experiment 1 and got similar performance metrics in lower time, refuting our concern that little data could be a bottleneck in the model.

However, for con gurations where ZeroBERTo was tested simulating real-life applications (Exp. 2 and 4), being exposed to new data, the performance was worse than XLM-R. The results suggest it occurs due to the inability of the embedded topic model to adequately represent new data as a composite of previously learned topics, over tting training data. This is clear from observing the high variance of the metrics among the k-folds. It allows us to conclude that, for now, the scenarios presented in experiments 1 and 3 are more suitable for using the model.

We have 0.54 of F1-score in the best scenario regarding the metrics obtained. Despite being a positive result considering that it is a multi-class classi cation, there is much room for improvement. The main reason to be pointed out for low performances is the use of multilingual models that were not ne-tuned in the Portuguese language, which is quite impressive.

A critical remark to be made is concerning the memory and time trade-off. For example, ZeroBERTo was more than 10x faster than XLM-R in Exp. 3. However, the topic model used by ZeroBERTo bases its clustering on the HDBSCAN method, which reduces time taken for data processing but increases the need for memory [McInnes and Healy, 2017], which XLM-R does not do. As the size of input data grows, processing may become unfeasible. XLM-R, on the other hand, does not use any interaction between data and can be processed in parallel and distributed without any negative effect on the nal result. It should be noted, however, that ZeroBERTo does not depend on BERTopic and can use other Topic Modeling techniques that address this issue more adequately in other scenarios.

A signi cant dif culty of this work was that, as far as the authors are aware of, there are no large benchmark datasets for multi-class text classi cation in Portuguese, nor general use datasets with semantically meaningful labels. In this sense, some future work directions involve the production of benchmark datasets for Portuguese text classi cation (as UTL-TC). It would also be interesting to produce Natural Language Inference datasets in Portuguese, which could, in addition to the existing ones [Fonseca et al., 2016, Real et al., 2020], enable ne-tuning of Transformers in Portuguese. Then, it would be possible to compare the performance of the models using BERTimbau (BERT-Portuguese) [Souza et al., 2020] both in clustering and classifying. It would also be worthwhile to test the proposed model in other domains: to name one, legislative data present similar challenges [Ferraz et al., 2021]. Another interesting future work would be to enable ZeroBERTo to deal with multi-label classi cation, where each document can have none, one or several labels.

Acknowledgments

This research was supported in partnership with Itaú Unibanco S.A. through the scholarship program Programa de Bolsas Itaú (PBI), and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001, CNPQ (grant 310085/2020-9), and USP-IBM-FAPESP Center for Arti cial Intelligence (FAPESP grant 2019/07665-4), Brazil. Any opinions, ndings, and conclusions expressed in this manuscript are those of the authors and do not necessarily re ect the views, of cial policy, or position of the nanciers.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020.

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus. In Proceedings of the 20th International Conference on Computational Linguistics, COLING '04, page 364 es. ACL, 2004.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. Importance of Semantic Representation: Dataless Classi cation. In Proceedings of the 23rd AAAI Conference on Arti cial Intelligence, volume 2, pages 830 835, 2008.

Qiuxing Chen, Lixiu Yao, and Jie Yang. Short text classi cation based on LDA topic model. 2016 International Conference on Audio, Language and Image Processing (ICALIP), pages 749 753. IEEE, 2016.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475 2485. Association for Computational Linguistics, 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco GuzmÆn, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440 8451, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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), pages 4171 4186, 2019.

Thomas Palmeira Ferraz, Alexandre Alcoforado, Enzo Bustos, AndrØ Seidel Oliveira, Rodrigo Gerber, Naíde Müller, AndrØ Corrêa d'Almeida, Bruno Miguel Veloso, and Anna Helena Reali Costa. DEBACER: a method for slicing moderated debates. Anais do XVIII Encontro Nacional de InteligŒncia Arti cial e Computacional, pages 667 678. SBC, 2021.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. ASSIN: Avaliacao de similaridade semantica e inferencia textual. In 12th International Conference on Computational Processing of the Portuguese Language, PROPOR, pages 13 15, 2016.

Maarten Grootendorst. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics., 2020.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2545 2568, 2021.

Paul S Jacobs. Joining statistics with NLP for text categorization. Third Conference on Applied Natural Language Processing, pages 178 185, 1992.

Yangfeng Ji and Jacob Eisenstein. Representation Learning for Text-level Discourse Parsing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13 24, 2014.

Daniel Z Korman, Eric Mack, Jacob Jett, and Allen H Renear. De ning textual entailment. Journal of the Association for Information Science and Technology, 69(6):763 772, 2018.

Xin Li, Yuhong Guo, and Dale Schuurmans. Semi-Supervised Zero-Shot Classi cation with Label Representation Learning. 2015 IEEE International Conference on Computer Vision (ICCV), pages 4211 4219. IEEE, 2015.

Lajanugen Logeswaran, Ann Lee, Myle Ott, Honglak Lee, Marc'Aurelio Ranzato, and Arthur Szlam. Few-shot sequence learning with transformers. arXiv preprint arXiv:2012.09543, 2020.

Leland McInnes and John Healy. Accelerated hierarchical density based clustering. 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pages 33 42. IEEE, 2017.

Dheeraj Mekala and Jingbo Shang. Contextualized weak supervision for text classi cation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 323 333, 2020.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. Text Classi cation Using Label Names Only: A Language Model Self-Training Approach. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9006 9017, 2020.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classi cation from labeled and unlabeled documents using EM. Machine learning, 39(2):103 134, 2000.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. The ASSIN 2 Shared Task: a quick overview. In International Conference on Computational Processing of the Portuguese Language, pages 406 412. Springer, 2020.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. Encyclopedia of database systems, 5:532 538, 2009.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In Advances in Neural Information Processing Systems, pages 935 943, 2013.

FÆbio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In Brazilian Conference on Intelligent Systems, pages 403 417. Springer, 2020.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In International Conference on Learning Representations, ICLR 2019, 2019.

Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112 1122, 2018.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking Zero-shot Text Classi cation: Datasets, Evaluation and Entailment Approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3914 3923, 2019.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In Advances in Neural Information Processing Systems, 33, 2020.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. Integrating Semantic Knowledge to Tackle Zero-shot Text Classi cation. In Proceedings of NAACL-HLT, pages 1031 1040, 2019.