The Role of Context in Detecting the Target of Hate Speech

Ilia Markov
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
i.markov@vu.nl

Walter Daelemans
CLiPS, University of Antwerp
Antwerp, Belgium
walter.daelemans@uantwerpen.be

Abstract

Online hate speech detection is an inherently challenging task that has recently received much attention from the natural language processing community. Despite a substantial increase in performance, considerable challenges remain and include encoding contextual information into automated hate speech detection systems. In this paper, we focus on detecting the target of hate speech in Dutch social media: whether a hateful Facebook comment is directed against migrants or not (i.e., against someone else). We manually annotate the relevant conversational context and investigate the effect of different aspects of context on performance when adding it to a Dutch transformer-based pre-trained language model, BERTje. We show that performance of the model can be significantly improved by integrating relevant contextual information.

1 Introduction

Hate speech detection models play an important role in online content moderation and promotion of healthy online debates (Halevy et al., 2020). This has motivated a considerable interest in the task within a variety of disciplines, including social sciences and the natural language processing (NLP) community.

Recent advances in the field of NLP, which include the use of deep learning and ensemble architectures, have led to the development of automated hate speech detection approaches with an increased performance (Kumar et al., 2020; Zampieri et al., 2020; Markov and Daelemans, 2021). However, the task remains challenging from multiple perspectives, e.g., the use of figurative language and cross-domain scalability, amongst others (van Aken et al., 2018; Vidgen and Derczynski, 2020; Pamungkas et al., 2021; Lemmens et al., 2021). These challenges constrain the performance and generalizability of hate speech detection models, and include the problem of integrating contextual information, that is, improving hate speech detection models by making them context aware (Pavlopoulos et al., 2020; Menini et al., 2021; Vidgen et al., 2021).

Modeling contextual information is indisputably important for developing robust hate speech detection systems (de Gibert et al., 2018; Pavlopoulos et al., 2020; Vidgen et al., 2021). For instance, the comment 'go back home' is clearly hate speech if it is posted under a news article about refugees and asylum seekers. However, previous work on detecting both the type and target of online hate speech has mostly focused on message content alone, without accounting for the context of the target comments (Risch and Krestel, 2020; Zampieri et al., 2020). This is partially related to the lack of contextual information in the vast majority of datasets annotated for hate speech, which implies that hate speech detection models cannot exploit the conversational context when they are trained on existing datasets (Vidgen and Derczynski, 2020).

More recent studies have specifically looked into the effect of context on hate speech detection. For instance, Pavlopoulos et al. (2020) experimented with various strategies for integrating contextual information into BiLSTM and BERT models, where context is limited to the preceding ('parent') comment in the Wikipedia conversations dataset. The authors report that though context significantly affects annotation process by both amplifying or mitigating the perceived toxicity of posts, they found no evidence that adding context leads to a large or consistent improvement in performance of the examined models. Menini et al. (2021) highlighted similar challenges: while showing that context affects annotation process (fewer tweets were annotated as abusive when context was provided to annotators), they report that when experimenting with different models (BERT, BiLSTM, SVM) and a context window ranging from one to all preceding tweets, contextual information did not lead to a better classifier performance. Vidgen et al. (2021) introduced

Proceedings of the 29th International Conference on Computational Linguistic, pages 37–42
October 12–17, 2022.

the Contextual Abuse Dataset composed of Reddit messages, where previous or several previous messages were considered as the context and "every annotation has a label for whether contextual information was needed to make the annotation". The authors report that 25-32% of content was labelled as context-dependent, and these messages are more challenging for detection, leaving integrating context for future work.

In this paper, we address hate speech target detection in hateful Dutch Facebook comments: whether the target of hateful content is the social group of interest, that is, migrants or someone else (see Section 2). While in previous work the preceding comment(s) in the discussion thread or the text of the post was used as context (Gao and Huang, 2017; Karan and Šnajder, 2019; Pavlopoulos et al., 2020; Menini et al., 2021; Lemmens et al., 2022), we manually annotate the relevant context, that is, we look specifically at the part of the prior conversation that provides the context, and use that annotation to demonstrate its utility in hate speech target prediction by adding relevant contextual information to a Dutch transformer-based pre-trained language model, BERTje (de Vries et al., 2019).

Hate speech is deeply contextual, and while most previous work ignores the conversational context, and more recent work, which looked into it based on previous comments or post, comes away concluding that such surface-level information is not helpful in prediction, this study shows and quantifies the impact that can be brought by relevant context on classifier performance when detecting the target of online hate speech.

2 Data

We used the LiLaH dataset, as in (Markov et al., 2021). The dataset is composed of Facebook posts by mainstream media outlets in Dutch (i.e., news articles that were published by the media outlets and are (re-)published or shared as Facebook posts) and readers' comments on these posts in a comment section, which were manually annotated by three trained annotators for fine-grained types and targets of hate speech (see below in this section) with a 'moderate' agreement. The annotations were performed in-context, that is, annotators first read entire comment threads and then labeled each comment.

We randomly selected a subset of the dataset composed of 35 posts and around 6,000 comments discussing the migrants topic and annotated this data for context dependency: if context influences the annotators' decision to assign a label to a comment, that is, if assigning hatefulness to a comment depends on understanding its context, or if the target of hate speech is not sufficiently clear without the context, the annotator marked the target comment as context-sensitive and indicated the ID of the post or the ID of the previous comment (not necessarily directly preceding) in the discussion thread that serves as the corresponding context. For example, the comment 'I would have served pork steaks' (*Ik zou varkenslapjes geserveerd hebben*) is hate speech directed against migrants if we take into account that the article is about a Muslim woman who was served alcohol at a show and was upset since this was against her religion. In this case, the annotator would mark the comment as context-dependent and indicate the ID of the post under which the comment was made.

We merged the fine-grained types of hate speech present in the data (e.g., violence, offensiveness, threat) into a single hate speech category, removing comments that belong to the non-hate speech class, which is the commonly used set-up for the hate speech target detection task (Zampieri et al., 2019a,b; Caselli et al., 2021), and used the binary target classes within the hateful messages. That is, we distinguish between migrants as the target of hate speech and merge all other fine-grained target classes into the 'other' category in order to have a sufficient amount of training and test examples per class. In more detail, the 'other' category consists of hate speech directed against (1) the article's author or the media spreading the article; (2) the author of another preceding comment under the same post; (3) other entities related to the migrants group, as they represent a positive attitude towards this group; and (4) people or institutions that do not belong to any of the above categories. For the binary target classes used in this study, the inter-annotator agreement was 'moderate' (Cohen's Kappa = 0.46).

We used training and test partitions splitting the dataset by post boundaries in order to avoid within-post bias, that is, all comments belonging to the same thread are in the same split. The splitting was done so that the distribution of 'migrants' and 'other' classes is as balanced as possible (roughly 40%–60%, respectively), while the proportion of 80% training and 20% test messages is preserved.

	Train (28 posts)		**Test** (7 posts)		**Total** (35 posts)	
	# messages	% context	# messages	% context	# messages	% context
Migrants	1,017	60.2	238	57.6	1,255	59.8
Other	1,660	28.9	431	38.3	2,091	30.8
Total	2,677	40.8	669	45.1	3,346	41.7

Table 1: Statistics of the dataset used in terms of the number of posts, number of comments per class and the percentage of messages annotated as context-dependent within each class.

The statistics of the dataset used in terms of the number of posts and comments in the training and test sets, as well as the percentage of context-dependent messages per class is provided in Table 1. We note that context-sensitive comments are frequent within both categories. Context-dependent messages within the 'migrants' category (59.8%; 750 messages) are more frequent than within the 'other' category (30.8%; 644 messages), which could be explained by the characteristics of the dataset used: it consists of discussion threads on the migrants topic, while in order to direct hate speech against someone else (e.g., previous commenter, article's author) hateful content creators would have to deviate from the original discussion topic by explicitly specifying the target of their hate speech. Out of 1,394 messages labeled as context-dependent, the vast majority (88%) refer to original post as the source of relevant context, 7% to previous comment and 5% to a comment located higher up in the discussion thread.

3 Experiments and Results

We use the monolingual Dutch transformer-based pre-trained language model, BERTje (de Vries et al., 2019), from the Hugging Face transformers library[1], which showed near state-of-the-art results in previous work on Dutch hate speech detection, e.g., (Caselli et al., 2021; Markov et al., 2022). The model was pre-trained using the same architecture and parameters as the original 768-dimensional BERT model (Devlin et al., 2019) on a dataset of 2.4 billion tokens.

We set the maximum sequence length parameter to 512 in order to account for the context, the other parameters have default values, and fine-tune the model for a single epoch. Following the approach proposed in (Pavlopoulos et al., 2020), we concatenate the context and the text of the target comment separated by BERTje's [SEP] token, as in the next sentence prediction task in BERTje's pre-training stage, and fine-tune the model on this data.

We use only the content of the target comment as the baseline and examine the following ways for adding contextual information: (1) adding the text of the post on which the comment was made (comment & post); (2) adding the preceding comment (if any) in the discussion thread (comment & preceding comment); (3) adding the preceding comment and the post (comment & preceding comment & post); and (4) adding the relevant annotated context (comment & context). Since BERTje is sensitive to random seeds, we report the results in terms of precision, recall and F1-score (macro) averaged over five runs, and standard deviations in Table 2.

The obtained results are in line with previous findings in the sense that adding the content of a preceding comment does not facilitate classifier performance (Karan and Šnajder, 2019; Pavlopoulos et al., 2020; Menini et al., 2021). However, we observe a moderate improvement by adding the content of the post (2 F1 points) and a significant improvement (according to McNemar's significance test (McNemar, 1947) with $\alpha < 0.05$) caused by pointing at the actual context in the discussion thread (6 F1 points). The results partially reflect the annotation process, described in Section 2, where most of the hateful messages contain the relevant context in the post text.

To further examine the importance of contextual information, we conducted an additional experiment using only the relevant context (while discarding the content of the target message), obtaining the following results: precision = 0.60 (±0.009), recall = 0.60 (±0.007), F1 = 0.60 (±0.009) (average over 5 runs). Considering that the majority baseline precision = 0.32, recall = 0.50, and F1 = 0.39, this experiment confirms that context contains useful information and can be used in isolation to predict the label of the target message.

The detailed results per class for one of the experiments reported in Table 2 for the baseline (comment only) and 'comment & context' strategies are presented in Table 3. We note that with the

[1]https://huggingface.co/GroNLP/bert-base-dutch-cased

	Precision	Recall	F1-score
Comment (baseline)	0.65 (±0.008)	0.66 (±0.008)	0.63 (±0.011)
Comment & post	0.66 (±0.008)	0.67 (±0.004)	0.65 (±0.008)
Comment & preceding comment	0.64 (±0.007)	0.65 (±0.004)	0.63 (±0.015)
Comment & preceding comment & post	0.64 (±0.004)	0.65 (±0.008)	0.63 (±0.000)
Comment & context	**0.69** (±0.005)	**0.71** (±0.008)	**0.69** (±0.008)

Table 2: Results for the baseline and examined strategies for adding contextual information averaged over five runs. The standard deviations are also reported. The best results are highlighted in bold typeface.

	Comment (baseline)			Comment & context		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Migrants	0.51	0.69	0.58	0.56	0.74	0.64
Other	0.79	0.63	0.70	0.83	0.68	0.74
macro avg	0.65	0.66	0.64	0.69	0.71	0.69

Table 3: Results per class for the baseline and 'comment & context' approaches.

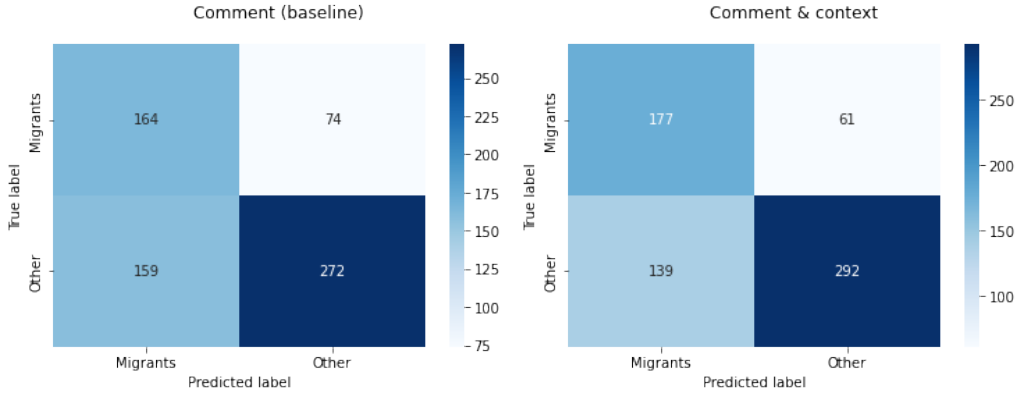

Figure 1: Confusion matrices for the baseline and 'comment & context' approaches.

additional contextual information, there is an improvement in performance for both 'migrants' and 'other' categories in terms of both precision and recall. In line with de Gibert et al. (2018) and Vidgen et al. (2021), we observed that context-sensitive messages are more challenging for classification: out of 302 context-dependent messages within the both categories 57% were identified correctly by the baseline approach, while out of 367 messages not dependent on the context, 71% were assigned the correct label. Integrating the relevant context lead to an improvement for both context-dependent and independent messages, resulting in 60% and 87% correctly-identified messages, respectively.

While for the 'migrants' class integrating the context lead to an improvement for the context-dependent messages (81% instead of 67% were identified correctly after adding the context), and no improvement was observed for the context-independent messages (65% vs. 71% without the context), for the 'other' class, the main source of improvement is the context-independent messages (84% instead of 72% were identified correctly), while the number of correctly-identified context-dependent messages within this category dropped from 48% to 42%. Zooming in on the fine-grained classes within the 'other' category, we note that the results are improved for all the classes, except for the hate speech directed towards article's author or media spreading the news, where only two more messages were misclassified after adding the contextual information.

The confusion matrices for this experiment, presented in Figure 1, demonstrate that integrating the context improves the results both in terms of false positives and false negatives, providing additional evidence that context plays an important role in detecting the target of online hateful comments.

4 Conclusions

Despite recent advances, there are multiple challenges that remain and limit the development of ro-

bust real-world hate speech detection systems. One of such challenges, addressed in this work, is to explicitly account for relevant conversational context when developing context-aware hate speech detection approaches.

While prior work has shown that the easy-to-obtain contextual information such as previous comment or post does not provide a large or consistent improvement, we demonstrated that if the model can zoom in on the relevant context, the performance increases significantly.

A limitation of this work is that we use hand-labeled contextual information, and thus report an upper bound of improvement in performance. Nonetheless, we believe that this study is an important step towards developing more robust and context-aware automated hate speech detection approaches.

Given the great potential for encoding contextual information and its significant effect on detecting the target of hate speech presented in this work, the directions for future work include detecting relevant context for a target comment automatically and exploring its effect on performance, as well as investigating the impact of context on detecting fine-grained types and targets of online hate speech.

Acknowledgements

This research has been supported by the Flemish Research Foundation through the bilateral research project FWO G070619N "The linguistic landscape of hate speech on social media". The research also received funding from the Flemish Government (AI Research Program). This research has also been supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: The Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, pages 54–66, Online. ACL.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 11–20, Brussels, Belgium. ACL.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv/1912.09582*.

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)*, pages 4171–4186, Minneapolis, MN, USA. ACL.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Alon Halevy, Cristian Canton Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2020. Preserving integrity in online social networks. *arXiv/2009.10311*.

Mladen Karan and Jan Šnajder. 2019. Preemptive toxic language detection in Wikipedia comments using thread-level context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134, Florence, Italy. ACL.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. ELRA.

Jens Lemmens, Tess Dejaeghere, Tim Kreutz, Jens Van Nooten, Ilia Markov, and Walter Daelemans. 2022. Vaccinpraat: Monitoring vaccine skepticism in Dutch Twitter and Facebook comments. *Computational Linguistics in the Netherlands Journal*, 11:173–188.

Jens Lemmens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16, Online. ACL.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. ACL.

Ilia Markov, Ine Gevers, and Walter Daelemans. 2022. An ensemble approach for Dutch cross-domain hate speech detection. In *Proceedings of the 27th International Conference on Natural Language and Information Systems*, pages 3–15, Valencia, Spain. Springer.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Kyiv, Ukraine (Online). ACL.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about NLP? The role of context in abusive language annotation and detection. *ArXiv*, abs/2103.14916.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, online. ACL.

Julian Risch and Ralf Krestel. 2020. Toxic comment detection in online discussions. *Deep Learning-Based Approaches for Sentiment Analysis*, pages 85–109.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *arXiv/1809.07572*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv/2004.01670*.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. ACL.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420. ACL.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. ACL.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). ICCL.