determine the threshold of how many words to analyse by selecting the top words using Equation 2. All targets sets used in our work are compiled from previous experiments (listed in Appendix C).
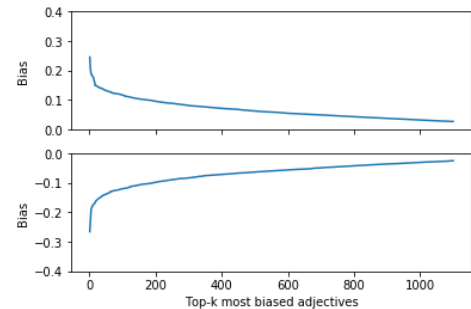
Figure 1: Bias distribution in adjectives in the r/TheRedPill

---

[2] Alternative bias definitions are possible here, such as the *direct bias* measure defined in (Bolukbasi et al. 2016). In fact, when compared experimentally with our metric in r/TheRedPill, we obtain a Jaccard index of 0.857 (for female gender) and 0.864 (for male) regarding the list of 300 most-biased adjectives generated with the two bias metrics. Similar results could also be obtained using the relative norm bias metric, as shown in (Garg et al. 2018).

[3] https:www.nltk.org

3.2 Sentiment Analysis

To further specify the biases we encounter, we take the sentiment polarity of biased words into account. Discovering consistently strong negative polarities among the most biased words towards a target set might be indicative of strong biases, and even stereotypes, towards that specific population[4]. We are interested in assessing whether the most biased words towards a population carry negative connotations, and we do so by performing a sentiment analysis over the most biased words towards each target using the `nltk` sentiment analysis python library (Hutto and Gilbert 2014)[5]. We estimate the average of the sentiment of a set of words W as such:

$$Sent(W) = \frac{1}{|W|} \sum_{w \in W} SA(w) \qquad (3)$$

where SA returns a value $\in [-1, 1]$ corresponding to the polarity determined by the sentiment analysis system, -1 being strongly negative and +1 strongly positive. As such, $Sent(W)$ always returns a value $\in [-1, 1]$.

Similarly to POS tagging, the polarity of a word depends on the context in which the word is found. Unfortunately, contextual information is not encoded in the pre-trained word embedding models commonly used in the literature. As such, we can only leverage the prior sentiment polarity of words without considering the context of the sentence in which they were used. Nevertheless, a consistent tendency towards strongly polarised negative (or positive) words can give some information about tendencies and biases towards a target set.

3.3 Categorising Biases

As noted, we aim to discover the most biased terms towards a target set. However, even when knowing those most biased terms and their polarity, considering each of them as a separate unit may not suffice in order to discover the relevant concepts they represent and, hence, the contextual meaning of the bias. Therefore, we also combine semantically related terms under broader rubrics in order to facilitate the comprehension of a community's biases. A side benefit is that identifying concepts as a cluster of terms, instead of using individual terms, helps us tackle stability issues associated with individual word usage in word embeddings (Antoniak and Mimno 2018) - discussed in Section 5.1.

We aggregate the most similar word embeddings into clusters using the well-known k-means clustering algorithm. In k-means clustering, the parameter k defines the quantity of clusters into which the space will be partitioned. Equivalently, we use the *reduction factor* $\in (0, 1)$, $r = \frac{k}{|V|}$, where $|V|$ is the size of the vocabulary to be partitioned. The lower the value of r, the lower the quantity of clusters and their average intra-similarity, estimated by assessing the average similarity between all words in a cluster for all clusters in a partition. On the other hand, when r is close to 1, we obtain more clusters and a higher cluster intra-similarity, up to when $r = 1$ where we have $|V|$ clusters of size 1, with an average intra-similarity of 1 (see Appendix A).

In order to assign a label to each cluster, which facilitates the categorisation of biases related to each target set, we use the UCREL Semantic Analysis System (USAS)[6]. USAS is a framework for the automatic semantic analysis and tagging of text, originally based on Tom McArthur's Longman Lexicon of Contemporary English (Summers and Gadsby 1995). It has a multi-tier structure with 21 major discourse fields subdivided in more fine-grained categories such as *People*, *Relationships* or *Power*. USAS has been extensively used for tasks such as the automatic content analysis of spoken discourse (Wilson and Rayson 1993) or as a translator assistant (Sharoff et al. 2006). The creators also offer an interactive tool[7] to automatically tag each word in a given sentence with a USAS semantic label.

Using the USAS system, every cluster is labelled with the most frequent tag (or tags) among the words clustered in the k-means cluster. For instance, *Relationship: Intimate/sexual* and *Power, organising* are two of the most common labels assigned to the gender-biased clusters of /r/TheRedPill (see Section 5.1). However, since many of the communities we explore make use of non-standard vocabularies, dialects, *slang* words and grammatical particularities, the USAS automatic analysis system has occasional difficulties during the tagging process. Slang and community-specific words such as *dateable* (someone who is *good enough* for dating) or *fugly* (used to describe someone considered very ugly) are often left uncategorised. In these cases, the uncategorised clusters receive the label (or labels) of the most similar cluster in the partition, determined by analysing the cluster centroid distance between the unlabelled cluster and the other cluster centroids in the partition. For instance, in /r/TheRedPill, the cluster *(interracial)* (a one-word cluster) was initially left unlabelled. The label was then updated to *Relationship: Intimate/sexual* after copying the label of the most similar cluster, which was *(lesbian, bisexual)*.

Once all clusters of the partition are labelled, we rank all labels for each target based on the quantity of clusters tagged and, in case of a tie, based on the quantity of words of the clusters tagged with the label. By comparing the rank of the labels between the two target sets and combining it with an analysis of the clusters' average polarities, we obtain a general understanding of the most frequent conceptual biases towards each target set in that community. We particularly focus on the most relevant clusters based on rank difference between target sets or other relevant characteristics such as average sentiment of the clusters, but we also include the top-10 most frequent conceptual biases for each dataset (Appendix C).

---

[4]Note that potentially discriminatory biases can also be encoded in a-priori sentiment-neutral words. The fact that a word is not tagged with a negative sentiment does not exclude it from being discriminatory in certain contexts.

[5]Other sentiment analysis tools could be used but some might return biased analyses (Kiritchenko and Mohammad 2018).

---

[6]http://ucrel.lancs.ac.uk/usas/, accessed Apr 2020

[7]http://ucrel-api.lancaster.ac.uk/usas/tagger.html, accessed Apr 2020

4 Validation on Google News

In this section we use our approach to discover gender biases in the Google News pre-trained model[8], and compare them with previous findings (Garg et al. 2018; Caliskan, Bryson, and Narayanan 2017) to prove that our method yields relevant results that complement those found in the existing literature.

The Google News embedding model contains 300-dimensional vectors for 3 million words and phrases, trained on part of the US Google News dataset containing about 100 billion words. Previous research on this model reported gender biases among others (Garg et al. 2018), and we repeated the three WEAT experiments related to gender from (Caliskan, Bryson, and Narayanan 2017) in Google News. These WEAT experiments compare the association between male and female target sets to evaluative sets indicative of gender binarism, including *career* Vs *family*, *math* Vs *arts*, and *science* Vs *arts*, where the first sets include a-priori male-biased words, and the second include female-biased words (see Appendix C). In all three cases, the WEAT tests show significant p-values ($p = 10^{-3}$ for *career/family*, $p = 0.018$ for *math/arts*, and $p = 10^{-2}$ for *science/arts*), indicating relevant gender biases with respect to the particular word sets.

Next, we use our approach on the Google News dataset to discover the gender biases of the community, and to identify whether the set of conceptual biases and USAS labels confirms the findings of previous studies with respect to arts, science, career and family.

For this task, we follow the method stated in Section 3 and start by observing the bias distribution of the dataset, in which we identify the 5000 most biased uni-gram adjectives and nouns towards 'female' and 'male' target sets. The experiment is performed with a reduction factor $r = 0.15$, although this value could be modified to zoom out/in the different clusters (see Appendix A). After selecting the most biased nouns and adjectives, the k-means clustering partitioned the resulting vocabulary in 750 clusters for women and man. There is no relevant average prior sentiment difference between male and female-biased clusters.

Table 1 shows some of the most relevant labels used to tag the female and male-biased clusters in the Google News dataset, where $R.Female$ and $R.Male$ indicate the rank importance of each label among the sets of labels used to tag each cluster for each gender. Character '-' indicates that the label is not found among the labels biased towards the target set. Due to space limitations, we only report the most pronounced biases based on frequency and rank difference between target sets (see Appendix B for the rest top-ten labels). Among the most frequent concepts more biased towards women, we find labels such as *Clothes and personal belongings*, *People: Female*, *Anatomy and physiology*, and *Judgement of appearance (pretty etc.)*. In contrast, labels related to strength and power, such as *Warfare, defence and the*

Table 1: Google News most relevant cluster labels (gender).

Cluster Label	R. Female	R. Male
Relevant to Female		
Clothes and personal belongings	3	20
People: Female	4	-
Anatomy and physiology	5	11
Cleaning and personal care	7	68
Judgement of appearance (pretty etc.)	9	29
Relevant to Male		
Warfare, defence and the army; weapons	-	3
Power, organizing	8	4
Sports	-	7
Crime	68	8
Groups and affiliation	-	9

army; weapons, *Power, organizing*, followed by *Sports*, and *Crime*, are among the most frequent concepts much more biased towards men.

We now compare with the biases that had been tested in prior works by, first, mapping the USAS labels related to *career, family, arts, science* and *maths* based on an analysis of the WEAT word sets and the category descriptions provided in the USAS website (see Appendix C), and second, evaluating how frequent are those labels among the set of most biased words towards women and men. The USAS labels related to *career* are more frequently biased towards men, with a total of 24 and 38 clusters for women and men, respectively, containing words such as 'barmaid' and 'secretarial' (for women) and 'manager' (for men). *Family*-related clusters are strongly biased towards women, with twice as many clusters for women (38) than for men (19). Words clustered include references to 'maternity', 'birthmother' (women), and also 'paternity' (men). *Arts* is also biased towards women, with 4 clusters for women compared with just 1 cluster for men, and including words such as 'sew', 'needlework' and 'soprano' (women). Although not that frequent among the set of the 5000 most biased words in the community, labels related to *science* and *maths* are biased towards men, with only one cluster associated with men but no clusters associated with women. Therefore, this analysis shows that our method, in addition to finding what are the most frequent and pronounced biases in the Google News model (shown in Table 1), could also reproduce the biases tested[9] by previous work.

5 Reddit Datasets

The Reddit datasets used in the remainder of this paper are presented in Table 2, where *Wpc* means average words per comment, and *Word Density* is the average unique new words per comment. Data was acquired using the Pushshift data platform (Baumgartner et al. 2020). All predefined sets of words used in this work and extended tables are included in Appendixes B and C, and the code to process the datasets and embedding models is available publicly[10]. We expect

---

[8]We used the Google news model (https://code.google.com/archive/p/word2vec/), due to its wide usage in relevant literature. However, our method could also be extended and applied in newer embedding models such as ELMO and BERT.

[9]Note that previous work tested for arbitrary biases, which were not claimed to be the most frequent or pronounced ones.

[10]https://github.com/xfold/LanguageBiasesInReddit

to find both different degrees and types of bias and stereotyping in these communities, based on news reporting and our initial explorations of the communities. For instance, /r/TheRedPill and /r/The_Donald have been widely covered as misogynist and ethnic-biased communities (see below), while /r/atheism is, as far as reporting goes, less biased.

For each comment in each subreddit, we first preprocess the text by removing special characters, splitting text into sentences, and transforming all words to lowercase. Then, using all comments available in each subreddit and using Gensim word2vec python library, we train a skip-gram word embeddings model of 200 dimensions, discarding all words with less that 10 occurrences (see an analysis varying this frequency parameter in Appendix A) and using a 4 word window.

After training the models, and by using WEAT (Caliskan, Bryson, and Narayanan 2017), we were able to demonstrate whether our subreddits actually include any of the predefined biases found in previous studies. For instance, by repeating the same gender-related WEAT experiments performed in Section 4 in /r/TheRedPill, it seems that the dataset may be gender-biased, stereotyping men as related to *career* and women to *family* (p-value of 0.013). However, these findings do not agree with other commonly observed gender stereotypes, such as those associating men with *science and math* (p-value of 0.411) and women with *arts* (p-value of 0.366). It seems that, if gender biases are occurring here, they are of a particular kind – underscoring our point that predefined sets of concepts may not always be useful to evaluate biases in online communities.[11]

5.1 Gender biases in /r/TheRedPill

The main subreddit we analyse for gender bias is The Red Pill (/r/TheRedPill). This community defines itself as a forum for the 'discussion of sexual strategy in a culture increasingly lacking a positive identity for men' (Watson 2016), and at the time of writing hosts around 300,000 users. It belongs to the online Manosphere, a loose collection of misogynist movements and communities such as pickup artists, involuntary celibates ('incels'), and Men Going Their Own Way (MGTOW). The name of the subreddit is a reference to the 1999 film The Matrix: 'swallowing the red pill,' in the community's parlance, signals the acceptance of an alternative social framework in which men, not women, have been structurally disenfranchised in the west. Within this 'masculinist' belief system, society is ruled by feminine ideas and values, yet this fact is repressed by feminists and politically correct 'social justice warriors'. In response, men must protect themselves against a 'misandrist' culture and the feminising of society (Marwick and Lewis 2017; LaViolette and Hogan 2019). Red-pilling has become a more general shorthand for radicalisation, conditioning young men into views of the alt-right (Marwick and Lewis 2017). Our question here is to which extent our approach can help

in discovering biased themes and concerns in this community.

Table 3 shows the top 7 most gender-biased adjectives for the /r/TheRedPill, as well as their bias value and frequency in the model. Notice that most female-biased words are more frequently used than male-biased words, meaning that the community frequently uses that set of words in female-related contexts. Notice also that our POS tagging has erroneously picked up some nouns such as *bumble* (a dating app) and *unicorn* (defined in the subreddit's glossary as a 'Mystical creature that doesn't fucking exist, aka "The Girl of Your Dreams"').

The most biased adjective towards women is *casual* with a bias of 0.224. That means that the average user of /r/TheRedPill often uses the word *casual* in similar contexts as female-related words, and not so often in similar contexts as male-related words. This makes intuitive sense, as the discourse in /r/TheRedPill revolves around the pursuit of 'casual' relationships with women. For men, some of the most biased adjectives are *quintessential*, *tactician*, *legendary*, and *genious*. Some of the most biased words towards women could be categorised as related to externality and physical appearance, such as *flirtatious* and *fuckable*. Conversely, the most biased adjectives for men, such as *visionary* and *tactician*, are internal qualities that refer to strategic game-playing. Men, in other words, are qualified through descriptive adjectives serving as indicators of subjectivity, while women are qualified through evaluative adjectives that render them as objects under masculinist scrutiny.

Categorising Biases We now cluster the most-biased words in 45 clusters, using $r = 0.15$ (see an analysis of the effect r has in Appendix A), generalising their semantic content. Importantly, due to this categorisation of biases instead of simply using most-biased words, our method is less prone to stability issues associated with word embeddings (Antoniak and Mimno 2018), as changes in particular words do not directly affect the overarching concepts explored at the cluster level and the labels that further abstract their meaning (see the stability analysis part in Appendix A).

Table 4 shows some of the most frequent labels for the clusters biased towards women and men in /r/TheRedPill, and compares their importance for each gender. *SentW* corresponds to the average sentiment of all clusters tagged with the label, as described in Equation 3. The *R.Woman* and *R.Male* columns show the rank of the labels for the female and male-biased clusters. '-' indicates that no clusters were tagged with that label.

Anatomy and Physiology, *Intimate sexual relationships* and *Judgement of appearance* are common labels demonstrating bias towards women in /r/TheRedPill, while the biases towards men are clustered as *Power and organising, Evaluation, Egoism, and toughness*. Sentiment scores indicate that the first two biased clusters towards women carry negative evaluations, whereas most of the clusters related to men contain neutral or positively evaluated words. Interestingly, the most frequent female-biased labels, such as *Anatomy and physiology* and *Relationship: Intimate/sexual*

---

[11]Due to technical constraints we limit our analysis to the two major binary gender categories – female and male, or women and men – as represented by the lists of associated words.

Table 2: Datasets used in this research

Subreddit	E.Bias	Years	Authors	Comments	Unique Words	Wpc	Word Density
/r/TheRedPill	gender	2012-2018	106,161	2,844,130	59,712	52.58	$3.99 \cdot 10^{-4}$
/r/DatingAdvice	gender	2011-2018	158,758	1,360,397	28,583	60.22	$3.48 \cdot 10^{-4}$
/r/Atheism	religion	2008-2009	699,994	8,668,991	81,114	38.27	$2.44 \cdot 10^{-4}$
/r/The_Donald	ethnicity	2015-2016	240,666	13,142,696	117,060	21.27	$4.18 \cdot 10^{-4}$

Table 3: Most gender-biased adjectives in /r/TheRedPill.

Female			Male		
Adjective	Bias	Freq (FreqR)	Adjective	Bias	Freq (FreqR)
bumble	0.245	648 (8778)	visionary	0.265	100 (22815)
casual	0.224	6773 (1834)	quintessential	0.245	219 (15722)
flirtatious	0.205	351 (12305)	tactician	0.229	29 (38426)
anal	0.196	3242 (3185)	bombastic	0.199	41 (33324)
okcupid	0.187	2131 (4219)	leary	0.190	93 (23561)
fuckable	0.187	1152 (6226)	gurian	0.185	16 (48440)
unicorn	0.186	8536 (1541)	legendary	0.183	400 (11481)

Table 4: Comparison of most relevant cluster labels between biased words towards women and men in /r/TheRedPill.

Cluster Label	SentW	R. Female	R. Male
Relevant to *Female*			
Anatomy and physiology	-0.120	1	25
Relationship: Intimate/sexual	-0.035	2	30
Judgement of appearance (pretty etc.)	0.475	3	40
Evaluation:- Good/bad	0.110	4	2
Appearance and physical properties	0.018	10	6
Relevant to *Male*			
Power, organizing	0.087	61	1
Evaluation:- Good/bad	0.157	4	2
Education in general	0.002	-	4
Egoism	0.090	-	5
Toughness; strong/weak	-0.004	-	7

Table 5: Comparison of most relevant cluster labels between biased words towards women and men in /r/dating_advice.

Cluster Label	SentW	R. Female	R. Male
Relevant to *Female*			
Quantities	0.202	1	6
Geographical names	0.026	2	-
Religion and the supernatural	0.025	3	-
Language, speech and grammar	0.025	4	-
Importance: Important	0.227	5	-
Relevant to *Female*			
Evaluation:- Good/bad	-0.165	14	1
Judgement of appearance (pretty etc.)	-0.148	6	2
Power, organizing	0.032	51	3
General ethics	-0.089	-	4
Interest/boredom/excited/energetic	0.354	-	5

(second most frequent), are only ranked 25th and 30th for men (from a total of 62 male-biased labels). A similar difference is observed when looking at male-biased clusters with the highest rank: *Power, organizing* (ranked 1st for men) is ranked 61st for women, while other labels such as *Egoism* (5th) and *Toughness; strong/weak* (7th), are not even present in female-biased labels.

Comparison to /r/Dating_Advice In order to assess to which extent our method can differentiate between more and less biased datasets – and to see whether it picks up on less explicitly biased communities – we compare the previous findings to those of the subreddit /r/dating_advice, a community with 908,000 members. The subreddit is intended for users to 'Share (their) favorite tips, ask for advice, and encourage others about anything dating'. The subreddit's About-section notes that '[t]this is a positive community. Any bashing, hateful attacks, or sexist remarks will be removed', and that 'pickup or PUA lingo' is not appreciated. As such, the community shows similarities with /r/TheRedPill in terms of its focus on dating, but the gendered binarism is expected to be less prominently present.

As expected, the bias distribution here is weaker than in /r/TheRedPill. The most biased word towards women in /r/dating_advice is *floral* with a bias of 0.185, and *molest* (0.174) for men. Based on the distribution of biases (following the method in Section 3.1), we selected the top 200 most biased adjectives towards the 'female' and 'male' target sets and clustered them using k-means ($r = 0.15$), leaving 30 clusters for each target set of words. The most biased clusters towards women, such as *(okcupid, bumble)*, and *(exotic)*, are not clearly negatively biased (though we might ask questions about the implied exoticism in the latter term). The biased clusters towards men look more conspicuous: *(poor)*, *(irresponsible, erratic, unreliable, impulsive)* or *(pathetic, stupid, pedantic, sanctimonious, gross, weak, nonsensical, foolish)* are found among the most biased clusters. On top of that, *(abusive)*, *(narcissistic, misogynistic, egotistical, arrogant)*, and *(miserable, depressed)* are among the most sentiment negative clusters. These terms indicate a significant negative bias towards men, evaluating them in terms of unreliability, pettiness and self-importance.

No typical bias can be found among the most common labels for the k-means clusters for women. *Quantities* and *Geographical names* are the most common labels. The most relevant clusters related to men are *Evaluation* and *Judgement of Appearance*, together with *Power, organizing*. Table 5 compares the importance between some of the most relevant biases for women and men by showing the difference in the bias ranks for both sets of target words. The table shows that there is no physical or sexual stereotyping of women as in /r/TheRedPill, and *Judgment of appearance*, a strongly female-biased label in /r/TheRedPill, is more frequently biased here towards men (rank 2) than women (rank 6). Instead we find that some of the most common labels used to tag the female-biased clusters are *Quantities*, *Language,*

Table 6: Comparison of most relevant labels between *Islam* and *Christianity* word sets for /r/atheism

Cluster Label	SentW	R. Islam	R. Chr.
Relevant to Islam			
Geographical names	0	1	39
Crime, law and order: Law and order	-0.085	2	40
Groups and affiliation	-0.012	3	20
Politeness	-0.134	4	-
Calm/Violent/Angry	-0.140	5	-
Relevant to Christianity			
Religion and the supernatural	0.003	13	1
Time: Beginning and ending	0	-	2
Time: Old, new and young; age	0.079	-	3
Anatomy and physiology	0	22	4
Comparing:- Usual/unusual	0	-	5

speech and grammar or *Religion and the supernatural*. This, in conjunction with the negative sentiment scores for male-biased labels, underscores the point that /r/Dating_Advice seems slightly biased towards men.

5.2 Religion biases in /r/Atheism

In this next experiment, we apply our method to discover religion-based biases. The dataset derives from the subreddit /r/atheism, a large community with about 2.5 million members that calls itself 'the web's largest atheist forum', on which '[a]ll topics related to atheism, agnosticism and secular living are welcome'. Are monotheistic religions considered as equals here? To discover religion biases, we use target word sets *Islam* and *Christianity* (see Appendix B).

In order to attain a broader picture of the biases related to each of the target sets, we categorise and label the clusters following the steps described in Section 3.3. Based on the distribution of biases we found here, we select the 300 most biased adjectives and use an $r = 0.15$ in order to obtain 45 clusters for both target sets. We then count and compare all clusters that were tagged with the same label, in order to obtain a more general view of the biases in /r/atheism for words related to the Islam and Christianity target sets.

Table 6 shows some of the most common clusters labels attributed to Islam and Christianity (see Appendix B for the full table), and the respective differences between the ranking of these clusters, as well as the average sentiment of all words tagged with each label. The '-' symbol means that a label was not used to tag any cluster of that specific target set. Findings indicate that, in contrast with Christianity-biased clusters, some of the most frequent cluster labels biased towards Islam are *Geographical names*, *Crime, law and order* and *Calm/Violent/Angry*. On the other hand, some of the most biased labels towards Christianity are *Religion and the supernatural*, *Time: Beginning and ending* and *Anatomy and physiology*.

All the mentioned biased labels towards Islam have an average negative polarity, except for *Geographical names*. Labels such as *Crime, law and order* aggregate words with evidently negative connotations such as *uncivilized, misogynistic, terroristic and antisemitic*. *Judgement of appearance*, *General ethics*, and *Warfare, defence and the army* are also found among the top 10 most frequent labels for Islam, aggregating words such as *oppressive, offensive and totalitarian* (see Appendix B). However, none of these labels are relevant in Christianity-biased clusters. Further, most of the words in Christianity-biased clusters do not carry negative connotations. Words such as *unitarian, presbyterian, episcopalian* or *anglican* are labelled as belonging to *Religion and the supernatural*, *unbaptized* and *eternal* belong to *Time* related labels, and *biological, evolutionary* and *genetic* belong to *Anatomy and physiology*.

Finally, it is important to note that our analysis of conceptual biases is meant to be more suggestive than conclusive, especially on this subreddit in which various religions are discussed, potentially influencing the embedding distributions of certain words and the final discovered sets of conceptual biases. Having said this, and despite the community's focus on atheism, the results suggest that labels biased towards Islam tend to have a negative polarity when compared with Christian biased clusters, considering the set of 300 most biased words towards Islam and Christianity in this community. Note, however, that this does not mean that those biases are the most frequent, but that they are the most pronounced, so they may be indicative of broader socio-cultural perceptions and stereotypes that characterise the discourse in /r/atheism. Further analysis (including word frequency) would give a more complete view.

5.3 Ethnic biases in /r/The_Donald

In this third and final experiment we aim to discover ethnic biases. Our dataset was taken from /r/The_Donald, a subreddit in which participants create discussions and memes supportive of U.S. president Donald Trump. Initially created in June 2015 following the announcement of Trump's presidential campaign, /r/The_Donald has grown to become one of the most popular communities on Reddit. Within the wider news media, it has been described as hosting conspiracy theories and racist content (Romano 2017).

For this dataset, we use target sets to compare *white last names*, with *Hispanic names*, *Asian names* and *Russian names* (see Appendix C). The bias distribution for all three tests is similar: the *Hispanic*, *Asian* and *Russian* target sets are associated with stronger biases than the *white names* target sets. The most biased adjectives towards white target sets include *classic*, *moralistic* and *honorable* when compared with all three other targets sets. Words such as *undocumented*, *undeported* and *illegal* are among the most biased words towards Hispanics, while *Chinese* and *international* are among the most biased words towards Asian, and *unrefined* and *venomous* towards Russian. The average sentiment among the most-biased adjectives towards the different targets sets is not significant, except when compared with Hispanic names, i.e. a sentiment of 0.0018 for white names and -0.0432 for Hispanics (p-value of 0.0241).

Table 7 shows the most common labels and average sentiment for clusters biased towards Hispanic names using $r = 0.15$ and considering the 300 most biased adjectives, which is the most negative and stereotyped community among the ones we analysed in /r/The_Donald. Apart from geographical names, the most interesting labels for Hispanic vis-à-

Table 7: Most relevant labels for Hispanic target set in /r/The_Donald

Cluster Label	SentW	R. White	R. Hisp.
Geographical names	0	-	1
General ethics	-0.349	25	2
Wanting; planning; choosing	0	-	3
Crime, law and order	-0.119	-	4
Gen. appearance, phys. properties	-0.154	21	10

vis white names are *General ethics* (including words such as *abusive, deportable, incestual, unscrupulous, undemocratic*), *Crime, law and order* (including words such as *undocumented, illegal, criminal, unauthorized, unlawful, lawful* and *extrajudicial*), and *General appearance and physical properties* (aggregating words such as *unhealthy, obese* and *unattractive*). All of these labels are notably uncommon among clusters biased towards white names – in fact, *Crime, law and order* and *Wanting; planning; choosing* are not found there at all.

6 Discussion

Considering the radicalisation of interest-based communities outside of mainstream culture (Marwick and Lewis 2017), the ability to trace linguistic biases on platforms such as Reddit is of importance. Through the use of word embeddings and similarity metrics, which leverage the vocabulary used within specific communities, we are able to discover biased concepts towards different social groups when compared against each other. This allows us to forego using fixed and predefined evaluative terms to define biases, which current approaches rely on. Our approach enables us to evaluate the terms and concepts that are most indicative of biases and, hence, discriminatory processes.

As Victor Hugo pointed out in Les Miserables, slang is the most mutable part of any language: 'as it always seeks disguise so soon as it perceives it is understood, it transforms itself.' Biased words take distinct and highly mutable forms per community, and do not always carry inherent negative bias, such as *casual* and *flirtatious* in /r/TheRedPill. Our method is able to trace these words, as they acquire bias when contextualised within particular discourse communities. Further, by discovering and aggregating the most-biased words into more general concepts, we can attain a higher-level understanding of the dispositions of Reddit communities towards protected features such as gender. Our approach can aid the formalisation of biases in these communities, previously proposed by (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018). It also offers robust validity checks when comparing subreddits for biased language, such as done by (LaViolette and Hogan 2019). Due to its general nature – word embeddings models can be trained on any natural language corpus – our method can complement previous research on ideological orientations and bias in online communities in general.

Quantifying language biases has many advantages (Abebe et al. 2020). As a diagnostic, it can help us to understand and measure social problems with precision and clarity. Explicit, formal definitions can help promote discussions on the vocabularies of bias in online settings. Our approach is intended to trace language in cases where researchers do not know all the specifics of linguistic forms used by a community. For instance, it could be applied by legislators and content moderators of web platforms such as the one we have scrutinised here, in order to discover and trace the severity of bias in different communities. As pernicious bias may indicate instances of hate speech, our method could assist in deciding which kinds of communities do not conform to content policies. Due to its data-driven nature, discovering biases could also be of some assistance to trace so-called 'dog-whistling' tactics, which radicalised communities often employ. Such tactics involve coded language which appears to mean one thing to the general population, but has an additional, different, or more specific resonance for a targeted subgroup (Haney-López 2015).

Of course, without a human in the loop, our approach does not tell us much about *why* certain biases arise, what they mean in context, or how much bias is too much. Approaches such as Critical Discourse Analysis are intended to do just that (LaViolette and Hogan 2019). In order to provide a more causal explanation of how biases and stereotypes appear in language, and to understand how they function, future work can leverage more recent embedding models in which certain dimensions are designed to capture various aspects of language, such as the polarity of a word or its parts of speech (Rothe and Schütze 2016), or other types of embeddings such as bidirectional transformers (BERT) (Devlin et al. 2018). Other valuable expansions could include to combine both bias strength and frequency in order to identify not only strongly biased words but also frequently used in the subreddit, extending the set of USAS labels to obtain more specific and accurate labels to define cluster biases, and study community drift to understand how biases change and evolve over time. Moreover, specific ontologies to trace each type of bias with respect to protected attributes could be devised, in order to improve the labelling and characterisation of negative biases and stereotypes.

We view the main contribution of our work as introducing a modular, extensible approach for exploring language biases through the lens of word embeddings. Being able to do so without having to construct a-priori definitions of these biases renders this process more applicable to the dynamic and unpredictable discourses that are proliferating online.

References

Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M.; and Robinson, D. G. 2020. Roles for computing in social change. In *Proc. of ACM FAccT 2020*, 252–260.

Antoniak, M., and Mimno, D. 2018. Evaluating the Stability of Embedding-based Word Similarities. *TACL 2018* 6:107–119.

Aran, X. F.; Such, J. M.; and Criado, N. 2019. Attesting biases and discrimination using language semantics. In *Re-*

sponsible Artificial Intelligence Agents workshop of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019).

Balani, S., and De Choudhury, M. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *ACM CHI 2015*, 1373–1378.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435*.

Beukeboom, C. J. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In Laszlo, J.; Forgas, J.; and Vincze, O., eds., *Social Cognition and Communication*. New York: Psychology Press.

Bhatia, S. 2017. The semantic representation of prejudice and stereotypes. *Cognition* 164:46–60.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? In *NeurIPS 2016*, 4349–4357.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Carnaghi, A.; Maass, A.; Gresta, S.; Bianchi, M.; Cadinu, M.; and Arcuri, L. 2008. Nomina Sunt Omina: On the Inductive Potential of Nouns and Adjectives in Person Perception. *Journal of Personality and Social Psychology* 94(5):839–859.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12(Aug):2493–2537.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. *ICWSM 2017* (Icwsm):512–515.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS 2018* 115(16):E3635–E3644.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. K. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74(6):1464.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2018. Beyond Distributive Fairness in Algorithmic Decision Making. *AAAI-18* 51–60.

Haney-López, I. 2015. *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. London: Oxford University Press.

Holmes, J., and Meyerhoff, M. 2008. *The handbook of language and gender*, volume 25. Hoboken: John Wiley & Sons.

Hutto, C. J., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *AAAI 2014*.

Kehus, M.; Walters, K.; and Shaw, M. 2010. Definition and genesis of an online discourse community. *International Journal of Learning* 17(4):67–86.

Kiritchenko, S., and Mohammad, S. M. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

LaViolette, J., and Hogan, B. 2019. Using platform signals for distinguishing discourses: The case of men's rights and men's liberation on Reddit. *ICWSM 2019* 323–334.

Marwick, A., and Lewis, R. 2017. Media Manipulation and Disinformation Online. *Data & Society Research Institute* 1–104.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS 2013*, 3111–3119.

Mountford, J. 2018. Topic Modeling The Red Pill. *Social Sciences* 7(3):42.

Nosek, B. A.; Banaji, M. R.; and Greenwald, A. G. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6(1):101.

Papacharissi, Z. 2015. Toward New Journalism(s). *Journalism Studies* 16(1):27–40.

Romano, A. 2017. Reddit just banned one of its most toxic forums. but it won't touch the donald. *Vox, November* 13.

Rothe, S., and Schütze, H. 2016. Word embedding calculus in meaningful ultradense subspaces. In *ACL 2016*, 512–517.

Sahlgren, M. 2008. The distributional hypothesis. *Italian Journal of Linguistics* 20(1):33–53.

Schrading, N.; Ovesdotter Alm, C.; Ptucha, R.; and Homan, C. 2015. An Analysis of Domestic Abuse Discourse on Reddit. (September):2577–2583.

Sharoff, S.; Babych, B.; Rayson, P.; Mudraya, O.; and Piao, S. 2006. Assist: Automated semantic assistance for translators. In *EACL 2006*, 139–142.

Summers, D., and Gadsby, A. 1995. Longman dictionary of contemporary english.

Swales, J. 2011. The Concept of Discourse Community. *Writing About Writing* 466–473.

van Miltenburg, E. 2016. Stereotyping and Bias in the Flickr30K Dataset. (May):1–4.

Watson, Z. 2016. Red pill men and women, reddit, and the cult of gender. *Inverse*.

Wetherell, M., and Potter, J. 1992. Mapping the language of racism: Discourse and the legitimation of exploitation.

Wilson, A., and Rayson, P. 1993. Automatic content analysis of spoken discourse: a report on work in progress. *Corpus based computational linguistics* 215–226.

Appendices

A   Further experiments on /r/TheRedPill

In this section we perform various analysis on different aspects on /r/TheRedPill subreddit. We analyse the effect of changing parameter r to modify partition granularity, analyse the model stability, and study the performance of two POS taggers on Reddit.

Partition Granularity   The selection of different r values for the k-means clustering detailed in Section 3.3 directly influences the number of clusters in the resulting partition of biased words. Low values of r result in smaller partitions and hence biases defined by bigger (more general) clusters, while higher values of r result in a higher variety of specific USAS labels allowing a more fine-grained analysis of the community biases at the expense of conciseness.
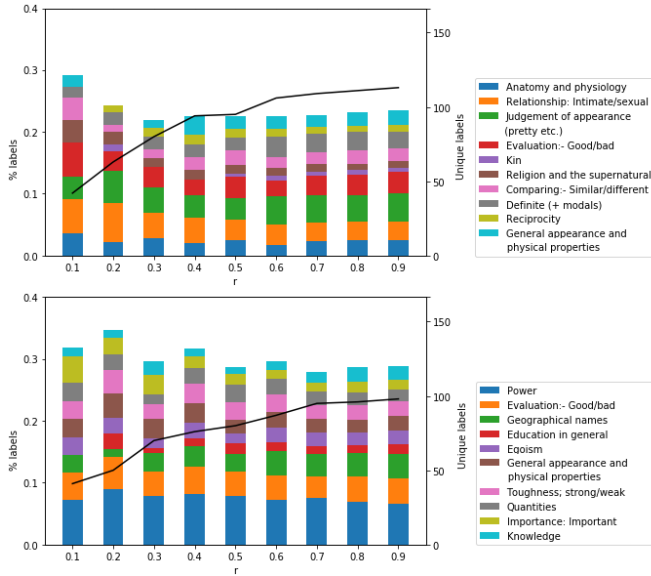

Figure 2: Relative importance (left axis) of the top 10 most frequent labels for women (top) and men (bottom), and number of unique labels (right axis) using different partition granularities (r) on /r/TheRedPill

Figure 2 shows the relative importance of the top 10 most frequent biases in /r/TheRedPill for women and men, presented in section 5.1, together with the quantity of unique USAS labels in each partition obtained for different values of r (see Section 3.3). Both figures show that most of the top 10 frequent labels for both women and men (see Section 5.1), have similar relative frequencies when compared with the total of labels in each partition for all values of r, with few exceptions such as *Reciprocity* and *Kin* labels for women and *Education in general* for men. This indicates that the set of the most frequent conceptual biases in the community is consistent among different partitions, usually aggregating on average between the 22 and 30% of the total of the clusters for women and men, despite the increase

in the quantity of clusters and unique labels obtained when using higher values of r. Even considering that the relative frequencies of the presented labels are similar between partitions, the different partitions share, on average, 7 out of the 10 most frequent labels for women and men. Among the top 10 most frequent labels for women in all partitions we find *Anatomy and Physiology*, *Relationship: Intimate/sexual* and *Judgement of appearance*. For men, some of the most frequent labels in all partitions contain *Power*, *Evaluation Good/bad* and *Geographical Names*, among others.

Stability analysis   To test the stability of our approach we created 10 bootstrapped models of /r/TheRedPill in a similar way as done by (Antoniak and Mimno 2018), including randomly sampling 50% of the original dataset and averaging the results over the multiple bootstrapped samples. Results show that the average relative difference between the ranks of the most frequent labels with respect to male and female-related target sets remains similar across all ten subdatasets. The results show the robustness of our approach to detect conceptual biases, and demonstrated that the biases were extended and shared by the community.

Frequency analysis   To test the effect that the word frequency threshold (when traininig the word embeddings model) has on the discovered conceptual biases of a community, we trained two new models for /r/TheRedPill changing the minimum frequency threshold to 100 (f_{100}) and 1000 (f_{1000}). First, as a consequence of the increase of the frequency threshold, the new models had relevant vocabulary differences when compared with the original f_{10} presented in Section 5.1: while the original model has a total of 3,329 unique adjectives, f_{100} has 1,540 adjectives (roughly 54% less), and f_{1000} has 548 adjectives (roughly 84% less). However, a quantitative analysis of the conceptual biases of the models show that the conceptual biases are almost the same for f_{10} and f_{100}, and very similar for f_{1000}: almost all top 10 labels most biased towards women and men in f_{10}, are also biased towards the same target set in f_{100} and f_{1000}. The only exception (1 out of the 20 labels for women and men) when comparing f_{10} and f_{100} models is label 'Evaluation:-Good/bad', a label slightly biased towards men in f_{10} but ranked in the same position for women and men in f_{100}. In f_{1000}, the figures are very similar too, but as there are many less words in the vocabulary (84% less), the resulting clusters do not have all labels present in f_{10}. However, and very importantly, all labels that do appear in f_{1000} (13 of 20) have the same relative difference as in f_{10}, with only two exceptions ('Quantities' and 'Knowledge').

Part-of-Speech (POS) analysis in Reddit   We performed an experiment comparing the tags provided by the *nltk* POS tagging method with manual annotations performed by us over 100 randomly selected posts from the subreddit /r/TheRedPill, following the same preprocessing presented in the paper, and focusing on adjectives and nouns. The results show that the manual POS tagger agrees with the *nltk* tagger 81.3% of the times considering nouns over 744

unique nouns gathered from 100 randomly selected comments. For adjectives, the manual tagger agrees with the *nltk* tagger 71.1% of the times, over 315 unique words tagged as adjectives by any of the two methods (manual, *nltk*) over the same set of comments. In addition, we also compared *nltk* with the *spacy* POS tagger using the same approach. The results show an agreement of 68.8% for nouns and 63.7% for adjectives, obtaining worse results than with the *nltk* library. Although the experiments are not conclusive, (a larger–scale experiment would be needed), the *nltk* library seems to indeed be helpful and be better suited than *spacy* to POS tag on Reddit.

B Most Frequent Biased Concepts

In this section we present the set of top 10 most frequent labels of the communities explored in this work, including all subreddits and Google News.

Top 10 most frequent labels in Google News (Section 4): Female: *Personal names, Other proper names, Clothes and personal belongings, People:- Female, Anatomy and physiology, Kin, Cleaning and personal care, Power, organizing, Judgement of appearance (pretty etc), medicines and medical treatment.* **Male**: *Personal names, Other proper names, Warfare, Power, organizing, Religion and the supernatural, Kin, Sports, Crime, Groups and affiliation, Games.*

Top 10 most frequent labels in /r/TheRedPill (Section 5.1): Female: *Anatomy and physiology, Relationship: Intimate/sexual, Judgement of appearance (pretty etc.), Evaluation:- Good/bad, Kin, Religion and the supernatural, Comparing:- Similar/different, Definite (+ modals), Reciprocity, General appearance and physical properties.* **Male**: *Power, organizing, Evaluation:- Good/bad, Geographical names, Education in general, Egoism, General appearance and physical properties, Toughness; strong/weak, Quantities, Importance: Important, Knowledge.*

Top 10 most frequent labels in /r/Dating_Advice (Section 5.1): Female: *Quantities, Geographical names, Religion and the supernatural, Language, speech and grammar, Importance: Important, Judgement of appearance (pretty etc.), Money: Price, Time: Period, Science and technology in general, Other proper names.* **Male**: *Evaluation:- Good/bad, Judgement of appearance (pretty etc.), Power, organizing, General ethics, Interest/boredom/excited/energetic, Quantities, Happy/sad: Happy, General appearance and physical properties, Calm/Violent/Angry, Helping/hindering.*

Top 10 most frequent labels in /r/atheism (Section 5.2): Islam: *Geographical names, Crime, law and order: Law and order, Groups and affiliation, Politeness, Calm/Violent/Angry, Judgement of appearance (pretty etc.), General ethics, Relationship: Intimate/sexual, Constraint, Warfare, defence and the army; weapons.* **Christian**: *Religion and the supernatural, Time: Beginning and ending, Time: Old, new and young; age, Anatomy and physiology, Comparing:- Usual/unusual, Kin, Education in general, Getting and giving; possession, Time: General: Past, Thought, belief.*

Top 5 most frequent labels in /r/The_Donald (Section 5.3): Hispanic: *Geographical names, General ethics,*

Wanting; planning; , Crime, law and order, Comparing:- Usual/unusual. **Asian**: *Geographical names, Government etc., Places, Warfare, defence and the army; , Groups and affiliation.* **Russian**: *Power, organising, Quantities, Evaluation:- Good/bad, Importance: Important, Sensory:- Sound.*

C Target and Evaluative Sets

The sets of words used in this work were taken from (Garg et al. 2018), and (Nosek, Banaji, and Greenwald 2002). For the WEAT test sets performed in Section 4, we used the same target and attribute word sets used in (Caliskan, Bryson, and Narayanan 2017). Below, we list all target words sets used.

Google News target and attribute sets From (Garg et al. 2018). *Female*: sister, female, woman, girl, daughter, she, hers, her. *Male*: brother, male, man, boy, son, he, his, him. *Career words*: executive, management, professional, corporation, salary, office, business, career. *Family*: home, parents, children, family, cousins, marriage, wedding, relatives. *Math*: math, algebra, geometry, calculus, equations, computation, numbers, addition. *Arts*: poetry, art, sculpture, dance, literature, novel, symphony, drama. *Science*: science, technology, physics , chemistry, Einstein, NASA, experiment, astronomy.

Google News set of USAS labels related with WEAT experiments: *Career*: Money & commerce in industry, Power, organizing. *Family*: Kin, People. *Arts*: Arts and crafts. *Science*: Science and technology in general. *Mathematics*: Mathematics.

/r/TheRedPill target sets From (Nosek, Banaji, and Greenwald 2002). *Female*: sister, female, woman, girl, daughter, she, hers, her. *Male*: brother, male, man, boy, son, he, his, him.

/r/atheism target sets From (Garg et al. 2018). *Islam words:* allah, ramadan, turban, emir, salaam, sunni, koran, imam, sultan, prophet, veil, ayatollah, shiite, mosque, islam, sheik, muslim, muhammad. *Christianity words*: baptism, messiah, catholicism, resurrection, christianity, salvation, protestant, gospel, trinity, jesus, christ, christian, cross, catholic, church

r/The_Donald target sets From (Garg et al. 2018). *White last names*: harris, nelson, robinson, thompson, moore, wright, anderson, clark, jackson, taylor, scott, davis, allen, adams, lewis, williams, jones, wilson, martin, johnson. *Hispanic last names*: ruiz, alvarez, vargas, castillo, gomez, soto, gonzalez, sanchez, rivera, mendoza, martinez, torres, rodriguez, perez, lopez, medina, diaz, garcia, castro, cruz. *Asian last names*: cho, wong, tang, huang, chu, chung, ng, wu, liu, chen, lin, yang, kim, chang, shah, wang, li, khan, singh, hong. *Russian last names*: gurin, minsky, sokolov, markov, maslow, novikoff, mishkin, smirnov, orloff, ivanov, sokoloff, davidoff, savin, romanoff, babinski, sorokin, levin, pavlov, rodin, agin