Social Impact of Natural Language Processing

Harshith Srinivas
Department of Computational Social Science
Paderborn University
33098 Paderborn, Germany
harshith@mail.uni-paderborn.de

Abstract

In this paper (Hovy and Spruit, 2016), we see How Natural Language Processing (NLP) has evolved significantly in the last couple of years by focusing on tasks involving statistical models which mostly involves anonymous corpora, with the goal of enriching linguistic analysis by development and deployment of larger Language Models, especially for English. BERT, its variants, GPT-2/3, and others

Paper also outlines several social implications of NLP and discuss their ethical significance, as well as ways to address them because the increase in usage of NLP in social media data can now have a direct impact on users lives. we outline how Larger Language Model plays role in social impact? By taking following into considerations: Environmental and Financial Costs, Unfathomable Training Data, Research Trajectories, Abusive Language and Synthetic Data. . . (Wohllebe, 2019)

1 Background

After the Nuremberg trials revealed the atrocities conducted by Nazis in medical sciences, International Review Board (IRB) was formed to incorporate the principles of biomedical ethics as a lingua franca of medical ethics (Beauchamp et al., 2001). This boards primary agenda was to prevent the direct exploitation on Human subjects. NLP and other Data Sciences have not been / less engaged in these discussions as IRB do not raise any Flag when working on existing corpora.

In another instance, when public outcry over the "emotional contagion" experiment on Facebook (Kramer et al., 2014) suggests that data sciences now affect human subjects in real time, and that we might have to reconsider the application of ethical considerations to our research (Puschmann and Bozdag, 2014).

The important ethical concern in data science till date is, 'privacy concerns' (Leverson et al., 2015) which also involves aspects like digital rights management/access control, policy making and security which is not concerned to NLP but has to be addressed in data science community as whole.

Authors in this paper believed that the field of ethics can contribute a more general framework and states the paper as an interdisciplinary collaboration between NLP and ethics researchers. To facilitate the discussion, author has also provided some of the relevant terminology from the literature on ethics of technology, namely the concepts of exclusion, over-generalization, bias confirmation, topic under- and overexposure, and dual-use problems.

2 Does NLP need Ethics?

When authors searched for 'Ethics' in Association for Computational Linguistics (ACL) anthology they found only three results, one of those papers (McEnery, 2002) turns out to be a panel discussion, another is a book review, and the final one was, who devote most of the discussion to legal and quality issues of data sets. Authors also got to know social implications which was addressed in some NLP curricula (Hector Mart ´ ´ınez Alonso, personal communication) with no practical rules. Main reason for this according to authors is that these technologies doesn't involve human-subjects(3Except for annotation: there are a number of papers on the status of crowdsource workers (Fort et al., 2011; Pavlick et al., 2014).Couillault et al. (2014) also briefly discuss annotators, but mainly in the context of quality control.) directly because earlier NLP was focused on enriching existing text which was not strongly linked to any author or human source(newswire). But due to increased use of Social Media data in recent times and used of NLP to improve the research where it can directly impact on individuals like traceability Couillault et al 2014 (i.e individual can be identified), discrimination e.g: minorities, gender, race ((Silverstein, 2003; Agha, 2005; Hovy and Johannsen, 2016) Johannsen et al., 2015), language is uttered in specific situation i.e. the texts we use in NLP carry latent information about the author and situation, albeit to varying degrees (Bamman et al. 2014). All these information is sufficient to predict individual or group characteristics from Text ((Rosenthal and McKeown, 2011; Ciot et al., 2013; Liu and Ruths, 2013; Plank and Hovy, 2015); Nguyen et al., 2011; Alowibdi et al., 2013; Volkova et al., 2014; Volkova et al., 2015; Preotiuc-Pietro et al., 2015a; Preot¸iuc-Pietro et al., 2015b), and these characteristics can be used in Language Models to influence them directly (Mandel et al., 2012; Volkova et al., 2013; Hovy, 2015). Due to development and use of these language-based technologies are increasing rapidly Authors in this research discipline urge to follow the importance of Ethical Implications in NLP research.

Language Model (LM):(Bender et al., 2021) (Bender and Koller, 2020) refers language model to systems which are trained on string prediction tasks. For example,what word comes——? what word [MASK] here?. That is, predicting the likelihood of a token (character, word or string) given either its preceding context or (in bidirectional and masked LMs) its surrounding context. Initially proposed by Shannon in 1949 (Shannon and Weaver, 1949), some of the earliest implemented LMs date to the early 1980s and were used as components in systems for automatic speech recognition (ASR), machine translation (MT), document classification, and more (Rosenfeld, 2000). Even since days of n-grams getting popular we have seen patterns achieving better score with increase data and increasing size of models until scores do not see the improvement and move to new architectures that can take advantage of increasingly large amount of data we have. With this increase in size, we also see the changes in types of tasks these LMs are used for like selecting among outputs of acoustical and translation models, LSTM-derived word vector was quickly replaced as efficient way to represent bag of words features in NLP tasks involving labelling and classification. Also, pre-trained LMs can be easily retrained s (few-shot, one-shot or even zero-shot learning) on small dataset to perform meaning-manipulative tasks like summarization, question-answering and the like. Author also show the rise in multilingual models that feed data from several language models into single language model. The idea behind this is using high resource language architecture as English to support low resource language architecture, recently around 100 languages were combined into single model leading to model-size reduction strategies like knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015) quantization (Shen et al., 2020; Zafrir et al., 2019), factorized embedding parameterization and cross-layer parameter sharing (Lan et al., 2019), and progressive module replacing (Cohn et al., 2020).
The Figure 1 below shows the recent trends in LMs training data-set size in gigabytes and number of parameter count. We see general trend starting from BERT (Devlin et al., 2018) in 2019 with few hundred million parameters up to recently in Switch-C (Fedus et al., 2021) in 2021 with trillion parameters and author expect this upper trend to continue.

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-Gen (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

Figure 1: Overview of recent large language models

Environmental and Financial Cost: Average person across the year is responsible for producing 5Tons of CO2 emissions per year whereas Strubell et al. benchmarked that training a Transformer(big) model (Vaswani et al., 2017) with neural architecture search produces 284Tons of CO2 emissions Training a single BERT base model (without hyperparameter tuning) on GPUs was estimated to require as much energy as a trans-American flight. Atilla Wohllebe, 2019 explains the CO2 emission per piece compared to a letter and e-mail (refer Figure 2 below).

Instrument		CO_2-emissions (Grams)
Letter		26
E-Mail	Standard	4
	With picture attached	50
	Spam	0.3
SMS		0.014

Figure 2: CO2 emissions of selected communication instruments at a glance (based on Selfmailer (n.d.), McAfee (2009),Goncalves (2009))

Strubell et al. also examine the cost of these LMs based on their accuracy gains (BLEU-score). For the task of Machine Translation authors estimate that an increase in 0.1 BLUE-score (English to German Translation) would result in increase of $ 150,000 in terms of computation cost, again this is inclusive of CO2 emissions.

Several recommendations are presented by authors for encouraging more equitable access to NLP research and reduce carbon footprint by retraining the model for downstream use to reduce training time and hyper parameter sensitivity. Authors mentions that government investment in computing clouds ensures that researchers have equitable access.

So, we must ask ourselves which researchers and which languages get to 'play' in this space and who is cut out?

Current Mitigation Efforts: Renewable energy sources are potential cost mitigation strategy but will still incur cost inform of infrastructure. For example: Trees are cleared for wind farms (https://www.heraldscotland.com/news/18270734.14m-trees-cut-scotland-make-waywind-farms/).

Another strategy which author mentioned is to prioritize computational efficient hardware and algorithms through SustainNLP (https://sites.google.com/view/sustainlp2020/organization) workshops and Schwartz et al. (Schwartz et al.) also encourages to promote Green-AI initiatives (Amodei and Hernandez).

In the sample papers from ACL NLP conference in 2018 and 2019 found that most research was concerned to accuracy improvements as primary contribution and none focused on efficiency improvement as primary since then works like (Henderson et al., 2020; Lottick et al., 2019) have produced online tools to help researchers to benchmark energy usage.

Author also outlines who is involved in these costs. Large Language Models, particularly those in English language and other high-resource languages leaving in big cities are ones who is benefiting more but marginalized communities (Adam et al., 2001; Bullard, 1993) around the world are most likely to face negative impact by climate change (Anthoff et al., 2010; Atallah et al., 2002) but these communities are rarely able to see the benefits of these larger LMs as it is not developed to support these regional languages (we do not intend to erase existing work on low-resource languages. One particularly exciting example is the Masakhane project (Nekoto et al., 2020), which explores participatory research techniques for developing MT for African languages. These promising directions do not involve amassing terabytes of data).

3 The social impact of NLP research

Authors state there are also societal impact factors of NLP arising from the interaction between language, society, and individuals: failing to recognize group membership (see Section), implying the wrong group membership (see Section), and overexposure (see Section). The following discussion talks about the sources of these problems in data, modeling, and research design, along with possible solutions.

3.1 Exclusion

As we saw that Language uttered in specific situation (Language is situated) Bamman et al, 2014 carries demographic information (i.e latent information). Overfitting due to this demographic bias in training data is caused by the i.i.d. assumption (model assumes all languages in sample data to be identical), which can result in training models that perform worse or fail entirely on data with different demographics.

A study applying demographic bias to training data will result in the exclusion or misrepresentation of demographic data, and this presents an ethical challenge for the conduct of research, thus threatening the objectivity and universality of scientific knowledge. For an instance, in some cases standard language technology is easier to use for white males from California (since they are considered when developing it) than for women or citizens of Latino or Arabic descent which in-turn reinforce the existing demographic bias making technology biased to specific individuals or groups.

Researchers have recently highlighted the effects of exclusion on NLP research, exemplified by Hovy and Søgaard (2015) and Jørgensen et al (2015): Compared to modeled demographics, state-of-the-art NLP models are significantly less accurate for young people and ethnic minorities. Creating awareness of these problems can help in preventing the problem of Exclusion.

Potential counter measures to demographic biased information are to downsampling the overrepresente d training data to even out distribution. Also, Mohammady and Culotta (2014) shows another approach using existing demographic statistics for supervising the later. In general, overfitting and imbalancing training data can be used to reduce demographic bias.

3.2 Overgeneralisation

In the previous section (Exclusion) we understood side-effect of Data. Now in this section we see modelling effect of data.

Consider, for instance, the automatic inference of attribute values of users, an interesting and common task in NLP, whose solution can also be used in many useful applications, such as recommendation engines and fraud or deception detection (Badaskar et al., 2008; Fornaciari and Poesio, 2014; Ott et al., 2011; Banerjee et al., 2014).

When cost of False-Positive is low may lead to bias confirmation and overgeneralisation. For an instance, consider a situation where you receive an e-mail conveying your retirement wishes on your 25th birthday. Here model used right training data but wrong target variable. Which arise a question "would a false answer be worse than no answer?". In this case we can handle the impact if models learn from rejection, introducing dummy variables, modelling the regularization, cost sensitive learning and varying of confidence thresholds.

3.3 Unfathomable Traing Data

It is easy to imagine that because the Internet is a large and diverse virtual space, datasets such as CommonCrawl must be broadly representative of the ways in which different people view the world. Upon closer examination, we see that there are several factors that limit Internet participation, limit the discussions which will be included by the crawling methodology, and finally limit the texts that will most likely be included after the crawled data has been filtered. In all cases, the voices of people most likely to hew to a hegemonic viewpoint are also more likely to be retained.

Author also talks about who has access to the internet and who is contributing to these discussions. And they found it was younger people (2018).; the Internet. , 2018) from more developed-cities around world contribute most. For an instance, GPT-2's training data is sourced by scraping outbound links from Reddit, and Pew Internet Research's 2016 survey reveals 67 % of Reddit users in the United States are men, and 64 % between ages 18 and 29.13 Similarly, recent surveys of Wikipedians find that only 8.8–15 % are women or girls (Barera, 2020).Although people who feel unwelcome in mainstream sites may set up different communication channels, these may be less likely to be included in language modeling training data. Take, for example, older adults in the US and UK. It was Lazar et al. who outlined how they individually and collectively articulate anti-ageist frames specifically through blogging (Lazar et al., 2017), which older adults prefer to more popular social media sites for discussing sensitive topics (Liu et al., 2019). Discussions in these forums often revolve around what constitutes age discrimination and its impact. However, a blogging community such as the one described by Lazar et al. is less likely to be found than other blogs that have more incoming and outgoing links. Jones (twi) documents (using digital ethnography techniques (Jones, 2020)) mentions another instance where Twitter moderation practices result in more accounts of users receiving death threats being suspended than those issuing death threats, leading to a decrease in participation among users from marginalised groups.

Finally while talking about current practices in filtering data, Colossal Clean Crawled Corpus (Raffel et al., 2019), used to train a trillion parameter LM in (Fedus et al., 2021), is cleaned, inter alia, by discarding any page containing one of a list of about 400 "Dirty, Naughty, Obscene or Otherwise Bad Words" (Bender and Koller, 2020, p.6). This list is overwhelmingly words related to sex, with a handful of racial slurs and words related to white supremacy (e.g. swastika, white power) included. While possibly effective at removing documents containing pornography (and the associated problematic stereotypes encoded in the language of such sites (Speer, 2017)) and certain kinds of hate speech, this approach will also undoubtedly attenuate, by suppressing such words as twink, the influence of online spaces built by and for LGBTQ people (Benjamin, 2019). If we filter out the discourse of marginalized populations, we fail to provide training data that reclaims slurs and otherwise describes marginalized identities in a positive light.

Static Data/Changing Social Views: Language Models run the risk of 'value stock', relying older, less-inclussive understandings.For instance, the Black Lives Matter movement (BLM) influenced Wikipedia article generation and editing such that, as the BLM movement grew, articles covering shootings of Black people increased in coverage and were generated with reduced latency (Twyman et al., 2017). Importantly, articles describing past shootings and incidents of police brutality were created and updated as articles for new events were created, reflecting how social movements make connections between events in time to form cohesive narratives (Polletta, 1998). More generally, Twyman et al. highlight how social movements actively influence framings and reframings of minority narratives.

Encoding Bias: Now we got to know that training data over-represent hegemonic views and also is subjected to biases (Blodgett et al 2020). Documentation of problem is important first step but not the possible solution. First, model auditing techniques typically rely on automated systems for measuring sentiment, toxicity, or novel metrics such as 'regard' to measure attitudes towards a specific demographic group (Sheng et al., 2019). But these systems themselves may not be reliable means of measuring the toxicity of text generated by LMs. For example, Studies of the Perspective API model have revealed a stronger link between toxicity and the identification of marginalized and specific groups in a sentence (Hutchinson et al., 2020; Prabhakaran et al., 2019). Second, auditing an LM for biases requires an a priori understanding of what social categories might be salient. The works cited above generally start from US protected attributes such as race and gender (as understood within the US). But, of course, protected attributes are not the only identity characteristics that can be subject to bias or discrimination, and the salient identity characteristics and expressions of bias are also culture-bound (Fiske, 2017; Sczesny et al., 2004). To be effective, components like toxicity classifiers need culturally relevant training data, and even then, if we don't know what to audit, we may miss marginalized identities.

Curation,Documentation and Accountability: Larger Language Models, we get a question "how big is too big?".It is not about exact size but practices of curation, documentation and accountability (Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019). Author also recommends to fix a budget for documentation at start of project and also collect only essential data which can be documented with available resources.The purpose of this documentation is to understand sources of bias and potential mitigating strategies because when we rely on ever larger datasets we risk incurring documentation debt, 18 (An undocumented dataset that is both too large and too undocumented to be documented by post-hoc methods.)

4 The Problem of Exposure

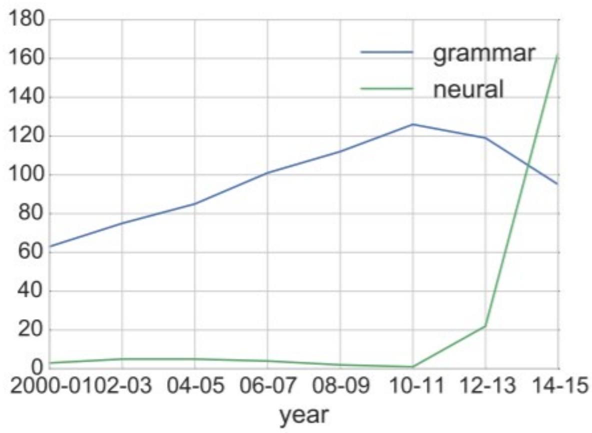

Figure 3: ACL title keywords over time

It is possible to address both exclusion and overgeneralization algorithmically, while topic overexposure resulted from research design, we can observe this effect in waves of research topics that receive increased mainstream attention, often to fall out of fashion or become more specialized, cf. ACL papers with "grammars" versus "neural" in the title (Figure 3). This kind of topic exposure leads to psychological statement called 'availability heuristics' (Tversky and Kahneman, 1973): things individual or groups familiar with, seem much more important than unfamiliar stuff. For an instance Farmer protest in India, Me-too movement worldwide are associated to specific groups or individual, the available heuristics become ethically charged when characteristics such as violence or negative emotions are more strongly associated with certain groups or ethnicities (Slovic et al., 2007). As a result, overexposure can result in biases that can in-

fluence decisions. In some ways, the frantic public discussion about AI risks is a result of overexposure (Sunstein, 2004). Hence, there are no easy solutions to this problem, which might only become apparent in hindsight.
Underexposure have negative impact on evaluation. Similar to Western, Educated, Industrialized, Rich, and Democratic research participants (so called WEIRD people) Henrich et al. (2010) in psychology. NLP tend to focus on INDO-EURO dataset sources rather than sources from small languages which creates imbalancing in available labelled data. Author also found that most of existing labelled data has very less set of languages or only English (majority) (Schnoebelen, 2013; Munro, 2013) resulting in low typological variety: both morphology and syntax of English are global outliers. When analyzing a random sample of Twitter data from 2013, we found that there were no treebanks for 11 of the 31 most frequent languages, and even fewer semantically annotated resources (the ACE corpus covers only English, Arabic, Chinese, and Spanish)(Thanks to Barbara Plank for the analysis!).
In order to develop NLP tools that can detect language outliers there are many approaches (Yarowsky and Ngai, 2001; Dasand Petrov, 2011; Søgaard, 2011; Søgaard et al., 2015; Agic et al., 2015). Research on other languages may be discouraged due to the need to develop basic models for them, so researchers are less likely to pursue them (other than English).

5 Research Trajectories

Author focus on fact that research time is very valuable resource and it is perhaps over allocated towards LMs and using them to achieve state-of-art scores on leader-boards particularly around Natural Language Understanding (NLU) tasks. But LMs have been shown when they do well due to spurious (Le Bras et al., 2020; Niven and Kao, 2019) dataset artifacts (Niver & Karo 2019,Bras et al 2020).Bender & Koller 2020 argue from a theoretical perspective, languages are systems of signs (Saussure et al., 1959), i.e. pairings of form and meaning. But the training data for LMs is only form; they do not have access to meaning.

6 Dual use of Problems

Text classification can detect slang or hidden message (Huang et al., 2013) but also have potential to be used for censorship. At same time NLP techniques can be used to detect fake news and also generate them in first place is recently shown by Hovy (2016).
In light of these examples, we should be more aware of how others use NLP technology. Despite the unprecedented scale and availability of NLP technologies, it is difficult to know what the consequences will be. Despite the fact that this decision is left to each individual researcher, the examples show that moral considerations extend beyond the immediate research project. In spite of not directly being held accountable for unintended consequences of our research, we can acknowledge the way in which NLP can enable morally questionable/sensitive practices, raise awareness, and inform the discussion.

7 Stochastic

From Linguistics and psychology we know that human-human interaction is co-constructed and leads to a shared model of world (Reddy 1979 and clark 1996). But a LM is system for haphazardly stitching together linguistic forms from its vast training data, without any reference to meaning - Stochastic Parrot.
We say seemingly coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs and intentions within context (Clark et al., 1983).

8 Potential Harms

Author states that if reader encounter a synthetic text that has got more hate-speech can experience stereotype threats or direct negative psychological impact: can boost extrimist recruiting (McGuffle & Newshouse 2020)on message boards.LMs can be probed to replicate training data for personal identification information (Carlini et al 2020).Also, Noble 2018 states that LMs can also be used as hidden components, for an example in internet search systems to influence to results without user attention which can again lead to many discrimination's.

From Linguistics and psychology we know that human-human interaction is co-constructed and leads to a shared model of world (Reddy 1979 and clark 1996). But a LM is system for haphazardly stitching together linguistic forms from its vast training data, without any reference to meaning -

Stochastic Parrot.
We say seemingly coherent because coherence is in fact in the eye of the beholder. Our human understanding of coherence derives from our ability to recognize interlocutors' beliefs and intentions within context (Clark et al., 1983).

9 Conclusion

In this paper, we outlined the potential social effects of NLP, and suggested ways for practitioners to address this. We also introduced exclusion, overgeneralization, bias confirmation, topic overexposure, and dual use and countermeasures for same. Also we discussed what is Language model and how NLP has characterised the usage of NLP in last few years and how it is impacted on environment and there cost of development. We also discussed how training data will help to improve state-of-art scores both in-terms for accuracy and ethical aspects. Finally we ended with discussing how big is too big for language model and there potential risks and harms.

Acknowledgments

The paper was published by seven authors, but some were required to remove their names by their employers. The remaining listed authors are extremely grateful to our colleagues for the effort and wisdom they contributed to this paper. In addition, in drafting and revising this paper, we benefited from thoughtful comments and discussion from many people: Alex Hanna, Amandalynne Paullada, Ben Hutchinson, Ben Packer, Brendan O'Connor, Dan Jurafsky, Ehud Reiter, Emma Strubell, Emily Denton, Gina-Anne Levow, Iason Gabriel, Jack Clark, Kristen Howell, Lucy Vasserman, Maarten Sap, Mark Díaz, Miles Brundage, Nick Doiron, Rob Munro, Roel Dobbe, Samy Bengio, Suchin Gururangan, Vinodkumar Prabhakaran, William Agnew, William Isaac, and Yejin Choi and our anonymous reviewers. epresents the work of seven authors, but some were required by their employer to remove their names. The remaining listed authors are extremely grateful to our colleagues for the effort and wisdom they contributed to this paper. In addition, this paper has benefited from thoughtful comments and discussion from many people: Alex Hanna, Amandalynne Paullada, Ben Hutchinson, Ben Packer, Brendan O'Connor, Dan Jurafsky, Ehud Reiter, Emma Strubell, Emily Denton, Gina-Anne Levow, Iason Gabriel, Jack Clark, Kristen Howell, Lucy Vasserman, Maarten Sap, Mark Díaz, Miles Brundage, Nick Doiron, Rob Munro, Roel Dobbe, Samy Bengio, Suchin Gururangan, Vinodkumar Prabhakaran, William Agnew, William Isaac, and Yejin Choi and our anonymous reviewers.

The authors would like to thank Joel Tetrault, Rachel Tatman, Joel C. Wallenberg, the members of the COASTAL group, and the anonymous reviewers for their detailed and invaluable feedback. The first author was funded under the ERC Starting Grant LOWLANDS No. 313695. The second author was funded by the Netherlands Organization for Scientific Research under grant number 016.114.625.

References

Pew. 2018. Internet/Broadband Fact Sheet. (2 2018).

Hussein M Adam, Robert D Bullard, and Elizabeth Bell. 2001. *Faces of environmental racism: Confronting issues of global justice*. Rowman & Littlefield.

Asif Agha. 2005. Voice, footing, enregisterment. *Journal of linguistic anthropology*, 15(1):38–59.

Dario Amodei and Danny Hernandez. Ai and compute, may 2018. *URL https://openai. com/blog/ai-and-compute*.

David Anthoff, Robert J Nicholls, and Richard SJ Tol. 2010. The economic impact of substantial sea-level rise. *Mitigation and Adaptation Strategies for Global Change*, 15(4):321–335.

Mikhail J Atallah, Victor Raskin, Christian F Hempelmann, Mercan Karahan, Radu Sion, Umut Topkara, and Katrina E Triezenberg. 2002. Natural language watermarking and tamperproofing. In *International workshop on information hiding*, pages 196–212. Springer.

Michael Barera. 2020. Mind the gap: Addressing structural equity and inclusion on wikipedia.

Tom L Beauchamp, James F Childress, et al. 2001. *Principles of biomedical ethics*. Oxford University Press, USA.

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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces*.

Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Robert D Bullard. 1993. *Confronting environmental racism: Voices from the grassroots*. South End Press.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

Herbert H Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2):245–258.

Trevor Cohn, Yulan He, and Yang Liu. 2020. Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

Susan T Fiske. 2017. Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on psychological science*, 12(5):791–799.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Dirk Hovy and Anders Johannsen. 2016. Exploring language variation across europe-a web-based tool for computational sociolinguistics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2986–2989.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

World Bank. 2018. Indiviuals Using the Internet. (2018).

Leslie Kay Jones. 2020. # blacklivesmatter: An analysis of the movement as social drama. *Humanity & Society*, 44(1):92–110.

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Amanda Lazar, Mark Diaz, Robin Brewer, Chelsea Kim, and Anne Marie Piper. 2017. Going gray, failure to hire, and the ick factor: Analyzing how older bloggers talk about ageism. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 655–668.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

JD Leverson, H Zhang, J Chen, SK Tahir, DC Phillips, J Xue, P Nimmer, S Jin, M Smith, Y Xiao, et al. 2015. Potent and selective small-molecule mcl-1 inhibitors demonstrate on-target cancer cell killing activity as single agents and in combination with abt-263 (navitoclax). *Cell death & disease*, 6(1):e1590–e1590.

Wendy Liu and Derek Ruths. 2013. What's in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kadan Lottick, Silvia Susai, Sorelle A Friedler, and Jonathan P Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability. *arXiv preprint arXiv:1911.08354*.

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*, pages 220–229.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 92–98.

Francesca Polletta. 1998. Contending stories: Narrative in social movements. *Qualitative sociology*, 21(4):419–446.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*.

Cornelius Puschmann and Engin Bozdag. 2014. Staking out the unclear ethical terrain of online social experiments. *Internet Policy Review*, 3(4):1–15.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Ronald Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772.

F de Saussure, C Bally, and A Seschehaye. 1959. Course in general linguistics, philosophical library. *New York*.

R Schwartz, J Dodge, NA Smith, and O Etzioni. Green ai. arxiv 2019. *arXiv preprint arXiv:1907.10597*.

Sabine Sczesny, Janine Bosak, Daniel Neff, and Birgit Schyns. 2004. Gender stereotypes and the attribution of leadership traits: A cross-cultural comparison. *Sex roles*, 51(11-12):631–645.

Claude E Shannon and Warren Weaver. 1949. The mathematical theory of com-munication. *Urbana: University of Illinois Press*, 96.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8815–8821.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & communication*, 23(3-4):193–229.

Rob Speer. 2017. Conceptnet numberbatch 17.04: better, less-stereotyped word vectors. *ConceptNet blog, April*, 24.

Marlon Twyman, Brian C Keegan, and Aaron Shaw. 2017. Black lives matter in wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pages 1400–1412.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Atilla Wohllebe. 2019. Dialogue marketing: Ecological sustainability of letter and e-mail in comparison in germany. *Journal of Environmental Sustainability*, 7(1):4.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. *arXiv preprint arXiv:1910.06188*.