EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D6.1: Recommendations on avoiding gender and other biases (T6.4)

Executive summary
Gender biases and other distortions are a key concern in text analytics and content creation systems. In this report, we review the literature on biases of linguistic models, discuss biases in journalism, and describe technical notions of bias. Building upon lessons from the literature, we conclude with recommendations for detecting and avoiding biases in the context of news.

Partner in charge: UH

Project co-funded by the European Commission within Horizon 2020 Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	-
RE	Restricted to a group specified by the Consortium (including the Commission Services)	-
CO	Confidential, only for members of the Consortium (including the Commission Services)	

This project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D6.1**
Deliverable full title:	**Recommendations on avoiding gender and other biases**
Deliverable short title:	**Avoiding Biases**
Document identifier:	**EMBEDDIA-D61-AvoidingBiases-T64-submitted**
Lead partner short name:	**UH**
Report version:	**final**
Report submission date:	**30/04/2019**
Dissemination level:	**PU**
Nature:	**R**
Lead author(s):	**Michael Mathioudakis (UH)**
Co-author(s):	**Carl-Gustav Linden (UH), Senja Pollak (JSI), Matthew Purver (QMUL), Anka Supej (JSI)**
Status:	**_ draft, _ final, x submitted (tick one)**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
04/03/2019	v1.0	Michael Mathioudakis (UH)	Produced first draft (introduction, research on linguistic biases, measuring bias, recommendations on avoiding biases).
26/03/2019	v1.1	Senja Pollak (JSI)	Added content related to research on linguistic biases.
27/03/2019	v1.2	Matthew Purver (QMUL)	Added content on research on linguistic biases in embeddings; revised entire document.
28/03/2019	v1.3	Carl-Gustav Linden (UH)	Added content on gender and other biases in journalism.
03/04/2019	v1.4	Senja Pollak (JSI), Anka Supej (JSI)	Edited content on research on linguistic biases; edited introduction.
05/04/2019	v1.5	Michael Mathioudakis (UH)	Added section for concusions; formatted the document according to the template; edited the entire document.
07/04/2019	v1.6	Marko Robnik-Šikonja (UL)	Internal review.
11/04/2019	v1.7	Michael Mathioudakis (UH)	Incorporated edits and addressed some of the comments by Marko Robnik-Šikonja.
15/04/2019	v1.8	Carl-Gustav Linden (UH)	Addressed remaining comments by M. Robnik-Šikonja.
19/04/2019	v1.9	Saturnino Luz (UEDIN)	Internal review.
22/04/2019	v2.0	Michael Mathioudakis (UH)	Addressed comments by S. Luz.
24/04/2019	v2.1	Matthew Purver (QMUL)	Minor language edits before quality check.
25/04/2019	v3.0	Nada Lavrac (JSI)	Report quality checked and finalised.
30/04/2019	final	all authors	Addressed comments by N. Lavrač. Final version for submission prepared by S. Pollak.
30/04/2019	submitted	Tina Anžič (JSI)	Report submitted.

Table of contents

List of abbreviations

EC	European Commission
DoA	Description of Action
IAT	Implicit Association Test

1 Introduction

There is a considerable amount of evidence that bias (explicit or implicit) is an inherent part of human cognition. One method of identifying bias in humans is through the so-called Implicit Association Test (IAT), developed by Greenwald et al. [GMS98]. IAT is an electronic test that lets the test subject group words according to categories. The assumption behind the IAT is that the subject will group the words, which he/she strongly associates with the category, faster than the words he/she does not or only weakly associates with the category. While the implicit attitudes are measured through reaction time, the subject is also asked to report his/her explicit attitude towards the topic in question. A study by Nosek et al. [NBG02a], which utilized IAT as a measure of implicit bias, showed that people exhibit preference of young vs. old, white vs. black, and follow gender-career stereotypes. Specifically, subjects exhibited a stronger association of male terms with science and female terms with arts. They also showed strong association of female terms (e.g., girl) and family terms (e.g. children), as well as male terms (e.g., boy) and career terms (e.g., executive). Other studies that employed IAT showed that female respondents do not associate themselves strongly with words related to math and science [NBG02b]. If people are biased, it should come as no surprise that language, in which humans think, read, and speak, contains biased attitudes as well. Indeed, the evidence showing that corpora capture semantics has been substantial, e.g., [BCZ16, BLE07].

In what follows, we use the term ‘bias’ to refer to negative predisposition towards a group of people, particularly when such a predisposition is based on biological or other features that are beyond one’s control (e.g., gender, race, or family’s socioeconomic status), and is thus considered unfair. Gender bias and racism are two instances of such biases. Texts and corpora convey all kinds of social phenomena, including political-, gender-, age-, or race-related bias. The use of statistical models that are built on such datasets is therefore likely to further amplify bias.

In this report, our discussion of bias takes place in the context of two tasks.

The first task is that of detecting bias in news content and user comments. Accurately detecting biases in news content is important in order to prevent the spread of stereotypes. The task is relevant to large and small news providers alike. Small news providers might lack the resources available to large news providers for thorough editing of their content to remove biases. On the other hand, while large providers with wide audience and better resources (e.g., the New York Times, the Economist, Le Figaro) may be able to avoid some problems of explicitly biased language, they are still often subject to other biases: framing bias due to their political positioning or effects of authors’ implicit biases; and even moderate biases can have large effects over time if persistent.

The task of detecting biases is also important in many settings that involve user-generated content. As one example, discussions on social media and online forums often contain biased language and automatic detection of biases can help understand the arguments being made; in some cases, bias can become extreme and take the form of sexist or racist comments, and automatic bias detection could then help moderate discussion and fight hate speech. Another example is that of Web search, as the results of search engines reflect possibly biased user queries and click behaviour. Biased language might also appear in mobile text messages. Smart keyboard apps that learn language models from text messages and provide word suggestions might generate word suggestions that reflect the biased language of their users. In our project, we will be focusing on detecting biases in user comments that are related to news content.

The second task of interest is that of automatic generation of bias-free news content. It is related to the task of bias detection, in the following sense. If one has good algorithmic methods to generate news content and detect biases in news content, then the two methods could be deployed sequentially to first

generate news content, and then detect and remove biases in the generated content. A better method would perform news generation that is itself unbiased.

However, a major issue for the automatic detection and generation methods mentioned above is the challenge of algorithmic bias, i.e., biased decisions and other unfair outcomes that are (at least partially) the result of computational processes. The above automatic methods typically learn statistical models of natural language from data, and such models are therefore prone to reproducing the biases of the humans producing the data. For example, detection models can learn to associate labels with factors that are statistically associated with them in the dataset used, but are otherwise causally unrelated (e.g., if the only mentions of Islam in a dataset happen to be within instances of hate speech, a statistical classifier might classify all its future mentions -- even entirely objective or positive ones -- as hate speech). News generation systems that use a linguistic model built from user-generated comments on social media can reproduce sexist or racist language used by social media users. Irrespective of the source of the data, even if the origin and quality might not be known or readily assessed, automatic news analysis and generation systems must make sure that the produced results are free of known pernicious biases.

In the rest of this deliverable, we discuss further certain issues that are related to the two aforementioned tasks. Our discussion covers related literature on the following topics: how gender bias is reflected into language (Section 2.1); efforts to detect bias in user-generated content (Section 2.2); how gender and other biases are preserved in widely used instances of machine-learned representations of language (Section 2.3); and biases of journalists and other professionals in the news business (Section 3). Moreover, we describe a framework for measuring bias of content generation systems and discuss related measures of bias (Section 4). Finally, we conclude with recommendations for detecting and avoiding biases in news reporting (Section 5) and lay out the directions for future work (Section 6).

2 Bias in language and word embeddings

2.1 Gender and language

Studies in the relation between language and gender have a long research tradition and have been studied from different perspectives, either by searching for differences in male and female language use, i.e. analysing characteristics of male and female discourse style(s), so called genderlects (early work by Lakoff [LAK73], Spender [SPE80], Tannen [TAN90]), but even more relevant to the notion of bias, by observing gender construction by its representation in language. Lakoff, with the foundational work on language and gender [LAK73], has suggested that the women's inequality is reflected both in ways women are expected to speak as well as how they are spoken of.

Document corpora have been the source of various studies, where even analysis of the words *woman* and *man* shows a large number of differences in their collocation environment. Pearce examined the representation of men and women in the British National Corpus (BNC)[1] in five different semantic domains and concluded that, to a large extent, collocations follow gender stereotypes [PEA08]. Baker used Sketch Engine to examine the differences in representation of gendered items [BAK14].

[1] The British National Corpus: http://www.natcorp.ox.ac.uk/.

2.2 Bias in news media data

Voigt et al. compiled five datasets of comments made by or addressed to persons of known gender [VJP18]. They include comments from large social media platforms Facebook and Reddit, but also text from Fitocracy (a fitness-related online social network) and TED talks.

Kiritchenko and Mohammad compiled a dataset (called the Equity Evaluation Corpus), on which they tested a large number of automatic sentiment analysis systems that took part in a recent shared task and found that several of the systems show statistically significant bias (slightly higher sentiment intensity predictions for one race or one gender) [KM18].

Potash et al. built a corpus, used for bias detection in news published during a conflict [PRR17]. The corpus is built based on user interactions on social media platforms – should only one side of the political spectrum interact with a certain article, the article is likely to contain bias. Based on this corpus, the researchers built a bias-classification model, which achieved high accuracy.

Recasens et al. [RDJ13] analysed a large corpus of edits of Wikipedia articles, in order to identify biased text. The rationale for their approach is that Wikipedia enforces a Neutral Point of View Policy (NPOV), which guides editors to edit biases out of the text. By analysing such edits, Recasens et al. were able to identify two types of biases: subjective (or 'framing' bias) and epistemological (relating to what is commonly accepted to be true). They found that certain linguistic cues (e.g., occurrence of assertive verbs) are closely associated to such biases and, building upon this finding, they built machine learning models to identify those words in sentences that introduce biases of the aforementioned types.

Chen et al. [CWA18] used a corpus of articles that are already labelled for political bias and, using neural autoencoders, built machine-learned models that map text of a given bias to text of opposite bias.

2.3 Biases in word embeddings

Many recent approaches to computational natural language processing involve the use of word embeddings, i.e. vector representations of words which capture useful information about their similarity or relatedness, and which can be learned from large text corpora without any annotation using methods such as word2vec [MSC13]. However, the fact that they are learned from word co-occurrence associations observed in human language means that they can preserve the biases reflected in that language. Caliskan et al. [CBN17] showed that this is the case: popular word embedding models like Glove [PSM14] and word2vec [MSC13] do indeed preserve such biases. Several authors examine the embeddings for words that are known to be associated with gender or racial stereotypes. In particular, Caliskan et al. [CBN17] set to find whether biased word associations that were previously discovered through the implicit association test IAT [GMS98] - e.g., male names more closely associated with career, female names more closely associated with family - are also present in embeddings.

Bolukbasi et al. [BCZ16] studied a standard word2vec model trained on a large set of Google News articles by Mikolov et al. [MSC13]. In particular, they studied gender-based analogues, i.e. pairs of words that represent the same notion for its male and female version. Examples of such analogues are the pairs (she, he), (her, his), and (woman, man). The paper finds that many such analogues correspond to pairs of words that exhibit a certain female-to-male direction in the space of word2vec representations. In addition, they find that (i) among analogues generated from the word2vec embeddings, a high proportion contain words that should be gender neutral, and (ii) particularly some words that represent occupations exhibit strong gender biases along that direction (e.g., *man* – *computer programmer*, *woman* – *homemaker*). Garg et al. [GSJ18] track gender biases encoded over time in the embeddings of words related to occupations. The word embeddings they use come from

corpora that span the past 100 years. They measure the gender bias of a word by comparing its distance to clearly male and female words (e.g., *he* vs *she*) and compare the word bias to the percentage of females in the corresponding occupation. They find that word embeddings reflect gender ratios in occupations. Furthermore, they use the same word embeddings to track bias towards ethnic and religious minorities in a similar manner: by comparing the embeddings of words that describe ethnic or religious groups to certain positively or negatively connotated adjectives. They find a consistently decreasing bias against ethnic Asians but increasing bias against Islam.

It is therefore important to take account of these effects in models and tools that use embeddings; otherwise, their biases can be incorporated into the tools' outputs (e.g., biasing sentiment analysis output against particular genders or ethnicities). [BCZ16] suggest simple methods to de-bias word embeddings along the gender direction -- essentially making sure that words that should be considered gender-neutral have representations that fall between clearly male and female word representations; however, some more recent work shows that this fails to remove all the implicit bias information in the embeddings vectors [GG19]. Other methods may fare better: [ZZW18] suggest a procedure which is incorporated into the method by which the embeddings are learned in the first place; and [ZLM18] propose a method for using adversarial learning to encourage learning without bias. Park et al. [PSF18] used a dataset of sexist [WH16] and abusive tweets [FDC18] and tested the influence of using debiasing word embeddings proposed in [BCZ16] in the detection model. This research is in its early stages; such methods show promise but must be used with care and should be evaluated carefully. Furthermore, it is not clear how the biases in the embeddings (actually differences in distances and positions in the numeric space) affect machine learning models produced for downstream tasks (e.g., detection of hate speech). These issues are currently unresearched.

3 Gender and other biases in journalism

The purpose of this section is to explore how journalists and news organisations deal with biases and what methods are used to avoid prejudices and predispositions to impact reporting. In news journalism, an objectivity norm has been the standard where reporters should not let personal values influence their work. However, there is confusion around the very meaning of objectivity [KRO07]. Moreover, in real life situations journalists might behave very differently for many reasons and some sources of biases are well known.

First, we analyse different foundations of these biases [DEA19]. Biases are, for instance, present in the culture and language of the society on which the journalist reports - possibly reflecting injustices, gender inequality, and a limited role of women in public life. It is an established fact that women are discriminated in certain domains of media, especially in sports reporting [LRT05] where a small proportion of coverage focus on female athletes [BIS03, EBI00]. In several cases, tracking male and female names have been automated. For instance, a Swedish programmer at the daily Dagens Nyheter has developed Gender Equality Tracker[2] that collects names and pronouns from media in several countries. The results show, for instance, that about 68% of the names and pronouns in UK news media are male. This translates to 2.1 mentions of males for every female. In a given week, the best performer was the Daily Mail (52% female names and pronouns) and the worst was the Times of London (26%) while the share fluctuates widely in the Financial Times (between 16% and 42% during a random period of four weeks). To reduce the gender bias, the Financial Times has plugged in a gender tracker to the editorial system that alerts writers when the share of female sources is too low [WAT18].

On a higher level, the enclosure of a national or 'western' context might hinder journalists from understanding that their biases as personal values are actually grounded in a larger cognitive structure

[2] http://www.prognosis.se/repr-monitor/UK/

[VDA12]. National strategies can have a strong framing effect on news media. One of the most recent examples is 'the war on terror' after the attacks in New York in 2001 that changed the public debate to totally focus on national security with many serious implications, notably on privacy issues, but also on global stability [RLE09, ZUB19]. Values also reflect journalists' background, for instance upbringing, education and place of residence [DEU05]. Biases are also derived from the news organisation's mission and business model. During the years after Donald Trump was elected the US president this has been openly visible: the TV-channel Fox News, for instance, engages the audiences by supporting the president while liberal news outlets such as the New York Times or Washington Post have based their business model on anti-Trump reporting, boosting earnings through a 'Trump-bump' [NFK17].

Biases can be traced to the sources that journalists use for reporting, for instance corporate promotional material, organisational preferences present in press releases that are the standard sources of news, or routine use of sources whose bias has already been determined by their organizations and institutions [BEN88, DAV00, DAV02]. Gender has a decisive role: in the British press men are more than twice as likely as women to be quoted as sources both on the national and the regional level [ROS07].

Biases can be drawn from the lack of diversity in the newsroom, which might not reflect the composition of the population. It is common that reporters are recruited from journalism schools that lack diversity in gender and even more so regarding people with an immigrant background. The newsroom culture privileges elite and other (white) male voices [FFI05, ROS07].

Journalists tend to recognize and select perspectives from reality that are consistent with the existing sources of biases. Therefore, they need to become conscious of their biases and decide when it is appropriate and useful to apply them in reporting, and when not; and when they may be useful, and when they are inappropriate.
Conclusion 1: *Journalists need to learn how to manage bias*.

There are biases that are considered appropriate such as belief in representative government, open government, human rights for all, and social equality. Bias may also serve to create a narrative texture or make a story understandable since news are often framed to better align with the audience's experiences or worldviews to increase engagement [SCH99]. Framing techniques are a matter of choice, what facts and perspectives to include and exclude. Draining a story of all bias can drain it of its humanity, its lifeblood. Since biases are part of daily work, newsrooms have developed different mechanisms to deal with them.
Conclusion 2: *Journalists already have ways to evaluate and manage bias*.

There is a tendency in the news media to join other news outlets and journalists in collective action, in so called 'pack journalism' in cases of group thinking where they deem there is an overall agreement that, for instance, a politician has behaved badly and should be removed from office. These are instances where normal considerations are set aside and 'animal spirits' are set free because everybody else seems to be thinking in the same way. Confirmation biases are not rare in journalism, but seldom reach the level of discussion as in Sweden during the important #MeToo campaign in 2017 where several high-profile men were falsely accused by news media of being abusive.
Conclusion 3: *There are instances when journalists fail to manage bias*.

We end this section with the conclusion that news is the result of editorial choices that are influenced by a wide range of factors, including individual, organisational, social, cultural, economic and technical forces. They fit within the editorial focus and audience expectations. These values can also be seen as biases that are built into editorial processes and are typically hard to detect, even more difficult to change, for several reasons, such as strong path dependency, that is instutionalised ways of handling

challenges [NOR90]. Recently, we have seen signs of movement in the field. For instance, the Belgian news agency Belga, funded by Google, is developing a news bias detection platform.[3] This will help journalists and publishers as well as information and communication professionals to monitor, research, evaluate and monetise news media content.

4 Measures of bias of content generation systems

As algorithms are used more and more extensively to make decisions in various aspects of life, the scientific community tries to address ethical questions that arise from the use of algorithms. Ethical questions go well beyond text and content generation systems -- for example, there are ethical questions that arise with the use of self-driving cars.

At a high level, there are three broad groups of ethical issues: *fairness, accountability, and transparency*. Fairness asks for equal treatment of subjects despite differences in certain protected attributes (e.g., gender). Accountability asks for algorithms that are guaranteed to make decisions that adhere to certain ethical desiderata (e.g., fairness or not hurting humans). Transparency asks for algorithms that offer interpretable explanations of their decisions and are subjectable to audits.

Among these three notions, 'fairness' is closely related to that of 'bias' - and, in the context of this document, the two notions are in many ways opposites. We will thus say that language is *gender-fair* if one gender is not more or less strongly associated with words of positive or negative connotations - while we will say that language is *gender-biased* if it is not gender-fair. The concepts extend similarly to other cases of fairness and bias (e.g., with respect to race).

While these notions sound intuitive, we will need specific measures of bias to evaluate algorithms with respect to gender (or other kind of) bias. In what follows, we provide formal definitions of gender bias, drawn from the literature on algorithmic fairness [KCP17, KLR17, MCP19], in the backdrop of a simple and generic linguistic setting.

Specifically, let us consider a corpus of sentences. Possible corpora could be built from articles written by journalists, Web user comments on those articles, or synthetic sentences generated by a content generation system. Each sentence included in the corpus mentions one entity E, that represents one person, together with a qualitative descriptor Q (e.g., an adjective), that has a clear connotation, either positive or negative. 'Roger Federer' and 'Hillary Clinton' are two examples of entities. To give some examples of qualitative descriptors, 'high-achieving', 'powerful', 'high-ranking', etc, are considered positive qualitative descriptors; on the other hand, 'under-achieving', 'powerless', 'low-ranking', etc, are considered negative qualitative descriptors.

In this setting, we are not interested in the exact instance of descriptor, but only on whether it is positive or negative. If the descriptor associated with an entity is positive, we will write Q = +1, while negative descriptors will be denoted with Q = -1. For example, the sentence 'Hillary Clinton is a powerful politician' mentions the entity 'Hillary Clinton' (and so we have E = 'Hillary Clinton') and associates the entity with the positive descriptor 'powerful' (Q = +1). Note that, depending on the context, we can use a third value Q = 0 to denote absence of descriptor. For example, for the sentence 'Hillary Clinton is a politician' we have the same entity, but absence of descriptor (Q = 0). Note also that the sentences in the corpus might disagree with each other on the qualitative descriptor they use to describe the same entity. For example, one sentence in the corpus might be 'Hillary Clinton is a powerful politician' (Q =

[3] https://newsinitiative.withgoogle.com/dnifund/dni-projects/Digitally-enable-bias-detection-in-news-articles/

+1), another ‘Hillary Clinton is a powerless politician’ (Q = -1), and another ‘Hillary Clinton is a politician’ (Q = 0).

In addition to the corpus of sentences, we consider a *ground truth database*, which contains information about all entities E. Each entity E is associated with three attributes of interest, which we divide into three categories.

The first category includes *protected* attributes A. In our setting, we consider a single protected attribute, namely the gender of the entity. We will write A = 0 for female entities and A = 1 for male entities.

The second category includes qualifying attributes Y. In our setting, we consider a single qualifying attribute Y, namely the true (positive or negative) connotation of descriptors associated with entity E. Note that, while for the corpus of sentences the same entity can appear with descriptors of any connotation (Q=+1 or Q=-1), in the ground truth database there can only be one connotation Y=+1 or Y=-1 for the entity, and it is considered to be the true/correct one.

For example, if the corpus contains the sentence ‘Roger Federer is an under-performing athlete’ (Q = -1), but the ground truth database tells us that the entity ‘Roger Federer’ is associated with descriptors of positive connotation, then we interpret this to mean that the sentence is *biased against* the entity. We will say that a sentence is *accurate* with respect to the entity E it contains only if Q = Y, i.e. if the qualitative descriptor it uses has a connotation that agrees with the ground truth database.

Finally, the third category contains other attributes X. In our setting, we consider a single other attribute, namely the occupation of the entity. For illustration, Table 1 below contains an example of a corpus of sentences and Table 2 contains an example of a ground truth database.

Table 1. Corpus of sentences (example).

Sentence	**E**	**Q**
Hillary Clinton is a powerful politician.	Hillary Clinton	+1
Roger Federer is an under-performing athlete.	Roger Federer	-1
Roger Federer is a tennis player.	Roger Federer	0
Roger Federer is the greatest tennis player of all time.	Roger Federer	+1
Hillary Clinton is a powerless politician.	Hillary Clinton	-1

Table 2. Ground truth database (example).

E entity	**A** gender	**Y** correct connotations	**X** occupation
Hillary Clinton	0	+1	politician
Roger Federer	1	+1	athlete
Adolf Hitler	1	-1	politician
Medea	0	-1	mythical character

4.1 Observational measures

We start by describing so-called *observational bias* measures - i.e. measures that are defined on quantities that are assumed observed or observable, and do not refer explicitly to the mechanism that generates the data.

We begin by discussing the condition of *demographic parity*, as it appears in [KLR17] and other works in the literature of algorithmic fairness. It is defined with the following mathematical expression.

$$\text{demographic parity:} \quad P(Q = 1 | A = 1) = P(Q = 1 | A = 0)$$

Demographic parity expresses the requirement that the number of positive qualitative descriptors that accompany entities of each protected-attribute value should be proportional to the population of the corresponding group of entities. For example, if the number of sentences in the corpus that refer to male entities (A = 1) are as frequent as those for female entities (A = 0), then for demographic parity to be satisfied, the positive descriptors (Q=+1) used for male entities should be as many as those for female entities. If the condition is not met, we say we have a case of *demographic bias* in the corpus. For example, if male entities are associated with positive descriptors (Q = +1) twice as frequently as females, even though male and female entities appear equally frequently in the corpus, then we have demographic bias against females in the corpus. One can quantify demographic bias in terms of the discrepancy between the two probabilities $P(Q = 1 | A = 1)$ and $P(Q = 1 | A = 0)$. In this manuscript, let us quantify *demographic bias* simply as the difference of the two probabilities.

$$\text{demographic bias} = P(Q = 1 | A = 1) - P(Q = 1 | A = 0)$$

Demographic bias as quantified above is a simple and intuitive measure of bias; however, it completely ignores the ground truth database. Continuing with the example above, it is possible that the male entities (A=1), which are more frequently associated with positive descriptors (Q=+1) in the corpus, are also more frequently associated with positive descriptors in the ground truth database (Y = 1). In other words, it is possible that male entities in a particular corpus are associated more frequently with positive descriptors like ‘successful’ and ‘high-achieving’ than females, and the ground truth database tells us that indeed such descriptors are accurate. Should such a situation be considered as biased against females? In some contexts, it would not seem fair to require that some males be associated with negative descriptors in text only to balance out the ones for females, as demographic parity requires.

To address this shortcoming of demographic bias, we consider the condition known as *equality of opportunity*, as it appears in [KLR17].

$$\text{equality of opportunity:} \quad P(Q = 1 | A=1, Y=1) = P(Q = 1 | A=0, Y=1)$$

Equality of opportunity expresses the requirement that the probability that a truly positive (Y=1) entity appears as positive (Q = 1) in the corpus should not depend on the gender of the entity (A = 0 or A = 1). Building upon the condition of equality of opportunity, one can define associated measures of bias in terms of the discrepancy between the two probabilities $P(Q = 1 | A=1, Y=1)$ and $P(Q = 1 | A=0, Y=1)$ that correspond to each gender. In this manuscript, let us define the measure of *opportunity bias* simply as the difference between the two probabilities.

$$\text{opportunity bias} = P(Q = 1 | A=1, Y=1) - P(Q = 1 | A=0, Y=1)$$

Opportunity bias does not have the shortcomings of demographic bias, but it can only be evaluated if we have access to the ground truth database. This might not be realistic in many cases, as the ground truth database might be difficult to build in the first place. To see why, consider that it might be difficult to arrive to a consensus as to whether an entity should be associated with positive or negative connotations in text. On the other hand, demographic bias can be evaluated directly from the corpus of sentences.

4.2 Measures based on causal reasoning

As we have just seen, the two bias measures above cover two extremes: demographic bias ignores the ground truth and can be seen as unjust in some settings; opportunity bias assumes that ground truth is accessible at evaluation time, which might be unrealistic. Many cases fall in the middle of the two extremes, where ground truth is difficult to define, and therefore opportunity bias is difficult to evaluate, but at the same time demographic bias is considered too coarse a measure. Luckily, in some of these cases, we have knowledge of the mechanism that generates the data - and we can define bias based on the mechanism directly. This is the case, for example, for open-sourced content generation systems.

Awareness bias, as it appears in [KLR17], is a measure of bias that is defined upon the data generation mechanism - and particularly on whether protected attributes are used in the generation of data. For example, let us consider the case where the corpus of sentences is generated by a content-generation system. The system is allowed to generate sentences according to one of the two mechanisms below.

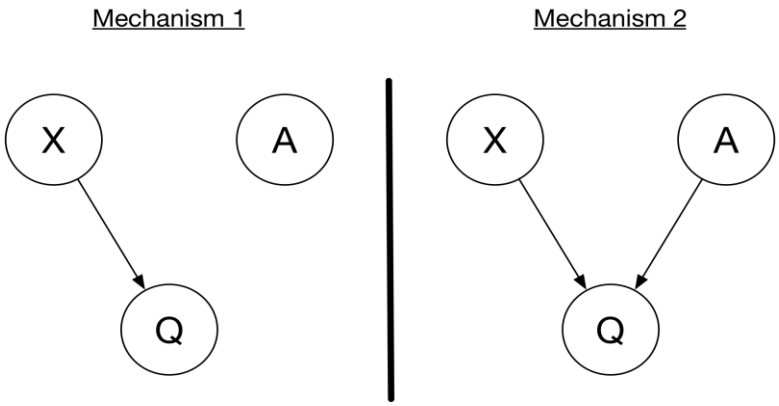

Figure 1. Two mechanisms for sentence generation. In both cases, X corresponds to the occupation of an entity, A to its gender, and Q to the type, positive or negative, of a qualitative descriptor.

The mechanism on the left of Figure 1 generates qualitative descriptors based only on the occupation of an entity. One possible instance of Mechanism #1 is to output a positive qualitative descriptor only for entities that are politicians.

Mechanism #1: if X = ‘politician’ then Q = +1; otherwise Q = -1

Compare it with Mechanism #2 on the right, which generates qualitative descriptors based on both the occupation and gender of an entity. One possible instance of Mechanism 2 is to output a positive qualitative descriptor only for those entities that are politicians and male.

Mechanism #2: if X = ‘politician’ and A = 1, then Q = +1; otherwise Q = -1

Mechanism 1 is said to avoid gender bias through unawareness of the gender (protected attribute A). In other words, it is ‘blind’ to gender. The opposite is true for Mechanism #2. Based on this intuition, we have the following definition for awareness bias.

awareness bias = 1 if mechanism uses protected attribute A, 0 otherwise

The problem with awareness bias is that, even if a mechanism is blind to the protected attribute A, A might still have dependencies with other attributes that the mechanism uses. To see why, consider again Mechanism #1 in Figure 1 above. It is possible that, even though the decision for Q does not depend on gender A directly, it does so indirectly. For example, it might be that gender determines occupation (e.g., if for whatever reason women are more likely to become nurses, while men are more likely to become politicians), as shown in Figure 2 below. By basing the decision for the qualitative descriptor (Q) on occupation (X), mechanism #1 is influenced by gender (A) *indirectly*.

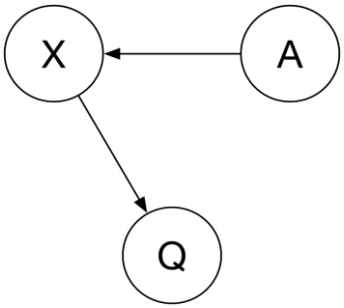

Figure 2. Although Mechanism #1 of Figure 1 does not consider gender (A) when it decides what qualitative descriptor (Q) to use for an entity, but only considers occupation (X), it is possible that the decision is indirectly based on gender, for example if occupation is determined by gender.

This shortcoming of awareness bias can be addressed with measures of *counterfactual bias*, which are based on the following question: does the mechanism produce the same results for cases that differ only in protected attribute A? Different variants for *counterfactual bias* measures exist, which require the language of causal calculus to define. This is beyond the scope of this deliverable, but the interested reader can see [KCP17, KLR17, MCP19]. We just note that, if the entire mechanism for data generation is available, then counterfactual fairness is straightforward to measure. However, difficulties arise when the mechanism is not fully specified. For example, in the case of Mechanism #1, the mechanism can be readily available as the source code of a content generation system. However, the dependency of occupation (X) on gender (A) requires expert knowledge that might not be available at the time when the mechanism is evaluated (for example, we might not know if females are more likely to become nurses and males more likely to become politicians). In such cases, one should consider different possibilities for the dependencies between the various quantities and explore how the evaluation of the mechanism would vary across them. The notion of counterfactual bias can also be seen as related to the more general notion of individual fairness [DHP12], which requires that similar cases should be treated similarly.

4.3 Trade-off between bias and accuracy

Consider a content generation system that is about to generate a sentence like the following: 'Roger Federer is an [blank] athlete'. In real time, the system has to decide whether to include a positive or negative descriptor (adjective) to describe what kind of athlete Roger Federer is. If the system had access to a ground-truth database like the one we assumed above, then the system's task would be simple: just use a descriptor with Q = Y that matches the value Y found in the database for the entity of Roger Federer. In reality, however, systems rarely have access to such a database in real time -- and the system would have to decide, with some uncertainty, about what descriptor to use.

In doing so, the system faces two kinds of desiderata: to be accurate and unbiased. Unfortunately, this leads to trade-offs: in many settings, it is impossible to have both. We saw this problem with demographic bias, discussed above: if we force male and female entities to appear with equal proportions of positive or negative descriptors in a corpus, then we will introduce inaccuracies (i.e., we will have to use negative descriptor for entities that should be associated with positive descriptors according to the ground truth database, or vice versa). The problem is more general though. Ground truth might be very difficult to define, e.g., if there is a general disagreement about whether descriptor should be positive or negative, even among experts. The issue is exacerbated if ground truth is defined based on user-generated content or is meant to represent public opinion. If such public opinion is inherently biased (e.g., if among the public women are considered inferior to men) then an accurate algorithm would replicate these biases. This is in direct conflict with the conditions for lack of bias, as captured by the measures described above.

We end the section noting that resolving the trade-off between accuracy and (lack of) bias is an active area of research at the time of this writing - see, for example, [CDG18, HLG19, HUM19, HPS16, ZVG17].

5 Recommendations on avoiding gender and other biases

Informed by the material presented above, we make the following recommendations for text analysts, journalists, and engineers of text generation systems.

Recommendation 1. Be aware of specific examples of gender and other biases that have been identified in English and other languages. Studies on word embeddings such as [BCZ16] have shown that language embodies gender biases that are associated with occupation, but also beyond that. If aware of these, journalists could avoid replicating them in articles and developers of content generation systems could test their systems to make sure their systems do not generate them in large proportions.

Recommendation 2. Use automated techniques such as that of [CWA18] to identify words that are possible sources of bias. Such tools can be used by journalists during proof-reading of news articles before publication, by discussion forum moderators to detect and justify the flagging of biased Web user comments (e.g., in case of hate speech), and by content-generation system developers to identify biases of generated text, when the system is deployed 'in-the-wild'.

Recommendation 3. Formalize notions of accuracy and bias in the setting at hand, as we did in the hypothetical setting of a sentence corpus in Section 4 above. When a large corpus of text is available, it is good to evaluate the bias in the qualitative descriptions used in the corpus according to a variety of measures. In particular, one should attempt to evaluate measures that are directly computable from the data at hand (e.g., demographic bias), as well as measures that are based on a ground truth when it is available (e.g., opportunity bias).

Recommendation 4. During the evaluation of content generation systems, explore the trade-off between accuracy and bias for different parameterizations of the system.

Recommendation 5. Investigate the use of automated de-biasing techniques such as that of [BCZ16, ZLM18, ZZW18] to help reduce the biases in word embedding models used as the basis of automated text analysis tools.

6 Conclusions and further work

The issue of bias has been studied in many settings and from many angles (e.g., journalistic, linguistic, and algorithmic bias). In the context of EMBEDDIA, the recommendations of Section 5 above will be considered throughout the project, particularly in the language technologies being developed in WP1-WP6. We also plan to perform specific studies, inspired by related work presented in WP2 and implementing the proposed frameworks of Section 4 with our news media data and testing the resulting ability to discover and quantify bias. The content of the deliverable was also already presented in the Tallinn workshop for EMBEDDIA researchers and media partners in March 2019.

Specifically, in WP3, we will apply this within T3.1 ("Cross-lingual context and opinion analysis") to investigate both algorithmic bias in our tools, and bias in content as authored by users. For the former,

we plan to study the effects of the use of different corpora for training, and for building word embeddings, on the outputs of the tools we develop for sentiment and opinion detection, in terms of their resulting observed demographic and opportunity bias behaviour. For the latter, in collaboration with WP1 T1.2 ("Context-dependent and dynamic embeddings"), we will investigate the ability of context-dependent word embeddings to reveal /framing/ bias in user-generated content, by analysing the differences in e.g. gender associations of words caused by their use in different lexical and sentential contexts.

In WP4 we will apply it within T4.3 ("Cross-lingual Identification of viewpoints and sentiment in news reporting"). We plan to perform a study, where at least one news corpus will be analysed from the perspective of demographic and political bias. This can be done either in terms of quantitative evaluation in relation to sentiment models, or by analysis of embeddings, in terms of analogies or similarities of selected concepts. For both tasks, collaboration with WP1 where the embeddings are trained and WP2 for identifying named entities will be considered in order to investigate the effect of different embedding and NER models on the resulting models.

Moreover, in the context of task T6.4, we will coordinate with WP5 to evaluate technology for Natural Language Generation (NLG) in terms of gender bias. We will publish the source code we develop for this evaluation, so that it is used or adapted by the scientific community for the evaluation of other NLG systems.

Finally, in D6.11, which is the second deliverable of task T6.4 and due at the end of the project (M36), we will summarize the findings of the aforementioned studies and discuss how to detect and avoid gender and other biases using the tools and technology developed within EMBEDDIA.

References

[BAK14] Baker, P. (2014). Using Corpora to Analyze Gender. London & New York: Bloomsbury Publishing.

[BCZ16] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

[BEN88] Bennett, L. W. (1988). News: The politics of illusion. New York: Longman.

[BIS03] Bishop, R. (2003). Missing in action: Feature coverage of women's sports in Sports Illustrated. Journal of Sport and Social Issues, 27(2), 184-194.

[CBN17] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

[CDG18] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

[CWA18] Chen, W. F., Wachsmuth, H., Al Khatib, K., & Stein, B. (2018). Learning to Flip the Bias of News Headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 79-88).

[DAV00] Davis, A. (2000). Public relations, business news and the reproduction of corporate elite power. Journalism, 1(3), 282-304.

[DAV02] Davis, A. (2002). Public relations democracy: Public relations, politics and the mass media in Britain. Manchester: Manchester University Press.

[DEA19] Dean, W. (2019). Journalism essentials. Retrieved on 6.3.2019 from the webpage https://www.americanpressinstitute.org/journalism-essentials/

[DEU05] Deuze, M. (2005). What is journalism? Journalism, 6(4), 442-464.

[DHP12] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226). ACM.

[EBI00] Eastman, S. T., & Billings, A. C. (2000). Sportscasting and sports reporting: The power of gender bias. Journal of Sport and Social Issues, 24(2), 192-213.

[FDC18] Founta, A.-M., Djouvas, D., Chatzakou, D., Leontiadis, I., Bleckburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. In AAAI.

[FFI05] Freedman, E., & Fico, F. (2005). Male and female sources in newspaper coverage of male and female candidates in open races for governor in 2002. Mass Communication & Society, 8(3), 257-272.

[GG19] Gonen, H. & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *CoRR*, https://arxiv.org/abs/1903.03862.

[GMS98] Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. Journal of Personality and Social Psychology 74 (6): 1464–1480.

[GSJ18] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644.

[HLG19] Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2019, January). A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 181-190). ACM.

[HPS16] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315-3323).

[HUM19] Hutchinson, B., & Mitchell, M. (2019, January). 50 Years of Test (Un) fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 49-58). ACM.

[KCP17] Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* (pp. 656-666).

[KLR17] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076).

[KM18] Kiritchenko, S. and Mohammad, S. M. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM), pages 43–53. New Orleans, June 5-6, 2018.

[KRO07] Kovach, B., & Rosenstiel, T. (2007). The elements of journalism: What newspeople should know and what the public should expect. New York: Three Rivers Press.

[LAK73] Lakoff, Robin (1973) Language and women's place. Language in Society 2: 45-79.

[LRT05] Len-Rios, M. E., Rodgers, S., Thorson, E., & Yoon, D. (2005). Representation of women in news and photos: Comparing content to perceptions. Journal of Communication, 55(1), 152-168.

[MCP19] Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019, January). Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 349-358). ACM.

[MSC13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (pp. 3111-3119).

[NBG02a] Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002a.) Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site. Group Dynamics: Theory, Research, and Practice 6 (1): 101–115.

[NBG02b] Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002b). Math = Male, Me = Female, Therefore Math ≠ Me. Journal of Personality and Social Psychology 83: 44-59.

[NFK17] Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., & Nielsen, R. K. (2017). Reuters institute digital news report 2017. Oxford: Reuters Institute.

[NOR90] North, D. C. (1990). Institutions, institutional change and economic performance. Cambridge: Cambridge University Press.

[PEA08] Pearce, M. (2008). Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. Corpora 3: 1-29.

[PRR17] Potash, P., Romanov, A., Rumshisky, A., and Gronas, M. (2017). Tracking Bias in News Sources Using Social Media: the Russia-Ukraine Maidan Crisis of 2013-2014. *Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism*: 13-18.

[PSF18] Park, J. H., Shin, J., and Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2799-2804.

[PSM14] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543).

[RDJ13] Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1650-1659).

[RLE09] Reese, S. D., & Lewis, S. C. (2009). Framing the war on terror: The internalization of policy in the U.S. press. Journalism, 10(8), 777-797.

[ROS07] Ross, K. (2007). The journalist, the housewife, the citizen and the press: Women and men as sources in local news narratives. Journalism, 8(4), 449-473.

[SCH99] Scheufele, D., A. (1999). Framing as a theory of media effects. The Journal of Communication, 49(1), 103-122.

[SPE80] Spender, D. (1980). Man Made Language. London; New York: Routledge & Kegan Paul.

[TAN90] Tannen, D. (1990). You Just Don't Understand: Women and Men in Conversation. New York: Ballantine Books.

[VDA12] Van Dalen, A. (2012). Structural bias in cross-national perspective: How political systems and journalism cultures influence government dominance in the news. The International Journal of Press/Politics, 17(1), 32-55.

[VJP18] Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., & Tsvetkov, Y. (2018). RtGender: A corpus for studying differential responses to gender. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

[WAT18] Waterson, J. (2018, November 14,). Financial Times tool warns if articles quote too many men. The Guardian. Retrieved on 11.3.2019 from https://www.theguardian.com/media/2018/nov/14/financial-times-tool-warns-if-articles-quote-too-many-men.

[WH16] Waseem, Z., and Hovy, D. (2016). Hateful symbols or hateful people? Predictive features of hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop: 88-93.

[ZLM18] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340). ACM.

[ZUB19] Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. New York: PublicAffairs.

[ZVG17] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1171-1180). International World Wide Web Conferences Steering Committee.

[ZZW18] Zhao, Zhou, Li, Wang, Chang (2018). Learning Gender-Neutral Word Embeddings.In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.*