OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1109/TKDE.2021.3139680

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Ferrer Aran, X., Van Nuenen, T., Criado, N., & Such, J. (in press). Discovering and Interpreting Biased Concepts in Online Communities. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*. https://doi.org/10.1109/TKDE.2021.3139680

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 18. Nov. 2026

Discovering and Interpreting Biased Concepts in Online Communities

Xavier Ferrer-Aran, Tom van Nuenen, Natalia Criado, Jose Such

Abstract—Language carries implicit human biases, functioning both as a reflection and a perpetuation of stereotypes that people carry with them. Recently, ML-based NLP methods such as word embeddings have been shown to learn such language biases with striking accuracy. This capability of word embeddings has been successfully exploited as a tool to quantify and study human biases. However, previous studies only consider a predefined set of biased concepts to attest (e.g., whether gender is more or less associated with particular jobs), or just discover biased words without helping to understand their meaning at the conceptual level. As such, these approaches can be either unable to find biased concepts that have not been defined in advance, or the biases they find are difficult to interpret and study. This could make existing approaches unsuitable to discover and interpret biases in online communities, as such communities may carry different biases than those in mainstream culture. This paper improves upon, extends, and evaluates our previous data-driven method to automatically discover and help interpret biased concepts encoded in word embeddings. We apply this approach to study the biased concepts present in the language used in online communities and experimentally show the validity and stability of our method.

Index Terms—I.2 Artificial Intelligence, I.2.7 Natural Language Processing, O.8.15 Social science methods or tools, O.9 Ethical/Societal Implications

1 Introduction

Linguistic biases have been the focus of human language analysis for quite some time [1], [2], [3]. Recently, it has been proven that machine learning approaches applied to corpora of human language, such as word embeddings, learn human-like semantic biases [4], [5]. Beyond the value this has towards creating *fairer* Artificial Intelligence [6], [7], [8], [9], [10], [11], it has also enabled compelling methods for social scientists to study human language and biases [12], [13], [14], [15], [16]. That is, by looking at the biases learned by a machine learning model, it is possible to study biases in datasets in more detail. For instance, by training word embeddings models and analyzing them, Garg et al. [15] were able to study the dynamics and the evolution of predefined biases related to gender and religion in the United States across 100 years.

Existing approaches to study biases in word embeddings (see Section 2 for details), however, have one of the following limitations. That is, either: i) they are only able to attest whether arbitrarily predefined biases exist or not, so they are unable to *discover* what the actual most salient biases might be in a model; or ii) they only discover biased words, which is of limited value when trying to make sense of the biases discovered and what they mean at a conceptual level. These limitations often make current approaches unsuitable to study biases in the language of online communities. This is because such communities often exhibit values and norms that are sometimes distinct from those held by the majority [17], [18], so anticipating potential biases in advance is difficult; and because understanding the meaning of those biases is needed to properly interpret and study them [19], [20], [21]. Studying the biases of online communities is very important to tackle the social problems they have been associated with, such as radicalization [18] and discrimination against some types of users, particularly those users with protected attributes such as gender, religion, ethnicity, social class, etc. [20], [22].

In this paper, we improve upon our previous, general and data-driven approach [23] to discover linguistic biases in word embeddings. Concepts of interest, e.g. those related to protected attributes, such as gender, religion and the like, are used to *discover* the most frequent and biased words towards them (e.g. frequent words biased towards 'women' rather than 'men'). Differently from our previous work, in this paper: i) we improve the automatic selection of biased words, using not just the strength of biases but also their frequency; ii) we present a novel automatic categorisation of the biases, through semantic clustering and semantic analysis; and iii) we present novel methods to help interpret the resulting biased concepts, using semantic labels and rankings. All of this allows our method to help make sense of the most important biases at a conceptual level and in a more detailed manner, using the vocabulary of the online community studied.

The paper is structured as follows: Section 2 discusses related work and details the gap this paper addresses. Section 3 introduces the notion of *bias strength* in word embeddings, and the preliminaries on which our method bases. We present our approach to discover and interpreting biased concepts, together with the different bias ranking in Section 4. We apply our method to discover biased concepts in models trained using the corpora of two online communities (/r/TheRedPill, /r/Atheism) on the English-speaking discussion platform Reddit in Section 5. We provide an extensive evaluation of our method in Section 6, showing the stability of the biases discovered and the effect of the different parameters of our model and those used to train the embeddings model, and demonstrating the validity of our method by applying it to the general-purpose Google News pre-trained model and comparing our method with arbitrarily predefined biases in that

• *This is the author's version of the paper accepted for publication in IEEE Transactions on Knowledge and Data Engineering (2021). The corresponding author is Xavier Ferrer, Department of Informatics, King's College London. E-mail: xavier.ferrer_aran@kcl.ac.uk*

model attested by previous work. Finally, we finish with some concluding remarks and pointers to the available datasets, code and demo of the tool developed based on our method in Section 7.

2 Related Work

In this section, we discuss two main streams in the most related work to this paper. One stream includes those works *attesting* predefined biased concepts, and the other stream includes those works *discovering* biased words in word embeddings.

Works *attesting* biased concepts in word embeddings measure the association between predefined concepts usually related to known stereotypes – for instance, whether men are more often associated with a professional career while women are more often associated with family [5], [15], [16], [24], [25], [26], [27], [28], [29]. An example of measure used to this end is the Word Embeddings Association Test (WEAT) [5], inspired by the Implicit Association Test (IAT) [30], widely used in psychology and the social sciences to study stereotypes [31]. In particular, all the concepts involved (e.g. men/women *and* career/family) are represented via sets of words, and WEAT compares distances in the word embeddings model between those sets of words using cosine similarity. While being able to attest biases between arbitrary, predefined concepts is of high value, it nonetheless requires all the involved concepts to be defined in advance, which may not be possible in online communities, as they may exhibit biases and stereotypes that are not based on those in mainstream culture [18]. In contrast, in our work, we only consider as input the *attribute* concepts of interest, e.g. men/women, and our method then discovers all other concepts that are most associated with these *attribute* concepts.

Works *discovering* biased words overcome some of the limitations of *attesting* by enumerating all the words that are biased towards others in a word embeddings model [4], [32], [33], [34], [35]. However, the resulting lists of biased words do not explain what these biases mean or to what extent they are important in the context of a community, both of which are needed to properly interpret and study the biases discovered in a community [19], [20], [21]. Previous work has demonstrated language innovation in the manosphere in order to understand how different groups within it view their own departure from mainstream society [36]. Socio-cultural linguistic work, which has long addressed the structure and meaning of specialised vocabularies, typically was based on smaller sets of data. The internet and social media have generated numerous communities, however, each producing vast amounts of content. Manual identification and analysis of the subcultures' specialised vocabularies has become impractical; automatic methods are needed to help with both the identification and understanding of community language.

Our previous work in this direction is presented in [23], in which we define a system to identify a set of words biased towards certain concepts, and organize them in categories in order to compare their biases. Although the approach gives an idea of what are the most biased words towards certain concepts in a community, and to which semantic categories they belong, it has several limitations. First, the bias measure used does not consider the frequency when selecting the set of biased words, which means that some of the biases may be rare and unrepresentative of the community. Second, the threshold used to select biased words is manually determined by analyzing the distribution of bias in the community, which makes it difficult to automate. Third, the analysis of biases performed does not allow for a link between the biased concepts and the actual language used by the community. This is important as smaller communities in larger societies frequently develop specialised vocabularies that reflect unique aspects of the identity, needs and realities of group members [37]. A further limitation is that the work neither considered the potential impact of stability, nor did it recognize the effect of different parameters in the model.

In this paper, we combine both frequency and bias to automatically identify the most salient words in a community; that is, words that are both frequently used and strongly biased towards the two concepts of interest. Afterwards, these words are aggregated based on their semantic similarity, by clustering them using the word embeddings model. The resulting semantically-similar clusters are analyzed in two ways. First, we categorize the major discourse field each of the clusters belongs to and provide a relative frequency of the fields. Second, clusters are ranked based on the strength of their bias, their frequency, and their sentiment to offer a more detailed view of the biases discovered and how they are expressed in the language of the community in a more or less biased, frequent, or positive/negative way. The semantic categorization and different rankings provides a context that offers both a general and more detailed view of the biases at the conceptual level over the different dimensions of bias and in an automatic manner.

3 Preliminaries

The first step to be able to discover biased concepts towards the attribute concepts of interest (e.g. men/women), is to discover the specific words that are biased towards the attribute concepts. To do this, we need to measure the bias between words in a word embeddings model. Given a word embeddings model built from a text corpora, and two sets of words representing the attribute concepts (e.g. men/women) one wants to discover biases towards, a common approach in the literature[1] is to leverage the cosine similarity between embeddings to identify words close to an attribute concept (e.g. men) and far from the other attribute concept (e.g. women) as we detail below.

Bias Strength. Let $T_1 = \{w_0, ..., w_i\}$ and $T_2 = \{w_0, ..., w_j\}$ be two sets of words that represent two different attribute concepts we want to discover biases towards — e.g. {*she, daughter, her, ..., mother*} and {*he, son, him, ..., father*} describing the concepts *women* and *men* — and c_1 and c_2 the centroids of T_1 and T_2 respectively, estimated by averaging the embedding vectors of all words in each attribute concept; we say that a word w is biased towards T_1 with respect to T_2 when the cosine similarity between the embedding of w is higher for c_1 than for c_2:

$$B(w, T_1, T_2) = cos(w, c_1) - cos(w, c_2) \quad (1)$$

Larger absolute values of B correspond to stronger biases, positive values of bias B indicate that word w is biased towards T_1, and negative values of B indicate that word w is biased towards T_2. By applying the bias strength equation to the entire community's vocabulary, one is able to discover all the words that are biased towards attribute concepts T_1 with respect to T_2 and vice versa.

1. Note that there are many alternative bias definitions in the literature [32], [38], such as the *direct bias* measure [4], the relative norm bias metric [15], and others, all with similar results as shown by previous research [15], [38]. When experimentally comparing the *direct bias* metric with ours in r/TheRedPill, we obtain a Jaccard index of 0.857 (for female) and 0.864 (for male) regarding the list of 300 most-biased adjectives generated with the two bias metrics.

4 Discovering and Interpreting Biased Concepts

Beyond discovering individually biased words, as stated in the previous section, our approach aims to discover and help understand the *concepts* behind the biased words, which prior work does not do. To this aim, we follow two main steps.

In the first step, described in detail in Section 4.1, we consider not only the strength of bias but also its frequency, in order to elicit the most common and frequently biased words in a community. After this, we aggregate all these resulting words into biased concepts, by clustering the most common biased words based on their embedding similarity.

In the second step, described in detail in Section 4.2, we focus on facilitating the interpretation of the biased concepts. We do so in two main ways: i) by classifying every biased concept into a major discourse field, which further abstracts the meaning of the biases and facilitates a general understanding of the types of biases present in the community; and ii) by creating rankings of the biased concepts based on the strength, frequency and sentiment of the biases discovered to offer a more detailed view and nuanced analysis.

4.1 Discovering Biased Concepts

We are not only interested in the strength of the bias of the biased words, but also if biased words are commonly used, hence frequent in the vocabulary. This has two main purposes: i) to focus on the representative biases of a community [19]; and ii) to avoid instability due to low-frequency words [39].

4.1.1 Bias Salience

Let V a set for words forming the vocabulary of a word embeddings model. We determine the salience Sal of a word $w \in V$ towards an attribute set T_1 w.r.t an attribute set T_2, by combining the normalized bias from Equation 1 with the normalized frequency rank[2] of the word w in the corpus—denoted by $R(w)$ and assigning the most frequent word in V rank 1 and the least frequent word rank $|V|$. Words with higher values of salience $Sal \in [0, 1]$ will be both frequent and biased towards the attribute set T_1 with respect to the attribute set T_2. Therefore, salience is computed as follows:

$$Sal(w, T_1, T_2) = \left(1 - \frac{R(w) - 1}{|V| - 1}\right) \cdot \left(\frac{B(w, T_1, T_2)}{\max\limits_{w_j \in V} B(w_j, T_1, T_2)}\right) \quad (2)$$

Even when knowing the strength and frequency of biased words, considering each of them as a separate unit is not enough to analyse and understand biases at a more conceptual level. There is a need to semantically combine related terms under broader rubrics in order to facilitate the comprehension of the biases. We start by selecting the most salient words, which are then aggregated through k-means clustering on the word embeddings.

4.1.2 Most Salient Words

We order all $w \in V$ by salience, using Equation 2, towards an attribute set T_1 with respect to an attribute set T_2 (e.g. the set of words representing women with respect to the set for words representing men), and select the most salient words to focus on the most prominent biases.

In order to do this, we consider the distribution of salience values for all the words in the vocabulary V (denoted by X) and we consider n standard deviations plus the mean of the salience distribution ($n * \sigma(X) + \bar{X}$) as the threshold to select the most salient words — we show later, in the extensive experiments conducted and described in Section 6.2, a characterization of how our method behaves with different n. That is, we select words biased towards each attribute set with salience greater or equal than $\delta = n * \sigma(X) + \bar{X}$. All the words in the vocabulary V with salience towards T_1 with respect to T_2 higher than a threshold δ form the set S_1. In a similar manner, we create the set of salient words S_2 by considering salience towards T_2 with respect to T_1. These words are then used for the semantic clustering, as detailed below.

4.1.3 Semantic Word Clustering

For each set of the most salient words in S_1 and S_2 and their embeddings, we use k-means to aggregate the most semantically similar words into clusters, based on the distance between word vectors. For each set of word vectors associated with the most salient words (denoted by S_1 and S_2), we apply k-means setting the number of clusters to all values within the $[2, \ldots, |V|]$ interval, and we select the partition that maximizes the silhouette score [41].[3] We repeat the clustering τ times in order to obtain the best partition by considering different k-means random initialisation variables —we show later experimentally how our method behaves with different τ. Notice that salient words S_1 and S_2 are clustered per separate hence resulting in two different partitions, K_1 and K_2 respectively.

After clustering, as we then have all the sets of words needed, we apply WEAT [5] between all clusters from both partitions K_1 and K_2 and attribute concepts T_1 and T_2, and only keep these clusters of each partition that return significant p-values[4] towards each respective attribute concept when compared with all clusters from the other partition. In this way, we make sure that the biased concepts represented in every cluster are relevant and strongly associated to each attribute concept.

The resulting sets of clusters for partitions K_1 and K_2 represent the most salient and representative biased concepts towards the attribute concepts T_1 and T_2 in the community.

4.2 Interpreting Biased Concepts

In this section, we focus on obtaining a general understanding of the clusters found in a given partition K to facilitate the

2. The reason to use the frequency rank instead of raw frequencies is that, since word frequencies are known to follow the Zipf's law [40], the difference in frequency is large between the most frequent words, which could then skew saliency towards frequency rather than having a balance between strength and frequency of the biases. Therefore, we use the word frequency rank instead of raw frequency to adequately smooth its importance in the equation. Note that this also ensures that when we select the most salient words later, we discard those that may have very similar frequencies but different ranks at the tail of the frequency distribution.

3. We use silhouette for its simplicity and visual clues, but our method is agnostic to the clustering algorithm and other alternatives, such as [42], [43], could also be used in conjunction with our method.

4. We correct for multiple tests to reduce Type I errors using Benjamini-Hochberg (BH) with 10% false discovery rate [44]. BH is considered more powerful than other too conservative methods like Bonferroni, which consider all p-values equally, leading to Type II errors. BH corrects p-values according to their ranking, with a better trade-off between Type I and II errors [45].

interpretation of the discovered biased concepts. Since the most salient words are grouped into semantically meaningful clusters, here we aim to organize and understand the meaning of the different clusters and facilitate the comparison between the biases towards each attribute concept. We do so in two main ways: by first categorizing the clusters via semantic tagging and analyzing the frequencies of the semantic tags, and second, by ranking the clusters of a partition based on different metrics. By combining these two ways, our method is able to offer both a *general* and a *detailed* view of the biased concepts of a community.

4.2.1 Categorising Biased Concepts

To give an overview of the biased concepts in a community, we conduct a semantic analysis. In particular, we tag every cluster in a partition with the most frequent semantic fields (domains) among its words. That is, for each cluster $C = \{w_0, \ldots, w_n\}$ in a partition K, we first obtain the semantic domain associated with each word $w \in C$ (denoted by $s(w) \subseteq D$, where D is the set of all semantic domains). We also obtain the multiset of semantic domains associated with the cluster as the sum of the semantic domains associated with its words, i.e. $SC = [s(w_0)] + \ldots + [s(w_n)]$, and then we select the semantic domain that is most frequent among all the words in the cluster as the cluster tag, i.e. $TC = \arg\max_{d \in D} m_{SC}(d)$, where $m_{SC}(d)$ denotes the multiplicity of semantic domain d in the multiset of semantic domains associated with cluster C. Regarding the specific semantic domains considered, we base on the UCREL Semantic Analysis System (USAS) [46], which has a multi-tier structure with 21 major discourse fields subdivided in more fine-grained semantic domains[5] such as *People*, *Relationships*, *Power*, *Ethics* and it has been extensively and successfully used for many tasks, such as the automatic content analysis of discourses [47] and as a translator assistant [48].

After the semantic tagging, we then analyze the frequency of the different semantic tags in a given partition $K = \{C_1, \ldots, C_n\}$. In particular, we compute the multiset formed by the tag of each cluster in the partition, denoted by $TK = [TC_1] + \ldots + [TC_n]$, and calculate the relative frequency of each semantic domain $d \in D$ as $r_d = \frac{m_{TK}(d)}{|TK|}$, where $|TK|$ is the total number of semantic domain tags in a partition K, and $m_{TK}(d)$ is the multiplicity of the semantic domain d in the multiset TK. This allows us to get a general idea of the biased concepts in a partition. For instance, this could show that for a given partition K (e.g. the clusters of the most salient biases towards men), the relative frequency of *Power* (r_{Power}) is higher than the relative frequency of *Relationships* ($r_{Relationships}$) — as we, in fact, see in one of the datasets (/r/TheRedPill) explored in the application of our method to Reddit communities later on, as described in Section 5.

4.2.2 Ranking Biased Concepts

While showing the relative frequencies of the categories of the biased concepts found is informative as an abstract, a birds-eye view of the biases in a community, the distinctions marked may be coarse-grained [49]. Following the example above, even though *Power* may be a frequent category, we do not actually know how this is expressed in the language of the community, nor how strongly or frequently biased these clusters are towards men. It also does not indicate whether the biases are typically expressed in a positive or negative way.

In this section, we present a more detailed analysis of the biased concepts found. Instead of looking at an aggregate view of categories of clusters as in the previous section, we focus on establishing different methods to rank each of the clusters discovered. In particular, we define three different metrics to prioritise the clusters of a partition K considering: i) the frequency of the bias; ii) the strength of the bias; iii) and the sentiment polarity of the bias. The rankings offer three different but complementary views to help understand the biases found in the community.

We define the rankings based on the clusters of each partition. In particular, given a cluster C in a partition K, we define the following metrics:

1) R_f, which measures the aggregated frequency in the text corpus of the words within cluster C, therefore establishing a ranking of the most common biases:

$$R_f(C) = \sum_{w \in C} F(w) \tag{3}$$

where $F(w)$ is the frequency of word w in the text corpus.

2) R_e, which orders the clusters based on the average bias strength of the words in the cluster with respect to the attribute concepts T_1 and T_2 (see Equation 1). This method assigns higher scores to the clusters that are more biased, and it is useful to identify relevant biases towards an attribute concept when compared to another:

$$R_e(C, T_1, T_2) = \sum_{w \in C} B(w, T_1, T_2) \cdot \frac{1}{|C|} \tag{4}$$

where $|C|$ is the size of the cluster.

3) R_s, which orders the clusters based on the average sentiment of the words in the cluster. We particularly consider rankings of both the most positive and most negative biases. Although strong negative polarities might be indicative of perilous biases towards a specific population, the fact that a cluster/word is not tagged with a negative sentiment does not exclude it from being discriminatory in certain contexts. We particularly define R_s as follows:

$$R_s(C) = \sum_{w \in C} SA(w) \cdot \frac{1}{|C|} \tag{5}$$

where SA returns a value $\in [-1, 1]$ corresponding to the sentiment polarity of a word w, with -1 being strongly negative and +1 strongly positive. Note that our model is agnostic to the sentiment analysis model used and different tools may be used [50].

5 Implementation on Reddit

In this section, we use our method to discover biases in textual corpora collected from two Reddit communities, /r/TheRedPill and /r/Atheism.

Although both communities are suspected to incorporate significant gender and religion biases respectively [51], less is known about the form these biases take – e.g., what the concepts are

5. http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

that are more biased towards women than towards men. Table 1 summarizes the datasets, including the protected attribute (*P.Attr*) we discover biases towards, the quantity of unique *Authors*, *Comments* and *Words*, words per comment (*Wpc*), and vocabulary *Density* (ratio of unique words to total number of words). The sets of words representing the attribute concepts gender and religion are taken from previous work [15], [52] and listed in Appendix B.

TABLE 1: Datasets of Reddit communities used in this paper.

Dataset	P.Attr	Years	Authors	Comments	Words	Wpc	Density
/r/TheRedPill	gender	2012-18	106,161	2,844,130	59,712	52.58	$3.99 \cdot 10^{-4}$
/r/Atheism	religion	2008-09	699,994	8,668,991	81,114	38.27	$2.44 \cdot 10^{-4}$

The datasets were collected using the Pushshift data platform [53]. The two Reddit models were trained using an Intel i5-9600K @3.70GHz with 32GB RAM, and an NVIDIA TitanXP GPU. To create an embedding model for each corpus, we first preprocess each comment by removing special characters, splitting text into sentences, and transforming all words to lowercase. Then, for each subreddit, we train a skip-gram word2vec[6] word embeddings model, using embeddings of 200 dimensions, words with at least 10 occurrences, a 4-word window and 100 epochs as training parameters, and $n = 4$ (four standard deviations) and $\tau = 200$ repetitions as parameters for our method. In Section 6.2, we offer an extensive analysis varying all these parameters, which shows similar results at the conceptual level regardless of the parameter choice. For the sentiment polarity, we used the `nltk` sentiment analysis system [56]. All our code is available publicly (see Section 7).

5.1 /r/TheRedPill

The /r/TheRedPill community defines itself as a forum for the 'discussion of sexual strategy in a culture increasingly lacking a positive identity for men' [51]. The name of r/TheRedPill 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. In response, men must protect themselves against a 'misandrist' culture and the feminising of society [18], [57]. More generally, the 'red pill' construct, found across many communities in the Manosphere, is an umbrella term for a set of beliefs that men and women are categorically different on the basis of a perception of clearly defined sex [58]. Communities such as The Red Pill are related to the online Manosphere [59], [60], a term used to describe a collection of predominantly web-based misogynist ideologies, which include deterministic views of masculinity and femininity [58]. Previous research shows that the Manosphere is characterized by a volume of hateful speech that is significantly higher than other Web communities [61], and we expect these biases to be present. We applied our method to /r/TheRedPill in order to be able to discover the exact biased concepts related to gender present in the community. More interestingly, our method enables the study of the shape these biases take, i.e., how these biases are actually expressed in the community's language. This is relevant as fringe communities such as these tend to produce concepts and expressions that are not in common use and represent the new realities, norms and values of subcommunities [36].

6. Using word2vec allows us to validate our method against the widely-used word2vec Google News pre-trained model (Section 6). However, our method can easily be extended to other models like ELMo [54] or BERT [55].

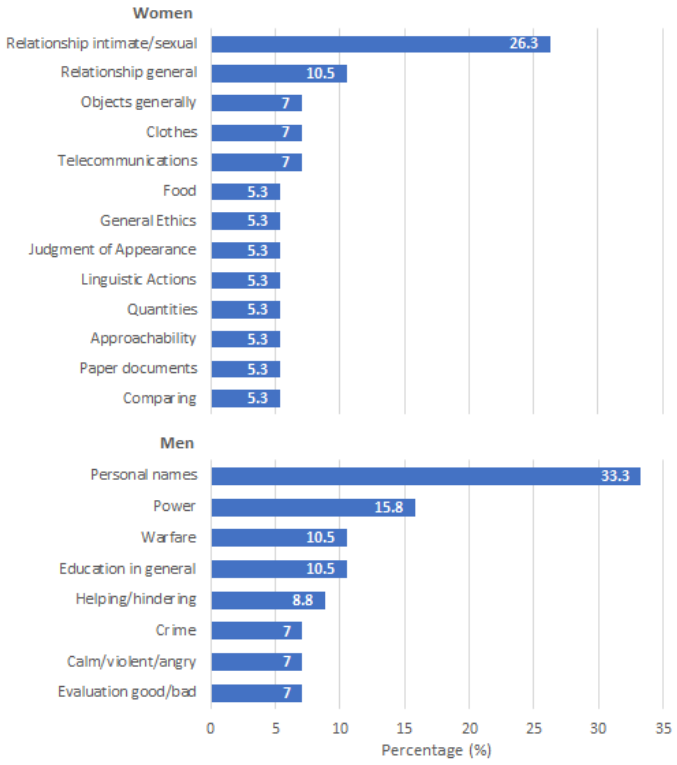

Fig. 1: Relative frequency of semantic categories for biased concepts towards *women* (top) and *men* (bottom) in /r/TheRedPill.

After applying our method to /r/TheRedPill to discover biases towards women and men considering nouns, adjectives and verbs, we obtain the most salient words towards both attribute concepts, with sizes 216 and 194 respectively. These were then clustered into 93 and 102 clusters, with a maximum of 10 and 8 words, a minimum of 1, and a mean of 2.32 and 1.90 words with a standard deviation of 1.78 and 1.33 for women and men, respectively. The WEATs performed were all with p-values between $1.2 \cdot 10^{-3}$ and $1.5 \cdot 10^{-2}$, passing Benjamini-Hochberg correction for multiple tests as introduced before.

Figure 1 shows the distribution[7] of the categories among clusters for *women* and *men*. The most frequent biases towards *women* refer to relationships (and particularly *sexual relationships*, used to tag the 26.3% of the total number of biased concepts discovered), appearance (Clothes and Judgement of Appearance, adding up to 12.3%) and objects (Objects generally, 7%). Following, the most frequent biases towards *men* refer to Personal Names (adding up to 33% of the conceptual clusters biased towards men, discussed below in the detailed analysis), Power (15.8%), Warfare (10.5%) and Violence/Crime (adding up to 14%, including the categories Calm/violent/angry and Crime). The semantic categorization of discovered biases shows a very different picture of the biases found towards the two genders: the most frequent semantic labels used to tag biased concepts towards women are predominantly related to objectification, i.e., appearance and sex, while men are predominantly described in relation to positions of power and strength, i.e., becoming successful agential subjects in the realm of dating and sex.

A more detailed view is shown in Figure 2, which compares

7. Note that for the sake of clarity and readability, only categories tagging at least 2% of the clusters are shown in the distribution.

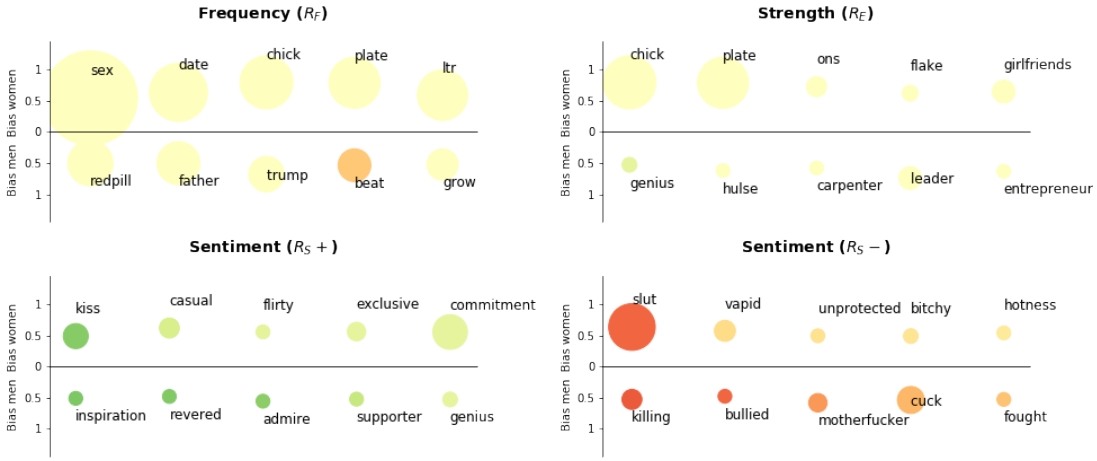

Fig. 2: Top-5 clusters, labelled with their most frequent word, biased towards *women* and *men* in /r/TheRedPill, ranked by most frequent (R_F), strong (R_E), sentimentally positive (R_S+), and sentimentally negative (R_S-) bias. Higher ranking clusters are shown ranked from left to right (x-axis), the area of each cluster corresponds to its frequency (quantity of times the words of the clusters were found in the dataset), the color with its average sentiment, and the y-axis shows the average salience of the words in the cluster. Finally, each of the clusters is tagged with its most frequent word.

the top-5 clusters for women and men in /r/TheRedPill, labeled using the most frequent word in the cluster, and ranked by the frequency, strength, and sentiment of the bias. In each figure, clusters are shown ranked from left to right starting with the highest ranked cluster and ignoring clusters with the same stem to show different biased concepts. The top part of the figure presents the biases towards women, while the biased concept towards men are shown on the bottom. The area of each cluster corresponds to the aggregated frequency of its words, that is, the quantity of times the words of the clusters were found in the subreddit. For instance, in r/TheRedPill, cluster 'sex' the most frequent biased concept for women while 'redpill' is the most frequent biased concept for men. Cluster colors correspond with its average sentiment, ranging from negative (red), to neutral (yellow), to positive (green), and the y-axis shows the average salience of the words in the cluster.

The most-frequent and strongest biases towards *women* appear in clusters labeled 'sex' or 'slut', which signal obvious biased language towards women through objectification. We can also see clusters with particular jargon such as 'ons' (one night stand), 'chick', 'plate' (a man or woman who is used for the purpose of sex), or 'flirty'. In the positively charged clusters, the appearance of the terms 'commitment' and 'exclusive' refers to issues of promiscuity and the "evolutionary" explanations for women's loyalty that previous studies have connected to the Red Pill community [60].

For *men*, the picture is quite different. The top clusters biased towards men stand out due to their focus on strength, power and violence, concurring with the general categories identified previously. Many of the men-biased top clusters, such as 'inspiration', 'genius', 'leader', and even 'trump' (found in the same cluster with other personal names associated with leadership, such as 'obama'), represent role models related with strength and power, while others clearly allude to violence like 'killing', 'beat', etc. Finally, the most frequent negative cluster for men contains the term 'cuck', signalling a conflict of masculinity identified in previous research, in which men struggle against one another for access to power within the dominant group [36], [59].

Our findings – i.e., that women are objectified whilst men engage in articulations of aggrieved manhood – are in line with and confirm previous social science studies on /r/TheRedPill regarding the hostile, misogynistic language in this community [36], [59], [60]. The fact that our models have isolated these issues indicates their reliability in replicating existing theory. It suggests that our method finds relevant biased concepts, and that it enables an exploration of the relations between words and concepts, as constructed by the community itself. We further evaluate the quantitative validity of our method in Section 6.3.

5.2 /r/Atheism

The /r/Atheism subreddit is a large community that calls itself 'the web's largest atheist forum', on which '[a]ll topics related to atheism, agnosticism and secular living are welcome'. Despite these goals, reading many of the comments on the subreddit leads to suspicions about biases towards different religions – particularly Islam vs Christianity – although these biases seem less evident than in the r/TheRedPill example. After applying our method considering nouns, adjectives and verbs, to /r/Atheism to discover biases towards *Islam* and *Christianity* (the two largest religions), we obtained the most salient words towards both attribute concepts, with sizes 516 and 381 respectively. These were then clustered into 188 and 178 clusters, with a maximum of 19 and 17 words, a minimum of 1 word, and a mean of 2.74 and 2.14 words per cluster with a standard deviation of 3.03 and 2.10 for *Islam* and *Christianity*, respectively. The WEATs performed were all with p-values ranging from $2.5 \cdot 10^{-7}$ to $3 \cdot 10^{-4}$, and passed Benjamini-Hochberg correction test as introduced before.

Figure 3 shows the distribution of the semantic categories among biased concepts for *Islam* and *Christianity*. The figure shows a clear difference between the two attribute concepts *Islam* and *Christianity*. The biased concepts for Islam are categorised in normative terms such as Warfare (20.8%), Calm/Violent/Angry (14.3%), and Crime (10.4%), together with names (aggregating Geographical names, Personal names and Other proper names, and adding up to 39% of all biased concepts for *Islam*; more on this later in the detailed analysis). Clusters about Christianity, on the other hand, fall mainly under the more descriptive and generic

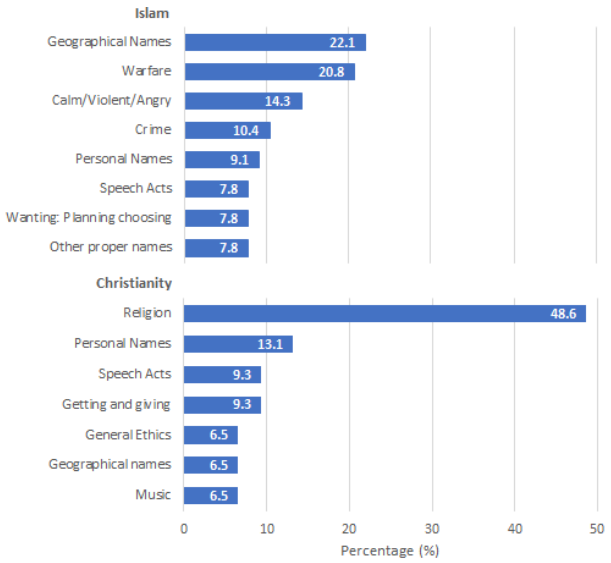

Fig. 3: Relative frequency of semantic categories for biased concepts towards *Islam* (top) and *Christianity* (bottom) in /r/Atheism.

category of Religion (with almost half of the biased concepts for *Christianity* being tagged with this semantic label), followed by Personal names (13.1%), Speech acts (9.3%) and General Ethics (6.5%). These broad groupings suggest a difference in evaluative orientation towards the two religions.

A more detailed view is shown in Figure 4, which compares the top-5 clusters for *Islam* and *Christianity*, ranked with the different rankings defined in Section 4.2.2. In each figure, Islam-biased clusters are shown on top and Christian-biased clusters on the bottom, the y-axis corresponds with average salience of the cluster, size corresponds with frequency (cluster 'evolution' is the most frequent in /r/Atheism, with more than 259K hits), and color with average sentiment. Observing the top most frequent biased clusters for both religions in the frequency ranking (R_F), the most frequent clusters biased towards *Christianity* contain general doctrinal concepts (e.g. 'heaven', 'sin', 'teachings'). Half of the top 5 most frequent clusters biased towards *Islam*, however, are related to violence ('violence', 'offensive', 'attack'). Moreover, the 'violence' cluster contains other strongly stereotyped terms such as 'terrorism', 'jihad', and 'extremism', suggesting anti-Muslim sentiments and Islamophobia. Further proof of this would need to be further substantiated through closer inspection of the context of these terms, but our approach does produce striking differences.

When we look at the strongest-biased clusters, we see that personal nouns describe the majority of the clusters biased towards *Islam*, such as 'ali', 'abdul', and 'omar'. These refer to public figures associated with socio-political issues of *Islam*. The cluster 'ali' also includes the terms 'hirsi' and 'ayaan', thus referring to activist Ayaan Hirsi Ali, who is known for her critical stance on *Islam* and practices such as forced marriage and honor violence. The cluster 'omar' also includes the terms 'sheikh' and 'ahmed', which likely refer to Ahmed Omar Saeed Sheikh, a British militant who was found guilty for the 1994 kidnappings of Western tourists in India. These names, then, signal a similar concern for issues of hostility and violence we saw in the previous ranking. For *Christianity*, most of the clusters again refer to more generic concepts about religion such as 'divinity', or 'nazareth'. This could indicate that the discourse on /r/Atheism, when dealing with *Islam*, revolves more around particular people and the news events they appear in, whereas *Christian* discourse is less topical and political, and revolves around theological concepts and concerns.

Finally, the distribution of sentiment across all clusters biased towards *Islam* is clearly more negative than *Christian*-biased clusters, and gives an idea on some of the most common biased concepts associated with both religions in /r/Atheism. The most negative clusters biased towards *Islam* are relatively frequent, biased and closely related (again) with violence, such as 'violent', 'attacks' or 'insult'. On top of that, we only find one positive cluster ('honour') for *Islam* (shown in R_s+ ranking), in contrast to the various positive clusters biased towards *Christianity*. Clustered words biased towards *Christianity* are slightly more positive, and refer to religious terms of reward or punishment, such as 'divinity', 'heaven', 'gift' and 'sin'. In general, the results suggest broad socio-cultural perceptions and stereotypes that characterize the discourse in /r/atheism community and that frequently associate Islam-biased clusters to negative connotations in contrast to Christian-biased concepts.

6 Evaluation

In this section, we provide an extensive evaluation of our method considering different dimensions: the stability of the biases discovered, the influence the model and learning parameters may have on them, and the validity of the biases discovered. We particularly analyze the stability of our models in Section 6.1, and the importance of the different parameters used for our method and when training the embedding models in Section 6.2. We finally evaluate our approach on discovering biased concepts biases in online communities in Section 6.3, by comparing the biased concepts discovered by our method with the arbitrary, predefined biases attested by previous work in the Google News pre-trained model.

6.1 Stability

Word embeddings may be unstable, particularly when trained with smaller corpora, so that small changes during training (e.g. new data) may result in different vector descriptions for the same word [39], [62]. This could be a problem, since different vector descriptions could result in different sets of biased concepts discovered and influence the analysis of biases of the online community. Although our method provides some stability mechanisms, such as the frequency of words and semantic aggregation into clusters, it is based on the same vector descriptions.

We tested the stability of the biases found in Sections 5.1 & 5.2, following [39], by training four new embedding models for each of the two datasets explored, randomly selecting 50% of the comments of the original models to reduce the training corpora (increasing the chance for instability), and using the same preprocessing and parameters as before. This resulted in eight new models trained with a random half of the comments: TRP_1 – TRP_4 for /r/TheRedPill, and ATH_1 – ATH_4 for /r/Atheism. For each of the new models, we applied our method as before.

Figure 5 shows the overlap coefficient [63], also known as Szymkiewicz–Simpson, of the sets of semantic categories for the 4 new models for /r/TheRedPill (TRP_1 to TRP_4) and its original model (TRP_{Orig}), per attribute concepts women W and men M (shown as superscripts), used by more than 1% of the clusters to improve readability. The higher the overlap, the closer to one, the lower the overlap, the closer to zero. The figure shows that the

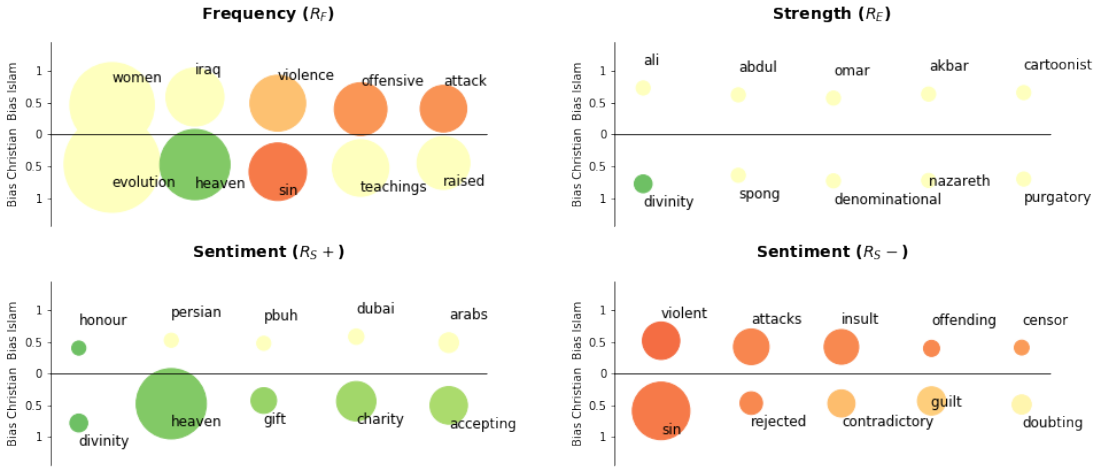

Fig. 4: Top-5 clusters, labelled with their most frequent word, biased towards *Islam* and *Christianity* in /r/Atheism, ranked by most frequent (R_F), strong (R_E), sentimentally positive (R_S+), and sentimentally negative (R_S-) bias.

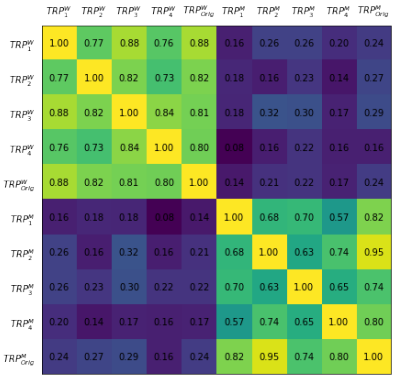

Fig. 5: Overlap coefficient between bootstrapped TRP_{1-4} and original TRP_{Orig} models for women (W) and men (M).

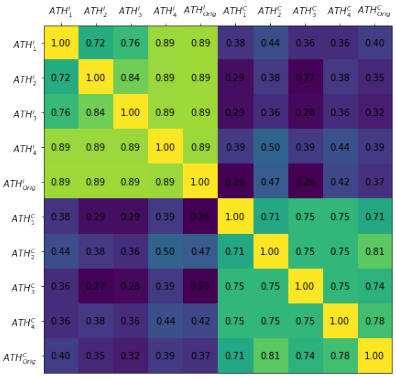

Fig. 6: Overlap coefficient between bootstrapped ATH_{1-4} and original ATH_{Orig} models for *Islam* (I) and *Christianity* (C).

overlap between *women*- and *men*-biased semantic tags is small, meaning that there is a clear difference between the semantic categories used to label the concepts biased towards both attribute concepts in all models. In addition, and when considering women and men attribute concepts per separate, the set of most frequent tags have an average overlap of 0.83 for both *women* and *men* with the original model TRP_{Orig}. That indicates that a very similar set of semantic tags was used to label women-biased concepts (and men-biased concepts, separately) across all models. Finally, the set of biases towards the same attribute concept are very similar in all new models ($TRP_{1...4}$).

Figure 6 shows the overlap coefficient with the same colour code as in the previous figurefor the 4 new models for /r/atheism (ATH_1 to ATH_4) and its original model(ATH_{Orig}), per attribute concept *Islam* (I) and *Christian* (C) (shown as superscripts), used by more than 1% of the clusters to improve readability. The figure shows that the overlap between Islam and Christian biased labels is small, keeping the differences between them across models. Also, the overlap coefficient between biases towards the same attribute concept is large, indicating that our approach is able to consistently identify similar biased concepts towards Islam and Christian across models in /r/Atheism. Therefore, given the results obtained, we can conclude that our method is able to pick up consistent biases, mitigating many stability issues associated with word embeddings.

6.2 Parameter Influence

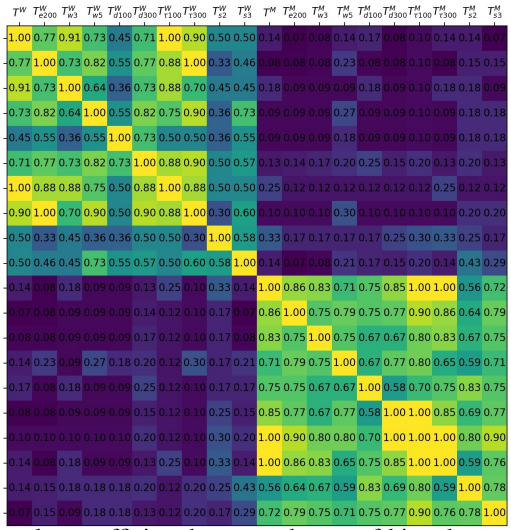

Fig. 7: Overlap coefficient between the set of biased concepts of the different models for /r/TheRedPill.

We now provide a more detailed analysis on the effects that different training parameters have on the models created and the final set of discovered biased concepts to complement the experiments on stability presented above. For this, and due to a lack of space, we focus on the /r/TheRedPill community, which has the least number of unique words and the highest vocabulary density, so it is the most prone to suffer from stability issues that can cause differences in the biased concepts discovered given changes in the dataset and/or the parameters used. In order to study the effect of all the parameters used, we trained five new models and ran nine new executions with different training and method parameters. When training the new models, only one parameter was changed at a time to study its influence *ceteris paribus*. That is, all of the other parameters were set to the default values used to train the original /r/TheRedPill model T (using a window size of 4, embedding dimension of 200, 100 epochs, $\tau = 200$, and using a salience threshold of $n = 4$ standard deviations – see Section 5 for all the details). The specific ranges of the parameters used in the different executions ran are presented below:

Window sizes (w): We trained two new models, T_{w3} and T_{w5}, utilizing window sizes of 3 and 5 (instead of the window size of 4 used in the original models), to evaluate the effect of window sizes in the resulting set of discovered biased concepts.

Embedding dimension (d): Models T_{d100}, and T_{d300}, were trained with 100 and 300 dimensions (instead of 200), respectively, to study the effect of embedding dimensionality.

Epochs (e): Model T_{e200} was trained using 200 epochs (instead of the 100 epochs of the original model), to study the effect of a longer training period.

Cluster repetitions (τ): Models $T_{\tau 100}$ and $T_{\tau 300}$ were trained using the same training parameters as the original model, but with τ 100 and 300 cluster iterations, to select the partition with higher silhouette score in order to study its effect in the final set of biased concepts (instead of the $\tau = 200$ used originally in Section 5).

Salience thresholds (s): Models T_{s2} and T_{s3} were trained with the same parameters as the original model, but using $n = 2$ and $n = 3$ standard deviations (instead of the $n = 4$ used originally in Section 5) to select the set of most salient words towards each attribute concept.

Figure 7 shows the overlapping coefficient between the sets of frequent biased concepts for women (represented with the superscript W) and men (represented with the superscript M) between the different models for /r/TheRedPill[8] and the original model presented in Section 5.1, represented as T. The results show that: i) an important overlap coefficient among the sets of biases towards the same gender for both women and men, indicating that biases towards women (and men) are similar among the different models, and 2) a strong difference between women and men biases in all models, indicating that the sets of discovered biased concepts towards the two genders are different in /r/TheRedPill. Focusing on the influence of specific parameters, the results suggest that, among the different training parameters tested (training epochs, window sizes, and embedding dimensions), low embedding dimensionality ($d = 100$) has the biggest impact when determining the set of biased concepts of a community, coinciding with previous studies that suggest there is a sweet spot for the dimensionality of word vectors [64]. Even in this case, the differences between women and men are still very marked. Importantly, and as one would expect, changing the n for the salience threshold also affects the resulting sets of biased concepts as a consequence of clustering a different selection of salient words. Again, proportions between women and men are kept, but the more salient words considered ($n = 2$) the more differences with the original. With lower values of n we are including potentially less salient biases, so n can act as a zoom in/out mechanism here as needed in the community of application, with higher values allowing to focus on the most salient (strong and frequent) biases of a community. Finally, cluster repetitions τ, different window sizes (w) and epochs (e) seem to have little impact.

6.3 Validation

Finally, to validate our method, we apply it to the large, widely-studied Google News pre-trained model, and compare the results with predefined biases that had been *attested* by previous work in this model [5], [15]. The aim was to see whether our method would discover, among others, those predefined biases attested by prior work to validate that it finds relevant biases. The Google News model contains 300-dimensional vectors for 3 million words and phrases, trained on part of the US Google News dataset[9]. Previous research on this model *attested* the predefined gender biases and stereotypes that associate women with *family* and *arts*, and men with *career*, *science* and *maths* [5], [15].

We applied our method to the Google News pre-trained model, following the steps we describe in Section 4. We first estimate the salience of the words in the vocabulary of the Google News model and select the most salient words towards *women* and *men*, creating $S1$ and $S2$ with sizes 1545 and 1271, respectively. The partition with higher silhouette score, using τ= 200, clustered women-salient words in 508 clusters, and men-salient words in 552. Using WEAT to compare all women clusters with men clusters returns significant p-values for all comparison combinations, with p-values ranging between $1.5 \cdot 10^{-4}$ and 0.008, passing the Benjamini-Hochberg correction test introduced before.

Here, we focus on the validation aspects, reporting whether our method also found, among all *discovered* biases, the arbitrary biases that had been *attested* by previous works. Note that comparing both approaches is not straightforward, since *discovering* the most biased words in a model is a different task than *attesting* whether some arbitrary word vectors are strongly correlated: even when the attestation of an specific arbitrary bias results in significance, it may not be one of the strongest biased associations in the model and therefore not selected - among the other many discovered and stronger biased associations - by our approach. Our method discovers biased concepts towards two target sets, in this example *men* and *women*, but these most biased concepts do not necessarily have to be the arbitrary biases tested in previous works. An example of this situation can be seen when observing the Google News analysis in Appendix A: surprisingly no-one has previously attested biases related to *physical appearance* in the Google News model, even though these biases are stronger and more pronounced (obtain more significant p-values when using WEAT) than the reported attested biased related to *career, family, arts, science and maths* in previous work.

Therefore, to compare both approaches, we map the word sets used by previous works related to *career, family, arts, science* and

8. Note that for the sake of clarity, we only considered labels used by more than 1% of the clusters in each partition, selecting 14 labels for each model and attribute concept on average, in order to compare relatively meaningful and frequent biases in the community.

9. https://code.google.com/archive/p/word2vec/

maths to the most similar USAS semantic categories, based on an analysis of the WEAT word sets and the category descriptions provided in the USAS website[10], and then count the number of clusters tagged with them for both target sets *women* and *men*. We consider that our method discovers similar biases attested by previous research if the associations reported in [5], [15] hold among the discovered biased concepts, that is, if we find more clusters tagged with labels related to *career, maths* and *science* among *men*-biased concepts than among *women*, and vice-versa for *family* and *arts*.

The association between the previously used WEAT word sets and USAS semantic categories goes as follows. We map *career words* to *Money & commerce in industry* and *Power and organizing* USAS categories; *family words* to *Kin and People*; words related to *arts* to *Arts and Crafts*, words related to *science* to *Science and technology in general*, and *mathematics* to the corresponding USAS label *Mathematics*.

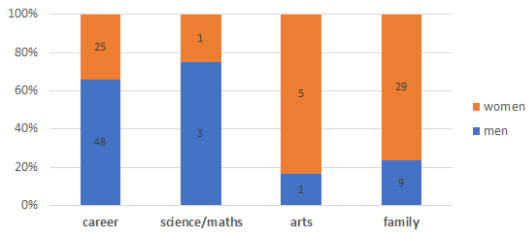

Fig. 8: Frequency of previously attested biased categories among discovered clusters in *Google News*.

Figure 8 shows that the semantic tags related to *career* are strongly biased towards men, as identified in previous works, with double the number of clusters for men than women, containing words such as 'boss' and 'magnate' (for men) and 'chairwoman', 'manageress', and 'governess' (for women). *Family*-related clusters are strongly biased towards women, with three times as many clusters for women than for men. Words include references to 'mothering', 'brides' and 'niece' (women), and also 'dad', 'patriarch' and 'nephew' (men). *Arts* is also strongly biased towards women, with 5 times more clusters for women compared than for men, including words such as 'knitting' and 'crochet' for women, and 'blacksmith' for men. Tags related to *science* and *maths* are not frequent among the set of most salient words in this model, but they are still more frequent among men-biased clusters than for women, with 3 and 1 clusters, respectively.

This analysis shows that: i) our method *discovers* relevant biases, as it discovers similar biases *attested* by previous research; ii) as detailed in Appendix A, our method discovers additional biases related to gender in Google News not reported by previous work.

7 Conclusions

In this paper, we improve and examine and validate a data-driven methodology for discovering biased concepts in linguistic corpora. Through the use of word embeddings and similarity metrics, which leverage the vocabulary used within specific communities, we are able to discover the biased concepts, in the form of clusters of words semantically related, towards different social groups when compared against each other. The resulting clusters abstract the inherent biases into more general (and stable) structures that allow a better understanding of the dispositions of these communities towards protected features. We discovered and analyzed gender and religion biases in two Reddit communities, and showed their stability. We also validated our method on the Google News dataset, confirming that it also discovers the arbitrary biases that had been attested by previous works.

Quantifying language biases through our approach has academic and societal advantages. For socio-cultural linguistics, manual identification and analysis of the subcultures' specialised vocabularies has become impractical; automatic methods that help with both the identification and understanding of community language. As a more general diagnostic, our method can help in understanding and measuring social problems and stereotypes towards certain populations in communities with more precision and clarity [19]. It can also promote discussions on the language used by particular communities. This is especially relevant considering the radicalisation of interest-based communities [18]. Our approach allows us to discover a broad categorical overview of biases, as well as a detailed analysis of these biases, as expressed in a community's own language. As such, we can trace language in cases where researchers do not know the specific linguistic forms employed in the community. This is all the more relevant given that online discourse communities are characterised by diachronic variability, with new users, topics and forms of dialect being introduced over time. As a bottom-up approach, our method can be used to monitor and account for these transformations. Finally, our method could underpin tools to help administrative bodies of web platforms to discover and trace biases in online communities to decide which do not conform to content policies.

It is finally important to note that, even though our method is automated, it is designed to be used with a human in the loop. This includes the need to consider in conjunction the more general and specific views provided by our method, and the wider context surrounding the biased concepts found, which is important to adequately interpret them. For instance, in /r/TheRedPill, the cluster *casual* contains words that carry positive sentiment score, but these words are part of a discourse about having casual relationships with women in order to conquer as many of them as possible – objectifying them in the process. This is something that becomes clear when looking at the relative frequency of semantic categories discovered, with the most frequent ones including Relationship Intimate/Sexual, Appearance and Objects, among others. These findings do not offer conclusive evidence of stereotyping, but they can be valuable assets for the work of social scientists and industry experts who are dealing with language biases in their communities. Finally, the target concepts should be defined with care, and the analysis based on them needs to be conducted with care as well. For instance, we considered gender as women vs men (as all the previous work related to gender biases in word embeddings to date did too). However, concepts may not always generalise in the same way and for every community, e.g., gender is increasingly considered fluid rather than binary [65], [66].

Code, Datasets and Tool

To facilitate follow-up work, all the code and datasets are available *here*.[11] The code repository has detailed instructions to replicate the experiments in the paper. Also, an interactive demo of our tool is *here*.[12]

10. A list of all USAS labels can be accessed at http://ucrel.lancs.ac.uk/usas/

11. https://github.com/xfold/DiscoveringAndInterpretingConceptualBiases

12. https://xfold.github.io/Web-DiscoveringAndInterpretingConceptualBiases/

ACKNOWLEDGMENTS

This research was funded by the EPSRC under grant EP/R033188/1, Discovering and Attesting Digital Discrimination (DADD) – https://dadd-project.github.io.

REFERENCES

[1] D. L. Hamilton and T. K. Trolier, "Stereotypes and stereotyping: An overview of the cognitive approach." in *Prejudice, discrimination, and racism.* San Diego, CA, US: Academic Press, 1986, pp. 127–163.

[2] S. A. Basow and S. A. Basow, "Gender : stereotypes and roles," Pacific Grove, Calif., 1992. [Online]. Available: http://catalog.hathitrust.org/api/volumes/oclc/24845451.html

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

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in Neural Information Processing Systems*, 2016, pp. 4349–4357.

[5] 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, 2017.

[6] S. Barocas, K. Crawford, A. Shapiro, and H. Wallach, "The problem with bias: from allocative to representational harms in machine learning. special interest group for computing," *Information and Society*, 2017.

[7] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[8] N. Criado and J. Such, "Digital discrimination," in *Algorithmic Regulation*. Oxford University Press, pp. 82–97.

[9] T. van Nuenen, X. Ferrer, J. Such, and M. Coté, "Transparency for whom? assessing discriminatory artificial intelligence," *Computer*, vol. 53, no. 11, pp. 36–44, 2020.

[10] X. F. Aran, T. van Nuenen, J. Such, M. Coté, and N. Criado, "Bias and discrimination in ai: a cross-disciplinary perspective," *IEEE Technology and Society*, vol. 40, no. 2, pp. 72–80, 2021.

[11] J. Such, "Privacy and autonomous systems," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 4761–4767.

[12] V. Bryson, *Feminist political theory*. Macmillan International Higher Education, 2016.

[13] B. Zhao, J. Ondrich, and J. Yinger, "Why do real estate brokers continue to discriminate? Evidence from the 2000 Housing Discrimination Study," *Journal of Urban Economics*, vol. 59, no. 3, pp. 394–419, 2006.

[14] C. van Miltenburg, "Stereotyping and bias in the flickr30k dataset," in *11th workshop on multimodal corpora: computer vision and language processing*, 2016.

[15] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *PNAS 2018*, vol. 115, no. 16, pp. E3635–E3644, 2018.

[16] A. C. Kozlowski, M. Taddy, and J. A. Evans, "The geometry of culture: Analyzing the meanings of class through word embeddings," *American Sociological Review*, vol. 84, no. 5, pp. 905–949, 2019.

[17] G. A. Fine and S. Kleinman, "Rethinking Subculture: An Interactionist Analysis," *American Journal of Sociology*, vol. 85, no. 1, pp. 1–20, aug 1979. [Online]. Available: http://www.jstor.org/stable/2778065

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

[19] R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson, "Roles for computing in social change," in *Procs. of the Conference on Fairness, Accountability, and Transparency (FAT)*, 2020, pp. 252–260.

[20] L. M. Tanczer, "Hacktivism and the male-only stereotype," *New Media & Society*, vol. 18, no. 8, pp. 1599–1615, 2016.

[21] A. Creese and A. Blackledge, "Stereotypes and chronotopes: The peasant and the cosmopolitan in narratives about migration," *Journal of Sociolinguistics*, 2019.

[22] X. F. Aran, J. Such, and N. Criado, "Attesting biases and discrimination using language semantics," in *Responsible Artificial Intelligence Agents workshop of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, 2019.

[23] X. Ferrer, T. van Nuenen, J. Such, and N. Criado, "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.

[24] A. Sutton, T. Lansdall-Welfare, and N. Cristianini, "Biased embeddings from wild data: Measuring, understanding and removing," in *International Symposium on Intelligent Data Analysis*. Springer, 2018, pp. 328–339.

[25] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Gender bias in contextualized word embeddings," in *Procs. of the Conference of the North American Chapter of the Association for Computational Linguistics)*, 2019, pp. 629–634.

[26] O. Papakyriakopoulos, S. Hegelich, J. C. M. Serrano, and F. Marco, "Bias in word embeddings," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 446–457.

[27] T. Manzini, L. Y. Chong, A. W. Black, and Y. Tsvetkov, "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings," in *Procs. of the Conf. of the North American Chapter of the Association for Comp. Linguistics*, 2019, pp. 615–621.

[28] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," in *Procs. of the Workshop on Gender Bias in Natural Language Processing*, 2019, pp. 166–172.

[29] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, "Understanding the origins of bias in word embeddings," in *International Conference on Machine Learning*, 2019, pp. 803–811.

[30] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, "Measuring individual differences in implicit cognition: the implicit association test." *J Pers Soc Psychol*, vol. 74, no. 6, p. 1464, 1998.

[31] A. K. Kiefer and D. Sekaquaptewa, "Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat," *Journal of experimental social psychology*, vol. 43, no. 5, pp. 825–832, 2007.

[32] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[33] N. Swinger, M. De-Arteaga, N. T. Heffernan IV, M. D. Leiserson, and A. T. Kalai, "What are the biases in my word embedding?" in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 305–311.

[34] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," in *Procs. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 609–614.

[35] M. H. Bodell, M. Arvidsson, and M. Magnusson, "Interpretable word embeddings via informative priors," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6324–6330.

[36] T. Farrell, O. Araque, M. Fernandez, and H. Alani, "On the use of jargon and word embeddings to explore subculture within the reddit's manosphere," in *12th ACM Conference on Web Science*, ser. WebSci '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 221–230. [Online]. Available: https://doi.org/10.1145/3394231.3397912

[37] K. M. MacQueen, E. McLellan, D. S. Metzger, S. Kegeles, R. P. Strauss, R. Scotti, L. Blanchard, and R. T. Trotter 2nd, "What is community? An evidence-based definition for participatory public health," *American journal of public health*, vol. 91, no. 12, pp. 1929–1938, dec 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11726368https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1446907/

[38] P. Badilla, F. Bravo-Marquez, and J. Pérez, "Wefe: The word embeddings fairness evaluation framework," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[39] M. Antoniak and D. Mimno, "Evaluating the Stability of Embedding-based Word Similarities," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 107–119, 2018.

[40] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[41] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[42] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Advances in neural information processing systems*, 2004, pp. 281–288.

[43] V. Garg and A. T. Kalai, "Supervising unsupervised learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 4991–5001.

[44] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[45] A. P. Diz, A. Carvajal-Rodríguez, and D. O. Skibinski, "Multiple hypothesis testing in proteomics: a strategy for experimental work," *Molecular & Cellular Proteomics*, vol. 10, no. 3, 2011.

[46] P. Rayson, D. Archer, S. Piao, and A. M. McEnery, "The ucrel semantic analysis system." in *In Procs of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, 4th Int. Conf. on Language Resources and Evaluation (LREC)*, 2004, pp. 7–12.

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

[48] S. Sharoff, B. Babych, P. Rayson, O. Mudraya, and S. Piao, "Assist: Automated semantic assistance for translators," in *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, 2006, pp. 139–142.

[49] P. Rayson, "From key words to key semantic domains," *International journal of corpus linguistics*, vol. 13, no. 4, pp. 519–549, 2008.

[50] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[51] Z. Watson, "Red Pill Men and Women, Reddit, And The Cult of Gender — Inverse," 2016. [Online]. Available: https://www.inverse.com/article/15832-red-pill-men-and-women-reddit-and-the-cult-of-gender

[52] B. A. Nosek, M. R. Banaji, and A. G. Greenwald, "Harvesting implicit group attitudes and beliefs from a demonstration web site." *Group Dynamics: Theory, Research, and Practice*, vol. 6, no. 1, p. 101, 2002.

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

[54] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of NAACL-HLT*, 2018, pp. 2227–2237.

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

[56] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

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

[58] R. Schmitz and E. Kazyak, "Masculinities in Cyberspace: An Analysis of Portrayals of Manhood in Men's Rights Activist Websites," *Social Sciences*, vol. 5, no. 2, p. 18, 2016.

[59] D. Ging, "Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere," *Men and Masculinities*, vol. 22, no. 4, pp. 638–657, 2019.

[60] J. Mountford, "Topic Modeling The Red Pill," *Social Sciences*, vol. 7, no. 3, p. 42, 2018.

[61] M. H. Ribeiro, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, S. Long, S. Greenberg, and S. Zannettou, "From Pick-Up Artists to Incels: A Data-Driven Sketch of the Manosphere," 2020. [Online]. Available: http://arxiv.org/abs/2001.07600

[62] D. Shiebler, L. Belli, J. Baxter, H. Xiong, and A. Tayal, "Fighting redundancy and model decay with embeddings," *arXiv preprint arXiv:1809.07703*, 2018.

[63] M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, no. 2, pp. 19–28, 2016.

[64] O. Melamud, D. McClosky, S. Patwardhan, and M. Bansal, "The role of context types and dimensionality in learning word embeddings," *arXiv preprint arXiv:1601.00893*, 2016.

[65] D. Tannen, "Gender and Discourse," Oxford and New York, 1994.

[66] C. Witt, *The Metaphysics of Gender*. Oxford and New York: Oxford University Press, 2012.

Xavier Ferrer Xavier Ferrer is a Research Associate in Digital Discrimination at the Department of Informatics, King's College London. Xavier obtained his PhD in Informatics in 2017 from the Artificial Intelligence Institute of the Spanish National Research Council (IIIA-CSIC) and the Universitat Autonoma de Barcelona (UAB). His research interests are at the intersection of natural language processing, machine learning, and fairness. Contact him on Twitter at @xaviferreraran, and at xavier.ferrer_aran@kcl.ac.uk.

Tom van Nuenen Tom van Nuenen is a Research Associate at the Department of Informatics, King's College London, where he investigates socio-cultural dimensions of digital discrimination. As a teacher of methodological courses that confront the history of hermeneutics with the practices of big data, he has held visiting positions at UC Berkeley, Copenhagen University, and Shandong University, among others. Tom has a special interest in the influence of online culture and datafication on the norms and epistemologies of global travel. Find him on Twitter and Medium at @tomvannuenen, and contact him at tom.van_nuenen@kcl.ac.uk.

Natalia Criado Dr Natalia Criado is Lecturer in Computer Science in the Department of Informatics at King's College London and a co-Director of the UKRI Centre for Doctoral Training in Safe and Trusted AI (www.safeandtrustedai.org). She is an expert in computational norms and normative multi-agent systems, as well as and in the application of multi-agent systems, data science and artificial intelligence to enhance cyber security and privacy. Contact her by email at natalia.criado@kcl.ac.uk.

Jose Such is Professor in the Department of Informatics at King's College London, and Director of the KCL Cybersecurity Centre. His research interests are at the intersection of Artificial Intelligence, Human-Computer Interaction and Cyber Security, with a strong focus on human-centred AI security, ethics, and privacy. He has been PI for a number of large projects funded by EPSRC, including the Discovering and Attesting Digital Discrimination (DADD) project, and the Secure AI Assistants (SAIS) project. Follow him on Twitter at @josemsuch, and contact him by email at jose.such@kcl.ac.uk.

APPENDIX A
GOOGLE NEWS ANALYSIS

Figure 9 shows the general distribution of biased concepts among all clusters for towards *women* and *men*, respectively, showing only labels most frequent than 1% of the total label frequency, and ignoring those clusters without a conceptual label for both women (273) and men (260). The general analysis of the community's biased concepts shows that some of the most common biases towards *women* in Google News are related with physical appearance (such as Clothes, Anatomy and Physiology, Personal Care, and Judgment of Appearance) and relationships. On the other hand, the most frequent men-biased labels are related with Power, Violence, Religion and sports.

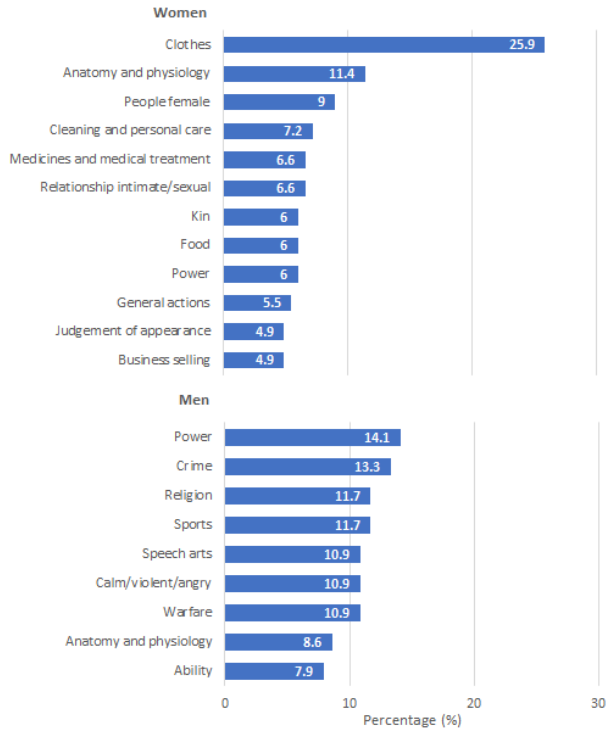

Fig. 9: Relative frequency of semantic categories for biased concepts towards *Women* (top) and *Men* (bottom) in Google News.

Figures 10 and 11 show the top 5 clusters for *women* and *men* in Google News by rank. Women-biased clusters are shown on top and men-biased clusters on the bottom, the y-axis corresponds with average salience of the cluster, size corresponds with frequency, and colour with average sentiment. Women-biased clusters are often related with physical appearance (for instance clusters 'modelling', 'lipsticks', 'beautiful', or 'cutest'), while men-biased clusters are often related with leadership, sports and strength (e.g. 'hero', 'duty', or 'wonderboy'). This picture is clear if we focus in the top most frequent biased clusters. For *women* we find words such as 'gal', 'sweetie' and 'teeny' (grouped in 'chick' cluster), 'curvy' and 'voluptuous' in the 'modelling' cluster, and words such as 'pastel', 'fuchsia', 'skanky' and 'slut' inside 'satin' and 'witch' clusters, respectively. For *men*, we find words such as 'general', 'veteran', 'pro' and 'astute' (in 'former' cluster), 'bumbling', 'weakling' and 'slacker' in 'hapless' cluster, and words related to sports and strength such as 'squad', 'pitch', 'fraternity' and 'skillfull' in the other clusters shown in Figure 11. Although not as strongly biased as in /r/TheRedPill, in Google News we find clusters with obvious negative connotations towards women such as 'witch', 'bitchy', 'gossips' and 'murderess', among the most frequent and negative women-biased rankings, respectively. On the other hand, the most negative clusters towards men are related to sports, being 'hooliganism', 'penalty', and 'disrespect' some of the most negative. Finally, the most positive rankings show that most of the *women* biased words are, again and excluding 'heroine' cluster, related to physical appearance: i.e. 'cutest', 'beautiful', 'gorgeous' and 'ravishing'. For *men* some of the most positive clusters contain words such as 'heroics', 'heroes', 'absolved' and 'jokester'.[13]

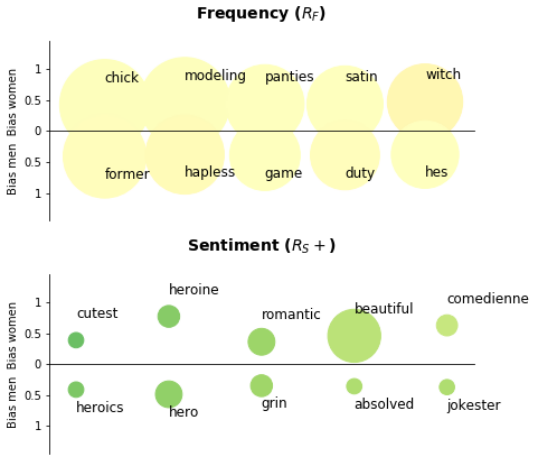

Fig. 10: Top-5 R_f and R_S+ clusters in GoogleNews.

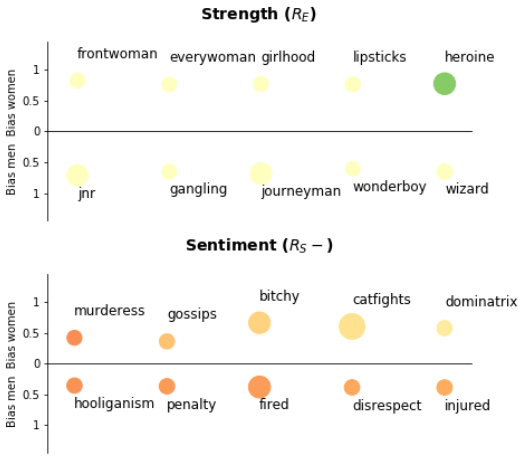

Fig. 11: Top-5 R_E and R_S- clusters in GoogleNews.

APPENDIX B
ATTRIBUTE AND OTHER SETS

/r/TheRedPill attribute concepts From [52]. *Female (women)*: sister, female, woman, girl, daughter, she, hers, her. *Male (men)*: brother, male, man, boy, son, he, his, him.
/r/Atheism attribute concepts From [15]. *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
Google News attribute and target sets From [15]. *Women*: sister, female, woman, girl, daughter, she, hers, her. *Men*: 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.

13. Note that all details of the biases discovered and biased concepts are publicly available in our Github repository, which is accessible at https://github.com/xfold/DiscoveringAndInterpretingConceptualBiases.