Comparative Evaluation of Label-Agnostic Selection Bias in Multilingual Hate Speech Datasets

Nedjma Ousidhoum, Yangqiu Song, Dit-Yan Yeung
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
nousidhoum@cse.ust.hk, yqsong@cse.ust.hk, dyyeung@cse.ust.hk

Abstract

Work on bias in hate speech typically aims to improve classification performance while relatively overlooking the quality of the data. We examine selection bias in hate speech in a language and label independent fashion. We first use topic models to discover latent semantics in eleven hate speech corpora, then, we present two bias evaluation metrics based on the semantic similarity between topics and search words frequently used to build corpora. We discuss the possibility of revising the data collection process by comparing datasets and analyzing contrastive case studies.

1 Introduction

Hate speech in social media dehumanizes minorities through direct attacks or incitement to defamation and aggression. Despite its scarcity in comparison to normal web content, Mathew et al. (2019) demonstrated that it tends to reach large audiences faster due to dense connections between users who share such content. Hence, a search based on generic hate speech keywords or controversial hashtags may result in a set of social media posts generated by a limited number of users (Arango et al., 2019). This would lead to an inherent bias in hate speech datasets similar to other tasks involving social data (Olteanu et al., 2019) as opposed to a *selection bias* (Heckman, 1977) particular to hate speech data.

Mitigation methods usually point out the classification performance and investigate how to debias the detection given false positives caused by gender group identity words such as "women" (Park et al., 2018), racial terms reclaimed by communities in certain contexts (Davidson et al., 2019), or names of groups that belong to the intersection of gender and racial terms such as "black men" (Kim et al., 2020). The various aspects of the dataset construction are less studied though it has recently been shown, by looking at historical documents, that we may somehow neglect the data collection process (Jo and Gebru, 2020). Thus, in the present work, we are interested in improving hate speech data collection with evaluation before focusing on classification performance.

We conduct a comparative study on English, French, German, Arabic, Italian, Portuguese, and Indonesian datasets using topic models, specifically Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We use multilingual word embeddings or word associations to compute the semantic similarity scores between topic words and predefined keywords and define two metrics that calculate bias in hate speech based on these measures. We use the same list of keywords reported by Ross et al. (2016) for German, Sanguinetti et al. (2018) for Italian, Ibrohim and Budi (2019) for Indonesian, Fortuna et al. (2019) for Portuguese; allow more flexibility in both English (Waseem and Hovy, 2016; Founta et al., 2018; Ousidhoum et al., 2019) and Arabic (Albadi et al., 2018; Mulki et al., 2019; Ousidhoum et al., 2019) in order to compare different datasets based on shared concepts that have been reported in their respective paper descriptions; and for French, we make use of a subset of keywords that covers most of the targets reported by Ousidhoum et al. (2019). Our first bias evaluation metric measures the average similarity between topics and the whole set of keywords, and the second one evaluates how often keywords appear in topics. We analyze our methods through different use cases which explain how we can benefit from the assessment[1].

Our main contributions consist of (1) designing bias metrics that evaluate hateful web content using topic models; (2) examining selection bias in eleven datasets; and (3) turning present hate speech

[1]Our code and data can be downloaded from https://github.com/HKUST-KnowComp/HS_Bias_Eval

corpora into an insightful resource that may help us balance training data and reduce bias in the future.

2 Related Work

Hate speech labeling schemes depend on the general purpose of the dataset. The annotations may include hateful vs. non hateful (Basile et al., 2019), racist, sexist, and none (Waseem and Hovy, 2016); as well as discriminating target attributes (ElSherief, Mai and Nilizadeh, Shirin and Nguyen, Dana and Vigna, Giovanni, and Belding, Elizabeth, 2018), the degree of intensity (Sanguinetti et al., 2018), and the annotator's sentiment towards the tweets (Ousidhoum et al., 2019). Besides English (Basile et al., 2019; Waseem and Hovy, 2016; Davidson et al., 2017; ElSherief, Mai and Nilizadeh, Shirin and Nguyen, Dana and Vigna, Giovanni, and Belding, Elizabeth, 2018; Founta et al., 2018; Qian et al., 2018), we notice a growing interest in the study of hate speech in other languages, such as Portuguese (Fortuna et al., 2019), Italian (Sanguinetti et al., 2018), German (Ross et al., 2016), Indonesian (Ibrohim and Budi, 2019), French (Ousidhoum et al., 2019), Dutch (Hee et al., 2015), and Arabic (Albadi et al., 2018; Mulki et al., 2019; Ousidhoum et al., 2019). Challenging questions being tackled in this area involve the way abusive language spreads online (Mathew et al., 2019), fast changing topics during data collection (Liu et al., 2019), user bias in publicly available datasets (Arango et al., 2019), bias in hate speech classification and different methods to reduce it (Park et al., 2018; Davidson et al., 2019; Kennedy et al., 2020).

Bias in social data is broad and addresses a wide range of issues (Olteanu et al., 2019; Papakyriakopoulos et al., 2020). Shah et al. (2020) present a framework to predict the origin of different types of bias including label bias (Sap et al., 2019), selection bias (Garimella et al., 2019), model overamplification (Zhao et al., 2017), and semantic bias (Garg et al., 2018). Existing work deals with bias through the construction of large datasets and the definition of social frames (Sap et al., 2020), the investigation of how current NLP models might be non-inclusive of marginalized groups such as people with disabilities (Hutchinson et al., 2020), mitigation (Dixon et al., 2018; Sun et al., 2019), or better data splits (Gorman and Bedrick, 2019). However, Blodgett et al. (2020) report a missing normative process to inspect the initial reasons behind bias in NLP without the main focus being on the performance which is why we choose to investigate the data collection process in the first place.

In order to operationalize the evaluation of selection bias, we use topic models to capture latent semantics. Regularly used topic modeling techniques such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have proven their efficiency to handle several NLP applications such as data exploration (Rodriguez and Storer, 2020), Twitter hashtag recommendation (Godin et al., 2013), authorship attribution (Seroussi et al., 2014), and text categorization (Zhou et al., 2009).

In order to evaluate the consistency of the generated topics, Newman et al. (2010) used crowdsourcing and semantic similarity metrics, essentially based on Pointwise Mutual Information (PMI), to assess the coherence; Mimno et al. (2011) estimated coherence scores using conditional log-probability instead of PMI; Lau et al. (2014) enhanced this formulation based on normalized PMI (NPMI); and Lau and Baldwin (2016) investigated the effect of cardinality on topic generation. Similarly, we use topics and semantic similarity metrics to determine the quality of hate speech datasets, and test on corpora that vary in language, size, and general collection purposes for the sake of examining bias up to different facets.

3 Bias Estimation Method

The construction of toxic language and hate speech corpora is commonly conducted based on keywords and/or hashtags. However, the lack of an unequivocal definition of hate speech, the use of slurs in friendly conversations as opposed to sarcasm and metaphors in elusive hate speech (Malmasi and Zampieri, 2018), and the data collection timeline (Liu et al., 2019) contribute to the complexity and imbalance of the available datasets. Therefore, training hate speech classifiers easily produces false positives when tested on posts that contain controversial or search-related identity words (Park et al., 2018; Sap et al., 2019; Davidson et al., 2019; Kim et al., 2020).

To claim whether a dataset is rather robust to keyword-based selection or not, we present two label-agnostic metrics to evaluate bias using topic models. First, we generate topics using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Then, we compare topics to predefined sets of key-

DATASET	KEYWORDS
Ousidhoum et al. (2019) Waseem and Hovy (2016) Founta et al. (2018)	ni**er, invasion, attack
Ousidhoum et al. (2019)	**FR** migrant, sale, m*ng*l **EN** *migrant, filthy, mong****d*
Albadi et al. (2018) Ousidhoum et al. (2019) Mulki et al. (2019)	**AR** مراة ، البعير، خنزير **EN** *woman, camels, pig*
Ibrohim and Budi (2019)	**ID** idiot, kafir, bego **EN** *idiot, infidel, stupid*
Sanguinetti et al. (2018)	**IT** invasione, basta, comunist **EN***invasion, enough, communist*
Fortuna et al. (2019)	**PT** discurso, odio, sapatao **EN** *speech, hate, romp*
Ross et al. (2016)	**DE** pack, aslyanten, rapefugees **EN** *pack, asylum seekers, rapefugees*

Table 1: Examples of keywords present in the predefined lists along with their English translations. The keywords include terms frequently associated with controversies such as *comunist* in Italian, slurs such as *m*ng*l* in French, insults such as خنزير in Arabic, and hashtags such as *rapefugees* in German.

DATASET	TOPIC WORDS
Founta et al. (2018) Ousidhoum et al. (2019) Waseem and Hovy (2016)	f***ing, like, know ret***ed, sh*t**le, c*** sexist, andre, like
Ousidhoum et al. (2019)	**FR** m*ng*l, gauchiste, sale **EN** *mon*y, leftist, filthy*
Albadi et al. (2018)	**AR** الشيعة ، اليهود ، المسيحية **EN** *Shia, Jewish, Christanity*
Mulki et al. (2019)	**AR** جبران، باسيل ، الله **EN** *Gebran, Bassil, God*
Ousidhoum et al. (2019)	**AR** البعير، الحريم ، خنازير **EN** *women (slang), camels, pigs*
Fortuna et al. (2019)	**PT** mulher, refugiados, contra **EN** *woman, refugees, against*
Sanguinetti et al. (2018)	**IT** migranti, roma, italia **EN** *migrants, Roma, Italy*
Ibrohim and Budi (2019)	**ID** user, orang, c*b*ng **EN** *user, person, t*dp*le*
Ross et al. (2016)	**DE** rapefugees, asylanten, merkel **EN** *rapefugees, asylum seekers, merkel*

Table 2: Examples of topics of length 3 generated by LDA. Non-English topics are presented along with their English translations.

words using a semantic similarity measure. We test our methods on different numbers of topics and topic words.

3.1 Predefined Keywords

In contrast to Waseem (2016), who legitimately questions the labeling process by comparing amateur and professional annotations, we investigate how we could improve the collection without taking the annotations into account. In other terms, how the data selection contributes to the propagation of bias and therefore, false positives during first, the annotation step, then the classification.

We define $\mathbf{B_1}$ and $\mathbf{B_2}$ assess how the obtained social media posts semantically relate to predefined keywords. The bias metric $\mathbf{B_1}$ measures this relatedness on average, while $\mathbf{B_2}$ evaluates how likely topics are to contain keywords. We use predefined sets of keywords that can be found in the hate speech resource paper descriptions (Waseem and Hovy, 2016; Ross et al., 2016; Sanguinetti et al., 2018; Founta et al., 2018; Albadi et al., 2018; Fortuna et al., 2019; Mulki et al., 2019), appeared on reported websites[2], or seen along with the corpus (Ibrohim and Budi, 2019; Ousidhoum et al., 2019).

Table 1 shows examples of keywords utilized to gather toxic posts. The list of keywords provided by Ibrohim and Budi (2019), which contains 126 words, is the largest we experiment with. The Portuguese, Italian, and German lists are originally small since they focus on particular target groups[3], whereas the remaining lists have been reduced slightly to meet the objectives presented in the descriptions of all the corpora we used.

3.2 Topic Models

Table 2 shows examples of topics that were generated from the chosen datasets. Although Founta et al. (2018) report collecting data based on controversial hashtags and a large dictionary of slurs, Waseem and Hovy (2016) on other hashtags, and Ousidhoum et al. (2019) on a different set of keywords, we can initially notice a recurring term in two English topics, and potentially more if we generate larger topics.

Moreover, Ousidhoum et al. (2019)'s Arabic dataset contains the word *pigs* used to insult people, a slang word, and the word *camels* as a part of a demeaning expression that means *"camels urine drinkers"* which is usually used to humiliate people from the Arabian Peninsula. The three words exist in the predefined list of keywords similarly to all French, Portuguese, Italian and most German and Idonesian topic words.

[2]Such as the HateBase`https://hatebase.org/`.

[3]The target groups are women, immigrants, and refugees.

Italian, German and Portuguese topics are composed of words related to immigrants and refugees as they correspond to the main targets of these datasets. The French topic also contains the name of a political ideology typically associated with more liberal immigration policies.

Other than slurs, named entities can be observed in Waseem and Hovy (2016)'s topic, which includes the name of a person who participated in an Australian TV show that was discussed in the tweets [4]; the German topic includes the name of the German Chancellor *Merkel* since she was repeatedly mentioned in tweets about the refugee crisis (Ross et al., 2016); Mulki et al. (2019)'s topic contains the name of the Lebanese political figure *Gebran Bassil* since they collected their dataset based on Twitter accounts of Syrian and Lebanese political figures; as well as names of religious groups in Albadi et al. (2018)'s topic in conformity with their collection strategy based on names of sects.

Despite their short length, the example topics can provide us with a general idea about the type of bias present in different datasets. For instance, topics generated from datasets in languages which are mainly spoken in Europe and the USA commonly target immigrants and refugees, in contrast to Arabic and Indonesian topics which focus on other cultural, social, and religious issues. Overall, all topics show a degree of potentially quantifiable relatedness to some predefined key concepts.

3.3 Bias Metrics

Mimno et al. (2011), Lau et al. (2014), and Röder et al. (2015) evaluate the quality of topics through coherence metrics that use Pointwise Mutual Information (PMI) and other similarity measures. Similarly, we would like to assess topic bias in hate speech based on the semantic similarity between high scoring words in each topic and the set of search keywords used to collect data.

Given a set of topics $\mathbf{T}=\{\mathbf{t_1},\ldots,\mathbf{t_{|T|}}\}$ generated by LDA, with each topic $\mathbf{t_i}=\{\mathbf{w_1},\ldots,\mathbf{w_n}\}$ composed of n words, and a predefined list of keywords $\mathbf{w'}$ of size m such as $\mathbf{w'}=\{\mathbf{w'_1},\ldots,\mathbf{w'_m}\}$, we define the two bias functions $\mathbf{B_1}$ and $\mathbf{B_2}$ based on $\mathbf{Sim_1}$ and $\mathbf{Sim_2}$, respectively.

$\mathbf{Sim_1}$ measures the similarity between two words $\mathbf{w_j} \in \mathbf{t_i}$ and $\mathbf{w'_k} \in \mathbf{w'}$ for $\mathbf{t_i} \in \mathbf{T}$, with $0 < i \leq |\mathbf{T}|$, such as:

$$\mathbf{Sim_1}(\mathbf{t_i}, \mathbf{w'}) = \frac{1}{n}\frac{1}{m}\sum_{j=1}^{n}\sum_{k=1}^{m}\mathbf{Sim}(\mathbf{w_j}, \mathbf{w'_k}) \quad (1)$$

$\mathbf{B_1}$ computes the mean similarity between each $\mathbf{w_j} \in \mathbf{t_i}$ and $\mathbf{w'_k} \in \mathbf{w'}$, then the mean given all generated topics, such as:

$$\mathbf{B_1}(\mathbf{T}, \mathbf{w'}) = \frac{1}{|\mathbf{T}|}\sum_{i=1}^{|\mathbf{T}|}\mathbf{Sim_1}(\mathbf{t_i}, \mathbf{w'}) \quad (2)$$

$\mathbf{Sim_2}$ measures the maximum similarity of each word $\mathbf{w_j} \in \mathbf{t_i}$ and keyword $\mathbf{w'_k} \in \mathbf{w'}$, w_j, such as $\forall \mathbf{w_j} \in \mathbf{t_i}$ and $\forall \mathbf{w'_k} \in \mathbf{w'}$ with $0 < j \leq n$ and $0 < k \leq m$:

$$\mathbf{Sim_2}(\mathbf{t_i}, \mathbf{w'}) = \max \mathbf{Sim}(\mathbf{w_j}, \mathbf{w'_k}) \quad (3)$$

Then, we calculate $\mathbf{B_2}$ similarly to $\mathbf{B_1}$:

$$\mathbf{B_2}(\mathbf{T}, \mathbf{w'}) = \frac{1}{|\mathbf{T}|}\sum_{i=1}^{|\mathbf{T}|}\mathbf{Sim_2}(\mathbf{t_i}, \mathbf{w'}) \quad (4)$$

Both $\mathbf{B_1}$ and $\mathbf{B_2}$ scores aim to capture how the word distribution of a given dataset can lead to false positives. $\mathbf{B_1}$ evaluates how the whole set of keywords $\mathbf{w'}$ semantically relates to the whole set of topics $\mathbf{T}$ by measuring their relatedness to each topic word $\mathbf{w_j} \in \mathbf{t_i}$, then to each topic $\mathbf{t_i} \in \mathbf{T}$. Whereas $\mathbf{B_2}$ verifies whether each topic word $\mathbf{w_j} \in \mathbf{t_i}$ is similar or identical to a keyword $\mathbf{w'_k} \in \mathbf{w'}$. In summary, $\mathbf{B_1}$ determines the average stability of topics given keywords, and $\mathbf{B_2}$ how regularly keywords appear in topics.

4 Results

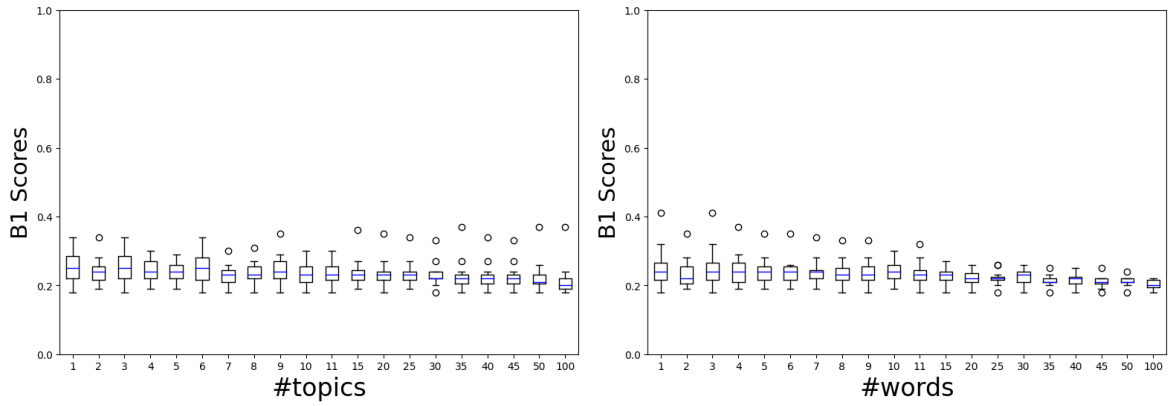
In this section, we demonstrate the impact of our evaluation metrics applied to various datasets and using different similarity measures.

4.1 Experimental Settings

The preprocessing steps we apply to all the datasets consist of (1) the anonymization of the tweets by changing *@mentions* to *@user*, then deleting *@user*s, and (2) the use of NLTK[5] to skip stopwords. Then, we run the Gensim (Řehůřek and Sojka, 2010) implementation of LDA (Blei et al., 2003) to generate topics. We vary the number of topics and words within the range [2,100] to take

[4] Waseem and Hovy (2016) report collecting tweets about *My Kitchen Rules (mkr)*.

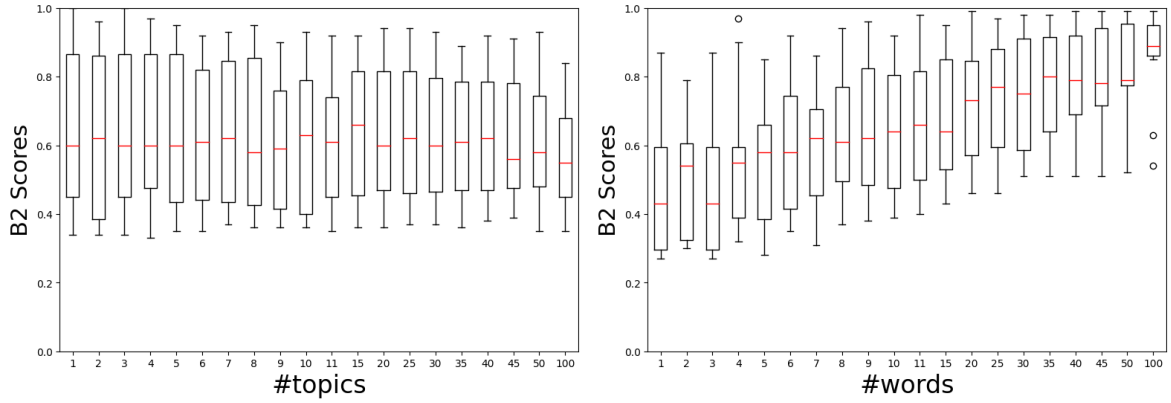
[5] https://www.nltk.org/



(a) $\mathbf{B_1}$ variations per number of topics.

(b) $\mathbf{B_1}$ variations per number of words.

Figure 1: Average $\mathbf{B_1}$ scores based on topic and word numbers in the interval [2, 100]. We fix the number of topics to 8 when we alter the number of words and similarly, we fix the number of words to 8 when we change the number of topics. We use the **multilingual Babylon embeddings** to compute the semantic similarity between words.



(a) $\mathbf{B_2}$ variations per number of topics.

(b) $\mathbf{B_2}$ variations per number of words.

Figure 2: Average $\mathbf{B_2}$ scores based on topic and word numbers in the interval [2, 100]. We fix the number of topics to 8 when we alter the number of words and similarly, we fix the number of words to 8 when we change the number of topics. We use the **multilingual Babylon embeddings** to compute the semantic similarity between words.

the inherent variability of topic models into account.

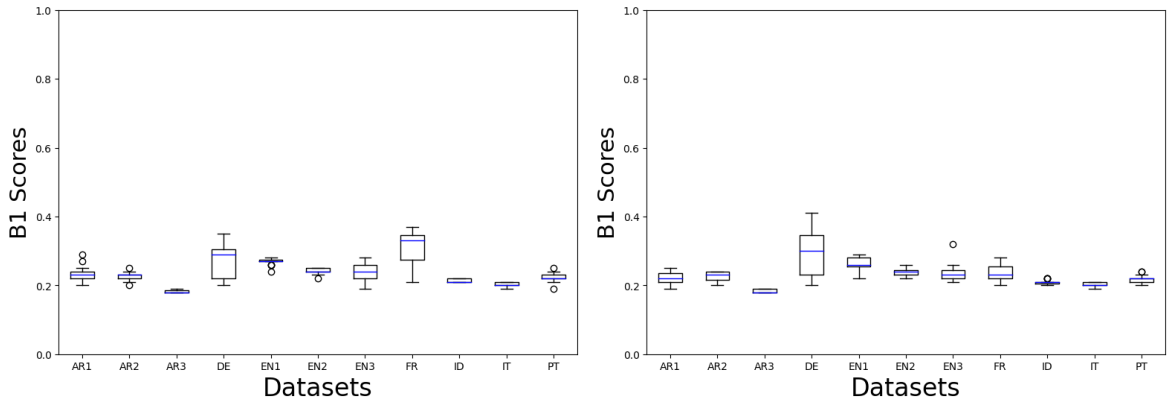
In the general cases presented in Figures 1, 2, 3, and 4, we fix the number of topics to be equal to 8 when we alter the number of topic words and likewise, we fix the number of topic words to be equal to 8 when we experiment with different numbers of topics. We define the semantic similarity measure $\mathbf{Sim}$ between each topic word and keyword to be the cosine similarity between their embedding vectors in the space of the multilingual pretrained Babylon embeddings (Smith et al., 2017) with respect to each of the seven languages we examine.

4.2 Robustness Towards The Variability of Topic Models

Figures 1 and 2 show the average $\mathbf{B_1}$ and $\mathbf{B_2}$ score variations given all the datasets. The scores are given numbers of topics and topic words within the range [2,100], respectively.

Despite $\mathbf{B_1}$ scores being similar on average, we notice that the larger the number of topics, the more outliers we observe. In parallel, the smaller the number of words, the more outliers we see. This is due to possible randomness when large topics are generated.

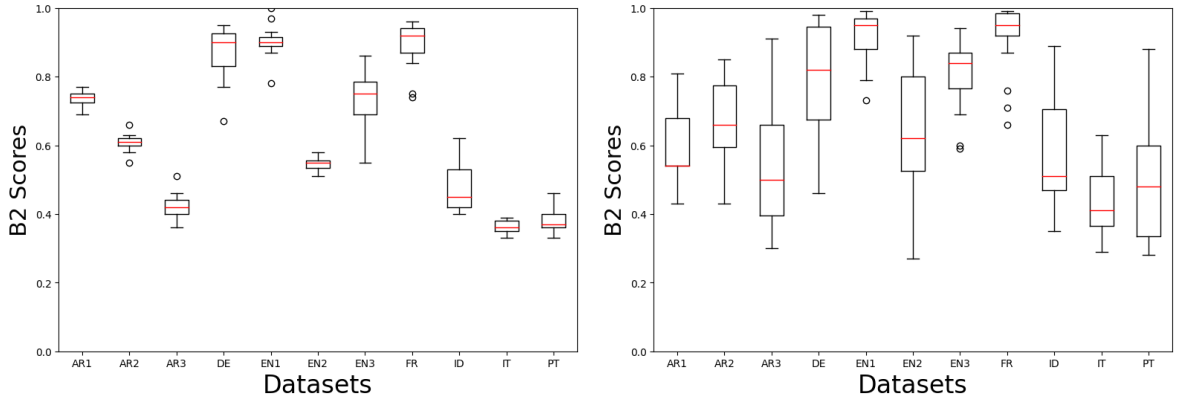
On the other hand, $\mathbf{B_2}$ scores are larger on average due to the high probability of keywords appearing in topics regardless of the dataset. This naturally translates to $\mathbf{B_2}$ showing more stability



(a) $\mathbf{B_1}$ scores per number of topics for different datasets.

(b) $\mathbf{B_1}$ scores per number of words for different datasets.

Figure 3: $\mathbf{B_1}$ score variations for different datasets. The numbers of topics and words in topics are in the range [2, 100]. We use **multilingual Babylon embeddings** to compute the semantic similarity between words. EN1,EN2,EN3 refer to Ousidhoum et al. (2019); Waseem and Hovy (2016); Founta et al. (2018); and AR1, AR2, AR3 to Ousidhoum et al. (2019); Albadi et al. (2018); Mulki et al. (2019) respectively.



(a) $\mathbf{B_2}$ scores per number of topics.

(b) $\mathbf{B_2}$ scores per number of topics.

Figure 4: $\mathbf{B_2}$ score variations for different datasets. The numbers of topics and words in topics are in the range [2, 100]. We use **multilingual Babylon embeddings** to compute the semantic similarity between words. EN1,EN2,EN3 refer to Ousidhoum et al. (2019); Waseem and Hovy (2016); Founta et al. (2018); and AR1, AR2, AR3 to Ousidhoum et al. (2019); Albadi et al. (2018); Mulki et al. (2019) respectively.

regarding the change in topic numbers in comparison to topic words.

4.3 Robustness of Keyword-based Selection

Figure 3 illustrates the variations of each dataset given the numbers of topics and topic words within the interval [2,100], respectively. In general, changes in $\mathbf{B_1}$ scores are small, such as the largest difference we observe is in the German dataset (Ross et al., 2016). In German, we reach the maximum 0.41 when the number of words in each topic equals 2, and the minimum when it equals 100. On the other hand, we observe the most noticeable changes when we vary the number of topics in French (Ousidhoum et al., 2019) such that $\mathbf{B_1} =$ 0.34 when $|\mathbf{T}| = 2$ versus 0.21 when $|\mathbf{T}| = 7$ and back to 0.37 when $|\mathbf{T}| = 100$.

However, we remark overall cohesion despite the change in topic numbers especially in the case of Italian and Portuguese caused by the limited numbers of search keywords, that equal 5 and 7 respectively.

Moreover, the account-based dataset of (Mulki et al., 2019), referred to as AR3 in Figures 3 and 4 shows more robustness towards keywords. Nevertheless, such a collection strategy may generate a linguistic bias that goes with the same stylistic features used by the targeted accounts, similarly to Waseem and Hovy (2016)'s user bias reported by Arango et al. (2019).

	Dataset	Orig	Rel	Gen
En	Founta et al. (2018)	0.94	0.80	0.80
	Ousidhoum et al. (2019)	0.92	0.79	0.96
	Waseem and Hovy (2016)	0.95	0.82	0.82
ar	Albadi et al. (2018)	0.70	0.72	0.75
	Mulki et al. (2019)	0.64	0.66	0.69
	Ousidhoum et al. (2019)	0.66	0.67	0.72

Table 3: $\mathbf{B_1}$ scores based on trained **hate speech embeddings** for 10 topics. We have manually clustered the keywords released by Ousidhoum et al. (2019) based on discriminating target attributes. For instance, the word *ni**er* belongs to the origin (Orig) category, *raghead* to religion (Rel), and *c**t* to gender (Gen). For normalization purposes, we skipped disability since we did not find Arabic keywords that target people with disabilities.

4.4 Hate Speech Embeddings

Besides using multilingual Babylon embeddings, we train hate speech embeddings with Word2Vec (Mikolov et al., 2013) to examine whether this can help us tackle the problem of out-of-the-vocabulary words caused by slang, slurs, named entities, and ambiguity.

Since we test on single French, German, Italian, Indonesian, and Portuguese datasets, we do not train embeddings on these languages due to the lack of data diversity. In contrast, we train English hate speech embeddings on Waseem and Hovy (2016), Founta et al. (2018)[6], the SEMEVAL data (Basile et al., 2019), and Ousidhoum et al. (2019)'s datasets. We train Arabic embeddings in the same way using Albadi et al. (2018)'s sectarian dataset, (Mulki et al., 2019)'s Levantine Arabic dataset, and Ousidhoum et al. (2019)'s heterogeneous dataset. The size of the data is relatively small but the different datasets are composed of tweets that have been collected for different purposes within more than one year apart.

We test on window sizes of 3, 5, 10, 15, and 50, embedding sizes of 50, 100, 200, and 300, and we manually classify keywords released by Ousidhoum et al. (2019) based on discriminating target attributes to analyze the metric $\mathbf{B_1}$.

The $\mathbf{B_1}$ scores reported in Table 3 are larger than the ones reported in Figures 1 and 3 resulting from the difference between the size of the embedding space of Babylon and hate speech embeddings. Our embeddings are trained on a limited amount of data but, we can still notice slight differences in the scores. Interestingly, $\mathbf{B_1}$ scores reveal potentially overlooked targets as in Albadi et al. (2018)'s sectarian dataset that is supposed to target people based on their religious affiliations, yet its $\mathbf{B_1}$ scores given all discriminating attributes are comparable.

[6]We use Tweepy http://docs.tweepy.org/en/latest/api.html to retrieve tweets that have not been deleted.

4.5 General versus Corpus-Specific Lists of Keywords

We consider two examples in the following use case: (1) Waseem and Hovy (2016) who report building their dataset based on hashtags such as *mkr, victim card, and race card*, and (2) Albadi et al. (2018) who report building their sectarian dataset based on religious group names such as *Judaism, Islam, Shia, Sunni and Christianity*. The initial list of predefined keywords such as the ones we have shown in Table 1 carries additional words in English and Arabic. Therefore, for these two datasets, we have measured bias using two predefined lists of keywords: the initial list and one that is specific to the dataset in question.

The scores given the general set of keywords are reported in Figures 3 and 4, such as AR2 refers to Albadi et al. (2018) and EN2 to Waseem and Hovy (2016). The $\mathbf{B_1}$ and $\mathbf{B_2}$ scores given corpus-specific lists of keywords are either the same or ± 0.01 the reported scores. We observed a maximum difference of 0.03, which is why reporting these scores would have been repetitive.

In conclusion, this is a symptom of high similarity in present English and Arabic hate speech datasets despite their seemingly different collection strategies and timelines.

4.6 WordNet and Targeted Hate Bias

In addition to word embeddings, we test our evaluation metrics on WordNet (Fellbaum, 1998)'s **WUP** (Wu and Palmer, 1994) similarity. **WUP** evaluates the relatedness of two synsets, or word senses, c_1 and c_2, based on hypernym relations. Synsets with short path distances are more related than those with longer ones. Wu and Palmer (1994) scale the depth of the two synset nodes by the depth of their Least Common Subsumer (LCS) or the most specific concept that is an ancestor of c_1 and c_2 (Newman et al., 2010).

In this use case, we aim to present a prospective label bias extension of our metrics by testing $\mathbf{B_1}$ on toxic tweets only. Consequently, we consider tweets that were not annotated normal or

DATASET	ORIG	REL	GEN
Founta et al. (2018)	0.27	0.27	0.26
Ousidhoum et al. (2019)	0.33	0.28	0.35
Waseem and Hovy (2016)	0.27	0.26	0.27

Table 4: $\mathbf{B_1}$ scores for English hate speech datasets using **WordNet** given 10 topics and keywords clustered based on origin (ORIG), religion (REL), and gender (GEN). The scores are reported for tweets that have not been labeled *non-hateful* or *normal*. Although we initially attempted to study the differences of pretrained word embeddings and word associations, we found that many $(\mathbf{w_j}, \mathbf{w'_k})$ pairs involve out-of-the-vocabulary words. In such cases $(\mathbf{w_j}, \mathbf{w'_k})$ would have a WordNet Similarity score $\mathbf{WUP} = 0$ which is why the scores are in the range [0.25, 0.35].

non-hateful. We question the present annotation schemes by computing $\mathbf{B_1}$ with $\mathbf{Sim{=}WUP}$.

Waseem and Hovy (2016), Founta et al. (2018) and Ousidhoum et al. (2019) report using different keywords and hashtags to collect tweets. However, the scores shown in Table 4 indicate that the datasets might carry similar meanings, specifically because $\mathbf{WUP}$ relies on hypernymy rather than common vocabulary use. The comparison of $\mathbf{B_1}$ scores given target-specific keywords also implies that the annotations could be non-precise. We may therefore consider fine-grained labeling schemes in which we explicitly involve race, disability, or religious affiliation as target attributes, rather than general labels such as *racist* or *hateful*.

4.7 Case Study

Figures 5(a) and 5(b) show bias scores generated for the German dataset (Ross et al., 2016) which contains 469 tweets collected based on 10 keywords related to the refugee crisis in Germany. We notice that $\mathbf{B_1}$ scores fluctuate in the beginning, reach a threshold, then get lower when the number of topics increases. $\mathbf{B_1}$ remains stable within different numbers of words as opposed to $\mathbf{B_2}$ scores that increase when more topic words are generated since eventually, all topics would include at least one keyword.

On the other hand, Figures 5(c) and 5(d) show bias scores generated for the Indonesian dataset (Ibrohim and Budi, 2019) which contains more than 13,000 tweets collected based on a heterogeneous set of 126 keywords. In such settings, $\mathbf{B_1}$ is almost constant for both the number of topics and topic words, contrary to $\mathbf{B_2}$ sores that arise when many topics are generated since new topics would include words that did not appear in the previously generated ones.

#TOPICS					
	$\mathbf{w'_{Sim}}$	#TWEETS	$\lvert\mathbf{w'}\rvert$	VOCAB	TWEET
$\mathbf{B_1}$	0.08	0.06	0.22	0.18	-0.03
$\mathbf{B_2}$	0.25	0.01	0.12	0.07	-0.14
#WORDS					
	$\mathbf{w'_{Sim}}$	#TWEETS	$\lvert\mathbf{w'}\rvert$	VOCAB	TWEET
$\mathbf{B_1}$	0.12	-0.08	0.23	0.20	-0.02
$\mathbf{B_2}$	-0.36	-0.19	0.10	-0.09	-0.04

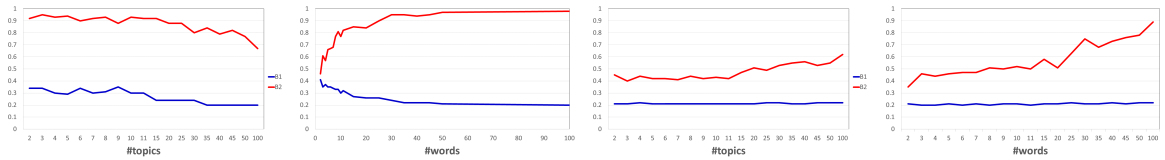
Table 5: Given the average $\mathbf{B_1}$ and $\mathbf{B_2}$ scores generated for each dataset, based on topics (**#TOPICS**) and topic words (**#WORDS**) in the interval [2,100], respectively, we compute Spearman's correlation scores between $\mathbf{B_1}$ and $\mathbf{B_2}$ and (1) the number of keywords $\lvert\mathbf{w'}\rvert$ and average cosine similarity between keywords $\mathbf{w'_{Sim}}$ given the language of the dataset; in addition to (2) the number of collected tweets **#TWEETS**, their average size **TWEET**, and size of vocabulary **VOCAB** in each dataset.

5 Discussion

We consider our bias evaluation metrics to be label-agnostic and tested this claim in the different use cases we presented in section 4. Table 5 reports the Spearman's correlation scores between the properties of each dataset and its average $\mathbf{B_1}$ and $\mathbf{B_2}$ scores given different numbers of topics and topic words. The correlation scores show that, on average, our metrics do not depend on summary statistics either. We observe low correlation scores between the different features and $\mathbf{B_1}$ scores. $\mathbf{B_1}$ correlates the best with the number of keywords and the vocabulary size whereas $\mathbf{B_2}$ correlates the best with the average cosine similarity between keywords.

Although our bias metrics do not take annotations into account, we notice a global trend of over-generalizing labels as presented in Section 4.6. Despite the fact that this is partly due to the absence of a formal definition of hate speech, we do believe that there could be a general framework which specifies several aspects that must be annotated.

Moreover, we notice recurring topics in many languages, such as those centered around immigrants and refugees which may later lead to false positives during the classification and hurt the detection performance. Hence, we believe that our evaluation metrics can help us recognize complementary biases in various datasets, facilitate trans-



(a) Variations of $\mathbf{B_1}$ and $\mathbf{B_2}$ scores given #topics in the German dataset.

(b) Variations of $\mathbf{B_1}$ and $\mathbf{B_2}$ scores given #words in the German dataset.

(c) Variations of $\mathbf{B_1}$ and $\mathbf{B_2}$ scores given #topics in the Indonesian dataset.

(d) Variations of $\mathbf{B_1}$ and $\mathbf{B_2}$ scores given #words in the Indonesian dataset.

Figure 5: Variations of $\mathbf{B_1}$ (in blue) and $\mathbf{B_2}$ (in red) scores on the German and Indonesian datasets.

fer learning, as well as enable the enhancement of the quality of the data during collection by performing an evaluation step at the end of each search round.

6 Conclusion

We proposed two label-agnostic metrics to evaluate bias in eleven hate speech datasets that differ in language, size, and content. The results reveal potential similarities across available hate speech datasets which may hurt the classification performance.

As unpreventable as selection bias in social data can be, we believe there is a way to mitigate it by incorporating evaluation as a step which directs the construction of a new dataset or when combining existing corpora.

Our metrics are extensible to other forms of bias such as user, label, and semantic biases, and could be adapted in cross-lingual contexts using different similarity measures.

Acknowledgements

We thank anonymous reviewers for their insightful comments. This work was supported by the Early Career Scheme (ECS, No. 26206717), General Research Fund (GRF, No. 16211520), Research Impact Fund (RIF, No. R6020-19), and Theme-based Research Scheme Project (T31-604/18-N) from the Research Grants Council (RGC) of Hong Kong.

References

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *Proceedings of ASONAM*, pages 69–76. IEEE Computer Society.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of SIGIR*, pages 45—54.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*, pages 512–515.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AIES*, pages 67—73.

ElSherief, Mai and Nilizadeh, Shirin and Nguyen, Dana and Vigna, Giovanni, and Belding, Elizabeth. 2018. Peer to peer hate: Hate instigators and their targets. In *Proceedings of ICWSM*, pages 42–51.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online (ALW3)*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael

Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of ICWSM*, pages 491–500.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of ACL*, pages 3493–3498.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of WWW '13 Companion*, pages 593—596.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of ACL*, pages 2786–2791.

James J Heckman. 1977. Sample selection bias as a specification error (with an application to the estimation of labor supply functions). Working Paper 172, National Bureau of Economic Research.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of RANLP*, pages 672–680.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of ACL*, pages 5491–5501.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306—-316.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of ACL*, pages 5435–5442.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*.

Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of NAACL*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of EACL*, pages 530–539.

Anqi Liu, Maya Srikanth, Nicholas Adams-Cohen, R Michael Alvarez, and Anima Anandkumar. 2019. Finding social media trolls: Dynamic keyword selection methods for rapidly-evolving online debates. *arXiv preprint arXiv:1911.05332*.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *J. Exp. Theor. Artif. Intell.*, 30(2).

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of WebSci '19*, pages 173–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Proceedings of Advances in Neural Information Processing Systems 26*, pages 3111–3119.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of EMNLP*, pages 262—-272.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of NAACL-HLT*, pages 100–108.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP-IJCNLP*, pages 4675–4684.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 446—457.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of EMNLP*, pages 2799–2804.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical cvae for fine-grained hate speech classification. In *Proceedings of EMNLP*, pages 3550—3559.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, pages 399—-408.

Maria Y. Rodriguez and Heather Storer. 2020. A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data*. *Journal of Technology in Human Services*, 38(1):54–86.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of LREC*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of ACL*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceeedings of ACL*, pages 5477–5490.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Comput. Linguist.*, 40(2).

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of ACL*, pages 5248–5264.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of ACL*, pages 1630–1640.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of ACL*, pages 133—-138.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of EMNLP*, pages 2979–2989.

Shibin Zhou, Kan Li, and Yushu Liu. 2009. Text categorization based on topic model. *Int. J. Comput. Intell. Syst.*, 2.