P7 - Problematic subgroup identification on text

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Eliana Pastor, Salvatore Greco

Project. This research aims to propose new methods for the identification and mitigation of disadvantaged subgroups in text classifiers.

Overview.

The identification of subgroups in which a model performs differently than overall behavior enables model understanding at the subgroup level and investigates their fairness and robustness. Existing solutions focus on tabular [5, 3, 1, 4] and speech data [2]. For speech data, they focus on the interpretable representation of utterances (e.g., the gender of the speaker, the speaking rate, and the level of noise). Few works focus on textual data, and most of them usually rely on template-generated evaluation data [6].

Goal.

The project involves identifying interpretable representations of textual data. A possible direction is to use LLM-prompts to derive categories (close to concept-based approaches). A definition of a hierarchy of categories could further enable the understanding. Once identified, we can leverage existing approaches for problematic subgroup identification.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing NLP fairness metrics, and techniques for the identification and evaluation of subgroups in textual data.
- **Identification of Research Gaps.** Identify key research gaps for identifying and evaluating subgroups performance, and optionally for mitigating performance disparities.
- **Implementation.** Select a specific research gap to address. Propose and implement a methodology to identify and evaluate subgroups' performance. This may involve 1) using LLM-prompting in zero or few-shot

learning for annotating metadata in texts or 2) deriving categories (e.g., concept-based). Then, propose or leverage existing techniques of subgroups' performance disparities mitigation.

- **Evaluation**. Assess the effectiveness and applicability of the newly implemented approach for at least two datasets.

References

[1] Yeounoh Chung et al. "Slice finder: Automated data slicing for model validation". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE. 2019, pp. 1550–1553.

[2] Alkis Koudounas et al. "Exploring subgroup performance in end-to-end speech models". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[3] Eliana Pastor, Luca De Alfaro, and Elena Baralis. "Looking for trouble: Analyzing classifier behavior via pattern divergence". In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 1400–1412.

[4] Svetlana Sagadeeva and Matthias Boehm. "Sliceline: Fast, linear-algebra-based slice finding for ml model debugging". In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 2290–2299.

[5] Nima Shahbazi et al. "Representation bias in data: a survey on identification and resolution techniques". In: *ACM Computing Surveys* 55.13s (2023), pp. 1–39.

[6] Tony Sun et al. "Mitigating gender bias in natural language processing: Literature review". In: *arXiv preprint arXiv:1906.08976* (2019).