12-12-2022

Addressing Biases in Text Classification

Dieter Gutschi
WU Vienna, dieter.gutschi@wu.ac.at

Margeret Hall
Wirtschaftsuniversität Wien, margeret.hall@wu.ac.at

Christian Haas
Vienna University of Economics and Business (WU), christian.haas@wu.ac.at

Patricia Klarner
WU Vienna, patricia.klarner@wu.ac.at

TREO
Technology, Research, Education, Opinion

Addressing Biases in Text Classification

Dieter Gutschi, WU Vienna, dieter.gutschi@wu.ac.at; Margeret Hall, WU Vienna, margeret.hall@wu.ac.at; Christian Haas, WU Vienna, christian.haas@wu.ac.at; Patricia Klarner, WU Vienna, patricia.klarner@wu.ac.at

Text classification tasks profited substantially from the advancements in the area of large language models (LLMs). Fine-tuning models such as BERT, GPT-2/3, and others, led to new benchmarks in natural language understanding (e.g., GLUE dataset). At the same time, Bender et al. (2021) outlined the potential risks of LLMs, which include the overrepresentation of majority viewpoints and the replication and/or amplification of encoded biases. However, researchers and practitioners often do not discuss this downside of LLMs when fine-tuning them, not only ignoring but propagating these same biases in the resulting representations and classification decisions.

Another important issue in this context is label bias. It can arise during data annotation due to different understandings and interpretations of annotators, especially in the case of labelling "fuzzy" concepts such as emotions. Disagreements between annotators are often resolved by majority voting, which levels out interpretation differences. While these differences can represent classification mistakes, they also reflect annotators' sociodemographic factors or moral values (Davani et al. 2022). As a consequence, researchers end up relying on an "artificial" golden standard of training supervised models that may not utilize the diversity of human interpretation.

These annotation challenges in creating reliable training datasets are well represented in the GoEmotions dataset (Demszky et al. 2020). GoEmotions represents an interesting advancement in the field of emotion recognition, but also highlights problems of the current state of the art in text classification. Three annotators assigned one or multiple of 27 emotion or neutral categories to Reddit comments. The difficulty of rating emotions is visible in the interrater correlation, which has a range of 0.162 to 0.645 with an average correlation of 0.278 (cohen's kappa: 0.095-0.749, mean=0.293).

Low or contradictory interrater agreement, such as evidenced in the GoEmotions dataset, can pose serious challenges to the validity and reliability of results generated using these models. Hence, we argue that research building upon LLMs needs to include a reflective view of their content risks and a sufficient discussion of the consequences of potential biases.

References

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Vol. 1), Association for Computing Machinery, pp. 610–623. (https://doi.org/10.1145/3442188.3445922).

Davani, A. M., Díaz, M., and Prabhakaran, V. 2022. "Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations," *Transactions of the Association for Computational Linguistics* (10), pp. 92–110. (https://doi.org/10.1162/tacl_a_00449).

Demszky, D., Jeongwoo, D. M., Cowen, A., Nemade, G., and Ravi, S. 2020. "GoEmotions: A Dataset of Fine-Grained Emotions," *ArXiv*.