Published in: Journal for Language Technology and Computational Linguistics
vol. 34 (2020) no. 1, pp. i-v.

Editorial

Recent years have seen a sharp increase in studies of offensive language (and related notions such as abusive language, hate speech, verbal aggression etc.) as well as of patterns of online behavior such as cyberbullying and trolling. Multiple efforts have been launched for the exploration of computational approaches and the establishment of benchmark datasets for various languages (Basile et al. (2019), Wiegand et al. (2018), Zampieri et al. (2019)).

Although many researchers start out with the intuition that 'I know it when I see it', it turns out that nailing down the boundaries of offensive language and implementing a computational approach for its recognition are abiding challenges. Not surprisingly, the inventory of categories used to classify units of text varies significantly across different papers and shared tasks. The most common division is between OFFENSIVE LANGUAGE (or ABUSE or HATE etc.) and a neutral or other class. Beyond binary classification, there are several ways to deepen the analysis. One of them consists in subdividing the overall offense class according to the targeted group, distinguishing, for instance, sexism and racism from a neither class (Waseem and Hovy (2016)). Another way of adding detail to the analysis is to distinguish sub-classes of offensive language. The 2018 GermEval Shared Task on Offensive Language (Wiegand et al. (2018)), for instance, included a fine-grained task that split up offensive language into the three sub-classes abuse, insult and profanity. A third option is to use a numeric measure of offensiveness (Ross et al. (2016)), foregoing class labels. Yet another dimension along which offensive language can be subdivided is the distinction between explicit and implicit cases (Waseem et al. (2017); Gao et al (2017); Struß el al (2019)). Further dimensions of fine-grained analysis are no doubt conceivable.

Beyond framing the task, sampling a data set is another key problem. As shown by Arango et al. (2019) and Wiegand et al. (2019), many current datasets for abusive language detection contain unwanted biases and artefacts that lead to overly optimistic assessments of the performance of classification systems.

Of the computational approaches taken in designing such systems, there exists a wide and rapidly evolving variety. Schmidt and Wiegand (2017) give a useful recent overview of the state of the art, also providing among other things an extensive discussion of feature-based classification. Wiegand et al. (2018) report that for the GermEval 2018 Shared Task feature-based supervised learning was competitive, even though many neural systems participated and performed well. Struß et al. (2019) report for the GermEval 2019 Shared Task in the following year that supervised classifiers using word embeddings, subword information and ensemble methods, also proved effective. However, similar effectiveness with less task-specific design could be achieved by classifiers based on the BERT model.

Against this background, we had issued a very broad call for contributions to this special issue of JLCL.[1] We are very happy to see that, between them, the contributions

[1] `https://easychair.org/cfp/JLCL_SIOL2020`

in this issue address a significant range of the topics that we had put forth in our Call for Papers.

The contribution by Palmer et al. presents a new **annotation scheme** for offensive language and hate speech, breaking the difficult annotation task down to four easier to answer questions. Notably, this scheme goes beyond the explicit cases of offensive language and also tackles several types of **complex, implicit, and/or pragmatically-triggered offensive language**. The authors' work also addresses the issue of **evaluation** by providing a new dataset for Evaluation of Complex Offensive Language Data in English (COLD-EN), which in future research can help diagnose systems' ability to appropriately classify instances of a set of special categories of (offensive) language. Among them are reclaimed slurs, non-slur offensive utterances containing pejorative adjectival nominalizations, and utterances conveying offense through linguistic distancing. The authors also conduct some experiments with state-of-the-art classifiers trained on different datasets to evaluate their diagnostic power when it comes to error analysis on offensive language detection.

Under the rubric of **Explainable AI**, the paper by Risch et al. explores how automatic classification systems can be equipped with mechanisms to explain *each individual* classification they make. Transparent and understandable explanations for decisions on which texts to allow or reject are needed to make such systems useful: both content moderators and users need to be able to understand the basis for classifications in order to be able to defend or take issue with them. As speakers' free speech rights hang in the balance and need to be weighed against others' rights to be protected from hateful and discriminatory speech, building explainability into automatic systems would go some way towards alleviating **ethical and legal concerns about automated offensive language detection**.

The third contribution by Shekhar et al. addresses offensive language in a **setting outside of the major Social Media platforms**, exploring the automation of comment moderation for news articles in two **less-resourced languages**, Estonian and Croatian. The authors create new datasets for both languages, labeled in the course process of real world human comment moderation. Owing to the focus on heterogeneous **real-world datasets** and the question of practical applicability, the authors address undesirable content besides offensive language, such as cases of deception and trolling, off-topic posts, copyright infringement, or pornography. The authors provide a systematic comparison of up-to-date classification approaches applied to the data, and propose a number of explanations for differences in performance resulting, for instance, from changes in comments and/or moderation policy over time.

While we could only briefly allude here to some of the contributions contained in these papers, we very much invite you to delve into them further for insight and inspiration. We would also like to thank the colleagues whose work made this special issue possible: the authors of the papers and the reviewers, who contributed to the quality of the published articles with careful and thoughtful feedback: Valerio Basile, Sara Tonelli, Paolo Rosso, Manfred Stede, Alexis Palmer, Torsten Zesch, Tatjana Scheffler, Sylvia Jaki, and Zeerak Waseem. Finally, we want to express our gratitude to the editors of

the Journal for Language Technology and Computational Linguistics for their support in putting together this issue, which it is our great pleasure to now release.

The guest editors, Josef Ruppenhofer, Melanie Siegel, Julia Maria Struß.

References

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (p. 45–54). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3331184.3331262 doi: 10.1145/3331184.3331262

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., ... Sanguinetti, M. (2019, June). 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* (pp. 54–63). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S19-2007 doi: 10.18653/v1/S19-2007

Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the eighth international joint conference on natural language processing, IJCNLP 2017, taipei, taiwan, november 27 - december 1, 2017 - volume 1: Long papers* (pp. 774–782). Retrieved from https://aclanthology.info/papers/I17-1078/i17-1078

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Eds.), *Proceedings of nlp4cmc iii: 3rd workshop on natural language processing for computer-mediated communication* (pp. 6–9). Retrieved from https://arxiv.org/pdf/1701.08118.pdf

Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10). Valencia, Spain: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-1101 doi: 10.18653/v1/W17-1101

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th conference on natural language processing (konvens 2019)* (pp. 352–363). Friedrich-Alexander-Universität Erlangen-Nürnberg.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017, August). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online* (pp. 78–84). Vancouver, BC, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-3012 doi: 10.18653/v1/W17-3012

Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88–93). San Diego, California: Association for Computational

Linguistics. Retrieved from https://www.aclweb.org/anthology/N16-2013 doi: 10.18653/v1/N16-2013

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019, June). Detection of Abusive Language: the Problem of Biased Datasets. In *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)* (pp. 602–608). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N19-1060 doi: 10.18653/v1/N19-1060

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language. In (pp. 1 – 10). Vienna, Austria: Austrian Academy of Sciences. Retrieved from http://nbn-resolving.de/urn:nbn:de:bsz:mh39-84935

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019, June). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 75–86). Minneapolis, Minnesota, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S19-2010 doi: 10.18653/v1/S19-2010