To Block or not to Block: Experiments with Machine Learning for News Comment Moderation

Damir Korenčić
Rudjer Boskovic Institute, Croatia
damir.korencic@irb.hr

Ipek Baris
University of Koblenz-Landau, Germany
ibaris@uni-koblenz.de

Eugenia Fernandez
euge.ft@gmail.com

Katarina Leuschel
katarina.leuschel@gmail.com

Eva Sánchez Salido
UNED, Spain
esanchez1751@alumno.uned.es

Abstract

Today, news media organizations regularly engage with readers by enabling them to comment on news articles. This creates the need for comment moderation and removal of disallowed comments – a time-consuming task often performed by human moderators. In this paper we approach the problem of automatic news comment moderation as classification of comments into *blocked* and *not blocked* categories. We construct a novel dataset of annotated English comments, experiment with cross-lingual transfer of comment labels and evaluate several machine learning models on datasets of Croatian and Estonian news comments.

1 Introduction

Comment sections are an important part of news sites, providing an opportunity for newsrooms to engage with their audience. Comment moderation aims to safeguard respectful conversation by blocking comments that are uncivil, disruptive or potentially unlawful. This is a complex task that balances legal implications and editorial guidelines. Common categories of blocked comments include: unsafe or illegal content (ex. defamation or hate speech), disruptive content (ex. trolling), advertisements, and copyrighted content (Risch and Krestel, 2018).

While newsrooms are becoming increasingly aware of the benefits provided by artificial intelligence and expect comment moderation to become more manageable, implementation of AI solutions is far from prevalent (Society of Editors, 2018; Beckett, 2019). Some newsrooms use custom automated comment moderation solutions developed in-house or third-party plugins to complement human moderation. Others rely on external companies that provide comment moderation performed by teams of contracted moderators (Society of Editors, 2018; Beckett, 2019; Woodman, 2013).

For most in-house and third-party solutions, the extent of use and details of the machine learning solutions are not publicly revealed. The stand-out third-party option is Perspective,[1] a free API developed by Google's Jigsaw, available in seven high-resourced languages (Beckett, 2019). To the best of our knowledge, there are no machine learning solutions suitable for comment moderation for under-resourced languages.

In the academic literature, the problem of comment moderation is commonly approached as a binary classification of comments into *blocked* and *not blocked* categories (Pavlopoulos et al., 2017; Risch and Krestel, 2018; Shekhar et al., 2020). In this paper, which reports the work done during the EMBEDDIA Hackashop hackathon[2] (Pollak et al., 2021), we approach the problem in the same manner and perform experiments with comment classification on datasets of Croatian and Estonian news comments (Shekhar et al., 2020).

Motivated by the lack of an English dataset of comments labelled as either blocked or not blocked, we construct such a dataset from existing datasets of news and social media comments. We then experiment with the cross-lingual transfer of English labels to Croatian and Estonian comment datasets by means of a multilingual BERT model (Pires et al., 2019; Ulcar and Robnik-Sikonja, 2020). Finally, we construct and evaluate several classification models trained on Croatian and Estonian datasets, analyze the results, and discuss the problem of automatic detection of blocked comments. We make the source code of the experiments freely available.[3]

[1] https://www.perspectiveapi.com
[2] http://embeddia.eu/hackashop2021/
[3] https://github.com/eugeniaft/embeddia-hackathon

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 127–133
April 19, 2021 © Association for Computational Linguistics

2 Related Work

Computational comment moderation includes tasks such as offensive language detection (Schmidt and Wiegand, 2017) and blocked comment detection (Risch and Krestel, 2018; Pavlopoulos et al., 2017; Napoles et al., 2017), which is the focus of our study. Most of the prior studies on comment filtering tackle the problem using text from high-resourced languages such as English (Napoles et al., 2017; Kolhatkar et al., 2019) and German (Risch and Krestel, 2018). There are only a few studies that focus on low-resourced languages (Shekhar et al., 2020; Pavlopoulos et al., 2017).

The methods for comment filtering vary from classical machine learning methods to deep learning approaches. Risch and Krestel (2018) classify comments with a logistic regression classifier using features computed from comments, news articles, and users. Deep neural networks such as RNN and CNN have also been applied (Pavlopoulos et al., 2017). Most recently, Shekhar et al. (2020) leverage Multilingual BERT (mBERT) (Devlin et al., 2019) for the moderation of news comments in Balto-Slavic languages.

3 English Dataset for Comment Moderation

There are multiple publicly available datasets in English with annotated comments that have been used in previous research about comment moderation. However, most of these datasets contain annotations of only a subset of the categories of blocked comments (Shekhar et al., 2020).

We construct a large corpus of comments containing different categories of blocked comments by unifying different datasets and defining a new label. Since comments in these datasets are not explicitly labeled as blocked, we created the *flagged* and *not flagged* labels instead. The idea is to identify comments that moderators should review and decide whether to block them or not. The *flagged* label therefore serves as an approximation of the blocking decision and classifiers that detect it automatically have the potential to save time and human effort.

3.1 Construction of the Dataset

We used five different datasets containing annotated comments from news articles, social media, and other fora. We included comments from platforms outside of news media since users are subject to a similar set of rules related to what content they can share.[4,5,6] Each dataset contains different annotations, including comments rated on a scale of toxicity, comments labelled for hateful speech and abuse, comments labeled for constructiveness and tone, etc. Our challenge was to define the labelling criteria for the binary labels *flagged* and *not flagged* and consistently apply them to the labels in the five datasets. Flagged comments are the comments most likely to require blocking based on the existing labels in the datasets, and are labeled according to the principles discussed in (Risch and Krestel, 2018) and guidelines for comment moderation in (Society of Editors, 2018) and (Woodman, 2013).

Data Source	# not flagged	# flagged	% flagged
SOCC	1,012	31	3%
YNACC	7,076	2,084	23%
DETOX	19,153	3,372	15%
Trawling	5,009	7,189	59%
HASOC	4,443	2,538	36%
Final dataset	36,693	15,214	29%

Table 1: Data source and class distribution statistics for the English dataset of flagged comments.

Our dataset consists of comments from the SOCC corpus (SFU Opinion and Comments Corpus) (Kolhatkar et al., 2019), YNACC corpus (The Yahoo News Annotated Comments Corpus) (Napoles et al., 2017), DETOX corpus (Wulczyn et al., 2017), Trawling corpus (Hitkul et al., 2020), and HASOC corpus (Hate Speech and Offensive Content Identification in Indo-European Languages) (Mandl et al., 2019). SOCC contains annotated comments from opinion articles. We used the *constructiveness* and *toxicity* labels and flagged comments whenever the toxicity level was *toxic* or *very toxic* and *not constructive*. YNACC contains expert annotated comments in online news articles. A comment was labeled flagged whenever a comment was *insulting*, *off-topic*, *controversial* or *mean* and *not constructive*. DETOX has comments from English Wikipedia talk pages. It contains annotations for *attack*, *aggression* and *toxicity*. A comment was labelled flagged whenever it was *toxic*, *aggressive* or if it contained an *attack*. We only included data from 2015. The Trawling data

[4] https://help.twitter.com/en/rules-and-policies/twitter-rules

[5] https://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines

[6] https://www.redditinc.com/policies/content-policy

Dataset	Example	Original Label
SOCC	This has to have been written by Chinese government sponsored propagandists.	Non-constr. & Toxic
YNACC	You and at least one other person are pretty dumb, huh? Unless you have two accounts, right, moron?	Mean & Off-topic
DETOX	You should block this idiot for life!	Aggressive
Trawling	So nowadays they let models have greasy unwashed hair and man hands?	Trolling
HASOC	Too many doctors on my fucking Facebook fuck off	Hateful or Offensive

Table 2: Examples of flagged comments.

includes samples of comments from Twitter, Reddit and Wikipedia talk pages. Comments are provided with the labels *Normal*, *Profanity*, *Trolling*, *Derogatory* and *Hate Speech*. A comment was labeled as *flagged* if it belonged to any of the categories except for *Normal*. Lastly, HASOC is composed of comments from Twitter and Facebook and has annotations on whether comments are *hateful*, *offensive* or neither. We included only the English comments and labelled them as *flagged* if they were either *hateful* or *offensive*.

The resulting dataset contains 51,907 labeled comments, 29% of those being flagged comments. Table 1 gives more details on the class distribution and Table 2 contains examples of comments from each dataset that have been labelled as flagged. The dataset can be easily reconstructed by using the code we make available and applying it to the individual sub-datasets which are freely available.

3.2 Classification Experiments

We run a set of experiments to evaluate the performance of classifiers on our dataset. We split our data into train, validation, and test sets using stratified sampling to account for class imbalance. In our first experiment, we trained a Logistic regression classifier and Support vector machine classifier with linear kernel. We later fine-tuned two different multilingual BERT models: CroSloEnBERT and FinEstEnBERT(Ulcar and Robnik-Sikonja, 2020). See Section 4.2 for more details about how the models were optimized and fine-tuned.

The results of the classification experiments are in Table 3. All trained models perform better than the baseline classifier that always chooses the minority class *flagged*. The non-neural classifiers have higher recall whilst the multilingual BERT models have higher F_1 score, accuracy, and precision. The classification results support the claim that the constructed *flagged* label is well-defined and consistent and that our dataset can be further used in research related to comment moderation.

Model	F_1	Prec.	Recall	Acc.
baseline	0.453	0.293	1.000	0.293
LogReg	0.732	0.710	**0.755**	0.838
SVM	0.728	0.725	0.730	0.840
BERT-CroSloEn	0.761	**0.871**	0.675	0.876
BERT-FinEst	**0.777**	0.841	0.722	**0.878**

Table 3: Classification results on English comments labeled as flagged or not flagged. F_1, precision and recall are reported for the class of flagged comments.

4 Automatic Comment Moderation Experiments

Next, we construct and evaluate classifiers that aim to detect blocked news comments. We experiment with EMBEDDIA multilingual BERT models (Ulcar and Robnik-Sikonja, 2020) fine-tuned for classification and with standard non-neural classifiers using n-gram features.

4.1 News Comment Datasets

We use the Ekspress dataset of Estonian news comments and the 24Sata dataset of Croatian news comments (Shekhar et al., 2020). Following Shekhar et al. (2020) we focus on the comments from 2019 that have labels of higher quality. The Estonian comments are simply labelled as either blocked or not blocked, while the blocked Croatian comments are further divided into eight subcategories. We remove the subcategories 2, 4 and 7 that contain either a negligible amount of comments or non-Croatian comments. We also remove all the non-Estonian comments from the Ekspress dataset. After cleaning, 816,131 Croatian and 865,022 Estonian comments remain. Both datasets are unbalanced – only 7.77% of Croatian and 8.99% of Estonian comments are labeled as blocked.

4.2 Classification Experiments

We solve the problem of binary classification of comments into *blocked* and *not blocked* categories. We train and evaluate the comment classifiers using

	24Sata dataset (Croatian)				Ekspress dataset (Estonian)			
Model	F_1	Precision	Recall	Accuracy	F_1	Precision	Recall	Accuracy
baseline	0.144	0.078	1.000	0.078	0.165	0.090	1.000	0.090
BERT-en	0.229	0.189	0.291	0.843	0.216	0.182	0.264	0.827
BERT-en-nat	0.514	**0.960**	0.350	0.948	0.479	0.782	0.345	0.933
BERT-native	**0.535**	0.904	0.379	0.949	0.459	**0.824**	0.319	0.933
LogReg-F1	0.502	0.828	0.360	0.944	**0.532**	0.712	0.425	0.933
LogReg-recall	0.384	0.311	**0.503**	0.875	0.236	0.149	**0.565**	0.671

Table 4: Classification results for the problem of detection of blocked comments.

stratified train/development/test subsets containing 80,000/15,000/15,000 comments.

First we experiment with the two multilingual BERT models CroSloEnBERT and FinEstEnBERT (Ulcar and Robnik-Sikonja, 2020), fine-tuned for classification. We rely on the Huggingface library (Wolf et al., 2020) and use the tokenizers embedded in the BERT models, limiting the number of tokens to 128. For each dataset, we build three fine-tuned BERT models. The first model, labeled *BERT-en* and also evaluated in Section 3.2, is fine-tuned only on English comments. The second model, labeled *BERT-nat*, is fine-tuned only on the target (native) language (Croatian or Estonian). The third model is produced by fine-tuning the English model on the dataset in the target language, and labeled as *BERT-en-nat*. We train the models by setting the batch size to 16 and number of epochs to 3, and perform optimization using Adam with weight decay (Loshchilov and Hutter, 2019). We select the models that exhibit the best accuracy in the training phase.

The second classification approach is based on two standard non-neural classifiers - Logistic regression and Support vector machine with linear kernel. Both classifiers are available as part of the scikit-learn[7] framework Buitinck et al. (2013). To perform model selection we vary both the regularization strength and the method of feature construction. We find the optimal model parameters by performing a grid search on separate train and test sets containing 40,000 and 10,000 comments. Two optimization criteria are used: F_1 score and recall. The search for a model with high recall is motivated by the observation that the majority of the models tend to favor high precision. We find that the Logistic regression offers better performance across both datasets, and that the best choice of features is the binary bag-of-words-and-bigrams vector.

The classification results are displayed in Table 4. The performance scores are modest in terms of F_1 and show sharp precision/recall tradeoffs. All of the models outperform the baseline classifier that always chooses the minority class. Accuracy scores are deceptively high due to the prevalence of the non-blocked comments in the datasets. BERT classifiers perform better on Croatian than on Estonian comments, possibly because of differences in the original multilingual BERT models. BERT models fine-tuned only on the English dataset of flagged comments have weak but above-baseline performance, which shows that a certain amount of cross-language knowledge transfer is achieved. The weak performance could be explained both by the language difference and the fact that the English dataset represents an approximation of the blocked comments class.

Shekhar et al. (2020) classify comments from the same datasets, train the models on data containing an equal share of blocked and not blocked comments, and report recall of 0.67, precision of 0.27, and F_1 of 0.38 for the Croatian comments. This result is in line with the sharp precision/recall tradeoffs we observe. Balanced training data in (Shekhar et al., 2020) is a possible reason for higher recall scores obtained (0.70 on the Croatian and 0.58 for the Estonian dataset).

Lastly, we examine the classifiers' performance on sub-categories of blocked Croatian comments detailed in (Shekhar et al., 2020). Table 5 contains recall scores achieved by the BERT-en model trained on the English dataset, BERT-native model trained on the Croatian dataset, the Logistic regression model, and the mBERT model of Shekhar et al. (2020) that is also trained on the Croatian dataset. The performances of the Logistic regression model and the mBERT model demonstrate the benefit of optimizing for recall. The BERT-en model achieves competitive results on the "Vulgarity" and "Abuse" categories, showing that detection of these types

[7] https://scikit-learn.org

Model	Disallowed	Hate Speech	Deception&Trolling	Vulgarity	Abuse	All Blocked
BERT-en	0.102	0.333	0.149	0.739	0.514	0.291
BERT-native	0.432	0.510	0.299	0.435	0.324	0.379
LogReg-recall	0.515	0.647	0.388	0.783	0.473	0.503
mBERT	0.642	0.722	0.546	0.881	0.723	0.673

Table 5: Recall on the subcategories of blocked Croatian comments.

of blocked comments was successfully transferred from the English dataset. Better results on other categories could be achieved by augmenting the English dataset with additional flagged comments containing deception and misinformation, as well as the spam and copyright infringement content pertaining to the “Disallowed” category.

5 Discussion

Automatic detection of blocked comments of the Croatian and the Estonian dataset is a hard problem. This claim is supported by modest F_1 scores and sharp precision/recall tradeoffs observed both in our experiments and in the experiments of Shekhar et al. (2020). While inclusion of non-textual comment features would probably lead to better results (Risch and Krestel, 2018), we hypothesize that the main problem is the poor quality of comment labelling.

The definition of sensible text categories and consistent annotation of texts with these categories falls within the domain of content analysis (Krippendorff, 2012). Ideally, the category definitions are discussed and fine-tuned, and the measure of inter-annotator agreement (IAA) is reported. In the case of the blocked comment detection, the precise process of category definition is unknown (Pavlopoulos et al., 2017; Risch and Krestel, 2018; Shekhar et al., 2020), while the IAA is either not available (Risch and Krestel, 2018; Shekhar et al., 2020), or modest (Pavlopoulos et al., 2017). Moreover, there are indications of inconsistencies in the definition of a blocked comment class. Shekhar et al. (2020) report that the varying blocking rates are probably caused by changes in moderation policy. Pavlopoulos et al. (2017) and Risch and Krestel (2018) report that a high influx of user comments, for example during high-interest events, causes more strict comment blocking. The mentioned problems should be tackled since the consistent labelling of the comments is key to building high-quality classifiers.

The binary classification approach might be in disconnect with the true needs of the comment moderators. An engineering perspective of a machine learning system can significantly differ from the end user’s perspective (Lee et al., 2017). We believe that studies including comment moderators are essential in order to define and evaluate the appropriate solution. For example, the amount of moderators’ time saved might prove as a useful metric, and the best application of classifiers might not be automatic blocking but flagging and pre-filtering of comments.

Additionally, moderators operate within boundaries set by in-house rules and practices and legal regulations. An investigation of the nature and impact of such restrictions would provide perspective on the role of automatic comment moderation. For example, in a scenario where the publisher can be held accountable for the comments containing hate speech, any automatic classifier would be required to achieve very high recall.

6 Conclusion and Future Work

We plan to further develop the dataset of flagged English comments, experiment with other classification models and to improve the BERT-based language transfer models. We also plan to examine multi-task learning approaches that can lead to state-of-art results on transferring knowledge among related tasks (Zhang and Yang, 2017).

We believe that more attention should be paid to the problem of comment labelling. This could lead to better classifiers, reliable inter-annotator agreement scores that can serve as upper bounds on performance, and to a better understanding of the semantics of the composite category of blocked comments.

In our view, an essential future work direction is design and implementation of studies with comment moderators that examine real-world scenarios and user needs. We believe that such studies would be invaluable and would lead to more realistic and usable machine learning comment moderation tools.

References

Charlie Beckett. 2019. New powers, new responsibilities: A global survey of journalism and artificial intelligence. *Polis, London School of Economics and Political Science. https://blogs. lse. ac. uk/polis/2019/11/18/new-powers-new-responsibilities*.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. 2013. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Hitkul, Karmanya Aggarwal, Pakhi Bamdev, Debanjan Mahata, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2020. Trawling for trolling: A dataset. *CoRR*, abs/2008.00525.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2019. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, pages 1–36.

Klaus Krippendorff. 2012. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Courtney Napoles, Joel R. Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *LAW@ACL*, pages 13–23. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *ALW@ACL*, pages 25–35. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *ACL (1)*, pages 4996–5001. Association for Computational Linguistics.

Senja Pollak, Marko Robnik Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, 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 Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose 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*. Association for Computational Linguistics.

Julian Risch and Ralf Krestel. 2018. Delete or not delete? semi-automatic comment moderation for the newsroom. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 166–176.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *SocialNLP@EACL*, pages 1–10. Association for Computational Linguistics.

Ravi Shekhar, Marko Pranjic, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: Benchmarking in croatian and estonian. *Special Issue on Offensive Language*, page 49.

Society of Editors. 2018. Moderation guide. page 42. [Online; accessed 21-February-2021].

Matej Ulcar and Marko Robnik-Sikonja. 2020. Finest BERT and crosloengual BERT - less is more in multilingual models. In *TDS*, volume 12284 of *Lecture Notes in Computer Science*, pages 104–111. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Emma Woodman. 2013. Online comment moderation: emerging best practices. a guide to promoting robust and civil online conversation. pages 45–46. the World Association of Newspapers (WAN-IFRA). [Online; accessed 21-February-2021].

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *WWW*, pages 1391–1399. ACM.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *CoRR*, abs/1707.08114.