Journal of Forensic Research and Criminal Investigation
JFRCI, 5(1): 160-165
www.scitcentral.com

ISSN: 2640-0846

Review & Perspective Article: Open Access

Natural Language Processing: Through the Looking Glass of Future Research

Eddie James Fraser*

[*]8 Feather Court Morayfield 4506 Queensland, Brisbane Australia.

Received March 04, 2024; Revised April 13, 2024; Accepted April 16, 2024

ABSTRACT

It is becoming commonplace for technology to comprehend text, language, and spoken and written words similarly to humans. However, we must address the biases that are present in natural language processing technologies. This paper intends to investigate these biases and provide recommendations for mitigating them and preventing them from occurring in the future. Although not thorough, this paper focuses on key features from the author's perspective.

Keywords: Natural language processing, AI, Artificial intelligence, Bias, Ablism, Sexism

INTRODUCTION

This paper heavily relies on the research conducted by Shrimai Prabhumoy and Dirk Hovy [1] in their publication "Five Sources of Bias in Natural Language Processing." The text analyses their impact on Natural Language Processing (NLP) and provides recommendations from the author on the prospective applications and opportunities of NLP in the future. Automated machine translation, chatbots, speech recognition, and sentiment analysis are domains that fall under this category. Nevertheless, this list is not comprehensive.

WHAT IS NATURAL LANGUAGE PROCESSING?

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence that focuses on providing computers with the ability to interpret text, language, and spoken and written words in the same manner as humans. Simply described, it is a field in which computers and artificial intelligence are used to analyze, derive meaning from, and comprehend human language. Like the application of human language, it is not an easy task to have something that can recognize human language, particularly within parts of speech such as grammar, usage exceptions, and variations in sentence structures, among many other competing factors. NLP can assist in this regard. Spell check, Siri, Alexa, Google Assistant, spam filters, and voice-to-text messaging are a few examples of NLP that we use every day. Positive applications of NLP are characterized by an ability to process vast quantities of language and data that are far beyond the human capacity to do so. This is an example of a potentially good application of NLP, as tasks that would take humans months or years to complete can be completed in significantly less time utilizing NLP [2].

NLP, ARTIFICIAL INTELLIGENCE, AND BIAS

These linguistic regularities may represent human biases such as ableism, sexism, and racism, to name a few. When analyzing artificial intelligence that captures linguistic regularities through collected data sets, it is possible that these linguistic regularities reflect human biases. There is no regulatory mechanism in place to audit and regulate potential threats to democracy, justice, and equity, but it is argued that such mechanisms should be considered, applied, and implemented as soon as possible. Moreover, it stands to reason that if datasets contain real natural language data that has not been altered or modified, and if data quality is used to enhance social group representation in algorithm analysis and how algorithms behave, then social group representation will be enhanced. This, in concert with the regulation of NLP, would satisfy the criterion for fairness regarding group traits when algorithms having the ability to make consequential choices (such as detecting whether a judicial judgement has aspects of prejudice) are used. It should be noted that the complexity of artificial intelligence and the flow-on effects of bias that have occurred have only begun to emerge in the last decade with the explosion of AI algorithms, social data, and computational power, and have

Corresponding author: Eddie James Fraser, 8 Feather Court Morayfield 4506 Queensland, Brisbane Australia, Tel: 0468617799; E-mail: eddiefraser91@gmail.com

Citation: Fraser EJ. (2024) Natural Language Processing: Through the Looking Glass of Future Research. J Forensic Res Criminal Investig, 5(1): 160-165.

become prevalent and accurately learned by NLP models that contain the language (that contains bias) that humans use; consequently, systems learn these historical patterns that contain these injustices and biases. In addition, it has been demonstrated in recent academic research that the use of any new technology has both intentional and unforeseen repercussions, and the application of NLP to any activity is no exception. In addition to bias, researchers in this field have identified the following unintended consequences: demographic bias, misidentification of speakers and their needs, and preconceptions. It is also noted that these biases occur in part because of the rapidly expanding field and the resulting incapacity and/or delays in the adaptation to new circumstances; this is in part due to the smaller data sets that were available at the onset of the concept, which has since been replaced by massive amounts of data and mass data production techniques. It is argued that NLP as it exists today is widespread and that its influence on people's lives and its prevalence in everyday life continue to grow [1]. It is interesting to notice that Hovy and Prabhumoye conclude that the emphasis has shifted from models as a tool for understanding to predictive models [1]. Clarity has been brought to bear on the fact that the tools produce excellent predictions, despite recognizing the difficulty in the analysis while completing the intended task and picking up on secondary aspects of language in the same breath, thereby opening the possibility of exploiting them to fulfil a perceived objective function [1]. The authors emphasize that language contains a great deal of secondary information about a speaker's membership, socio-demographic categories, and self-identification. Suppose I make a statement such as "This judicial decision is biased" or "This judge is friends with the prosecutor." What additional information about the speaker can be gleaned from the statement or sentence? In the context of a conversation, the active use of information can determine the potential age, gender, and regional or social class of the speaker. For example, in the first statement, I could infer (note that this is an inference and not a conclusion) that the speaker is an educated male with some legal knowledge, whereas in the second statement, I could infer that the speaker is a male from a low socioeconomic background with limited legal knowledge. It follows that NLP techniques fail to account for potential variations in language, such as demographic, cultural, and gender differences, and instead assume that all language adheres to a data-encoded norm [1]. This raises the question of who this "standard language" is, as well as the perception of bias: is its standard English or standard French? The response would depend on the tool's creator and the particular purpose being proposed.

BIAS

The notion of bias in this context involves fairness in algorithms by capturing or failing to catch hidden data properties, and it is claimed that bias in machine learning might be characterized as a potentially damaging property of the data [1]. Overgeneralization, issue exposure, and data modelling and research design concerning demographic bias are highlighted as three qualitative sources of bias [1]. An exhaustive survey of the manner in which bias is conducted revealed flaws in the design of research, which led to the recommendation of groundwork in the analysis of bias in NLP systems in an effort to comprehend why the behaviors within the system can be harmful and to whom they would be harmful, as well as engaging in conversations with the communities that are affected or involved by and by NLP systems 17]. At this juncture and in the following part, it would be beneficial to define some of these biases. Due to the limited scope of this work, this section is designed to provide background information and a first perspective but is not thorough. I examine five of the biases included in NLP tools. The purpose of this non-exhaustive list is to highlight to the reader that these should be considered when utilizing NLP tools to assist in research or its application as a research approach. These five biases are regarded as the most prevalent in NLP models. Importantly, these biases have the potential to exacerbate and contribute to the current gaps between users, which could have significant repercussions. The five sorts of bias that should be addressed the first is data bias and the selection of data. Secondly, annotation bias, which is labels used for training and techniques for utilizing and annotating, introduces selection bias. Selection bias is introduced by the samples picked for training or testing a particular NLP model. The third is that of input representational bias. The fourth, the choice of representation, such as the machine learning methods or the model employed, introduces the issue of model-based bias amplification. The fifth is bias from study design, which is the complete design process, which can lead to bias, especially when researchers choose an NLP model pipeline without due consideration [8]. There will be some insight into how this bias happens, as well as some solutions for mitigating it (**Figure 1).**

The figure above that was taken from the work of Hovy and Prabhumoye [1] is a schematic representation of the five bias sources that will be discussed in the following section of this paper.

A REVIEW OF FIVE BIASES

This section will now examine in greater detail these five biases, as well as their effects and mitigation in the context of NLP models. Data is the first of these.

DATA BIAS

It is hypothesized that a comparable sample of upper-middle-class, middle-aged, and college-educated men generate the majority of data sets derived from sources in a particular domain [1]. Modern syntactic analysis tools are still trained on newswire data from the 1980s and 1990s, with the result that this includes the language used by journalists at that

time. However, it is well known that language has evolved since then, and expressions that may not have been acceptable in those eras may be acceptable today due to the evolution of online and internet sources; therefore, it is proposed [1].

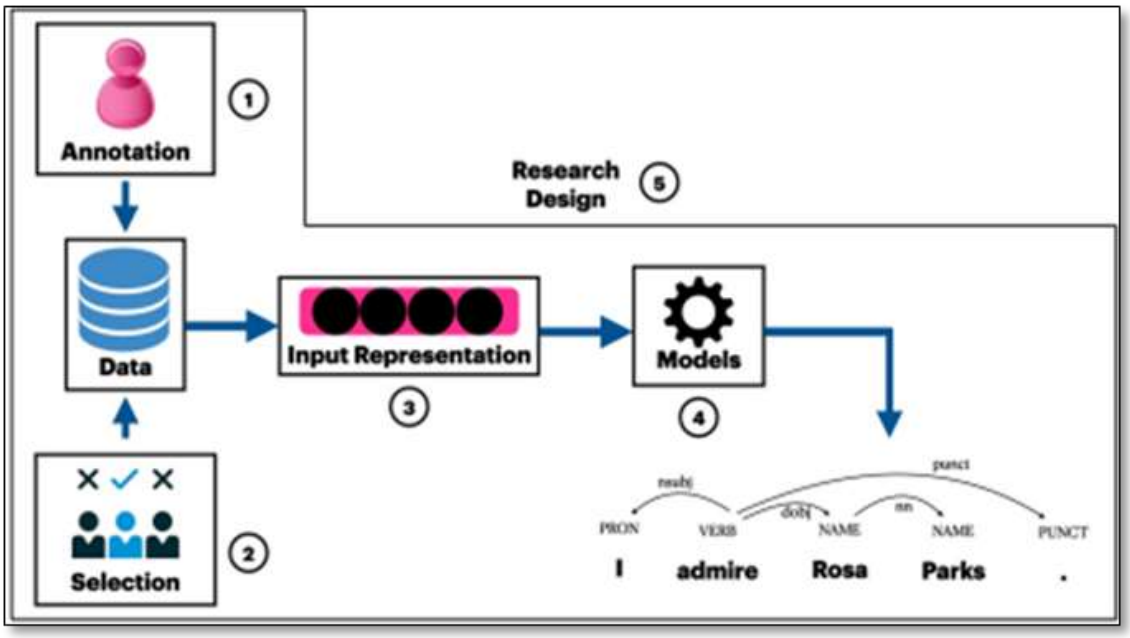

Figure 1. A Schematic Representation of The Five Bias Sources.

Even simple tasks, such as identifying nouns and verbs in speech tagging, give findings that are sexist, agist, or racist, as well as biased against the relevant user groups; this is known as selection bias, which is inherent in the data [1]. If, for instance, a person grew up speaking exclusively French as a dialect, it is not surprising that he or she may have difficulty understanding Hindi, given that data sets are subject to demographic bias. Most data sets will contain bias [1]. This is unavoidable; however, they are normally dormant and only become problematic when the data negatively affects certain groups or favors others disproportionately. Furthermore, within biased data sets, statistical models will contain specific linguistic signals that are unique to the dominant group, so they will not work as well for other groups [1]. Hovy and Prabhumoye note that measures to counter selection bias can be simple and that clarification and attention to what went into the construction of the data set, including the data collection process and underlying demographics, will provide future researchers with a method to evaluate the impact of any bias that becomes apparent when using that data [1]. It is noted that an additional benefit is gained by considering how that data is made up. The authors note that addressing data bias is not a one-time event, but rather a task that requires consistent and ongoing monitoring throughout the life cycle of data sets; alternatively, additional data can be collected to balance existing data sets to account for misrepresentation or exclusions in the data set used by the researcher [1].

ANNOTATION BIAS

Annotation bias can be introduced in numerous ways via annotator population data and is also known as label bias [1]. When evaluating how annotations create bias, it is because of the distracted, idle, or disinterested annotators who are assigned to the annotation task, resulting in the selection of the incorrect label. It is emphasized that while this may be true, the more troublesome bias is one that is created from well-intentioned annotators who systematically disagree, which occurs when more than one label is or is valid [1].

Hovy and Prabhumoye offer "social media" as an example of a lexicalized term that can generate an analytical result of a noun phrase made of an adjective and a noun or a noun compound composed of two nouns, depending on the annotator's interpretation of the lexicalized term [1]. To illustrate, if "social media" was viewed as fully lexicalized, the annotator would label it as a noun compound. However, if the annotator believed that the process was ongoing and the phrase was analytical, he or she would mark it as an adjective-noun compound [1]. With those opposing views in mind, the two annotators would systematically label “social” as an adjective or a noun respectively and in noting this whilst the disagreement is visible it cannot be said that either of those is either malicious or wrong [1]. A label bias can also result from author and annotator incongruence with their linguistic and social norms, as they reflect social and demographic differences, such as annotators rating utterances of different ethnic groups differently and harmless banter

being misinterpreted as hate speech due to the original speakers' unfamiliarity with communication norms [1]. There has been a shift towards using annotations from crowdsourcing rather than qualified experts, and while this is a cheaper alternative in principle, it has been found to induce a number of biases [1]. It has been demonstrated that the demographic composition of these crowdsourced annotators is not as representative as was previously believed. It might be stated that crowdsourcing is easier to scale and could cover more backgrounds than groups of professional annotators, but it is more difficult to connect with and teach crowdsourced annotators, and their incentives may not be aligned with the projects that one cares about [1]. Suppose, for example, that crowd workers are asked to annotate concepts such as dogmatism, microaggressions, or hate speech, with their responses including their own societal perspective of these concepts. The result would likely be multiple perspectives that are whole, or these annotations may have fragmented results due to the selection, which can produce either positive or negative bias depending on the selection [1]. The issue is that these many perspectives may not be what is desired if, for example, there is a theoretically well-defined and motivated way to annotate, and the concept of crowdsourcing and its related expenses pose numerous other ethical concerns [1]. It is noted that malicious annotators can be relatively easy to identify, with the use of multiple annotations per item and an annotation model to help identify biased annotators to account for human disagreement over a label, although this implies that there is a single correct label for human disagreement [1]. If more than one accurate response is conceivable, disagreement information should be used in the updating process of models by encouraging the models to adjust. If human annotators misunderstand categories, such as adjectives and nouns within a noun compound, as in the example of "social media," then regular updates should be made if they are mutually exclusive categories, such as verbs and nouns. Noting that the only way to address discrepancy in linguistic norms is to focus on the selection of annotators, such as by matching them to the author population in terms of linguistic norms, or by providing annotators with exclusive training, and that this position should be frequently considered despite the costly and time-intensive implications, this effort could result in less bias and better labels [1].

INPUT REPRESENTATION BIAS

Even well-labelled and balanced data sets include bias, and the most prevalent NLP system word embeddings have been proven to pick up on gender and racial prejudices, such as "man" being connected with "programmer" in the same manner that "woman" is associated with "homemaker" [1,3]. The authors argue that embeddings capture society opinions and that these societal biases are resistant to numerous corrective approaches, resulting in semantic prejudice [1]. It is observed that these biases apply not only to contextual representations of pretrained language models that are widely used in various NLP systems, but also to word embeddings, despite the fact that their widespread accessibility via the internet renders them more susceptible to societal biases [1]. The applicability of debiased embedding methods typically disguises or does not completely remove these biases, and it is noted that even if it were possible to remove the bias in embeddings, there is a lack of clarity regarding its central utility, as the issue is language model training objectives to predict the most probable following term given the previous one [1]. This hypothetical objective, while capturing semantic qualities, may not contribute to the development of impartial embedding since it maintains a normative view that represents the world as it is rather than a descriptive view that represents the world as we would like it to be [1]. A recommended practice while employing embeddings is to be aware of their biases, which aids in identifying the relevance of these embeddings in relation to the researcher's specific domain and activities [1]. As an example, such as using a model that is not directly applicable to data sets that contain legal articles or legal terminology such as “judicial objectivity”, “bias and the law”. In their research, Hovy and Prabhumoye observe that other techniques, data sets, and metrics have been presented for measuring the inherent bias in large language models and their sentence completion [1,4].

MODEL BIAS

As languages and language use continue to evolve, a representative sample may only be viewed as a snapshot or a temporary solution within a model. As a result, these biases cause startling performance discrepancies between the various user groups of these models [1]. Hovy and Prabhumoye note that it has been demonstrated that systems trained on biased data create an exacerbation of that bias when that data is applied to new data, and sentiment analysis tools raise societal prejudices that lead to different outcomes based on the different demographic groups, noting that the classification of the sentence differed simply by changing the gender of a pronoun [1,5].

It is noted that machine translation systems changed the perception of user demographics to make samples more male in translation which is known as bias overamplification which is essentially embedded in the models themselves [1]. Ultimately, this gives the correct results, but for the wrong reasons, and it is emphasized that this behavior is difficult to trace, that is, until a persistent instance of bias is identified [1]. In addition, the design of machine learning models is such that they always make a forecast. This occurs, for example, when the model is uncertain or when the answer is unknown. Hovy and Prabhumoye remark on the machine translation system case study "If a machine translation technology translates the gender-neutral Turkish phrase 'O bir doktor, o bir hemşire' into 'He is a doctor, she is a nurse,' it may reveal societal norms" [1]. In addition, it is mentioned

that this influences unintended outcomes and that, ideally, the models should tell the user that the translation could not be finished instead of predicting and producing the wrong translation. It is widely acknowledged and proposed that testing model systems on multiple data sets is preferable to focusing on a single designated test and moving away from pure performance metrics, as well as considering the robustness and conduct of a model in terms of how specially designed cases can provide additional insight and noting that the exploration of constraints to be specified on the model outcomes [1]. These restrictions will ensure that the proportion of anticipated labels for each user group are similar or approximately related [1]. Hovy and Prabhumoye admit that Kennedy and others propose using a sample-based approach to investigate the impact of individual words on a word's classification [1]. It is observed that if the number of explainable AI policy changes increases, the potential for such methods to become useful for legal redress and detecting bias will increase dramatically. Demographic explicit models will likely yield less biased and more personalized translations [1,6].

RESEARCH DESIGN BIAS

It is undeniable that the field of NLP is gaining popularity, particularly around cross-lingual and multilingual work. However, the majority of this research is still conducted in and on the English language, with a focus on Indo-European text and data sources rather than smaller language groups or other languages, despite the potential wealth of such data available from smaller language groups and languages [1]. As it stands, the majority of NLP tools are shaped around the English language, and as a result of this underexposure, researchers avoid working in areas where those languages may not have many resources. Instead, they choose to work in areas where data is readily available, which generates additional data in that language [1]. The universal dependency project aimed to address the lack of syntactic resources; however, research indicates that most conferences continue to prioritize well-resourced languages while excluding less-resourced languages. As a result, research tools for English are simplified and a heuristic bias is created due to overexposure [1,7].

Due to the unique structure of English, the only problem encountered with other languages is the n-gram or contiguous sequence of items from a sample of text or speech approach, which is particularly effective. The authors note that economic incentives to work on and in other languages have reignited interest in languages other than English, and that new neutral methods have made cross-lingual and multi-lingual method approaches much more feasible, including but not limited to zero-shot learning through multi-lingual representations [1]. The reality is that English remains the most frequently spoken language, and so represents the largest market for NLP tools. It follows that there are more commercial incentives to operate within the English language as opposed to other languages, thereby prolonging the presence of overexposure [1]. The composition of research groups, which does not necessarily mirror the demographic compositions of user bases, is cited as one of the causes of cultural and language inequality in research [1]. Consequently, the disenfranchised populations of speaker groups do not have proportional or equal representation. NLP programmes and models are beginning to address this issue, but it is evident that there is a great deal of potential for development in this area [1]. The authors note that a lack of analysis of the behavior of any model or failure to disclose can be harmful, even if the omission is not done with malice or ill will. It is noted that this is typically a result of the pressure to publish; one example given is that this would result in bias of not understanding the intended use of the model and, consequently, how those models can be misused, which ethics reviews and ethical consideration in NLP have sought to prevent [1]. It is well-known that in the social sciences and medical sciences, experiments and research require the approval of ethics committees revolving around the safety of the subjects and associated benefits, harms, and protection; however, not all of these categories are easily translated when considering NLP; however, considering some or all of these categories can aid in the framing of decision making within particular schools of philosophical thought [1]. It is accepted that there is no simple remedy to design bias, with the caveat that this prejudice may only become obvious in retrospect. However, it is evident that any metric or activity that affords the ability to reflect on a particular project can aid in overcoming inherent biases [1]. It has been suggested that specifically stating what language is being worked on (even if that language is English), so making overexposure bias more evident, might help ensure that results may not necessarily apply to all languages (even though it is perfectly acceptable to research) [1]. During the research design phase, it may also be beneficial to explore issues such as whether the findings would be valid and accepted in another language and whether the study would be done if the data were not easily accessible [1]. An evaluation Using various evaluation metrics and criteria to analyze any ethical flaws within a system or model can aid in the elimination of this form of bias by assessing the direction in which the research itself falls or feeds into an existing bias [1,8].

In addition, it is recommended that the prioritizing of equality and stakeholders from disadvantaged groups be considered in the use of NLP models and tools, and that the necessity for bias-aware methodology in NLP be reviewed and acknowledged [1]. Ultimately researchers need to be cognizant of the entire research design and consequent process, the annotation schemes or labelling procedures followed, the data sets chosen, the algorithms chosen for tasks, how they select to represent data and the evaluation methods used with respect to the automated system or model whilst globally considering the real-world application of their

work consciously deciding to assist marginalized community through technology [1,9].

CONCLUDING REMARKS ON NLP

This article has discussed the work of Hovy and Prabhumoye in relation to the usage of an NLP tool by scholars who continue to investigate and advance this discourse, but it is in no way restricted to this discourse. It has been a brief review of, in the opinion of Hovy and Prabhumoye, the most prevalent sources of bias in NLP models, focusing on research design factors. I have cautioned that these are not the only biases that must be considered and that there is a growing body of algorithmic and methodological approaches aimed at mitigating the biases that exist in all sources. The most challenging bias to address would be that of the research design because it demands the researcher to engage in self-examination and systematic analysis of their built-in predetermined blind spots and preconceived preconceptions. This is a subfield of linguistics that is under constant development. Future work on the grammar of judicial bias or other research topics around this may benefit from the use of NLP tools; but, as suggested, this should be done with caution and consideration for the possibility of bias. This is a recommendation for future researchers who wish to explore discourses in relation to a substantial body of work in their respective fields, whether to draw conclusions or recommendations in other areas that focus on using a similar method to the analysis of case law, statutes, legal data and material, and non-legal material alike.

REFERENCES

1. Hovy D, Prabhumoye S (2021) Five sources of bias in natural language processing. Lang Linguist Compass 15(8): e12432.
2. Bender E (2019) The BenderRule: On naming the languages we study and why it matters. The Gradient. Available online at: https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/
3. Dinan E, Fan A, Williams A, Urbanek J, Kiela D, et al. (2020) Queens are powerful too: Mitigating gender bias in dialogue generation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp: 8173-8188. Available online at: https://www.aclweb.org/anthology/2020.emnlp-main.656
4. Dixon L, Li J, Sorensen J, Thain N (2018) Measuring and mitigating unintended bias in text classification. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES 18: 67-73.
5. Goldstein DG, Gigerenzer G (2002) Models of ecological rationality: The recognition heuristic. Psychol Rev 109: 75-90.
6. Gonen H, Goldberg Y (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Long and Short Papers 1: 609-614. Available online at:

   https://www.aclweb.org/anthology/N19-1061
7. Hovy D (2015) Demographic factors improve classification performance. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Long Papers 1: 752-762. Available online at: https://www.aclweb.org/anthology/P15-1073
8. Hovy D, Spruit SL (2016) The social impact of natural language processing. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2: 591-598. Berlin, Germany: Association for Computational Linguistics. Available online at: https://www.aclweb.org/anthology/P16-2096
9. Munro R (2013) NLP for all languages. Idibon Blog. Available online at: http://idibon.com/nlp-for-all