

CoRAL: a Context-aware Croatian Abusive Language Dataset

Ravi Shekhar¹, Mladen Karan¹, Matthew Purver^{1,2}

¹Cognitive Science Research Group, School of Electronic Engineering & Computer Science,
Queen Mary University of London, UK

²Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
{r.shekhar, m.karan, m.purver}@qmul.ac.uk

Abstract

In light of unprecedented increases in the popularity of the internet and social media, comment moderation has never been a more relevant task. Semi-automated comment moderation systems greatly aid human moderators by either automatically classifying the examples or allowing the moderators to prioritize which comments to consider first. However, the concept of inappropriate content is often subjective, and such content can be conveyed in many subtle and indirect ways. In this work, we propose CoRAL¹ – a language and culturally aware Croatian Abusive dataset covering phenomena of implicitness and reliance on local and global context. We show experimentally that current models degrade when comments are not explicit and further degrade when language skill and context knowledge are required to interpret the comment.

1 Introduction

The growing volume of user-generated content – from social media to online forums and comments under news articles – implies a growing need for moderation of this content to counter abuse and the spread of misinformation. Automatic and semi-automatic moderation systems can greatly aid human moderators, making their work quicker, easier, and more accurate; however, most of this work focuses on English, ignoring smaller, less-resourced languages (Vidgen and Derczynski, 2020). This situation is improving with the advent of multilingual contextual language models, as they enable cross-lingual transfer learning: recent work shows that comment moderation models with reasonable performance for less-resourced languages can be produced using zero- or few-shot transfer learning after pre-training on majority language datasets (Pelicon et al., 2021a,b).

It is not always sufficient to identify whether a comment is inappropriate or not; further sub-categorization helps build measures to counter it. Previous work has taken a range of approaches to sub-categorizing inappropriate content. Waseem et al. (2017) divided abusive language into two orthogonal categories – directed/generalized and implicit/explicit. A very similar approach is taken by Zampieri et al. (2019). More fine-grained approaches include very specific topics such as *homophobia*, *cyberbullying* or *racism* (e.g., Mollas et al., 2022), and the annotation of community-specific extreme hate speech with targets from multiple countries (Maronikolakis et al., 2022); we refer to Poletto et al. (2021) for a comprehensive list. Recently, a unified taxonomy of abusive language categories has been proposed by Banko et al. (2020), a systematic division of slurs by Kurrek et al. (2020), and another taxonomy by Fortuna et al. (2019). Röttger et al. (2021, 2022) provide a detailed empirical analysis of model performance across different example categories. All of these approaches divide comments primarily on the basis of how/whom they insult. In contrast, we are interested in categorizing how such comments can be difficult to classify or interpret automatically due to their use of linguistic and cultural context.

Our goal is to create a dataset and accompanying annotation schema to quantify what categories (primarily related to linguistic and cultural context) of abuse are being used by people and how well NLP models handle these different categories. To this end, we identified three context dependency categories (CDC): Implicitness, Global Context, and Local Context. These CDCs are further sub-divided according to implicitness (explicit/implicit), use of (global/local) language alterations, and use of (global/local) external knowledge; see Section 2 for details. The closest related work in this vein is that of Wiegand et al. (2021), who give a systematic overview of various ways in which examples can be

¹The CoRAL dataset can be found [here](#).

difficult (e.g., sarcasm, dehumanization, inference required, multimodality, etc.). However, [Wiegand et al. \(2021\)](#) only focused on implicit abuse in English without any empirical analysis.

We focus on the Croatian language, a less-represented language in Natural Language Processing research. We annotated 2,240 Croatian comments from the 24sata newspaper² with our proposed CDCs. We experimented with four transformer-based models ([Devlin et al., 2019](#); [Ulčar and Robnik-Šikonja, 2020](#); [Ljubešić and Lauc, 2021](#); [Conneau et al., 2020](#)). Our experimentation shows that models do not perform equally well on all CDCs. The easiest CDC is explicit expression (e.g., cursing or using slurs), confirming the findings of [Wiegand et al. \(2019\)](#). More difficult CDCs are those that require global or local context for their interpretation, via language disguise or external knowledge.

The contribution of this paper is twofold. First, we present a publicly available schema and the *Context-aware Croatian Abusive Language Dataset* (CoRAL) comprised of Croatian news comments annotated for different CDCs. Second, we provide a quantitative and qualitative comparison of comment moderation models, revealing the limitations of different cross-lingual models when handling difficult examples and which CDCs are generally the most challenging.

2 Dataset

When building CoRAL, we aim to have annotated examples with the CDC’s they exhibit. Moreover, we focus on devising CDCs that would reflect the challenges models face when accounting for cultural context (global or local). By manual inspection, we identified three main CDCs of blocked comments on which cross-lingual models tend to fail: *Implicitness*, *Global context*, and *Local context*, which are further divided as follows.

- **Implicitness:** Defines whether examples express abuse directly or indirectly.
 - **Explicit Expression:** directly use abusive words, e.g., derogation, threatening language, slurs, profanity. (e.g. “*Retardiran si.*” [*You are retarded.*])
 - **Implicit Expression:** use indirect ways to express abuse, usually through vague

statements implying abuse without stating it, e.g., sarcastic compliments. (e.g. “*Pametan si ko panj.*” [*You’re as smart as a stump.*])

- **Global Context:** Defines if general background knowledge is required.
 - **Language Independent Disguise:** Linguistic alterations applicable in any language. E.g., adjacent character swap, missing characters/word boundaries, extra spaces, etc. (e.g. “*J**i se.*” [*F**k you.*])
 - **World Knowledge-Based:** The comment requires world/global knowledge (e.g., globally known characters, events, or facts) to be fully understood.(e.g. “*Adolf je bio u pravu.*” [*Adolf was right.*])
- **Local Context:** Defines if Croatia-specific background knowledge is required.
 - **Croatian Specific Disguise:** Linguistic alterations specific to the Croatian language. E.g., ad-hoc constructed words that are understandable to locals, missing/wrong diacritics, using dialects, etc. (e.g. “*Promijenit ću ti lični opis.*” [*I will change your personal description. – I will break your face.*])
 - **Croatia Knowledge-Based:** The comment requires Croatia specific knowledge (e.g., local characters, events, or facts) to be fully understood. (e.g. “*Treba tebe u Vrapče.*” [*You need to be put into Vrapče – Vrapče is a famous mental asylum in Croatia.*])
- **Other:** Anything else not covered above

To the best of our knowledge, CoRAL is the first dataset with annotations on which category of local/global context is required for interpretation.³

Dataset Annotation: We use the publicly available 24sata newspaper comment dataset ([Shekhar et al., 2020](#)).⁴ The dataset contains comments moderated by 24sata’s moderators based on the newspaper’s policy: rules include the removal of hate

³See Appendix 1 for examples of each CDC.

⁴Available at <https://clarin.si/repository/xmlui/handle/11356/1399> ([Pollak et al., 2021](#))

²<https://www.24sata.hr/>

	# Vote				Majority Votes		κ
	0	1	2	3	w Expl.	w/out Expl.	
Explicit Expression	506	425	484	825	1,309	-	0.45
Implicit Expression	1,297	567	275	101	376	363	0.25
Language Independent Disguise	1,941	95	78	126	204	99	0.70
World Knowledge-Based	1,136	571	357	176	533	163	0.31
Croatian Specific Disguise	1,642	312	193	93	286	146	0.30
Croatia Knowledge-Based	2,155	55	26	4	30	14	0.40
Others	1,866	198	103	73	176	175	0.47
Total	-	-	-	-	2,240	931	-

Table 1: Dataset Statistics: First, we report the number of annotators voted(0-3) for CDCs. Then we report with/without Explicit Expression CDC and inter-annotator agreement (Fleiss’ κ), based on the majority votes(i.e., 2 or 3 votes) . The “w/out Explicit” columns for all cases when it is not labeled as Explicit.

	# disagreements	sample size	# ambiguous	# majority ok
Explicit Expression	909 (40.6%)	91	70 (76.9%)	79 (86.8%)
Implicit Expression	842 (37.6%)	85	62 (72.9%)	64 (75.3%)
Language Independent Disguise	173 (7.7%)	50	36 (72.0%)	41 (82.0%)
World Knowledge-Based	81 (2.6%)	50	47 (92.2%)	43 (84.3%)
Croatian Specific Disguise	928 (41.4%)	93	70 (75.3%)	73 (78.5%)
Croatia Knowledge-Based	505 (22.5%)	51	47 (92.1%)	38 (74.5%)
Others	301 (13.4%)	50	40 (76.9%)	43 (82.7%)

Table 2: Analysis of data ambiguity. Columns are (1) number of examples with disagreement for a CDC, (2) size of the sample we annotated, (3) number of examples from the sample annotated as ambiguous (4) number of examples from the sample where the fourth annotator agrees with the majority CDC label of the remaining three.

speech, abusive statements, threats, obscenity, deception & trolling, vulgarity, and comments that are not in Croatian. We refer to [Shekhar et al. \(2020\)](#) for more details, reproduced here in Appendix 2.

We randomly selected 2,240 blocked comments from 2019 related to abuse only (i.e., 24sata’s abuse, hate speech, obscenity, and vulgarity categories). We take a *multi-label* approach: annotators were asked to select all (possibly multiple) CDCs they think apply to the comment; if none applies, then select *Other* and provide an explanation. Three annotators annotated each comment from a total of 7 annotators we had available. All annotators are university students and paid on an hourly basis. Each annotator was provided training and feedback during three pilots.

Dataset Statistics: In Table 1, we present the statistics of the dataset based on the majority CDC label. More than 58% of blocked comments is from *Explicit Expression* CDC, followed by *Croatian Specific Disguise* (23%). To further gain insight into the data, we remove all comments marked *Explicit Expression* CDC. In that case, most comments were from the *Implicit Expression* CDC, followed by *Croatian Specific Disguise*. The *World*

Knowledge-Based comments were less than 1.5%, which might be due to a small volume of world-related articles on the 24sata newspaper.

Inter-Annotator Agreement: The inter-annotator agreement, measured by Fleiss’ κ (Fleiss, 1971) is moderate or better (≥ 0.4) for 4/7 CDCs and fair (≥ 0.2) for the rest (see Table 1). We get the lowest agreement on the *Implicit Expression* CDC (0.25), likely due to this CDC being very subjective. On the other hand, the best agreement is on *Language Independent Disguise* (0.70), which is the most clearly defined CDC.

To further explore agreement, we divided the data into 4 subsets for every CDC, based on the number of annotators who gave a positive vote. 0 and 3 therefore correspond to perfect agreement between the three annotators, while 1 and 2 are disagreement. In Table 1, we provide the statistics of this division. To gain additional insight into the structure of disagreements we sampled 10% (but no fewer than 50) of examples with disagreement for each CDC (see Table 2). One of the authors then annotated these examples with a fourth “expert” CDC label. This additional label matched the majority label in more than 75% of cases for each

CDC label (Table 2, majority column). This indicates that many disagreements could be resolved by additional annotation or use of majority voting; but also that many examples with disagreement are genuinely ambiguous with no clear-cut obviously “correct” choice for the CDC label (multiple choices were all valid to an extent). Consequently, we opted not to force resolution of disagreements, but rather to leave them as part of the data.⁵ We next explore this ambiguity in more detail.

Some tasks are inherently subjective/ambiguous, and their disagreements can never be completely resolved — see (Uma et al., 2021) for a survey — and we believe our task is in this category. To confirm this, we further annotated examples from Table 2 as to their ambiguity (whether multiple choices seemed valid; see Table 2, ambiguous column). We find that for all CDCs, more than 70% of examples with disagreement are indeed ambiguous, explaining the relatively low values shown by traditional agreement measures that assume clear-cut decisions about assigning CDC labels (Table 1). The ambiguity problem is further exacerbated by the multi-label nature of the task, increasing the number of possible CDC label combinations and potential for disagreement. However, much recent work (Pavlick and Kwiatkowski, 2019; Basile et al., 2021; Leonardelli et al., 2021) shows it is possible (and also important) to design NLP models and evaluation measures that take task ambiguity into account. Consequently, we believe that CoRAL will be valuable for future research.

To get a better perspective on comments to which the majority of annotators assigned the *Other* label, an author manually inspected randomly selected 50 examples labeled with the *Other* CDC and 50 examples labeled with some other CDC. Examples labeled as *Other* were mainly spam or non-offensive (mislabeled) comments. In contrast, different CDC examples were mostly offensive, fitting well into one or more of the main six CDC categories. The latter case accounts for the majority of examples.

3 Results and Discussion

3.1 Experimental Set-up

For binary classification (i.e., *Abuse* vs. *Non-abuse*), we used the dataset from Pelicon et al. (2021b). We removed comments blocked for spam, deception & trolling and use of a language other

⁵We release all individual annotations, not only the majority vote based decisions.

than Croatian, giving 4750/518/580 data points for training/validation/testing, respectively. We used four transformer-based models; two pre-trained on 100+ languages, namely mBERT (Devlin et al., 2019) and XLM-RoBERTa base (Conneau et al., 2020) and two pre-trained on Croatian and 2-3 similar languages, namely cseBERT (Ulčar and Robnik-Šikonja, 2020) and BERTić (Ljubešić and Lauc, 2021). We fine-tuned all models for the binary comment moderation task using default hyperparameters for ten epochs, and selected the best model based on validation F1 score.⁶

3.2 Quantitative Results

Our primary goal is to study how models perform on fine-grained CDCs, and we report accuracy on CoRAL in Table 3. This number represents the proportion of comments from CoRAL that a classification model (*Abuse* vs. *Non-abuse*) classified as *Abuse* (by construction, all examples in CoRAL should belong to *Abuse*). We present the overall accuracy of each annotated CDC with and without the *Explicit expression* CDC. There are multiple insights from the results. For all CDCs except *Other*, cseBERT and BERTić perform best. We confirm this using a permutation test (Nichols and Holmes, 2002): for all CDCs except *Other* the differences between the better of cseBERT/BERTić and the better of mBERT/XLM-RoBERTa, are statistically significant ($p \leq 0.05$). This again shows that a small multilingual Masked Language Model (MLM) with similar languages beats a massively multilingual MLM, similar to Pelicon et al. (2021b).

Among all the CDCs, all models can easily identify the *Explicit Expression* examples. Comparatively, *Implicit Expression* is one of the most challenging CDC, with more than 40% difference between it and *Explicit Expression*. This shows that it is hard for any model to identify implicit expression. At the same time, the *Language Independent Disguise* CDC is easier for models than the *Croatian Specific Disguise* CDC, with more than 7% difference in the performance. On the *Croatian Knowledge-Based* comments, cseBERT and BERTić outperform mBERT and XLM-RoBERTa by a minimum 11%. This, again, indicates that smaller multilingual MLM has comparatively more cultural information encoded.

⁶On the corresponding test set, our model achieved macro F1 scores of 75.14, 76.72, 79.82, and 80.97 for mBERT, XLM-RoBERTa, cseBERT, and BERTić, respectively, which is similar to previously reported results (Pelicon et al., 2021b).

CDC Includes Explicit Expression	mBERT		XLM-RoBERTa		cseBERT		BERTiĆ	
	Yes	No	Yes	No	Yes	No	Yes	No
Overall	45.04	23.85	44.24	19.76	<u>56.70</u>	28.46	59.64	<u>26.53</u>
Explicit Expression	60.12	-	61.65	-	<u>76.78</u>	-	83.19	-
Implicit Expression	21.54	21.76	16.49	16.53	26.86	26.45	26.86	<u>25.90</u>
Language Independent Disguise	58.33	49.49	59.80	50.51	<u>74.51</u>	65.66	77.94	65.66
World Knowledge-Based	33.33	21.43	13.33	7.14	50.00	28.57	<u>46.67</u>	<u>21.43</u>
Croatian Specific Disguise	49.72	31.29	50.28	26.38	<u>64.92</u>	<u>40.49</u>	70.73	42.33
Croatia Knowledge-Based	40.91	23.29	38.81	15.75	<u>51.40</u>	<u>23.29</u>	55.59	27.40
Others	11.93	11.43	8.52	8.00	<u>11.36</u>	<u>10.86</u>	6.25	5.71

Table 3: Accuracy of the abusive comment on different CDCs. A_1/A_2 where A_1 is accuracy on the unmodified test set and A_2 after removing *Explicit Expression* examples. The best model is **bold** and second best underlined.

CDC	XLM-RoBERTa				BERTiĆ			
	0	1	2	3	0	1	2	3
Explicit Expression	13.64	27.06	47.93	69.70	16.60	38.35	71.07	90.30
Implicit Expression	57.75	31.75	18.91	9.90	74.79	46.74	29.45	19.80
Language Independent Disguise	42.09	54.74	58.97	60.32	57.24	69.47	83.33	74.60
World Knowledge-Based	38.73	49.56	47.34	56.25	51.94	64.62	69.47	73.30
Croatian Specific Disguise	45.13	44.55	39.90	36.56	60.54	58.65	52.33	62.37
Croatia Knowledge-Based	44.78	40.00	15.38	0.00	59.91	56.36	42.31	75.00
Others	50.21	19.70	9.71	6.85	69.40	15.15	4.85	8.22

Table 4: Performance of XLM-RoBERTa & BERTiĆ based models per CDC based on number of annotator’s votes.

To better understand the effect of the *Explicit Expression* comments, we removed all data points assigned the *Explicit* CDC label; results in Table 3. Overall performance drops by $\geq 22\%$, with a larger drop for cseBERT and BERTiĆ ($\geq 28\%$). For both *Local Context* CDCs, there is a larger drop in performance ($\geq 26\%$). This suggests we must find a better way to incorporate cultural knowledge into models. Furthermore, in Table 4 we report the performance based on the number of annotator’s votes, and show that our main observations still hold and are even more pronounced when considering data with high agreement.

3.3 Qualitative Results

Manual inspection of errors reveals some interesting patterns. Cases where all models fail almost always contain two or more CDCs simultaneously, e.g., “*Severaca moze glumiti jedino na camcu*” [*The only place where Severaca can act is a boat.*] – deliberate misspelling, reference to famous person, reference to local event).⁷ Moreover, examples where cseBERT and BERTiĆ outperform mBERT and XLM-RoBERTa mostly require local context: e.g., “*Opet. Retardesničaru.*” [*Again. You retarded right-wing extremist.*] – specific local word, wordplay only possible in Croatian). Finally,

⁷Severaca refers to Severina, a regionally famous singer who was in a leaked explicit video taking place on a boat. The comment implies her acting skills are limited to pornography.

we find that examples on which all models perform well mostly contain explicit abuse with no misspelling, e.g., “*Retard*” [*Retard*], which is in line with our empirical results.

4 Conclusion

We present the *Context-aware Croatian Abusive Language Dataset* (CoRAL), a dataset annotated with context dependency categories (CDC) of problematic examples for Croatian comment moderation. We annotated 2240 blocked comments for Explicitness, Implicitness, Language Independent Disguise, World Knowledge-Based, Croatian Specific Disguise, and Croatia Knowledge-Based. We found that only 58.44% had explicit expressions of abuse. This indicates that almost half the remaining examples are challenging (Croatian Specific Disguise alone accounting for $\approx 24\%$). This shows that addressing these categories of examples is very practically relevant. We tested four transformer-based models and found that explicit comments are the easiest and local context ones are hardest. We also found that language-specific multilingual language models better identify Croatian-specific blocked comments. Finally, we believe that CoRAL will help design better models for Croatian comment moderation, build a foundation for creating similar datasets in other languages, and develop novel methods by incorporating local context.

Ethical Consideration

Our proposed dataset and models are to support more accurate and robust detection of online abuse. We anticipate that the high-quality and fine-grained CDC labels in the dataset will advance research on online hate for low-resource languages. The dataset and models we present could, in principle, be used to train a generative hate speech model, but this is already possible using much larger datasets. Alternatively, the dataset and models could be used to understand current detection tools' limitations better and then attack them. However, we believe malicious actors are already manually employing similar attack methods to bypass the content rules of different platforms. Therefore, we believe that it is essential to understand how to attack the models and that our dataset will help the community fight such behavior by creating a more diverse dataset that leads to more robust models.

Acknowledgements

We thank anonymous reviewers for their valuable feedback, and acknowledge financial support from several sources: the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103) and the project Sovrag (Hate speech in contemporary conceptualizations of nationalism, racism, gender and migration, J5-3102); the UK EPSRC via the project Sodestream (Streamlining Social Decision Making for Improved Internet Standards, EP/S033564/1); and via the project RobaCOFI (Robust and adaptable comment filtering), which indirectly received funding from the European Union's Horizon 2020 research and innovation action programme via the AI4Media Open Call #1 issued and executed under the AI4Media project (Grant Agreement no. 951911). This paper reflects only the authors' views; the EC and the AI4Media project are not responsible for any use that may be made of the information contained herein.

References

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the fourth workshop on online abuse and harms*, pages 125–137.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. [We need to consider disagreement in evaluation](#). In *Proceedings of*

the 1st Workshop on Benchmarking: Past, Present and Future, pages 15–21, Online. Association for Computational Linguistics.

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Paula Fortuna, João Rocha da Silva, Juan Soler-Company, L. Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled portuguese hate speech dataset. *Proceedings of the Third Workshop on Abusive Language Online*.
- Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539.
- Nikola Ljubešić and Davor Lauc. 2021. [BERTiC - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian](#). In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.
- Antonis Maronikolakis, Axel Wisioerek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schuetze. 2022. [Listening to affected communities to define extreme speech: Dataset and experiments](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104, Dublin, Ireland. Association for Computational Linguistics.
- Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16.

- Thomas E Nichols and Andrew P Holmes. 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlić, Matthew Purver, and Senja Pollak. 2021a. [Zero-shot cross-lingual content filtering: Offensive language and hate speech detection](#). In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 30–34, Online. Association for Computational Linguistics.
- Andraž Pelicon, Ravi Shekhar, Blaž Škrlić, Matthew Purver, and Senja Pollak. 2021b. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.
- Senja Pollak, Marko Robnik-Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjic, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlić, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose G. Moreno, Antoine Doucet, and Hannu Toivonen. 2021. [EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions](#). In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 99–109, Online. Association for Computational Linguistics.
- Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. [Multilingual Hate-Check: Functional tests for multilingual hate speech detection models](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 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)*, pages 41–58.
- Ravi Shekhar, Marko Pranjic, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: Benchmarking in croatian and estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).
- M. Ulčar and M. Robnik-Šikonja. 2020. [FinEst BERT and CroSloEngual BERT: less is more in multilingual models](#). In *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.
- Zeerak Waseem, Thomas Davidson, NY Ithica, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *ACL 2017*, page 78.
- Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language—what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. 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)*, pages 602–608.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*, pages 1415–1420.

Appendix 1: Dataset Categories Examples

In this section we provide some examples for the different categories.

Explicit Expression

- Use of Derogation (*ti si nitko i ništa – You are a nobody.*)
- Threatening Language (*saznat ću gdje živiš – I will find out where you live.*)
- Slur (*retard – retard*)
- Profanity (*peder – fag*)

Implicit Expression :

- Abuse expressed using negated positive statements (*“Gej je okej” je krivo – “Gay is ok” is wrong.*)
- Abuse phrased as a question (*Zašto moramo tolerirati imigrante? – Why do we have to tolerate immigrants?*) Abuse phrased as an opinion (*Staljin je imao pravi pristup. – Stalin had the right approach.*)

Language Independent Disguise

- Swaps of adjacent characters (*jeib se – f*ck you*)
- Missing characters (*jbi se*)
- Missing word boundaries (*jebise*)
- Missing word boundaries (*jebise*)
- Added spaces between chars (*j ebi se*)
- Added spaces between chars (*j ebi se*)
- Added spaces between chars (*j ebi se*)
- substituting characters with “*”, “.” or similar. (*je*i se*)
- Leet speak spellings (*j3b1 5e*).

World Knowledge-Based

- Momentary (knowledge of characters/events important at this point in time or for a relatively limited time) - e.g., *Will Smith oscar slap, Brexit*
- Long-term (more stable general knowledge) - e.g., *The Pope, Berlin wall, The Beatles*

Croatian Specific Disguise

- ad-hoc constructed words that are understandable to locals (*Svi prekodrinaci su ološ – All X are scum.*, where X = *prekodrinac*, an ad hoc invented word from *preko* ("across") and *Drina* (name of a river) denoting someone living across the Drina river – i.e., Serbs),
- misspelt important words in a way that is specific for Croatian, mostly diacritics missing or wrong, like *dj/dz* for *đ/dž*, *djubre/dubre* (instead of *đubre* - piece of shit), *cetnik* (instead of *četnik* - member of a very unpopular military group),
- using dialects (some abuse can sound very different in some dialects, containing words like *Flundra, Droca, Štraca* – easy woman),
- idioms specific for Croatian (*Promijenit ću ti lični opis. – I will change your personal description. i.e., I will break your face.*).
- other ways of using non-abusive words to create abusive context (which requires language knowledge to properly decipher) - sarcasm, substituting slurs for similarly sounding non-slurs, inventing abusive comparisons without abusive words on the spot - e.g., *bistar si ko mocvara* (*your thinking is clear as a swamp*), *u gnjurac* (*gnjurac* is a bird, but sounds similar to *kurac* – *d*ck*), referring to a person from the sea side as *Tovar* (literal meaning is Donkey)

Croatia Knowledge-Based

- Momentary (knowledge of characters/events/facts important at this point in time or for a relatively limited time), e.g. *Vili Beroš* (health minister during the Covid 19 pandemic), *Uspinjača na sljeme* (controversial building project),
- Long-term (more stable local knowledge) - e.g., *HDZ* (a political party around for a long time), *‘91* (year of the Croatian war of independence), *Vrapče* (one of the most widely known Psychiatric institutions)

Appendix 2: Rule Description

We have reproduced rule description from [Shekhar et al. \(2020\)](#) in Figure 1.

Rule ID	Description	Definition	Severity
1	Disallowed content	Advertising, content unrelated to the topic, spam, copyright infringement, citation of abusive comments or any other comments that are not allowed on the portal	Minor
2	Threats	Direct threats to other users, journalists, admins or subjects of articles, which may also result in criminal prosecution	Major
3	Hate speech	Verbal abuse, derogation and verbal attack based on national, racial, sexual or religious affiliation, hate speech and incitement	Major
4	Obscenity	Collecting and publishing personal information, uploading, distributing or publishing pornographic, obscene, immoral or illegal content and using a vulgar or offensive nickname that contains the name and surname of others	Major
5	Deception & trolling	Publishing false information for the purpose of deception or slander, and “trolling” - deliberately provoking other commentators	Minor
6	Vulgarity	Use of bad language, unless they are used as a stylistic expression, or are not addressed directly to someone	Minor
7	Language	Writing in other language besides Croatian, in other scripts besides Latin, or writing with all caps	Minor
8	Abuse	Verbally abusing of other users and their comments, article authors, and direct or indirect article subjects, calling the admins out or arguing with them in any way	Minor

Table 1: Annotation schema for blocked comments, 24sata.

Figure 1: Rule description, reproduced from [Shekhar et al. \(2020\)](#)