Abusive Language Detection Using Auto-Machine Learning for Multiple Languages

Mackenzie Jorgensen and Minho Choi

Abusive language online has sparked debates about how to define and characterize it and to determine its impacts on users and businesses (3, 5, 6, 9). Publications on abusive comment moderation use both binary (7) and multi-class classifications (2, 8). Binary classifications label data as either abusive or not abusive. Multi-class classifications include data that have more than two label options for an element. We aim to explore the benefits of multi-class classification for efficient and precise labeling of abusive online comments.

How we analyze these data is another question, and further, what sort of semi-automatic system can learn to identify abusive language? Academic research (1, 2, 4, 7, 8, 10) and industry point toward machine learning (ML, a field in which statistical models learn from data sets to predict a target) to solve this problem. Google's Perspective program has shown promising results in working with different newspaper companies (https://www.perspectiveapi.com/#/home). Nevertheless, the component of ML concerning abusive language detection remains understudied. Auto-Machine Learning (Auto-ML) was first developed for non-technical experts to utilize ML technology more easily. However, Auto-ML also expedites data processing, model selection, and model parameter selection for researchers.

In this study, we selected data sets containing tweets from multi-class research conducted in English and German. The English-language dataset from Davidson, Warmsley, Macy, and Weber includes 24,802 tweets and is labeled as follows: 5% hate speech, 76.6% offensive, and 16.6% neither (4). The German-language dataset from Wiegand, Siegel, and Ruppenhofer labeled tweets as 21.0% abuse, 11.4% insult, 1.3% profanity, and 66.1% other (10). Once we clean the data sets (make all letters lowercase and restrict the length of the text to a specific range), we will split the data into training and testing data sets with a 9:1 ratio. We will then give our Auto-ML model the training data set to learn and subsequently we will evaluate the model on our testing data set. After we train and test the model, we will compare the results of the Auto-ML with the machine learning to results from current literature. We will explore whether a similar Auto-ML model works best for both English and German abusive language data sets or if specific models worked better for English rather than German data sets.

ACKNOWLEDGEMENTS

The authors would like to thank the MODERAT! (reduction of the moderation effort of user comments with the help of automation by text-analytical methods) project for providing inspiration for this research endeavor.

FUNDING INFORMATION

This research was supported by a 2019 DAAD RISE program to fund Mackenzie and Minho at European Research Center for Information Sciences Headquarters, in Münster, Germany.

REFERENCES

1. Bishop, C.M. 2006. Pattern recognition and machine learning. Springer.
2. Bourgonje, P., Moreno-Schneider, J., Srivastava, A., and Rehm. G. 2017. "Automatic classification of abusive language and personal attacks in various forms of online communication," in International Conference of the German Society for Computational Linguistics and Language Technology, Springer, Cham, pp. 180-191.
3. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. 2017. "Mean birds: Detecting aggression and bullying on twitter," in Proceedings of the 2017 ACM on web science conference, ACM, pp. 13-22.
4. Davidson, T., Warmsley, D., Macy, M., and Weber, I. 2017. "Automated hate speech detection and the problem of offensive language," in Eleventh International AAAI Conference on Web and Social Media.
5. Guberman, J., and Hemphill, L. 2017. "Challenges in modifying existing scales for detecting harassment in individual tweets," in Proceedings of the 50th Hawaii International Conference on System Sciences.
6. Niemann, M. 2019. Abusiveness is Non-Binary: Five Shades of Gray in German Online News-Comments, in Proceedings of the

21st IEEE Conference on Business Informatics (CBI 2019), IEEE, Moscow, Russia.

7. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. 2016. "Abusive language detection in online user content," in Proceedings of the 25th international conference on world wide web, International World Wide Web Conferences Steering Committee, pp. 145-153.
8. Park, J. H., and Fung, P. 2017. "One-step and two-step classification for abusive language detection on twitter," arXiv preprint arXiv:1706.01206.
9. Pater, J. A., Kim, M. K., Mynatt, E. D., and Fiesler, C. 2016. "Characterizations of online harassment: Comparing policies across social media platforms," in Proceedings of the 19th International Conference on Supporting Group Work, ACM, pp. 369-374.
10. Wiegand, M., Siegel, M., and Ruppenhofer, J. "Overview of the germeval 2018 shared task on the identification of offensive language," in 14th Conference on Natural Language Processing (KONVENS 2018), September 21, 2018, pp. 1-10.

Author

Mackenzie Jorgensen

Mackenzie Jorgensen studies Computer Science and Philosophy and has been conducting research since her first year at Villanova. Through the Match Research Program for First Year Students, she conducted human-computer-interaction research. With a 2017 NSF REU-D3 award, she used machine-learning cluster-algorithms to analyze big data on preterm births in Puerto Rico. Through Germany's 2018 DAAD RISE program, she conducted AI research, programming multi-agents to communicate and problem-solve. Through a second DAAD RISE-funded summer in 2019, she utilized Auto-Machine Learning to detect abusive language on Twitter. All of her summer research endeavors resulted in publications. After graduation, Mackenzie will pursue a PhD in Computer Science.

Author

Minho Choi

Minho Choi is majoring in Computer Science and Mathematics at Lewis & Clark College. Minho conducted two major research projects from his first-year of college. In 2018, Minho worked on comparative power analysis between complete randomization and biased coin design through John S. Rogers Science Program. In 2019, Minho explored the application of Auto-Machine Learning in abusive language detection through DAAD RISE program. After graduation, Minho will pursue his PhD in Mathematical Statistics.

Mentor

Marco Niemann

Marco Niemann (marco.niemann@ercis.uni-muenster.de) is a research assistant and Ph.D. candidate at the European Research Center for Information Systems (ERCIS), University of Muenster, Germany. His research focuses on data-driven and analytics-based systems as well as on information systems engineering. In the context of the project MODERAT! Marco currently researches, how an analytics-driven platform can support community managers of online newsmedia in moderating web debates. The central aspect is the provision of accurate, yet accessible, interpretable, and actionable ML and NLP-driven decision support.