NEATCLasS 2023: The 2nd Workshop on Novel Evaluation Approaches for Text Classification Systems

Björn Ross[1], **Roberto Navigli**[2], **Agostina Calabrese**[1], **Sheikh Muhammad Sarwar**[3]

[1]The University of Edinburgh, Edinburgh, United Kingdom
[2]Sapienza University of Rome, Rome, Italy
[3]Amazon.com
b.ross@ed.ac.uk, navigli@diag.uniroma1.it, a.calabrese@ed.ac.uk, smsarwar@amazon.com

Abstract

The use of machine learning algorithms for the detection of abusive language online has a massive impact on the timely and proactive removal of harmful content from the Web. However, such systems are far from perfect. One critical bottleneck towards the development of more effective technologies is the lack of proper evaluation procedures. If unreliable systems achieve astonishing scores with traditional metrics, how do we recognise progress when we see it? The workshop on Novel Evaluation Approaches for Text Classification Systems aims to encourage a wider conversation about how misinformation detection, abusive language detection, sentiment analysis and related tasks should be evaluated. It promotes the design of new evaluation metrics and benchmarks that better capture a system's performance, and the development of new systems that perform better on these novel metrics. The workshop was held for the second time this year. This introduction to the workshop proceedings provides a summary of contributions.

Introduction

The automatic or semiautomatic analysis of textual data is a key approach to analyse the massive amounts of user-generated content online, from the identification of sentiment in text and topic classification to the detection of abusive language, mis-information or propaganda. However, the development of such systems faces a crucial challenge. Static benchmarking da-tasets and performance metrics are the primary method for measuring progress in the field, and the publication of research on new systems typically requires demonstrating an improvement over state-of-the-art approaches in this way. Yet, these per-formance metrics can obscure critical failings in current models. There is clearly a need to rethink performance evaluation for text classification and analysis systems to be usable and trustable.

Recently there have been attempts to address this problem. For example, in abusive language detection, there are both static datasets of hard-to-detect examples (Röttger et al. 2021) and dynamic approaches for generating such examples (Calabrese et al. 2021). On the platform DynaBench (Kiela et al. 2021), benchmarks are dynamic and constantly updated with hard-to-classify examples, avoiding overfitting a predetermined dataset.

The goals of the ICWSM workshop on Novel Evaluation Approaches for Text Classification Systems (NEATCLasS) are to facilitate discussion around such new evaluation approaches, to stimulate the development of new and refinement of existing ones, to promote the use of novel metrics for abuse detection, sentiment analysis and similar tasks within the community, in order to better be able to measure whether models really improve upon the state of the art, and to encourage a wide range of models to be tested on these new metrics.

The second iteration of the workshop took place on 5 June 2023, in conjunction with the International AAAI Conference on Web and Social Media (ICWSM-2023).

Talks

The workshop started with a keynote talk from Dr Kiran Garimella, Assistant Professor in the School of Communication and Information at Rutgers University, on fact-checking on encrypted platforms.

Crowdsourced annotations, especially for subjective tasks, are often characterised by high levels of disagreement. Mestre et al. (2023) show this on the example of argumentative relations in political debates. Their research highlights the need for more robust evaluation metrics for subjective crowdsourcing tasks.

Jafari et al. (2023)'s work is on the task of detecting whether a given social media post agrees or disagrees with false information (misinformation alignment detection). They present a dataset, COVID-Myths, of 3528 annotated tweets and a model, CoMID, that is extensively evaluated.

Strathern and Pfeffer (2023) present a classification scheme with 11 classes of online misogyny and apply it to a case from social media when Amber Heard received online abuse after making allegations of domestic violence against Johnny Depp. The researchers show that a lot of implicit online misogyny is not captured by the Perspective API which is widely used for detecting toxic language.

In their demo paper, Althenayan et al. (2023) present Toxicity Inspector, which addresses the challenge of subjective and perceptions of toxicity by employing an iterative, human-in-the-loop feedback process. They also present a

case study to show the effectiveness of this feedback mechanism.

Finally, Assenmacher et al. (2023) investigate how abusive language and social bot detection researchers share Twitter data. They show that the strategy of sharing Twitter IDs and asking recipients to "rehydrate" the data by re-requesting them from the Twitter API leads to data loss because tweets may have been removed in the meantime. This approach is further complicated by recent API changes. The researchers propose recommendations for dataset creators. Based on the review scores, this contribution was honoured with the NEATCLasS 2023 Best Paper Award.

Group Activity

The workshop included an interactive group activity where participants could use a notebook provided by the organisers to qualitatively evaluate the reasoning capabilities of a large language model (i.e., ChatGPT). By using the ETHICS dataset (Hendrycks et al. 2021) participants tested ChatGPT's ability to provide moral judgments that align with human values. In case of wrong predictions, participants were asked to investigate the presence of potential flaws in the model's reasoning process by querying ChatGPT for an explanation. The activity therefore touched two topics that are core to the workshop: 1) evaluation and model diagnostic, and 2) annotations for tasks involving subjective judgements. In the discussion that followed the activity, participants highlighted how most wrong predictions were still supported by a fair reasoning and often justified by a lack of context. Interestingly, all explanations provided by ChatGPT were coherent with the predicted moral judgements.

Discussion

Text classification tasks based on social media data and other online interactions, such as the detection of abusive and hateful language, have arguably long had issues with benchmarking and evaluation due to issues with data access (Bruns et al. 2018), data sharing (Assenmacher et al. 2022) and subjective annotations (Waseem 2016). However, the last few years have seen further dramatic changes in computational social science and natural language processing that have compounded these issues:

The rapid pace at which new and ever bigger language models are released raises important questions for academic research. In a time where it is easier than ever to apply a language model for an ad-hoc task in the absence of any labelled data, effective evaluation strategies are sorely needed.

There is increasing awareness of biases in large language models (Navigli, Conia, and Ross 2023). Clearly, it is not sufficient to merely measure the performance of a system before deploying it, but it also needs to be checked how susceptible it is to such biases. A great deal of thought needs to go into such checks to avoid publishing a system that unwittingly harms groups of users who were not thought about or involved in its design (Ungless, Ross, and Belle 2023).

Yet, perhaps the dominant topic of conversation at ICWSM this year was the sudden announcement from Twitter to make fundamental changes to its API and its pricing model. Five years after Facebook tightened API restrictions as a result of the Cambrige Analytica scandal, this has reminded us all once again that access for researchers to relevant online conversations is not a given, but that unless we find more permanent solutions, we are at the mercy of corporations granting access to such data.

It is clear that evaluation in our field faces more and more challenges and that the questions do not have simple answers. The NEATCLasS workshop provides an opportunity to discuss and publish work that addresses these questions, and each of the research papers presented at this year's workshop tackled some of the issues that the field is facing. We look forward to continuing these important conversations and to keeping the topic of evaluation on the agenda of computational social science researchers.

Workshop Organisation

This workshop was organised by the following group.

Björn Ross is Lecturer (Assistant Professor) in Computational Social Science at the University of Edinburgh School of Informatics, in Edinburgh, United Kingdom,

Roberto Navigli is a full professor in the Department of Computer, Control and Management Engineering at the Sapienza University of Rome, in Rome, Italy,

Agostina Calabrese is a third-year Ph.D. student at the UKRI Centre for Doctoral Training in Natural Language Processing, in Edinburgh, United Kingdom,

Sheikh Muhammad Sarwar is an Applied Scientist at Amazon, Seattle, United States of America.

Acknowledgements

The NEATCLasS workshop was supported by an RSE Saltire Facilitation Network Award, award no. 1901.

References

Althenayan, H.; Bahlas, R.; Alharbi, M.; Alsuwailem, L.; Aldayel, A.; and ALahmadi, R. 2023. Toxicity Inspector: A Framework to Evaluate Ground Truth in Toxicity Detection Through Feedback. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Assenmacher, D.; Sen, I.; Fröhling, L.; and Wagner, C. 2023. The End of the Rehydration Era – The Problem of Sharing Harmful Twitter Research Data. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Assenmacher, D.; Weber, D.; Preuss, M.; Valdez, A. C.; Bradshaw, A.; Ross, B.; Cresci, S.; Trautmann, H.; Neumann, F.; and Grimme, C. 2022. Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem. *Social Science Computer Review*, 40(6): 1496–1522.

Bruns, A.; Bechmann, A.; Burgess, J.; Chadwick, A.; Clark, L. S.; Dutton, W. H.; and Zimmer, M. 2018. Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*, 25.

Calabrese, A.; Bevilacqua, M.; Ross, B.; Tripodi, R.; and Navigli, R. 2021. AAA: Fair Evaluation for Abuse Detection Systems Wanted. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, 243–252. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383301.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2021. Aligning AI With Shared Human Values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jafari, N.; Sarwar, S. M.; Allan, J.; Sung, K.; Dori-Hacohen, S.; and Rattigan, M. 2023. CoMID: COVID-19 Misinformation Alignment Detection Using Content and User Data. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Kiela, D.; Bartolo, M.; Nie, Y.; Kaushik, D.; Geiger, A.; Wu, Z.; Vidgen, B.; Prasad, G.; Singh, A.; Ringshia, P.; Ma, Z.; Thrush, T.; Riedel, S.; Waseem, Z.; Stenetorp, P.; Jia, R.; Bansal, M.; Potts, C.; and Williams, A. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4110–4124. Online: Association for Computational Linguistics.

Mestre, R.; Ryan, M.; Middleton, S. E.; Gomer, R.; Gheasi, M.; Zhu, J.; and Norman, T. J. 2023. Benchmark Evaluation for Tasks with Highly Subjective Crowdsourced Annotations: Case study in Argument Mining of Political Debates. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality*, 15(2).

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58. Online: Association for Computational Linguistics.

Strathern, W.; and Pfeffer, J. 2023. Identifying Different Layers of Online Misogyny. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Ungless, E. L.; Ross, B.; and Belle, V. 2023. Potential Pitfalls With Automatic Sentiment Analysis: The Example of Queerphobic Bias. *Social Science Computer Review*.

Waseem, Z. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. Austin, Texas: Association for Computational Linguistics.