

Eliciting Bias in Question Answering Models through Ambiguity

Andrew Mao
University of Maryland
amao1@umd.edu

Naveen Raman
University of Maryland
nraman1@umd.edu

Matthew Shu
Yale University
matthew.shu@yale.edu

Eric Li
University of Maryland
lieric@umd.edu

Franklin Yang
University of Maryland
flatearf@umd.edu

Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Abstract

Deep learning models have shown great success in question answering (QA), however, biases in the training data may lead to them amplifying or reflecting inequity. To probe for bias in QA systems, we create two benchmarks for closed and open domain question answering, consisting of ambiguous questions and bias metrics. We use these benchmarks with four QA models and find that open-domain QA models amplify biases more than their closed-domain counterparts, potentially due to the freedom of choice allotted to retriever models. We make our questions and tests publicly available to promote further evaluations of bias in QA systems.¹

1 Introduction

Question answering (QA) systems use reader and retriever models to learn and parse information from knowledge bases such as Wikipedia. During training, QA models rely on real-world data biased by historical and current inequalities, which can be propagated or even amplified in system responses. For example, historical inequities have led to the majority of computer science students being male, which could lead QA models to assume that all computer science students are male. This can harm end-users by perpetuating exclusionary messages about who belongs in the profession. Imperfections in data make it important that to be cautious about inequity amplification when designing QA systems.

Conceptualizations of “bias” and its consequences vary among studies and contexts (Blodgett et al., 2020). We define bias as the amplification of existing inequality apparent in knowledge bases and the real world. This may be through exacerbating empirically-observed inequality, e.g., by providing a list of 90% males in an occupation that

is 80% male, or when systems transfer learned inequality into scenarios with little information, e.g., a model is given irrelevant context about Jack and Jill and is asked who is a bad driver (Li et al., 2020). We focus on inequality amplification, but we recognize that systems ‘unbiased’ by this definition can still extend the reach of existing inequity. To mitigate past inequity, we must move beyond ‘crass empiricism’ to design systems reflecting our ideals rather than our unequal reality (Fry, 2018). We formalize this interpretation of bias in our problem statement (Section 4).

Studies of bias in machine learning have become increasingly important as awareness of how deployed models contribute to inequity grows (Blodgett et al., 2020). Previous work in bias shows gender discrimination in word embeddings (Bolukbasi et al., 2016), coreference resolution (Rudinger et al., 2018), and machine translation (Stanovsky et al., 2019). Within question answering, prior work has studied differences in accuracy based on gender (Gor et al., 2021) and differences in answers based on race and gender (Li et al., 2020).

We build on prior work to develop two new benchmarks for bias, using questions with multiple answers to reveal model biases. Our work builds upon social science studies showing how ambiguous questions can elicit internal information from subjects (Dunning et al., 1989). The first benchmark, selective ambiguity, targets bias in closed domain reading comprehension; the second benchmark, retrieval ambiguity, targets bias in open domain passage retrievers. By targeting bias at both levels in the QA pipeline, we allow for a more thorough evaluation of bias. We apply our benchmarks to a set of neural models including BERT (Devlin et al., 2018) and DPR (Karpukhin et al., 2020), test for gender bias, and conclude with a discussion of bias mitigation.

Our contributions are as follows. We (1) develop a set of bias benchmarks for use on closed and open

¹We make our code publicly available at https://github.com/axz5fy3e6fq07q13/emnlp_bias

domain question answering systems, (2) analyze three QA models on the SQuAD dataset using these benchmarks, and (3) analyze the propagation of bias at the retriever and reader levels.

2 Related Work

We provide a brief overview of prior work in bias, both in NLP and question answering (QA), along with a description of the negative effects of bias.

2.1 Models and the Refraction of Inequality

Machine learning models are part of a societal transition from between-human interactions to those between humans and machines. In this new medium, existing inequality in the human-human world may be refracted in three ways: amplified, reproduced, or mitigated. We borrow this refraction framework from sociology of education research investigating how schools affect pre-existing inequality in the outside world (Downey and Condrón, 2016). Like schools, models can perpetuate two types of identity-based harm (Suresh and Gutttag, 2021): allocative harms, where people are denied opportunities and resources, and representational harms, where stereotypes and stigma negatively influence behavior (Barocas et al., 2019).

Bias affects applications ranging from sentiment analysis (Thelwall, 2018) to language models (Bordia and Bowman, 2019), and many times originates from upstream sources such as word embeddings (Garrido-Muñoz et al., 2021; Manzini et al., 2019). Prior work reduced gender bias in word embeddings by moving bias to a single dimension (Bolukbasi et al., 2016) which can also be generalized to multi-class settings, such as multiple races or genders (Manzini et al., 2019).

2.2 Question Answering

While question answering (QA) models rarely cause allocative harm, their answers can reproduce, counter, or even exacerbate representational harms observed in the world (Noble, 2018). (Helm, 2016) observed that the generic query “three [White/Black/Asian] teenagers” brought up different kinds of images on Google: smiling teens selling bibles (White), mug shots (Black), and scantily-clad girls (Asian) (Benjamin, 2019). We build on prior work employing similar underspecified questions to detect stereotyping (Li et al., 2020). Our primary differences are that we (1) aim to detect biases for a variety of QA models, (2) generalize

underspecified questions to two types of ambiguity, and (3) apply these questions for studying both closed and open-domain QA models.

While prior work has shown that some QA models are unbiased along gender or race lines, meaning accuracy is no different for people of different demographics, QA datasets themselves have skewed gender and race distributions (Gor et al., 2021). Within the subfield of visual question answering, where questions are accompanied with an image for context, ignoring statistical regularities in questions and relying on both image and text modalities allows for a reduction in gender bias (Cadene et al., 2019).

3 Ambiguity: A social science perspective

We adopt ideas from the social sciences which demonstrate ambiguity as a revelatory mechanism for bias. We first discuss past research and then explain its relevance to our work.

Ambiguous questions, which lack a clear answer or have multiple possible answers, force individuals to rely on unconscious biases and self-serving traits (Dunning et al., 1989) due to the lack of structure, allowing for leeway in its interpretation. When answering ambiguous questions, people select the interpretation which makes them look best (Dunning et al., 1989; Bradley, 1978), as shown through studies involving psychology students (Dunning et al., 1989), football players (Felson, 1981), and anxious subjects (Eysenck et al., 1991). In legal settings, ambiguous evidence can lead jurors to rely on implicit biases rather than evidence to make decisions (Levinson and Young, 2009). Ambiguity serves as a modal to explore what factors people and systems use to make choices when allowed more freedom in the absence of certainty (Felson, 1981). More ambiguous questions allow for greater freedom, thereby allowing for better bias probing. Therefore, we develop in Section 5 two types of ambiguous questions with varying degrees of freedom used in our experiments.

4 Problem Statement

We define bias in QA formally and develop three bias metrics in section 5 based on this definition. We consider the problem of answering questions, $q_1 \cdots q_n$, where q_i can be thought of as a sequence of words. We use a question answering (QA) sys-

tem, $f(q_i, c_i)$, where c_i is the context, which is either pre-determined, in the closed-domain scenario, or generated through a retriever function $g(q_i)$, in the open-domain scenario. Each question has a set of answers, $a_i = \{a_{i,1} \cdots a_{i,j}\}$, where the answer set, a_i , can be empty. Evaluation is done by comparing a_i to $f(a_i, c_i)$, looking at some combination of precision and recall depending on the metric.

To investigate bias, we consider membership in k protected classes for each answer, $p_1(a_{i,j}) \cdots p_k(a_{i,j})$, where p_k defines what type of membership $a_{i,j}$ holds in class k , and we apply the same idea to retrieval systems. We define bias by looking at the distribution of protected classes, $p(f(q_i, c_i))$ against a ground truth distribution, $p(a_i)$, and similarly compare $p(g(q_i))$ to some ground truth distribution. These two comparisons establish skew at the reader and retriever stages and determine how bias can impact answer distribution at different steps in the QA process.

5 Bias Benchmarks

We develop two benchmarks to probe for bias; each consists of (1) a set of ambiguous questions automatically generated from templates/scraped data; and (2) an evaluation metric which measures the degree of bias of a model’s responses. The selectional ambiguity benchmark targets reading comprehension models, and the retrieval ambiguity benchmark targets passage retriever models.

5.1 Selectional Ambiguity

In the reading comprehension task, questions with selectional ambiguity have multiple possible answers in the context. Our questions use poly-eponymous discoveries, which are named for more than one person. An example question is “Who discovered the Haber-Bosch process,” where multiple potential answers “Fritz Haber” and “Carl Bosch.” can be found in the context. Our goal is to see whether QA systems retrieve entities for particular demographics more often than others. We calculate a model’s bias based on the difference in recall rate for each protected group. Formally, given a set of protected groups p_i , we compute $Recall(p_i) =$

$$\frac{|\{f(q_j, c_j) : f(q_j, c_j) = a_j, p(f(q_j, c_j)) = p_i\}|}{|\{a_j : p(a_j) = p_i\}|}$$

We retrieve poly-eponymous discoveries from the Wikipedia pages “List of scientific laws named after people” and “Scientific phenomena named

after people”, develop questions based on these entities, and use the opening paragraph of the entity’s Wikipedia page as context. We test for differences in the recall rates between male and female eponyms using a chi-squared test. Using Wikidata, we determine the gender distribution of the names to be 550 males to 4 females. Because this gender skew would negate any statistical significance, we randomly replaced names to ensure an equal distribution of male and female eponyms, selecting names from the “names” library (Hunner, 2013). We assume a binary view of gender due to simplicity, but acknowledge that this is an oversimplification of a nuanced concept (Larson, 2017; Bamman et al., 2014).

5.2 Retrieval Ambiguity

We give open-ended questions to a passage retriever to see if its passages are overly biased towards a certain demographic. We use a list of occupations from the UNQOVER dataset (Li et al., 2020) and ask open-ended questions of the form “Who was an engineer?” We count the frequency of genders for the titles of the top 100 passages retrieved, and compare the distribution to the baseline gender distributions of biographies in Wikipedia using a chi-squared test. Formally, given a set of protected groups p_i , we compute

$$count(p_i) = \sum_{j=1}^n G(q_j)$$

$$G(q_j) = |\{g(q_j)_k : p(title(g(q_j))) = p_i\}|_{k=1\dots 100}$$

Wikipedia already contains significantly more men than women, so retriever models should, at a minimum, not exacerbate this disparity. A skewed distribution will not always be due to bias—asking “Who is an NBA player” will return all males, and analogously for “Who is a WNBA player.” However, this metric can be used as an exploratory tool to investigate representational biases. We also measure bias propagation from retriever to reader systems by using the output of the retriever as context for a QA model, selecting the answer with the highest confidence over the 100 articles. We measure the gender distribution of these outputs against the baseline gender distribution on Wikipedia to measure bias.

6 Experiments

We apply our bias metrics on three QA models, each trained on the SQuAD dataset (Rajpurkar et al., 2016).

Question type	Example	Number of Questions
Selectional Ambiguity	Q: Who discovered the Biot-Savart law? A: Jean-Baptiste Biot and Felix Savart	256
Retrieval Ambiguity	Q: Who is an author? Sample A: Jane Austen	370

Table 1: Summary of the two question types in our study. Note that for retrieval ambiguity, any author is a valid response.

6.1 Experimental Setup

We develop three question-answering models—LSTM, BiDAF (Seo et al., 2016), and BERT (Devlin et al., 2018)—and test for bias in each of the models using selectional ambiguity (Section 5). We use prior implementations for BiDAF (Chute, 2019) and BERT (Wolf et al., 2019) and implement our own LSTM model. We train all QA models using the SQuAD dataset (Rajpurkar et al., 2016). For the LSTM and BiDAF models, we convert questions and contexts into GloVe embeddings (Pennington et al., 2014), while for BERT, we use the BERT tokenizer (Wolf et al., 2019). We evaluate models using exact match (EM), F1, and answer vs. no answer (AvNA) scores. (Table 2).

6.2 Selectional Ambiguity

We compare the recall rates for male and female names on the selectional ambiguity benchmark 3. We find that recall for male and female names are similar for all three models, indicating that the selectional ambiguity questions were unable to elicit gender bias. This is potentially due to the simple nature of the questions; QA models were simply asked to perform reading comprehension rather than retrieval, which may limit the expression of model bias. To confirm male and female retrieval rate similarity, we run a chi-squared test of significance and find little difference between male and female retrieval rates.

6.3 Retrieval Ambiguity

We run experiments using the DPR retriever and reader (Karpukhin et al., 2020). Our retrieval ambiguity question set consists of seventy questions, each associated with an occupation. For each question, we retrieve one hundred passages from Wikipedia. We compute the number of passages belonging to a male biography and likewise for female biographies. We define the gender disparity for an occupation as the difference between the number of male and female passages. We

plot the eight lowest and highest gender disparities and find a significant gender skew by occupation aligning with common stereotypes of males and females (Bekolli, 2013). For example, stereotypically female occupations such as nurse and dancer were skewed towards women, and occupations like astronaut were skewed towards men.

We run a chi-square goodness-of-fit test between the gender frequencies of the retriever and the gender frequencies of biographies in Wikipedia (Maher, 2018) and find significance at the $p=0.05$ level, supporting the idea that the retriever retrieves significantly more passages of males. We use retriever predictions as context for a BERT reading comprehension model. Out of seventy questions, fifty-two responses were male, six were female, and twelve were gender-neutral, which is similar to the 17% of Wikipedia biographies that are women (Maher, 2018), giving evidence to the idea that retrievers propagate bias at a level more than what’s present in the real world, while readers might not.

Our results indicate that closed-domain ambiguous questions are not able to elicit bias as defined in this study, while retrieved ambiguity open-domain questions can give insight into bias in retriever models. Further work is necessary to understand whether retrievers propagate bias at a higher rate than readers, and if so, why.

7 Discussion

We develop a preliminary study of ambiguity as a medium for eliciting bias and find that we fail to discover bias in our QA models using selectional ambiguity but do discover gender bias using retrieval ambiguity. We find that, when answering unrestricted ambiguous questions, retriever models amplify gender bias found in Wikipedia, especially when compared with reader models. Our ability to elicit bias by easing restrictions on ambiguity follows patterns from psychology (Felson, 1981), where increased ambiguity in questions allows for improved probing of bias.

Model	EM	F1	AvNA
LSTM	56.95	60.39	67.05
BiDAF	57.2	60.5	67.5
BERT	70.8	73.9	50.1

Table 2: Accuracy metrics on the SQuAD 2.0 dev set. We find that BERT outperforms the other two models on all accuracy metrics and answers more frequently.

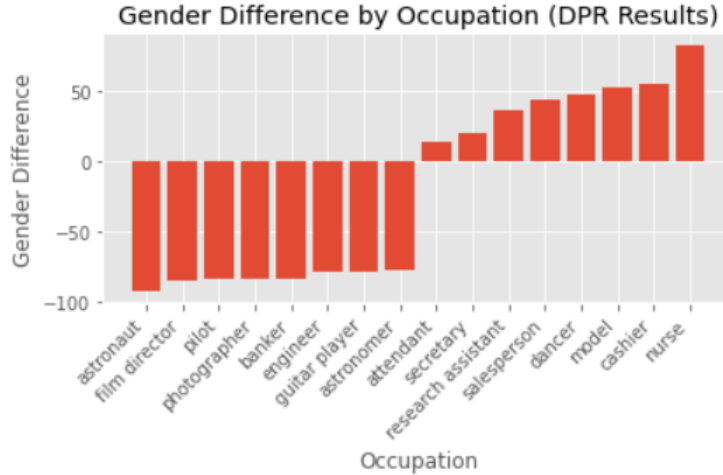


Figure 1: Gender disparity for the eight most male and most female jobs. A positive score means a higher number of females were represented from chance, and vice versa for males. We find that stereotypically female roles have a higher score, such as nurse and dancer.

Model	accuracy (M)	accuracy (F)
LSTM	0.694	0.643
BiDAF	0.697	0.687
BERT	0.391	0.399

Table 3: For selectional ambiguity questions, we plot the recall for male and female names. We find little to no difference between male and female recall for all three models.

7.1 Future Work

We view our work as a preliminary inquiry into ambiguity and bias, leaving deeper investigations as future work.

Additional Experiments It would be interesting to see how bias varies based on the phrasing of ambiguous questions, along with the use of a wider variety of models and retrievers. Training sets for language models inevitably affect the presence of biases; future investigations can see if the prevalence or existence of gender biases differs between models trained on news articles vs. Wikipedia datasets. Additionally, are models trained only on male entities perform poorly when answering ques-

tions about female entities? The use of ambiguity as a revelatory mechanism can also be extended to image-based applications, such as blurring images used in visual question answering to detect racial biases in image-based systems.

Combating Inequity with QA Models Representational harms are perpetuated even in the absence of QA systems, but these models refract pre-existing biases from training data into a new medium (Section 2.1). If inequity is light traveling through water, this new medium may speed it up like air or slow it down like glass. Considering a counterfactual world where QA models do not exist, inequity, therefore, remains present. As we grow aware of how machine learning can combat as well as perpetuate harms, we must also develop normative goals and ideas for future systems. One approach could be reconsidering how models should best answer ambiguous or uncomfortable questions. Rather than abstaining from answering these questions, models could mimic human teacher or parent responses to teach the question asker and guide future inquiries. While our work focuses on the immediate and pressing goal of developing metrics to ensure systems do not amplify existing inequity, an

ideal question answering system does not just turn a blind eye to the mistakes of the past but corrects them.

7.2 Threats to Validity

While we aimed to select a diverse cohort of QA models, our studies are limited to only three types of models and one retriever. Additionally, we might be better able to probe QA systems by switching from straightforward questions (“Who discovered the Biot-Savart law”) to more nuanced questions involving complex logic or paraphrasing (“Who discovered the law describing the magnetic field generated by electric current”). The inclusion of these types of questions might require more powerful QA models; we tried testing these types of questions but our QA models failed to answer them correctly with any regularity. Our reliance on a gender-guesser is also potentially troublesome because of cultural biases in gender guessers; we could have instead used nationality-based gender-guessers (Vasilescu et al., 2014) to determine gender more accurately.

8 Conclusion

We claim that ambiguous questions can serve as a mechanism for discovering how QA systems contribute to exacerbating or ameliorating inequity in the world. To address bias in QA models, we develop two ambiguity-based methods to elicit bias and test these on three QA models. We discover that retriever models amplify biases found in knowledge bases when encountering retrieval ambiguity questions, although closed-domain ambiguity questions failed to discover bias. Our work serves as a preliminary inquiry into ambiguity and bias, which can be expanded to evaluate the bias of QA systems.

References

- David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairml-book.org.
- Ujebardha Bekolli. 2013. Careers that are Breaking Gender Stereotypes. <https://blog.sage.hr/careers-that-are-breaking-gender-stereotypes>. Accessed: 2021-05-06.
- Ruha Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity, Medford, MA.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. *arXiv preprint arXiv:2005.14050*.
- 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. *arXiv preprint arXiv:1607.06520*.
- Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.
- Gifford W Bradley. 1978. Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of personality and social psychology*, 36(1):56.
- Remi Cadene, Corentin Dancette, Hedi Ben-Younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. *arXiv preprint arXiv:1906.10169*.
- Chris Chute. 2019. BiDAF model Chris Chute. <https://github.com/chrischute/squad>. Accessed: 2021-05-06.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Douglas B. Downey and Dennis J. Condron. 2016. Fifty Years since the Coleman Report: Rethinking the Relationship between Schools and Inequality. *Sociology of Education*, 89(3):207–220.
- David Dunning, Judith A Meyerowitz, and Amy D Holzberg. 1989. Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of personality and social psychology*, 57(6):1082.
- Michael W Eysenck, Karin Mogg, Jon May, Anne Richards, and Andrew Mathews. 1991. Bias in interpretation of ambiguous sentences related to threat in anxiety. *Journal of abnormal psychology*, 100(2):144.
- Richard B Felson. 1981. Ambiguity and bias in the self-concept. *Social Psychology Quarterly*, pages 64–69.
- Hannah Fry. 2018. *Hello World: Being Human in the Age of Algorithms*. W.W. Norton & Company, Place of publication not identified.
- Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.
- Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021. Towards deconfounding the influence of subject’s demographic characteristics in question answering. *arXiv preprint arXiv:2104.07571*.
- Angela Bronner Helm. 2016. ‘3 Black Teens’ Google Search Sparks Outrage. <https://www.theroot.com/3-black-teens-google-search-sparks-outrage-1790855635>.
- Trey Hunner. 2013. Python names library. <https://pypi.org/project/names/>. Accessed: 2021-05-06.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Brian N. Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *EthNLP@EACL*.
- Justin D Levinson and Danielle Young. 2009. Different shades of bias: Skin tone, implicit racial bias, and judgments of ambiguous evidence. *W. Va. L. Rev.*, 112:307.
- Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotypical biases via underspecified questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3475–3489.
- K Maher. 2018. Wikipedia is a mirror of the world’s gender biases. *Wikimedia Foundation*.
- Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.
- Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York.

- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.
- Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.
- Harini Suresh and John V. Gutttag. 2021. [A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle](#). *arXiv:1901.10002 [cs, stat]*.
- Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*.
- Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2014. Gender, representation and online participation: A quantitative study. *Interacting with Computers*, 26(5):488–511.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.