A Sentiment Analysis Benchmark for Automated Machine Learning Applications and a Proof of Concept in Hate Speech Detection

Marília Costa Rosendo Silva, Vitor Augusto de Oliveira, Thiago Alexandre Salgueiro Pardo

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

marilia.costa.silva@usp.br,
vitor.augusto.oliveira@alumni.usp.br,
taspardo@icmc.usp.br

Abstract. *Automated Machine Learning (AutoML) is a relevant research endeavor as it allows for speeding up and easing the development of new applied solutions using Artificial Intelligence. This paper addresses the challenge of providing standardized datasets for sentiment analysis in English and proposes an AutoML benchmark, resulting in 46 preprocessed datasets. More than this, a proof of concept is carried out for the hate speech detection task to present the potentialities of the proposed benchmark.*

Resumo. *O Aprendizado de Máquina Automático (AutoML) é uma área de pesquisa relevante, pois permite acelerar e facilitar o desenvolvimento de novas soluções aplicadas usando Inteligência Artificial. Este artigo aborda o desafio de fornecer conjuntos de dados padronizados para análise de sentimentos em inglês e propõe um benchmark de AutoML, resultando em 46 conjuntos de dados pré-processados. É realizada uma prova de conceito para a tarefa de detecção de discurso de ódio para apresentar as potencialidades do benchmark proposto.*

1. Introduction

Natural Language Processing (NLP) aims at enabling machines to deal with human languages. The tasks of Sentiment Analysis (SA) are among the most useful and challenging ones, with interest of academic, commercial, and government areas.

In SA research, Machine Learning (ML) techniques have been the dominant approach. Developing an ML solution, however, can be complex for non-experts. For this reason, Automated Machine Learning (AutoML) has gained importance, providing resources to speed up tuning and making ML approaches more accessible [Guyon et al. 2016]. There are a few dozen available AutoML frameworks/systems, and a system that performs well on some tasks may have a lower performance on others [Škrlj et al. 2021]. Therefore, standardized comparison practices, such as benchmarks, can contribute to the traceability of the literature.

In this context, we explore AutoML for SA tasks. This work brings two core contributions: it furnishes 46 preprocessed datasets for different SA tasks; and, as Proof of Concept (PoC), some experiments with statistical evaluation to support the empirical findings comprising hate speech detection datasets and AutoML Systems.

2. Related Works

There are several initiatives on AutoML and on benchmarking some areas and tasks, but there are limited efforts focused on SA. [Blohm et al. 2021] used 13 text datasets for classification tasks, including polarity classification, with only a general evaluation. RAFT [Alex et al. 2021] is a Few-Shot Learning benchmark and uses news articles, domain-specific datasets, one Hate Speech Dataset, and another with complaints on Twitter.

Regarding comparative evaluations and statistical tests, there are several approaches in the literature and, not rarely, limited understanding of the appropriate metrics. [Demšar 2006] recommended the non-parametric Friedman test when assessing multiple classifiers in multiple datasets, and the post-hoc Nemenyi test to assess pairwise differences when the null hypothesis is rejected.

3. Dataset Collection and Preprocessing

To produce a benchmark for the SA area, it is necessary to collect and preprocess datasets, for later selecting and applying AutoML techniques, and standardizing experiment setups. Part of the procedures was based on [Pineau et al. 2020].

The data sources included UC Irvine, GitHub, Hugging Face, Kaggle, SemEval, TensorFlow, OpenML, and research articles. Our work only considered datasets without synthetically generated instances. The authors managed to split all the datasets into two non-overlapping sets (training and test sets).

The fields with text data and the one with the target feature were renamed as "text" and "label", respectively. This standardized denomination facilitates large-scale experiments. Moreover, the preprocessing steps were customized for each dataset. In addition, the text could have more than one language. Nevertheless, this work addressed data exclusively in English that was identified with the use of fastText [Bojanowski et al. 2017]. Instances that combined English and another language were kept and Regular Expressions were used to remove Cyrillic, Chinese, or Arabic characters.

After the preprocessing steps, instances smaller than three characters or that were duplicates were excluded. Appendix A lists all the datasets and their corresponding tasks. The adopted preprocessing steps do not harm the performance of most NLP tasks.

4. AutoML Experiments

For our experiments, the task of Hate Speech Detection was chosen. It brings all the challenging characteristics of SA tasks (as subjectivity, different writing styles, and high dependence on text genre and domain) and represents a severe disease in our modern society. Handling such scientific and social problem is of upmost importance.

Thirteen preprocessed Hate Speech Detection datasets were used (see Appendix A for the citations). They were selected due to technical considerations: 10 datasets were binary, and they had different and contrasting degrees of class imbalance.

Three AutoML systems and one classifier were adopted. These systems were selected based on previous works and their underlying assumptions. The foundations of AutoGluon [Erickson et al. 2020], Auto-Sklearn [Feurer et al. 2015], and TPOT [Olson and Moore 2016] are ensembles, meta-learning, and genetic programming, respectively. The classifier was Logistic Regression (LR) combined with Random Search (RS). All the approaches had the same conditions: wall time of 15 minutes per dataset, Python Version 3.9.13, Ubuntu 22.04.2, and up to 4GB RAM.

After preprocessing all the hate speech datasets, this work embedded the text with

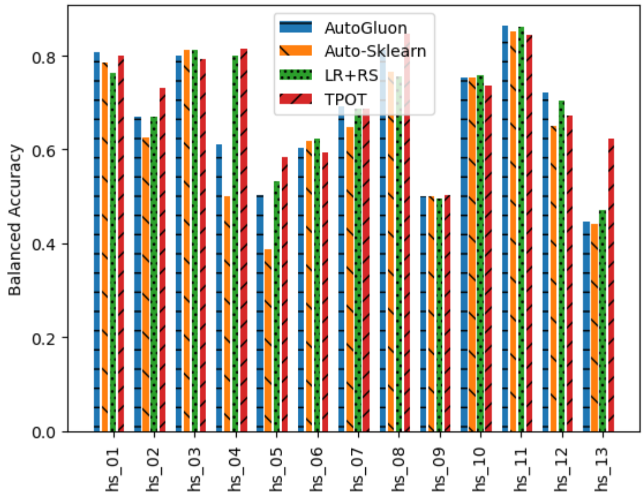

Figure 1. Bar Chart with Balanced Accuracy per Dataset

	Friedman		Kruskal-Wallis	
	χ_{n-1}	$p-value$	H	$p-value$
Values	5.0859	0.1656	1.4216	0.7005

Table 1. Statistics of the Non-Parametric Tests

Sentence BERT [Reimers and Gurevych 2019][1]. Figure 1 presents the bar charts corresponding to the AutoML systems and LR and their balanced accuracy per dataset.

On average, the best AutoML method is TPOT, with a mean of 71.01% of balanced accuracy. Auto-Sklearn has the largest standard deviation (14.97%). TPOT has the highest difference between the mean and median (2.24%), followed by Logistic Regression (1.82%), AutoGluon (1.60%), and Auto-Sklearn (0.61%). The maximum balanced accuracy is 86.50% (AutoGluon), whereas the smallest is 38.64% (Auto-Sklearn).

The results were evaluated with non-parametric hypothesis testing (Friedman and Kruskal-Wallis Tests). The statistics and p-values are presented in Table 1. With $\alpha < 0.05$, it is not possible to reject the null hypothesis, which means that there is no suggestion of statistically significant difference among the systems. This is a very interesting finding as it shows that different AutoML systems may prove to have similar results.

It is also interesting to evaluate the potentiality of the AutoML approaches when compared to the original results produced for the datasets. Unfortunately, it was not viable to perform comparisons for all the datasets (e.g., some of them did not have train-test splits by default). However, for three datasets, it was possible to supply fair comparisons.

Tables 2, 3 and 4 display performance metrics for three different datasets (*hs_03* [de Gibert et al. 2018], *hs_04* [Jigsaw 2018], and *hs_07* [Zampieri et al. 2019]) using AutoML systems and manual hyperparameter tuning from related literature (the evaluation metrics are the ones of the corresponding related works). The tables compare these methods based on various metrics, using different criteria for fair comparisons. Notably, there is a discernible difference between the best results from AutoML and the reported literature, with the latter performing about 9.3% better on average. However, considering that AutoML offers a more general solution and does not require specific tuning, it proves

[1] $"sentence-transformers/all-MiniLM-L6-v2"$

beneficial by relieving users of the task of building ML solutions from scratch. Despite the performance gap, the results achieved by AutoML are deemed satisfactory, suggesting that further investment in this approach is worthwhile.

Dataset	AutoGluon	Auto-Sklearn	LR+RS	TPOT	**[de Gibert et al. 2018]**
hs_03	0.817	0.798	**0.819**	0.803	**0.892**

Table 2. F1-Score for Benchmarking hs_03 [de Gibert et al. 2018]

Dataset	AutoGluon	Auto-Sklearn	LR+RS	TPOT	**[Jigsaw 2018]**
hs_04	0.9229	0.9060	**0.9278**	0.8572	**0.9885**

Table 3. Accuracy for Benchmarking hs_04 [Jigsaw 2018]

Dataset	AutoGluon	Auto-Sklearn	LR+RS	TPOT	**[Zampieri et al. 2019]**
hs_07	0.67	0.67	**0.71**	0.70	**0.80**

Table 4. F1-Score Macro for Benchmarking hs_07 [Zampieri et al. 2019]

5. Final Remarks

This paper introduces a benchmark dataset for SA, performing a PoC on detecting hate speech, showing that AutoML is a challenge that is worthy to follow. Overall, 46 preprocessed datasets are proposed. To the best of our knowledge, this is the first work that accomplishes this.

The proposed benchmark can be expanded with new datasets. They should have the same rationale – train-test splits, single class per instance, the same classes in the training and test sets, using the same codification, providing Python implementation with all the preprocessing steps (e.g., regular expressions and sorting functions), among others – and an available BibTex to furnish use in academia and by practitioners. These criteria can ensure sustainable growth and an update of the proposed benchmark.

Some limitations of this work are that it comprised only English datasets and that some datasets requiring credentials (e.g., using Twitter API to retrieve posts based on identifiers) might lose instances due to social media policy violations. Future research opportunities include improving algorithm initialization and evaluating other classification strategies.

A. Datasets

The next paragraph summarizes the tasks and datasets. They are split into six SA tasks[2].

Emotion Detection *(ed)*: [Strapparava and Mihalcea 2007], [Saravia et al. 2018], [Demszky et al. 2020], [Chakravarthi 2020], [Sosea et al. 2022]; **Fake News Detection** *(fn)*: [Wang 2017], [Pérez-Rosas et al. 2018], [Torabi Asr and Taboada 2018], [Torabi Asr and Taboada 2018], [Thorne et al. 2018], [Abu Salem et al. 2019], [Thorne et al. 2019], [Shahi and Nandini 2020], [Weinzierl and Harabagiu 2022], [Weinzierl and Harabagiu 2022]; **Hate Speech Detection** *(hs)*: [Waseem and Hovy 2016], [Davidson et al. 2017], [de Gibert et al. 2018], [Jigsaw 2018], [Founta et al. 2018], [Basile et al. 2019], [Zampieri et al. 2019], [Hugging Face 2019], [Gautam et al. 2020], [Mollas et al. 2020], [Grosz and Conde-Cespedes 2020], [Kaggle 2020c], [Mathew et al. 2021]; **Polarity Classification** *(pc)*: [Pang and Lee 2005], [Go et al. 2009], [Maas et al. 2011], [McAuley and Leskovec 2013], [Rosenthal et al. 2014], [Zhang et al. 2015], [Kaggle 2020d], [Bastan et al. 2020], [Sheng and Uthus 2020]; **Stance Detection** *(sd)*: [Kiesel et al. 2019], [Kiesel et al. 2019], [Kawintiranon and Singh 2021], [Kawintiranon and Singh 2021]; **Utility Analysis** *(ua)*: [Grano et al. 2017], [Gräßer et al. 2018], [Kaggle 2020b], [Keung et al. 2020], [Kaggle 2020a].

[2]https://github.com/marilia-cr-silva/nlp_datasets

References

[Abu Salem et al. 2019] Abu Salem, F. K., Al Feel, R., Elbassuoni, S., Jaber, M., and Farah, M. (2019). FA-KES: A Fake News Dataset around the Syrian War.

[Alex et al. 2021] Alex, N., Lifland, E., Tunstall, L., Thakur, A., Maham, P., Riedel, C., Hine, E., Ashurst, C., Sedille, P., Carlier, A., Noetel, M., and Stuhlmüller, A. (2021). RAFT: A Real-World Few-Shot Text Classification Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, pages 1–12.

[Basile et al. 2019] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

[Bastan et al. 2020] Bastan, M., Koupaee, M., Son, Y., Sicoli, R., and Balasubramanian, N. (2020). Author's Sentiment Prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615.

[Blohm et al. 2021] Blohm, M., Hanussek, M., and Kintz, M. (2021). Leveraging Automated Machine Learning for Text Classification: Evaluation of AutoML Tools and Comparison with Human Performance. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pages 1131–1136.

[Bojanowski et al. 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[Chakravarthi 2020] Chakravarthi, B. R. (2020). HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

[Davidson et al. 2017] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515.

[de Gibert et al. 2018] de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 11–20.

[Demšar 2006] Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, pages 1–30.

[Demszky et al. 2020] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

[Erickson et al. 2020] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. *arXiv:2003.06505*, pages 1–28.

[Feurer et al. 2015] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, page 2755–2763.

[Founta et al. 2018] Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *11th International Conference on Web and Social Media, 2018*, pages 491–500.

[Gautam et al. 2020] Gautam, A., Mathur, P., Gosangi, R., Mahata, D., Sawhney, R., and Shah, R. R. (2020). #metooma: multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of International AAAI Conference on Web and Social Media*, pages 209–216.

[Go et al. 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, pages 1–6.

[Grano et al. 2017] Grano, G., Di Sorbo, A., Mercaldo, F., Visaggio, C. A., Canfora, G., and Panichella, S. (2017). Software Applications User Reviews.

[Gräßer et al. 2018] Gräßer, F., Kallumadi, S., Malberg, H., and Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In *Proceedings of the 2018 International Conference on Digital Health*, page 121–125.

[Grosz and Conde-Cespedes 2020] Grosz, D. and Conde-Cespedes, P. (2020). Automatic Detection of Sexist Statements Commonly Used at the Workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: 2020 Workshops*, page 104–115.

[Guyon et al. 2016] Guyon, I., Chaabane, I., Escalante, H. J., Escalera, S., Jajetic, D., Lloyd, J. R., Macià, N., Ray, B., Romaszko, L., Sebag, M., Statnikov, A., Treguer, S., and Viegas, E. (2016). A brief review of the chalearn automl challenge: Any-time any-dataset learning without human intervention. In *Proceedings of the Workshop on Automatic Machine Learning*, pages 21–30.

[Hugging Face 2019] Hugging Face (2019). Tweets Hate Speech Detection. Accessed: 2022-04-05.

[Jigsaw 2018] Jigsaw (2018). Toxic Comment Classification Challenge. Accessed: 2022-04-06.

[Kaggle 2020a] Kaggle (2020a). Samsung Internal SSD Reviews. Accessed: 2022-04-06.

[Kaggle 2020b] Kaggle (2020b). Amazon Musical Instruments Reviews. Accessed: 2022-04-06.

[Kaggle 2020c] Kaggle (2020c). Terrorism And Jihadism Speech Detection. Accessed: 2022-04-06.

[Kaggle 2020d] Kaggle (2020d). Apple Twitter Sentiment Texts. Accessed: 2022-04-06.

[Kawintiranon and Singh 2021] Kawintiranon, K. and Singh, L. (2021). Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735.

[Keung et al. 2020] Keung, P., Lu, Y., Szarvas, G., and Smith, N. A. (2020). The Multilingual Amazon Reviews Corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4563–4568.

[Kiesel et al. 2019] Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., and Potthast, M. (2019). SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

[Maas et al. 2011] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

[Mathew et al. 2021] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14867–14875.

[McAuley and Leskovec 2013] McAuley, J. and Leskovec, J. (2013). Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, page 165–172.

[Mollas et al. 2020] Mollas, I., Chrysopoulou, Z., Karlos, S., and Tsoumakas, G. (2020). ETHOS: an Online Hate Speech Detection Dataset. *arXiv: 2006.08328*, pages 1–16.

[Olson and Moore 2016] Olson, R. S. and Moore, J. H. (2016). Tpot: A tree-based pipeline optimization tool for automating machine learning. In *Proceeding of the ICML 2016 AutoML Workshop*, pages 66–74.

[Pang and Lee 2005] Pang, B. and Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124.

[Pérez-Rosas et al. 2018] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

[Pineau et al. 2020] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). pages 1–22.

[Reimers and Gurevych 2019] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

[Rosenthal et al. 2014] Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment Analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 73–80.

[Saravia et al. 2018] Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.

[Shahi and Nandini 2020] Shahi, G. K. and Nandini, D. (2020). FakeCovid - A Multilingual Cross-domain Fact Check News Dataset for COVID-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, pages 1–9.

[Sheng and Uthus 2020] Sheng, E. and Uthus, D. (2020). Investigating Societal Biases in a Poetry Composition System. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106.

[Škrlj et al. 2021] Škrlj, B., Martinc, M., Lavrač, N., and Pollak, S. (2021). autoBOT: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, 110(5):989–1028.

[Sosea et al. 2022] Sosea, T., Pham, C., Tekle, A., Caragea, C., and Li, J. J. (2022). Emotion analysis and detection during COVID-19. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6938–6947.

[Strapparava and Mihalcea 2007] Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 70–74.

[Thorne et al. 2018] Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

[Thorne et al. 2019] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2019). The FEVER2.0 Shared Task. In *Proceedings of the Second Workshop on Fact Extraction and VERification*, pages 1–6.

[Torabi Asr and Taboada 2018] Torabi Asr, F. and Taboada, M. (2018). The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity. In *Proceedings of the First Workshop on Fact Extraction and VERification*, pages 10–15.

[Wang 2017] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 422–426.

[Waseem and Hovy 2016] Waseem, Z. and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

[Weinzierl and Harabagiu 2022] Weinzierl, M. and Harabagiu, S. (2022). VaccineLies: A Natural Language Resource for Learning to Recognize Misinformation about the COVID-19 and HPV Vaccines. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6967–6975.

[Zampieri et al. 2019] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420.

[Zhang et al. 2015] Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, pages 1–9.