Invitation to GermEval 2021 Shared Task: Identification of Toxic, Engaging, and Fact-Claiming Comments

Julian Risch[1], Anke Stoll[2], Lena Wilms[2], and Michael Wiegand[3]

[1]Hasso Plattner Institute, University of Potsdam
[1]julian.risch@hpi.de
[2]Department of Social Sciences, Heinrich Heine University Düsseldorf
[2]anke.stoll@hhu.de, lena.wilms@hhu.de
[3]Digital Age Research Center, Alpen-Adria-Universität Klagenfurt
[3]michael.wiegand@aau.at

Abstract

In this outline, we would like to present to you this year´s shared task of GermEval 2021 on the identification of toxic, engaging, and fact-claiming comments. The shared task comprises three binary classification subtasks with the goal to identify: toxic comments (e.g, insults or threats), engaging comments (e.g., expression of respect or mediation), and comments that include indications of the need for fact-checking, here referred to as fact-claiming comments. Building on the two previous GermEval shared tasks on the identification of offensive language in 2018 and 2019, we extend the task definition to mirror that moderators also need to highlight engaging comments, which foster respectful and encourage in-depth discussions and check facts that lines of arguments rely on. The dataset, which will be provided, includes 3,300 Facebook posts extracted from the page of a German news broadcast. A theoretical framework and additional reliability tests during the data annotation process ensure particularly high data quality. As an interdisciplinary team with computer science, computational linguistics, and social science backgrounds, we would like to encourage researchers from different research fields to participate in what we envision to be a practical and holistic approach to sustainably improve moderation of online communities.

1 Motivation and Task Overview

The demand for approaches to identify harmful user content online is unchanged while the research field of Natural Language Processing is constantly evolving. The two previous GermEval shared tasks on the identification of offensive language (Struß et al., 2019; Wiegand et al., 2018) mark important references for research teams from both science and industry that use the datasets to evaluate their frameworks. With this year´s shared task, we want participants to go beyond the identification of offensive comments. To this end, we extend the focus to two other classes of comments that are highly relevant to moderators and community managers on online discussion platforms: engaging comments and fact-claiming comments, meaning comments that should be considered as a priority for fact-checking. This shift aims to bridge the gap between the theoretical view on comment classification and the practical needs of discussion moderators. Teams can participate either in all three subtasks or just one ore two of the following subtasks.

Subtask 1: Toxic Comment Classification (Binary Classification Task) The issue of toxic, offensive, or hateful language in social media and online discussion platforms has not lost any of its actuality. Still, the detection of such content remains challenging and new approaches are constantly being demanded and developed. With this subtask we aim to continue the series of previous GermEval Shared Tasks on Offensive Language Identification (Struß et al., 2019; Wiegand et al., 2018).

Subtask 2: Engaging Comment Classification (Binary Classification Task) In addition to the detection of toxic language, community managers and moderators increasingly express interest in identifying particularly valuable user content, for example, to highlight them and to give them more visibility (Risch and Krestel, 2020). Especially rational, respectful, and reciprocal comments can encourage readers to join the discussion, increase positive perceptions of discussion providers, and can enhance more fruitful and less violent exchange (Price and Cappella, 2002; Prochazka et al., 2018).

Subtask 3: Fact-Claiming Comment Classification (Binary Classification Task) Beyond the challenge to ensure non-hostile debates, platforms

and moderators are under pressure to act due to the rapid spread of misinformation and fake news. Platforms need to review and verify posted information to meet their responsibility as information providers and distributors. As a result, there is an increasing demand for systems that automatically identify comments that should be fact-checked manually. Note that this subtask is not about the fact-checking itself or the identification of fake news. However, the identification of fact-claiming comments is a pre-processing step for manual fact-checking.

2 Data & Resources

We provide an annotated dataset of Facebook user comments that have been labeled by four trained annotators. The dataset is drawn from the Facebook page of a German news broadcast, including user discussions from February till July 2019. The dataset will be shared with registered participants and will be made accessible for non-commercial, academic research purposes after the shared task. The dataset is provided in anonymized form: user information and comment ID's will not be shared.

For annotation, we provide an extensive, theory-based annotation scheme, which is especially suitable for the detection of deliberative and uncivil text features in online discussions (available on request). We annotated 3,245 Facebook comments by multiple categories indicating toxic comments, engaging comments, and fact-claiming comments. High data quality is ensured by intensive annotator training as well as intercoder reliability testing using Krippendorff's alpha.[1] For the individual subtasks, the annotations define three class labels. Figure 1 shows example comments of each class.

Toxic Comments Toxic comments comprise uncivil forms of communication that can violate the rules of polite behavior, such as insulting discussion participants, using vulgar or sarcastic language or implied volume via capital letters (annotator agreement is in the range $0.73 < \alpha < 0.89$). Additionally, incivility can be characterized as a violation of democratic discourse values, e.g., by verbally attacking basic democratic principles or making it difficult for others to participate (Papacharissi, 2004). It includes discrimination or discrediting of participants as well as the accusation of lying or threats of violence (annotator agreement is in the range $0.83 < \alpha < 0.90$).

"Na, welchem tech riesen hat er seine Eier verkauft..?" *TOXIC*

"Ich macht mich wütend, dass niemand den Schülerinnen Gehör schenkt" *NOT TOXIC*

(a) Subtask 1: Identification of toxic comments.

"Wie wär's mit einer Kostenteilung. Schließlich haben beide Parteien (Verkäufer und Käufer) etwas von der Tätigkeit des Maklers. Gilt gleichermassen für Vermietungen. Die Kosten werden so oder so weiterverrechnet, eine Kostenreduktion ist somit nicht zu erwarten." *ENGAGING*

"Die aktuelle Situation zeigt vor allem eines: viele Kinder mussten erkennen, dass ihre Mütter bestenfalls das Niveau Grundschule, Klasse 3 haben." *NOT ENGAGING*

(b) Subtask 2: Identification of engaging comments.

"Kinder werden nicht nur seltener krank, sie infizieren sich wohl auch seltener mit dem Coronavirus als ihre Eltern - das ist laut Ministerpräsident Winfried Kretschmann (Grüne) das Zwischenergebnis einer Untersuchung der Unikliniken Heidelberg, Freiburg und Tübingen." *FACT-CLAIMING*

"hmm...das kann ich jetzt nich nachvollziehen..." *NOT FACT-CLAIMING*

(c) Subtask 3: Identification of fact-claiming comments.

Figure 1: Example comments and their class labels.

Engaging Comments Engaging comments include communication features that are in line with deliberative principles, such as rationality, reciprocity, and mutual respect (Gutmann and Thompson, 1998). The first category covers communication features, such as justification, solution proposals, or the sharing of personal experiences (annotator agreement is in the range $0,71 < \alpha < 0,89$). The second category covers empathy with regard to other users' standpoints (annotator agreement is in the range $0,79 < \alpha < 0,91$). The third category is present when the comment is in line with rules of polite interaction or includes the expression

[1] Krippendorff's alpha corrects for random agreement between coders by relating the observed mean deviation to the assumed mean deviation of a random agreement (Krippendorff, 2018).

of mutual respect as well as mediation (annotator agreement is in the range $0,85 < \alpha < 1$).

Fact-Claiming Comments Comments that contain the assertion of facts and the provision of evidence by cited external sources fall into the class of fact-claiming comments (annotator agreement is in the range $0,73 < \alpha < 0,84$).

3 Evaluation

Following in the footsteps of the GermEval 2019 Shared on Hierarchical Classification of Blurbs (Remus et al., 2019) and the GermEval 2020 Shared Task on the Classification and Regression of Cognitive and Motivational Style (Johannßen et al., 2020), we will use the platform `codalab.org` for evaluation. The evaluation uses precision, recall, and macro-average F1-score as metrics. Macro-average F1-scores give equal importance to each class, which is suited because class labels in our dataset are not uniformly distributed but are equally important to identify.

4 KONVENS 2021 online

With the organization of this shared task, we would like to enable research in the direction of semi-automatic comment moderation to support moderators with natural language processing. Our interdisciplinary team of task organizers includes computer scientists, social scientists, and computer linguists from University of Potsdam, Heinrich Heine University Düsseldorf, and University of Klagenfurt. We therefore strongly encourage the participation of scientists from different disciplines sharing our common interest in offensive language detection, moderation, and online discussions. KONVENS 2021 will be hosted by Heinrich Heine University Düsseldorf. We are looking forward to your participation.

References

Amy Gutmann and Dennis F Thompson. 1998. *Democracy and disagreement*. Harvard University Press.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Scheffer. 2020. Germeval 2020 task 1 on the classification and regression of cognitive and motivational style from text: Companion paper. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 1–10. CEUR-WS.org.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283.

Vincent Price and Joseph N Cappella. 2002. Online deliberation and its influence: The electronic dialogue project in campaign 2000. *it & Society*, 1(1):303–329.

Fabian Prochazka, Patrick Weber, and Wolfgang Schweiger. 2018. Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism studies*, 19(1):62–78.

Steffen Remus, Rami Aly, and Chris Biemann. 2019. Germeval 2019 task 1: Hierarchical classification of blurbs. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 280–292. German Society for Computational Linguistics and Language Technology (GSCL).

Julian Risch and Ralf Krestel. 2020. Top comment or flop comment? predicting and explaining user engagement in online news discussions. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 579–589. AAAI Press.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 352–363. German Society for Computational Linguistics and Language Technology (GSCL).

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 1–10. Austrian Academy of Sciences.