An Overview of Fairness in Data – Illuminating the Bias in Data Pipeline

Senthil Kumar B [1], **Chandrabose Aravindan** [1], **Bharathi Raja Chakravarthi** [2]
[1] Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu
[2] National University of Ireland Galway
senthil@ssn.edu.in

Abstract

Data in general encodes human biases by default; being aware of this is a good start, and the research around how to handle it is ongoing. The term 'bias' is extensively used in various contexts in NLP systems. In our research the focus is specific to biases such as gender, racism, religion, demographic and other intersectional views on biases that prevail in text processing systems responsible for systematically discriminating specific population, which is not ethical in NLP. These biases exacerbate the lack of equality, diversity and inclusion of specific population while utilizing the NLP applications. The tools and technology at the intermediate level utilize biased data, and transfer or amplify this bias to the downstream applications. However, it is not enough to be colourblind, gender-neutral alone when designing a unbiased technology – instead, we should take a conscious effort by designing a unified framework to measure and benchmark the bias. In this paper, we recommend six measures and one augment measure based on the observations of the bias in data, annotations, text representations and debiasing techniques.

1 Introduction

Irrespective of the data sources, the majority of the bias prevails in data itself (Lam et al., 2011; Fine et al., 2014; Jones et al., 2020). The inherent bias in data affects the core NLP tasks such as Part-Of-Speech (POS) tagging, POS Chunking (Manzini et al., 2019), and dependency parsing (Garimella et al., 2019). Other than data bias, the techniques used to represent the data also pose a threat to NLP systems (Bolukbasi et al., 2016; Caliskan et al., 2017; Ethayarajh et al., 2019; Zhou et al., 2019). Eventually the bias magnifies itself due to these biased data representation in downstream applications such as Named Entity Recognition (NER) (Manzini et al., 2019), coreference resolution (Zhao et al., 2017; Rudinger et al., 2018; Zhao et al., 2019), sentiment analysis (Kiritchenko and Mohammad, 2018), machine translation (Stanovsky et al., 2019; Escudé Font et al., 2019), social data analysis (Waseem and Hovy, 2016; Davidson et al., 2017; Sap et al., 2019; Hasanuzzaman et al., 2017), and bio-medical language processing (Rios et al., 2020). However,a very little attention is given to data collection and processing. Wikipedia seems like a diverse data source but fewer than 18% of the site's biographical entries represents women(Wagner et al., 2015).

Olteanu et al. (2019) observed that the dataset extracted from social media data has its own bias in various aspects such as age, gender, racism, location, job, and religion. Existing research classified biases at various levels, including the bias in data - source itself, the bias in the data analysis pipeline, and the biased data in building the systems. For example, using a biased abusive language detection system may result in discrimination against a group of minority peoples such as African-American (Waseem and Hovy, 2016).

The survey by Blodgett and O'Connor (2017) on bias in NLP systems found that works related to bias failed to explain why the system behaviours are 'biased', harmful, what kind of behaviours lead to bias, in what ways, to whom and why. The paper also discusses the need to conceptualise bias better by linking the languages and social hierarchies in the society. Another survey by Sun et al. (2019) analyses the bias framework and issues with the current debiasing methods. Their study reveals that the gender bias in NLP is still at a nascent level and lacks unified metrics and benchmark evaluation methods. This paper audits or surveys the present situation in analysing the bias at various levels of data.

This paper has been divided into four section. In

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 34–45
April 19, 2021.

section 2, we analyse the bias in language resources or corpora, in word representations, in pre-trained language models. In section 3, based on the observation of bias in corpora level, social data, text representations, metric to measure and mitigate bias, we infer six recommendations and one measure augmenting the existing one. In section 4, we conclude with the need of practicing standard procedures across the data pipeline in NLP systems to support ethical and fairness usage of data.

2 Data Bias

Language resources are present in two forms: 1) Online digital data 2) Annotated corpus. In fact language technologies are data-driven methods based on statistical learning techniques. There needs to be a framework to monitor the flow of data from the source to the NLP system and to measure the bias emanating at each level. The dataset can be representative of a specific genre or domain and a balanced one to avoid selectional bias as noted by Søgaard et al. (2014). The bias also emanates from labelling the corpus by annotators. Dickinson and Meurers (2003) observed bias in widely-used Penn Treebank annotated corpora which contain many errors. Apart from the linguistic experts, using non-experts in the annotation process through crowd-sourcing platforms also leads to considerable bias in data (Waseem, 2016). Figure.1 shows the effects of bias in data which leads to bias in gender, bias in syntactic and shallow parsing across domains, bias due to the disparity in language resources represented by tree hierarchy.

2.1 Fallacies in Language resources or corpora

As language resources or corpora are crucial for any NLP systems, the following show the gender and domain bias in data itself, occurred due to tagging and parsing, annotation issues in corpus creation.

Bias in Diversity and Inclusion of languages:

- The quantitative analysis by Joshi et al. (2020) reveals the disparity amongst language resources. The taxonomy of languages in the aspect of resource distributions (Joshi et al., 2020) shows that among six classes of languages, the Class-0, which contains the largest section of languages – 2,191 languages – are still ignored in the aspect of language technologies, and there is no resource present. This non-availability of language resources aggravates the disparity in language resources.
- Hence the Diversity and Inclusion (DI) score should be recommended to measure the diversity of NLP system methods to apply for different languages and the system's contributions to the inclusivity of poorly resourced or less represented languages Joshi et al. (2020).
- As languages are dynamic, Hovy and Søgaard (2015) warned the consequence of using age-old limited training data to train a model which could be evaluated on new data.

Bias across domain:

- Fine et al. (2014) investigated bias across five annotated corpora: Google Ngram, BNC (written, spoken), CELEX (written, spoken), Switchboard (Penn Treebank) using psycholinguistic measures identified domain bias in the corpora. Google appears to be familiar with terminology dealing with adult material and technology related terms. Similarly, BNC is biased towards Briticisms. The Switchboard corpus overestimates how quickly humans will react to colloquialisms and backchannels during telephone conversations.

Bias in annotating social data:

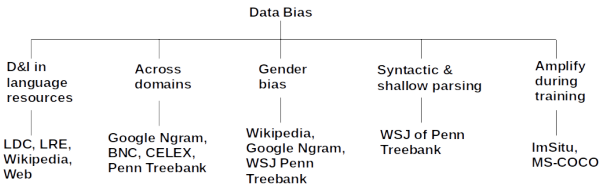
- Davidson et al. (2017) found a substantial racial bias that exists among all datasets and recommended the way to measure the origin of bias during collection and labelling.
- The bias in social data cripples during the labelling process as noted by Waseem (2016). The author criticized the decision by annotators by labelling the words from African-American English (AAE) as offensive since it is being used widely among its users. The models built on expert annotations perform comparatively better than the amateur annotators as they are more likely to label an item as hate speech. Indeed annotating task for social data is complex if the task is to categorize the abusive behaviour, as there is no standard procedure, and what qualifies as abusive is still not clear (Founta et al., 2018).
- The dialect-based bias in the corpora was observed by Sap et al. (2019). Amazon Mechanical Turk annotators were asked to find whether

Twitter user's dialect and race are offensive to themselves or others. The result showed that dialect and race are less significant to label AAE tweet as offensive to someone. Thus the racial bias emanates from skew annotation process can be removed by not having skewed demographics of annotators.

Gender bias in data / corpus:

- After 2010, Wikipedia is aimed at closing the gender gap by rolling out several projects, which includes WikiProject women, WikiProject gender studies, Women in Red. To determine the type and quality of material that is acceptable in Wikipedia articles, the content on Wikipedia must be written from a neutral point of view (NPOV) [1].
- But still, there is a large gender gap among the Wikipedia editors. Findings by Lam et al. (2011) indicate a male-skewed gender imbalance in English Wikipedia editors. Females are more likely to exit from Wikipedia than males before accumulating many edits. Hence the proportion of males who become administrators is more than females. Based on eight interest areas, females and males are focused on different content, females concentrate more on People and Art areas, while males focus more on Geography and Science.

Figure 1: Data bias tree hierarchy



- How gender bias has developed in the Wikipedia was analysed by Schmahl et al. (2020) using four categories: Career, Science, Family and Arts. The stereotypical gender bias in the categories family and science is decreasing, and art-related words are becoming more biased towards females. Their findings also reveal that the selection of corpus for word embedding depends on the task. For example, to have a gender-neutral word embedding related to Science, one may best use the corpus of 2018.

[1] https://en.wikipedia.org/wiki&/Wikipedia:Neutral_point_of_view

- Analysis on the word embedding representation of the Google Ngram corpus shows that stereotypical gender associations in languages have decreased over time but still exists (Jones et al., 2020).

Bias in syntactic and shallow parsing:

- Apart from the labelling bias in Penn Treebank, Garimella et al. (2019) observed a gender bias in syntactic labelling and parsing for the Wall Street Journal (WSJ) articles from Penn Treebank with gender information. The POS taggers, syntactic parsers trained on any data performed well when tested with female-authored articles. Whereas the male writings performed better only on sufficient male-authored articles as training data. This highlights the grammatical differences between gender and the need for better tools for syntactic annotation.
- Off-the-shelf POS taggers show differences in performance for languages of the people in different age groups, and the linguistic differences are not solely on lexical but also grammatical (Hovy and Søgaard, 2015).

2.2 Fallacies in Word Representations

Bias is pervasive in word embeddings and neural models in general. Apart from analysing the language resources and corpora, the bias in the language can also be studied by using the word embeddings – distributed representation of words in vector form. Biases in word embeddings may, in turn, have unwanted consequences in downstream applications.

Bias in word representations:

- A pioneer work in the detection of bias on word embeddings by Bolukbasi et al. (2016) proposed two debiasing techniques: hard and soft-debias. They showed that gender bias could be captured by identifying the direction in embedding subspace and could be neutralised. The hard-debiasing effectively reduce gender stereotypes when compared with the soft-debiasing technique.
- Caliskan et al. (2017) measured the bias in the Word2Vec embeddings on Google News corpus and pre-trained GloVe using WEAT, WEFAT score. The authors noted that the gender association strength of occupation words is highly correlated.

- Using the notion of bias subspace, the WEAT shows systemic error in overestimating the bias (Ethayarajh et al., 2019). Even though Bolukbasi et al. (2016) shown that removing bias by subspace projection method, there is no theoretical guarantee that the debiased vectors are entirely unbiased or the method works for embedding models other than the skip-gram with negative sampling (SGNS). Proposed the measure relational inner product association (RIPA) that analyses how much gender association in embedding space is due to embedding model and training corpus. For further notion on the impact of bias, subspace and debiasing refer Ethayarajh et al. (2019).

Representation bias in social data:

- The Twitter data was analysed using demographic embeddings (location, age, location), which shows that the gender variable has a small impact when compared with the location variable in classifying the racism tweets (Hasanuzzaman et al., 2017).

- The Word2Vec trained on L2-Reddit shows bias in the dataset using multi-class for a race, gender and religion (Manzini et al., 2019). Embeddings are debiased using hard, and soft-debias and its downstream effect shows decreased performance by POS tagging and an increase in NER and Chunking.

Representation bias in applications:

- To detect an unintended bias in word embeddings Sweeney and Najafian (2019) proposed a Relative Negative Sentiment Bias (RNSB) framework. WEAT score shows that the Word2vec and GloVe word embeddings are biased with respect to national origin, whereas RNSB measure the discrimination with respect to more than two demographics within a protected group. The study also reveals that the unintended bias is more in GloVE embeddings.

Representation bias in language resources:

- Brunet et al. (2019) used PPMI, Word2Vec and GloVe embeddings to train the two corpora Wikipedia and NYT, used WEAT metric to quantify the amount of bias contributed by the data in training.

- Using Word2vec trained on Google News, GloVe on NYT and COHA embeddings for temporal analysis Garg et al. (2018) found the occupational and demographic shifts over time. The embeddings are leveraged to quantify the attitude towards women and ethnic minorities.

- Zhao et al. (2019) reported that bias gets amplified while training the biased dataset. For example, some of the training set verbs have small bias and are heavily biased towards females after training. Kitchen and technology-oriented categories in MS-COCO are aligned towards females and males, respectively that have been amplified during training.

2.3 Fallacies in Pre-trained Language Models

Instead of using the word embeddings directly from the models for classification, the embeddings from pre-trained language models (LM) such as GloVe, ELMo, BERT, GPT can be used to train the task-specific datasets using transfer learning.

Pre-trained LM in MT:

- Escudé Font et al. (2019) found gender bias in the translation of English-Spanish in the news domain. The baseline system is trained with GloVe and use hard-debias along with GN-GloVE embeddings for debiasing. The BLEU score increased for the pre-trained embedding and improved slightly for the debiased model using transfer learning which means the translation is preserved while enhancing the gender debias.

- Embeddings of gendered languages such as Spanish and French contain gender bias. For English, Zhou et al. (2019) used the fastText monolingual embeddings pre-trained on Wikipedia corpora. For word translation, the bilingual embeddings ES-EN and FR-EN from MUSE aligns the fastText monolingual embeddings. The bias in bilingual embeddings is observed using MWEAT and mitigated using Shift_EN method.

Pre-trained LM in Coreference:

- Zhao et al. (2019) used ELMo pre-trained embeddings to represent OntoNotes to train the LSTM model. The bias is evaluated using the WinoBias dataset. Data augmentation and

Data	Resource/Task/Techniques	Work by
Bias in Diversity, Inclusivity of Language Resources	Class-0: 2191 languages virtually no unlabelled data to use	(Joshi et al., 2020)
Bias in Language Resources	Wikipedia	(Lam et al., 2011)
	Google Books Ngram Corpus	(Schmahl et al., 2020)
		(Jones et al., 2020)
Bias in Domain	Google Ngram, BNC(written,spoken), CELEX(written, spoken), Switchboard(Penn Treebank)	(Fine et al., 2014)
Bias amplify in training	Bias in embeddings gets amplified while training the datasets – imSitu	(Zhao et al., 2017)
Lexicons to detect dialects in social data	AAE	(Davidson et al., 2017), (Sap et al., 2019), (Blodgett and O'Connor, 2017)
	HateBase	(Davidson et al., 2017)
	Slur DB, ethnic slur by Wikipedia	(Hasanuzzaman et al., 2017)
Data design/schema-level constructs	WinoBias-gender bias in Coref.	(Zhao et al., 2018, 2019)
	Winogender-gender bias in Coref.	(Rudinger et al., 2018)
	WinoMT - gender bias in MT	(Stanovsky et al., 2019)
	EEC-gender/race bias in SA	(Hasanuzzaman et al., 2017)

Table 1: Survey on the research done in the bias in data

Neutralization are used to fine-tune the embeddings, which reduces bias.

Pre-trained LM in Social data analysis:

- Park and Fung (2017) trained the Twitter dataset on CNN, GRU and Bi-directional GRU with attention models using FastText, Word2vec pre-trained embeddings. The gender bias on models trained with sexist and abusive tweets was mitigated by debiased word embeddings, data augmentation and fine-tuning with a large corpus. Result concluded that debiased word embeddings alone do not effectively correct the bias, while fine-tuning bias with a less biased source dataset greatly improves the performance with drop-in bias.

3 Observations and Inferences

The biases that are much extensively experimented can be categorized as: fine-grained such as gender, racism, religion, and location; coarse-grained such as demographic and gender+occupation. Most experiments or analysis are performed on observing gender bias and its stereotypes from data and through word representations. The other types of bias, such as racism, sexism, and religion, are extensively studied in social data. The other biases from intersectional bias attributes such as demographic (age, gender, location), gender+racism, and gender+location+job are not extensively studied. Much attention is needed to study the other types and bias based on the intersectional view.

3.1 Bias in data / corpora

A standard process has to be defined on how data are collected, processed and organized. Gebru et al. (2020) recommended that every dataset should have datasheets that are metadata that documents its motivation, composition, collection process, recommended uses, and so on. Every training dataset should be accompanied by information on how the data were collected and annotated. If data contain information about people, then summary statistics on geography, gender, ethnicity and other demographic information should be provided. Specifically, steps should be taken to ensure that such datasets are diverse and do not represent any particular groups. The model performance may hinder if the dataset reflects unwanted biases or its deployment domain does not match its training or evaluation domain. Table 1 shows the analysis of

Labelling / Annotations	Resource/Task/Techniques	Work by
Bias in labelling	POS Tag, Dependency parse in WSJ Penn Treebank	(Garimella et al., 2019)
Bias in social data annotations	Inter-Annotator agreement for Twitter data collection	(Waseem, 2016),(Sap et al., 2019)
	Misleading label in financial tweets by annotators	(Chen et al., 2020)

Table 2: Analysis of bias in labelling / annotation

bias in data.

1. **Datasheets & Data Statements**: Datasheet for data should be documented well, to avoid the misuse of data sets for unmatched domains and tasks (Gebru et al., 2020). Bender and Friedman (2018) proposed data statements to be included in all NLP systems' publication and documentation. These data statements will address exclusion and bias in language technologies that do not misrepresent the users from others.

2. **D&I metric for corpora**: The metric to measure the language diversity and inclusivity is necessary for each corpus that helps to measure the language support by the corpora and its power to include poor-resource languages (Joshi et al., 2020).

3. **Data design – Schema templates**: The schema-level constructs such as WinoBias, Winogender for coreference systems, WinoMT for MT and EEC for sentiment analysis are used to measure and mitigate the bias for task / application-specific purposes. These schemas provide semantic cues in sentences apart from syntactic cues to detect the bias.

4. **Metric for subset selection** - The subset selection is a fundamental task that does not consider the metric to select the set of instances from a more extensive set. Mitchell et al. (2019) formalized how these subsets can be measured for diversity and inclusion for image datasets. There should be a similar kind of metrics that should be formulated for text datasets too.

5. **Bias through an Intersectional view** – Bias also emanates due to the cross-section of the gender and racism. Solomon et al. (2018) demands the study of bias through the intersection of race and gender, which would alleviate the problem experienced by Black Women. This is done by analysing multiple identities through an intersectional lens. Research shows that Black or women are not inclusive of Black women's unique perspectives.

Our recommendations based on the above:

- **Recommendation-1:** Research on various semantic-level constructs specific to applications is needed. Like data design practiced in database systems to capture the semantics of data using schema representations, schema-level constructs are needed for task-specific applications to mitigate bias.

- **Recommendation-2:** The need for Taxonomy of bias for LT w.r.t EDI – which focus other than binary gender such as transgender, LGBTQ, and bias through an intersectional view.

- **Recommendation-3:** Biases defined for linguistic resources can not be used or may unmatch for the data from other domains such as bio-medical, science and law. Need of domain-wise bias study as it may differ in its own perspectives.

3.2 Bias in corpus labelling or annotations

If the data labelling is done through crowdsourcing, then necessary information about the crowd participants should be included, alongside the exact request or instructions that they were given. To help identify sources of bias, annotators should systematically label the content of training data sets with standardised metadata. Table 2 shows the analysis of bias in labelling corpora, social data annotations.

1. **Is Social data fair enough?** – Some bias is inherent in social data, which are unique. Many researchers have warned

Representation / Embedding	Technique / Model	Work by
Commonly used word representations to measure bias	Word2Vec, GloVe, ELMo, PPMI, TF-IDF, Character n-grams	-
Bias in word representations	Word2Vec GloVe	(Bolukbasi et al., 2016) (Caliskan et al., 2017)
Bias among word representations	Word2Vec Vs. GloVe GloVe, Word2Vec Vs. ConceptNet	(Brunet et al., 2019) (Sweeney and Najafian, 2019)
Bias in bilingual embeddings	ES-EN, FR-EN	(Zhou et al., 2019)
Word representations to neutralize bias	HistWords – Embed on historic data GN-GloVe – Gender-neutral embed GeoSGLM – demographic embed	(Garg et al., 2018) (Zhao et al., 2018) (Escudé Font et al., 2019) (Hasanuzzaman et al., 2017)
Pre-trained LM	ELMo, BERT, GPT	(Park and Fung, 2017) (Escudé Font et al., 2019) (Zhao et al., 2019) (Liang et al., 2020)

Table 3: Analysis of bias in text representations

against the naive use of social data in NLP tasks (Jørgensen et al., 2015). Olteanu et al. (2019) categorised four biases that occur at the source level and how biases manifest themselves. Applications using social data must be validated and audited, effectively identifying biases introduced at various levels in a data pipeline. Other harmful blind spots along a data analysis pipeline further require better auditing and evaluation frameworks.

Our recommendation based on the above:

- **Recommendation-4:** The need for the study on bias in social data by NLP research community: Social data contains diverse dialects, which is not a norm in language resources. There are methodological limitations, and pitfalls (Jørgensen et al., 2015) as well as ethical boundaries and unexpected consequences (Olteanu et al., 2019) in social data. Hence the propagation and amplification of bias in the data pipeline of social data differ, which needs to be addressed.

3.3 Bias in text representations

Word embeddings contain bias and amplify bias present in the data, such as stereotypes. Word embeddings trained on different models produce different results. Bias in word embeddings is noted by (Bolukbasi et al., 2016) in Word2Vec, which represents a stereotypical bias among the pair of words using word analogy. Recently word embeddings specifically to neutralize the gender in embeddings – GN-GloVe, embeddings on demographic attributes – GeoSGLM are used to detect and mitigate bias in word representations. Table 3 shows the analysis of bias in text representations for monolingual, bilingual and embeddings to neutralize the bias.

1. **Limitations in Transformer Models** - Limitations of this models are noted in computational abilities (Hahn, 2020), multilingual embeddings produced by BERT (Singh et al., 2019), mBERT (Wu and Dredze, 2020), bilingual embeddings by MUSE (Zhou et al., 2019).

2. **Fairness in pre-trained LM** – The model cards proposed by Mitchell et al. (2019) are recommended to the pre-trained LM that can substantiate the context in which the models can be used and can provide a benchmarked evaluation on various bias types. Currently, the GPT-2 model card [2] have not mentioned the type of bias in the dataset used to train and the model fitness for specific applications.

3. **Is pre-trained LM bias-free?** – Dale (2021)

[2] https://github.com/openai/gpt-2/blob/master/model_card.md

Measure/Mitigate	Metric/Techniques	Work by
Metric proposed to evaluate bias	WEAT – for SGNS model	(Caliskan et al., 2017)
	Relative norm distance	(Garg et al., 2018)
	WEAT1, WEAT2 – bias b/w embeddings	(Brunet et al., 2019)
	SEAT – bias in sentence embeddings	(May et al., 2019)
	RIPA – for any embedding model	(Ethayarajh et al., 2019)
	MWEAT – bias in MT	(Zhou et al., 2019)
	RNSB – bias in Sentiment Analysis	(Sweeney and Najafian, 2019)
Debiasing techniques proposed	Hard-debias, Soft-debias	(Bolukbasi et al., 2016)
	RBA	(Zhao et al., 2017)
	Auxiliary dataset	(Zhao et al., 2018)
	Named Entity anonymization	
	Data Augmentation	(Zhao et al., 2019)
	Neutralization	
	SENT-DEBIAS	(Liang et al., 2020)

Table 4: Metric/Methods used to evaluate bias and debias

found that the pre-trained LM such as GPT-3 has the output embody all the biases that might be found in its training data.

4. **Transfer learning is ecological?**– Training the language models is costly to train financially and environmentally, which uses more carbon dioxide than the manufacturing and lifetime use of a car (Strubell et al., 2019).

Our recommendation and augmentation based on the above:

- **Recommendation-5:** A semantic-aware neural architecture to generate debiased embeddings for monolingual, cross-lingual and multi-lingual applications.
- **Augmentation-1:** We augment Gebru et al. (2020) to adopt and extend Datasheets for the language resources, annotated corpora and Model cards by Mitchell et al. (2019) for the algorithms used in pre-train LM and techniques for debiasing.

3.4 Measuring and mitigating bias

Since the plurality in languages inherently carries bias, it is essential to analyze the normative procedure's bias. Because of the complexity and types of bias, the detection and quantification of bias are not always possible by using formal mathematical techniques. Table 4 shows the metrics or methods used to measure and mitigate the bias.

1. **Fairness in bias metric** – Gonen and Goldberg (2019) observed that the bias emanating from word stereotypes and learned from the corpus is ingrained much more deeply in the word embeddings. Ethayarajh et al. (2019) found that the commonly used WEAT does not measure the bias appropriately for the embeddings other than the SGNS model. May et al. (2019) proposed SEAT to measure bias in sentence embeddings. This implies the need of a specific algorithm or method that can be used across all the embeddings to measure the bias (Gonen and Goldberg, 2019; Davidson et al., 2017).

2. **Application specific bias** – The bias can also be measured specific to the applications. For example, RNSB, TGBI (Cho et al., 2019) are proposed to measure the bias in Sentiment Analysis, MT, respectively.

3. **Fairness in debiasing** – The experiment on GN-GloVE and hard-debias reveals the presence of systemic bias in the embeddings independent of gender directions measure the gender association (Gonen and Goldberg, 2019). Even though many systems use hard-debias or soft-debias as de-facto standards to mitigate bias, they do not effectively remove it from the word representations. This requires a standardized framework that effectively measures and mitigate the bias across the domains and applications.

Our recommendation based on the above:

- **Recommendation-6:** An Unified and End-to-End Framework – there is a need for a unified

framework to measure and mitigate the bias based on benchmark metrics and methods at various levels data pipeline.

4 Conclusion

The above analysis at various levels reveals that most NLP systems consider bias at a nascent level. Computer scientists should strive to develop algorithms that are more robust to biases in the data. Various approaches are being pursued. Such debiasing approaches are promising, but they need to be refined and evaluated in the real world. All need to think about appropriate notions of fairness in data. Should the data be representative of the world as it is, or of a world that many would aspire to? To address these questions and evaluate the broader impact of training data and algorithms, machine-learning researchers must engage with social scientists, and experts in other domains. Based on the observations of methods used at various levels, we recommend six measures and augment one measure to support ethical practice and fairness in the usage of data. Practising and adopting various normative procedures across the data pipeline in NLP systems would enhance the Equality, Diversity and Inclusion of different subgroups of peoples and their languages.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

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

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6106–6110, Marseille, France. European Language Resources Association.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Robert Dale. 2021. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118.

T. Davidson, Dana Warmsley, M. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Markus Dickinson and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, page 107–114, USA. Association for Computational Linguistics.

Joel Escudé Font, Marta Costa-jussa, and R. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Alex B. Fine, Austin F. Frank, T. Florian Jaeger, and Benjamin Van Durme. 2014. Biases in predicting the human language model. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 7–12, Baltimore, Maryland. Association for Computational Linguistics.

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. *CoRR*, abs/1802.00393.

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):E3635–E3644.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. Datasheets for datasets.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic word embeddings for racism detection on Twitter. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 926–936, Taipei, Taiwan.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

Jason Jeffrey Jones, M. Amin, Jessica Kim, and S. Skiena. 2020. Stereotypical gender associations in language have decreased over time. *Sociological Science*, 7:1–35.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. 2011. Wp:clubhouse? an exploration of wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, page 1–10, New York, NY, USA. Association for Computing Machinery.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

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

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada. Association for Computational Linguistics.

Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. Quantifying 60 years of gender bias in biomedical research with word embeddings. In *Proceedings of*

the 19th SIGBioMed Workshop on Biomedical Language Processing, pages 1–13, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Katja Geertruida Schmahl, Tom Julian Viering, Stavros Makrodimitris, Arman Naseri Jahfari, David Tax, and Marco Loog. 2020. Is Wikipedia succeeding in reducing gender bias? assessing changes in gender bias in Wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 94–103, Online. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

A. Solomon, D. Moon, A. L. Roberts, and J. E. Gilbert. 2018. Not just black and not just a woman: Black women belonging in computing. In *2018 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, pages 1–5.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Claudia Wagner, D. García, M. Jadidi, and M. Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *ICWSM*.

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, Austin, Texas. Association for Computational Linguistics.

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, San Diego, California. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

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 the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.