CS4248 Natural Language Processing
Assignment 4: NLP Ethics

Distributed on 12 Apr 2021
Due in LumiNUS `Files` by 22 Apr 2021 11:59 PM SGT

This assignment contributes 5 marks towards your final mark for the class, and is graded out of a rubric of 100 points.

Integrity Note. *Since this assignment is similar to other assignments for Natural Language Processing courses at other institutions, there are (undoubtedly) solutions posted somewhere. Under the NUS Code of Conduct, you must follow class policy in working on this individual assignment. When in doubt of whether an action would constitute a violation of policy, please ask us on* `Slack` *on the* `general` *channel or by private Direct Message to Min,* **before** *attempting the action.*

Acknowledgements. This assignment is adapted from lists of readings from Prof. Emily Bender of the University of Washington, whose permission was explicitly secured for our course to use this assignment.

Welcome to the last assignment in CS4248! At the end of this module, we'd like to reinforce that ethics has everything to do with our development of technology. As natural language processing (and other AI systems) improve, it becomes increasingly important that practitioners exercise appropriate judgement in defining the technology, as the gap between practitioners and policymakers and legislation will continue to increase.

We hope you'll reflect on this issue of ethics that we bring up here during your future work and study, as the systems we build influence ourselves and our fellow peoples. By right, these activities are better discussed and shared among the community to engender discussion, so you are welcomed and encouraged to discuss your issues further on Slack in the `#assignment-4` channel.

1 Essay Question

In this final lecture, we focused on some of the dangers specific to natural language processing application in the real world, and elaborated on in the subsequent *Readings* section below. These include:

- Embeddings and Language Behavior as Truth
- Exclusion / Discrimination and Bias
- Language Variation and Emergent Bias

1. Write a short, 1–2 page (500–1000 word) essay, discussing an application scenario of natural language processing (NLP) of your choice, drawn from i) a case study, ii) your project work, or iii) an academic paper, where one of the ethical scenarios discussed in lecture applies. In the discussion of the scenario of your choice, link in a discussion of (at least) one of the papers in the *Readings* and how it applies to your scenario. You are encouraged to address some of the reading questions applicable to each of the areas below. As the areas are somewhat overlapping, you may want to survey the other areas' readings and questions as well. Fit your response into a coherent narrative.

   Grading Rubric. Grading will be based on your essay's: clarity of writing (30%), thoughtfulness (40%), connection to literature (20%), and proper English use and submission formatting (10%). Submissions must be self-contained to a maximum of 2 pages and 1000 words. Where appropriate, you may include references on an additional page, along with your *Declaration of Independent Work*.

2 Readings

The readings and discussion questions below are non-exhaustive and we note that this area of NLP is rapidly changing. It is strongly suggested that you also search for other applicable work on your own, possibly leveraging the search on the *ACL Anthology*.

You may want to start off with the **Introductory Paper**: Hovy, D. & Spruit, S. (2016) The Social Impact of Natural Language Processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, and/or papers from the Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, held in 2017.

2.1 Exclusion/Discrimination and Bias

Reading Questions:

- What went wrong?
- Who was harmed?
- Who benefited?
- What (if anything) is offered as a way to mitigate such harm in the future?
- What (if any) analogies do you see to the kind of NLP tasks your project worked on?

Papers:

You may also find useful papers in the Proceedings of the Second Workshop on Gender Bias in Natural Language Processing from 2020.

- Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In Proceedings of the first ACL workshop on ethics in natural language processing (pp. 1-11). Valencia, Spain: Association for Computational Linguistics.
- Rudinger, R., May, C., & Van Durme, B. (2017). Social bias in elicited natural language inferences. In Proceedings of the first ACL workshop on ethics in natural language processing (pp. 74-79). Valencia, Spain: Association for Computational Linguistics.

2.2 Embeddings and Language Behavior as Truth

Reading questions:

- How do the word embedding readings relate to the distributional hypothesis?
- What, if any, bias did the authors discover? What impacts do they describe following from that bias?
- What, if any, means of mitigating the bias do they authors propose? How are they evaluated?
- How do the scenarios described relate to the issue of using descriptive models prescriptively?

Papers:

- (*Only Section 6 on Broader Impacts.* Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). Language Models are Few-Shot Learners. ArXiv, abs/2005.14165.

- Ananya, Parthasarthi, N., Singh, S. (2019). GenderQuant: Quantifying mention-level genderedness. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (long and short papers) (pp. 2959-2969). Minneapolis, Minnesota: Association for Computational Linguistics.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in neural information processing systems 29 (pp. 4349-4357). Curran Associates, Inc.
- Daumé III, H. (2016). Language bias and black sheep. (Blog post, accessed 15 Apr 2021)
- Gonen, H., Goldberg, Y. (2019). https://arxiv.org/abs/1903.03862. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, pp. 609-614. Minneapolis, Minnesota: Association for Computational Linguistics.

2.3 Language Variation and Emergent Bias

Reading Questions:

- In what ways did the language vary?
- What social categories did it vary with?
- How did the language variation affect system performance?
- How would that differential performance lead to ethical or social problems in the world?
- How could the system be made more robust to language variation?
- How could such a system be deployed more responsibly?

Papers:

- Tan, S., Joty, S., Varshney, L.R., & Kan, M.-Y. (2020) Mind Your Inflections! Improving NLP for Non-Standard Englishes with Base-Inflection Encoding. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.
- Tan, S., Joty, S., Kan, M.-Y. & Socher, R. (2020) It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations. In Proceedings of the 2020 Annual Meeting of the Association of Computational Linguistics (ACL '20).
- Garimella, A., Banea, C., Hovy, D., & Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 3493-3498). Florence, Italy: Association for Computational Linguistics.
- Hovy, D., & Søgaard, A. (2015). Tagging performance correlates with author age. In Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers) (pp. 483-488). Beijing, China: Association for Computational Linguistics.
- Huang, X., & Paul, M. J. (2019). Neural user factor adaptation for text classification: Learning to generalize across author demographics. In Proceedings of the eighth joint conference on lexical and computational semantics (*SEM 2019) (pp. 136-146). Minneapolis, Minnesota: Association for Computational Linguistics.
- Jørgensen, A., Hovy, D., & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In Proceedings of the workshop on noisy user-generated text (pp. 9-18). Beijing, China: Association for Computational Linguistics.

- Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 51-57). Vancouver, Canada: Association for Computational Linguistics.
- Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In Proceedings of the first ACL workshop on ethics in natural language processing (pp. 53-59). Valencia, Spain: Association for Computational Linguistics.

Students, you must include the text of the two statements below in your submitted work and digitally sign your homework using your Student ID number (starting with `A...`*; N.B., not your NUSNET email identifier). Make sure you have attached this statement to your submission either in written or typed form.*

Delete (and where appropriate, fill in) one of the two forms of Statement 1:

1A. Declaration of Original Work. *By entering my Student ID below, I certify that I completed my assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, I am allowed to discuss the problems and solutions in this assignment, but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify my answers as per the Pokémon Go rule.*

1B. Exception to the Class Policy. *I did not follow the CS4248 Class Policy in doing this assignment. This text explains why and how I believe I should be assessed for this assignment given the circumstances explained.*

Signed, [Enter your A... Student ID here]

2. References *I give credit where credit is due. I understand need not (but am encouraged to) reference papers already mentioned in this assignment specification. I acknowledge that I used the following websites or contacts to complete this assignment:*

- *Sample. Website 1, for following mathematical proofs.*
- *Sample. My friend, A000000X, whom helped me figure out the course deadlines*