Improving Hate Speech Type and Target Detection with Hateful Metaphor Features

Jens Lemmens and **Ilia Markov** and **Walter Daelemans**
CLiPS, University of Antwerp
Lange Winkelstraat 40
2000, Antwerp (Belgium)
firstname.lastname@uantwerpen.be

Abstract

We study the usefulness of hateful metaphors as features for the identification of the type and target of hate speech in Dutch Facebook comments. For this purpose, all hateful metaphors in the Dutch LiLaH corpus were annotated and interpreted in line with Conceptual Metaphor Theory and Critical Metaphor Analysis. We provide SVM and BERT/RoBERTa results, and investigate the effect of different metaphor information encoding methods on hate speech type and target detection accuracy. The results of the conducted experiments show that hateful metaphor features improve model performance for the both tasks. To our knowledge, it is the first time that the effectiveness of hateful metaphors as an information source for hate speech classification is investigated.

1 Introduction

In this paper, the usefulness of hateful metaphors used as features for detecting the type and target of Dutch online hate speech comments is investigated. Although both hate speech and metaphor detection have been researched widely (e.g., MacAvaney et al., 2019; Basile et al., 2019; Leong et al., 2018, 2020), and figurative language used in hateful content has been identified as one of the main challenges in (implicit) hate speech detection (MacAvaney et al., 2019; van Aken et al., 2018), the question whether detecting (hateful) metaphors and using them as features improves hate speech detection models has remained unstudied in previous research. Therefore, it is the goal of the present paper to address this question.

In order to achieve this goal, we used the Dutch LiLaH[1] corpus which consists Facebook comments on online newspaper articles related to either migrants or the LGBT community. The comments were annotated for the type of hate speech and the target of hate speech, and for "hateful metaphors", i.e., metaphors that express hate towards a specific target (e.g., "het parlement is een circus!"; *the parliament is a circus*). We investigate whether features based on these manual annotations can improve Natural Language Processing (NLP) models that predict the type (e.g., violence, offense) and target (e.g., migrants, LGBT, journalist) of hateful content. Our experimental setup is therefore different from the commonly-used one in the sense that we are focusing only on the fine-grained hate speech categories and not on classification of hateful and non-hateful content. We hypothesize that hateful metaphors contain valuable information for type and target classification, especially in cases of implicit hate speech, and can therefore improve classification accuracy when used as features.

Prior to the classification experiments, a linguistic analysis of the annotated metaphors is conducted in the framework of Conceptual Metaphor Theory and Critical Metaphor Analysis. We would like to warn that for clarity of exposition, randomly chosen examples of hate speech from our corpus will be provided in this paper, and that some readers could find those offensive.

2 Related research

Hate speech detection Hate speech – frequently defined as a form of communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other (Nockleby, 2000) – has been extensively researched in the field of NLP. Pretrained language models such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa) (Devlin et al., 2019; Liu et al., 2019) provide the best results for hate speech detection, including type and target classification (Basile et al., 2019; Zampieri et al., 2019b, 2020), while shallow machine learning models (e.g., Support Vector Ma-

[1] https://lilah.eu/

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 7–16
June 6, 2021. ©2021 Association for Computational Linguistics

chines (SVM)) can achieve a near state-of-the-art performance (MacAvaney et al., 2019).

Examples of successful machine learning models include the winning teams of both subtasks A (binary hate speech detection) and B (binary target classification) of task 5 of SemEval 2019: multilingual detection of hate speech against women and immigrants on Twitter (Basile et al., 2019). These teams all used SVM-based approaches for both languages provided (English and Spanish) with the exception of the winner of task B for Spanish, who used various other classifiers and combined them by means of majority voting. For English, the winning teams obtained an F1-score of 65% for task A and an EMR score of 57% for task B.

Examples of effective neural approaches can be found in OffensEval 2020 (Zampieri et al., 2020). This shared task consisted of three subtasks: (A) offensive language identification, (B) categorization of offensive types and (C) target identification for multiple languages. For English, each of the top 10 teams for all three tasks used pretrained language models such as BERT and RoBERTa. The highest macro F1-scores obtained for task A, B, and C were 92%, 75% and 71%, respectively.

Figurative and implicit language in hate speech In their hate speech detection survey, MacAvaney et al. (2019) highlight current challenges in hate speech detection. One of the main challenges mentioned is the use of figurative and implicit language such as sarcasm and metaphors, which can lead to classification errors, as evidenced by their experiments. An SVM classifier with TF-IDF weighted character n-gram features was used to perform hate speech detection on the Stormfront, TRAC, HatEval and HatebaseTwitter datasets (de Gibert et al., 2018; Kumar et al., 2018; Basile et al., 2019; Davidson et al., 2017). An error analysis of the misclassified instances showed that sarcastic and metaphorical posts were the main causes of misclassifications, next to too little context (posts containing fewer than 6 tokens) and aggressive statements occurring in posts that were not annotated as "hateful".

Similar findings were observed by van Aken et al. (2018). An ensemble of machine learning and deep learning models was used for multi-class classification of toxic online comments and an error analysis of the incorrect predictions showed that metaphors can lead to classification errors because the models require significant world knowledge to process them.

To address the problem of implicit language in hate speech, more recent studies have used datasets that distinguish between implicit and explicit hate speech, such as AbuseEval v1.0 (Caselli et al., 2020). This dataset was created by annotating the OLID/OffensEval dataset (Zampieri et al., 2019a) for implicitness/explicitness. The authors of AbuseEval v1.0 provide results with BERT for binary classification (abusive, non-abusive) and multi-class classification (non-abusive, implicit abuse, explicit abuse) for the same train/test split and show that the binary classification task (71.6% macro F1-score) becomes substantially more complex when distinguishing between implicit and explicit abusive language (61.4% macro F1-score). Additionally, they show that the results for implicit hate speech detection (24% precision, 23% recall) are substantially lower than for explicit hate speech detection (64% precision, 51% recall).

Metaphors The foundations of the state-of-the-art way of thinking about metaphors is presented in "Metaphors We Live By" (Lakoff and Johnson, 1980), in which metaphors are defined as utterances that describe a target concept in terms of a source concept that is semantically distinct from the target concept, this includes idiomatic expressions and dead metaphors such as "the *body* of a paper" and "the *foot* of a mountain". The authors argue that specific metaphorical expressions can be traced back to more abstract metaphor schemes that overarch similar metaphors. This is what they call "Conceptual Metaphor Theory" (CMT). Examples are utterances such as "he *attacked* my arguments" and "I *destroyed* him during our discussion" which can be traced back to the conceptual metaphor *argument is war*.

In Charteris-Black (2004), Critical Metaphor Analysis (CMA), an integration of various linguistic disciplines such as cognitive linguistics, corpus linguistics and discourse analysis, is applied to CMT. According to CMA, metaphors highlight certain aspects of the target concept while hiding other aspects. At the same time, they uncover the speaker's thought patterns and ideological views. Therefore, metaphors – this includes dead metaphors used subconsciously – provide insights into how a speaker or community perceives the target domain. In short, metaphors reveal speaker bias. This is particularly valuable in the present study, since the toxicity that is spread through hateful metaphors resides in the source domains, more precisely in the aspect of the source domain that is highlighted by the metaphor.

Metaphor detection in NLP Recent advances in NLP-related metaphor studies can be found in the 2020 VUA and TOEFL metaphor detection shared task (Leong et al., 2020). The participating models showed substantial improvements compared to previous research, such as the 2018 VUA metaphor detection shared task (Leong et al., 2018), due to the effectiveness of (pretrained) transformer and language models. More than half of the participants used BERT (or related) models and all participating teams obtained higher F1-scores on the VUA metaphor corpus than the best-performing approach that participated in the 2018 shared task (65.1% F1-score). Further, the 2020 winning model, which consists of transformer stacks with linguistic features such as part-of-speech (PoS) tags, outperformed its predecessor of 2018 by more than 10% (76.9% F1-score, Su et al., 2020).

Contributions To our knowledge, we are the first to use hateful metaphor features for hate speech detection. We provide SVM and BERT/RoBERTa results and show the impact of using hateful metaphors as features on predicting the type and target of hateful content. In addition, the qualitative analysis of the annotated metaphors provide insights into what linguistic strategies are used to convey hate towards specific target groups.

3 Data

3.1 Corpus description

The Dutch LiLaH corpus consists of approximately 36,000 Facebook comments on online news articles related to migrants or the LGBT community mined from three popular Flemish newspaper pages (HLN, Het Nieuwsblad and VRT)[2]. The corpus, which has been used in several recent studies on hate speech detection in Dutch, e.g., (Markov et al., 2021; Ljubešić et al., 2020), was annotated for the type and target of hateful comments following the same procedure and annotation guidelines as presented in (Ljubešić et al., 2019), that is, with respect to the type of hate speech, the possible classes were violent speech and offensive speech (either triggered by the target's personal background, e.g., religion, gender, sexual orientation, nationality, etc., or on the basis of individual characteristics), inappropriate speech (without a specific target), and appropriate speech. The targets, on the other hand, were divided into migrants and the LGBT community, people related to either of these communities (e.g., people who support them), the journalist who wrote or medium that provided the article, another commenter, other targets and no target. The comments were labeled by two trained annotators (both Master's students and native speakers of Dutch) and the final labels were determined by a single expert annotator (PhD student and native speaker of Dutch).

As mentioned, our analysis deviates from the more "standard" experimental setup in hate speech research, namely classifying comments into hate speech or non-hate speech. In contrast, we consider only the fine-grained hate speech categories, i.e., discarding the non-hate speech classes (i.e., "inappropriate speech" and "appropriate speech" for the type class; "no target" for the target class) and focusing only the type and target of hateful content. Additionally, the four hate speech type categories (violent-background, violent-other, offensive-background, offensive-other) were converted to binary classes (violent, offensive).

The statistics of the hate speech comments used for our metaphor analyses are shown in Table 1. For the machine learning experiments, we selected a balanced subset in terms of the number of comments per class and the number of literal and non-literal comments per class (whenever possible). The statistics of the train/test partitions used for these machine learning experiments are shown in Table 2. In the subsets used, Cohen's Kappa equals 0.46 for the target classes and 0.54 for the type classes, indicating a "moderate" agreement between the two annotators for both the type and target annotations.

3.2 Hateful metaphor annotations

All hateful metaphors in our corpus were annotated by the same expert annotator mentioned above. For this task, the definition of a metaphor presented in Lakoff and Johnson (1980), described in Section 2, was adopted. More specifically, we define *hateful* metaphors as metaphorical utterances (including similes) that express hate towards a specific target, and therefore occur in hate speech comments, that

[2] https://www.facebook.com/hln.be; https://www.facebook.com/nieuwsblad.be; https://www.facebook.com/vrtnws

are not used to refer to someone else's opinion or previous comments, and that are written in Dutch.

We found that 2,758 (14.7%) out of all 18,770 hateful comments in our corpus contain at least one hateful metaphor. In those comments, 282 were LGBT-related, whereas all other 2,476 non-literal comments were related to migrants. In other words, 15.7% of all hate speech comments on LGBT-related news articles (1,797 in total) contain one or more hateful metaphor(s), whereas 14.6% of all hate speech comments on migrants-related news articles (16,973 in total) contain one or more hateful metaphor(s). See Table 1 for more fine-grained information on (non-)literal comments per type/target.

A qualitative analysis showed that many similar metaphors occurred in the corpus (in line with CMT). Therefore, we manually determined the source domains of the metaphors in a bottom-up fashion. If only one variation of a metaphor occurred for a certain source domain, it was added to the category "other". A list of the source domains, the number of comments in our corpus that contain them, a Dutch example, and its English translation can be found below together with a linguistic analysis in line with CMT and CMA.

- **Animals** (646), e.g., "migranten zijn bruine apen" (*migrants are brown apes*)
- **Dirt and personal hygiene** (529), e.g., "de EU is een beerput" (*the EU is a cesspool*)
- **Body parts** (299), e.g., "bij jouw geboorte hebben ze de baby weggegooid en de moederkoek gehouden" (*when you were born, they threw away the baby and kept the placenta*)
- **Disease and illness** (228), e.g., "jij bent vergif" (*you're poison*)
- **History** (192), e.g., "die minister is Hitler" (*that minister is Hitler*)
- **Food** (147), e.g., "bootvluchtelingen zijn vissoep" (*boat refugees are fish soup*)
- **Fiction** (139), e.g., "de Bijbel is een sprookjesboek" (*the Bible is a collection of fairy tales*)
- **Mental conditions** (119), e.g., "ik dacht dat het internetuurtje in het gekkenhuis al voorbij was" (*I thought that internet time in the madhouse was already over*)
- **Products** (107), e.g., "migranten zijn importbelgen" (*migrants are imported Belgians*)
- **Children** (80), e.g., "politici zijn kleuters" (*politicians are toddlers*)
- **Carnival and circus** (75), e.g., "politici zijn clowns" (*politicians are clowns*)
- **Home and kitchen linen** (68), e.g., "hoofddoeken zijn keukenhanddoeken" (*head scarfs are kitchen towls*)
- **Sight** (65), e.g., "je draagt paardenkleppen" (*you're wearing horse blinkers*)
- **Religious mythology** (44), e.g., "het paard van Troje is al binnen" (*the Trojan horse is already inside*, referring to migrants)
- **Sand** (24), e.g., "die migranten moeten terug naar hun zandbak" (*those migrants should return to their sand boxes*)
- **Tourism** (19), e.g., "oorlogsvluchtelingen zijn gewoon citytrippers" (*war refugees are just on a citytrip*)
- **Machines** (14), e.g., "IS strijders zijn moordmachines" (*IS warriors are murder machines*)
- **Physical conditions** (7), e.g., "trans-atleten zijn paralympiërs" (*trans-athletes are paralympians*)
- **Lottery** (4), e.g., "die migranten denken dat ze de Euromillions gewonnen hebben zeker?" (*those migrants must think that they've won Euromillions*)
- **Other** (349), e.g., "migranten zijn geleide projectielen" (*migrants are guided missiles*)

In our corpus, the source domains in metaphors that express hate towards migrants frequently refer to animals, especially pests (e.g., "parasites", "cockroaches") and primates (e.g. "apes"), commodities (e.g., "import Belgians/criminality") and food (e.g., "rotten apples", "boat refugees are fish soup"). These findings are in line with previous work on English and cross-lingual hate speech (Demjen and Hardaker, 2017; Dervinyté, 2009). Given the persuasive, ideological nature of metaphors (cf. CMA), the usage of these metaphors suggests that the speaker wishes for migrants and their "species" to be "exterminated", "kept in the zoo", "returned to sender", "thrown in the bin", and to stop "breeding".

Conversely, the source domains that were found in hateful metaphors that target the LGBT community often refer to diseases, and mental and physical conditions. This indicates that the user of these metaphors believes that the LGBT community should be "cured", "hospitalized"or "internalized". Other hateful metaphors that target the LGBT community highlight aspects such as appearance and therefore refer to carnival or the circus,

Task	Class	Literal	Non-literal	All
Type	Violence	394	80	474
	Offensive	15,618	2,678	18,296
Target	Migrants/LGBT	5,184	723	5,907
	Related	558	84	642
	Journalist/medium	544	90	634
	Commenter	2,946	574	3,520
	Other	6,780	1,287	8,067
Total		**16,012**	**2,758**	**18,770**

Table 1: Statistics of all hateful comments in our corpus, including the number of hateful comments per type/target class, and the number of literal and non-literal comments (in total and per class).

		Training set			Test set			
Task	Class	Literal	Non-literal	Both	Literal	Non-literal	Both	Total
Type	Violence	311	63	374	83	17	100	474
	Offensive	1,000	1,000	2,000	250	250	500	2,500
	All	1,311	1,063	**2,374**	333	267	**600**	**2,974**
Target	Migrants/LGBT	200	200	400	50	50	100	500
	Related	333	67	400	83	17	100	500
	Journalist/medium	328	72	400	82	18	100	500
	Commenter	200	200	400	50	50	100	500
	Other	200	200	400	50	50	100	500
	All	1,261	739	**2,000**	315	185	**500**	**2,500**

Table 2: Statistics of the subsets used in the type and target classification experiments, including the number of comments in the train/test splits for the type and target prediction tasks, the number of comments per class, and the number of literal and non-literal comments.

such as “de Antwerp Pride is een carnavalsstoet” (*the Antwerp Pride is a carnival parade*).

Journalists or newspapers, on the other hand, are often described as “linkse” (*left-wing*) or “rechtse” (*right-wing*) “ratten” (*rats*) that need to be “uitgeroeid” (*exteriminated*). Other metaphors often refer to dirt and personal hygiene such as “strontgazet” (literally *“excrement newspaper”*), “rioolgazet” (literally *“sewer newspaper”*), and “riooljournalist” (literally *“sewer journalist”*) highlighting the quality of journalism.

Other social media users and commenters are metaphorized in a variety of ways in our corpus, depending on the context and on what aspect the speaker wants to highlight. Examples are “vuile hond” (*dirty dog*), “domme geit” (*stupid goat*), “schaap” (*sheep*), “mongool” (*person with Down syndrome*), “kleuters” (*toddlers*), and “middeleeuwers” (*people who live in the middle ages*).

Finally, the “other” category is complex, due to its variety of target groups that it contains. Politicians, for example, are often metaphorized as left-wing or right-wing “rats”, similar to how journalists, newspapers, other social media users, and the followers of those political parties are occasionally metaphorized as well. Further, religious institutions are often characterized as a circus or a hospital for the mentally ill, whereas religion itself is described as a fairytale or a disease.

4 Classification experiments

4.1 SVM

An SVM model was established with Sklearn (version 0.23.1, Pedregosa et al., 2011) by using token 1- and 2-grams with TF-IDF weights fed into a linear SVM, henceforth referred to as “SVM”. Grid search under 10-fold cross-validation was conducted to determine the optimal settings for the “C”, “loss”, and “penalty” parameters[3]. Then, the following methods were used to integrate the hateful metaphor features:

Generic metaphor features which do not take into account the source domains of the metaphors.

- **N tokens** – the number of hateful metaphorical tokens was counted and appended to the feature vectors.
- **N expressions** – the number of hateful metaphorical expressions was counted and appended to the feature vectors.

[3] Since the classes are not distributed equally in the subset used for type classification, the “class weight” parameter was also optimized in the type prediction task.

- **Suffix** – a suffix in the form of the placeholder[4] “MET” was added at the end of all hateful metaphorical tokens before vectorization, e.g., “You’re a pigMET.” This way, the model distinguishes between a hateful, non-literal token and the same token used literally and in a non-hateful way (e.g., “That farmer bought a pig”).
- **Tokens** – the token “MET” was added after all metaphorical tokens before vetorization, e.g., “You’re a pig MET”. This allows the model to see similarities between a word form used literally and the same word form used figuratively, yet distinguish between them because of the placeholder that follows.
- **Tags** – all subsequent metaphorical tokens were enclosed in tags, such as in “You’re a MET dirty pig MET”. This method allows the model to focus on the on- and offset tokens of the metaphorical expressions.
- **All features** – the combination of all feature sets described above. For example, this encoding method would transform the utterance "migrants are a Trojan Horse" into “migrants are a MET trojanMET MET horseMET MET" and append the numerical features ("2" and "1" in this case) to its feature vector after vectorization to represent the number of hateful metaphorical tokens and expressions in the text, respectively.

Source domain metaphor features Since the source domains of the hateful metaphors could contain useful information for the predictions of the type and target of hate speech, because they highlight certain aspects of the target domain and reflect the way that the speaker perceives it (as described in Section 2), all methods described above were also used to encode hateful metaphor information while considering the source domains of the metaphors. More specifically, when using in-text metaphor information encoding methods, the "MET" placeholder was replaced with the first three characters of the names of the source domain of the metaphor (e.g., "ANI" for animal, "HIS" for history, etc.). For the numerical features, on the other hand, 20-dimensional vectors were used to count the number of metaphorical tokens/expressions in each comment (each dimension representing one of the 20 source domains that were observed in the linguistic analysis of the metaphors).

	CV		Test set		
Approach	**F**	**Std**	**Pre**	**Rec**	**F**
SVM	55.9	2.5	56.6	56.5	56.4
+n tokens	56.9	2.8	58.0	57.9	57.5
+n expressions	**57.3**	2.9	56.6	56.4	56.1
+suffix	55.6	2.2	57.4	57.9	57.3
+tokens	56.9	2.1	56.6	56.6	56.3
+tags	57.0	2.2	57.2	57.3	57.0
+all	56.4	2.4	**59.0**	**58.9**	**58.8**
BERTje	-	-	**63.1**	**62.8**	**62.4**
+tags	-	-	61.2	61.2	61.1
RobBERT	-	-	**61.9**	**61.8**	**61.8**
+tags	-	-	60.9	60.8	60.8

Table 3: 10-fold cross-validation and test set performances (%) on the **target** prediction task with **generic** metaphor features (best results in bold).

	CV		Test set		
Approach	**F**	**Std**	**Pre**	**Rec**	**F**
SVM	71.5	3.5	68.8	79.9	72.3
+n tokens	73.8	3.2	74.0	81.6	**76.9**
+n expressions	**74.1**	3.2	**74.2**	80.4	76.7
+suffix	71.3	2.9	68.5	80.9	72.2
+tokens	73.4	3.4	71.0	**82.4**	74.8
+tags	73.1	3.1	71.2	81.0	74.6
+all	73.6	3.2	73.8	80.6	76.5
BERTje	-	-	80.2	78.5	79.3
+tags	-	-	**82.7**	**80.0**	**81.2**
RobBERT	-	-	81.1	74.8	77.4
+tags	-	-	**82.0**	**77.2**	**79.3**

Table 4: 10-fold cross-validation and test set performances (%) on the **type** prediction task with **generic** metaphor features (best results in bold).

4.2 BERTje and RobBERT

Predictions for both tasks were made with BERTje and RobBERT (de Vries et al., 2019; Delobelle et al., 2020; the Dutch versions of BERT and RobBERTa) using HuggingFace 4.0.0 (Wolf et al., 2020). In an attempt to improve these models, the "tags" method described above was used, but with the “<met>” (onset) and “</met>” (offset) placeholders for generic features and the same more fine-grained placeholders as described above when using source domain features. This tagging method is frequently used to highlight textual features or external knowledge in sequence-to-sequence tasks such as machine translation and named entity recognition (e.g., Chatterjee et al., 2017; Li et al., 2018). Four epochs were used for training and all other parameters were set to default. The experiments were conducted five times with different seeds and we report the median of these runs.

[4] In order to ensure that the placeholders were not confused with actual text, all text was lowercased and all placeholders were uppercased before training.

5 Results

5.1 Quantitative results

The 10-fold cross-validation and test results of the SVM model[5], BERTje and RobBERT without additional features, with generic features or with source domain features for both tasks can be found in Table 3, 4, 5 and 6, respectively.

No additional features Without using additional features, it can be observed that BERTje performed best for both the target and type prediction tasks, closely followed by RobBERT and finally the SVM classifier. It can also be observed that target prediction accuracy is substantially lower than type prediction accuracy for all the models.

Generic features Regarding the SVM model, all proposed feature implementation methods improved the performance of the SVM classifier, with the exceptions of the token labels and number of metaphorical expressions for the target prediction task, and the suffix labels for the type prediction task. The best SVM-based approach for target predictions used the combination of all features, which showed a 2.4% F1-score improvement over the SVM classifier without additional features. For the type prediction task, the number of hateful metaphorical tokens used as feature improved the SVM baseline by 4.6% F1-score. Further, the performance of both BERTje and RobBERT improved by 1.9% when adding metaphor features to the text data for the type prediction task. Adding these labels before training on the target prediction task, however, did not improve the performance.

Source domain features With respect to the SVM approach, all feature implementation methods improved its performance for both the type and target prediction tasks, with the exception of the suffix features used for the type prediction task. Amongst the different types of source domain features, both numerical features (number of metaphorical tokens and number of metaphorical expressions) improved the SVM approach the most for type predictions (4% in F1-score). Conversely, adding the source domains after all hateful metaphors as tokens improved target prediction with SVM the most (1.6% in F1-score). On the other hand, the performance of the language models could only be improved marginally: when adding in-text features before training RobBERT on the type prediction task, its performance increased by 0.1% in F1-score.

	CV		Test set		
Approach	**F**	**Std**	**Pre**	**Rec**	**F**
SVM	55.9	2.5	56.6	56.5	56.4
+n tokens	**57.5**	2.7	57.8	57.6	57.4
+n expressions	57.3	2.9	58.2	58.0	57.8
+suffix	55.6	2.5	57.2	57.5	57.0
+tokens	56.9	2.0	**58.2**	**58.4**	**58.0**
+tags	57.0	1.7	57.6	57.9	57.4
+all	56.1	1.7	57.6	57.6	57.3
BERTje	-	-	**63.1**	**62.8**	**62.4**
+tags	-	-	61.2	61.4	61.2
RobBERT	-	-	**61.9**	**61.8**	**61.8**
+tags	-	-	61.2	61.7	61.4

Table 5: 10-fold cross-validation and test set performances (%) on the **target** prediction task with **source domain** metaphor features (best results in bold).

	CV		Test set		
Approach	**F**	**Std**	**Pre**	**Rec**	**F**
SVM	71.5	3.5	68.8	79.9	72.3
+n tokens	**74.3**	4.2	73.7	80.1	**76.3**
+n expressions	74.0	3.1	73.7	80.1	**76.3**
+suffix	71.0	3.3	68.4	80.2	72.0
+tokens	72.9	3.6	69.7	**82.9**	73.7
+tags	73.0	3.9	70.9	81.8	74.5
+all	73.3	4.1	**74.3**	77.2	75.6
BERTje	-	-	80.2	**78.5**	**79.3**
+tags	-	-	**81.6**	77.1	79.0
RobBERT	-	-	**81.1**	74.8	77.4
+tags	-	-	79.8	**75.8**	**77.5**

Table 6: 10-fold cross-validation and test set performances (%) on the **type** prediction task with **source domain** metaphor features (best results in bold).

Overall Substantial improvements up to 4.6% and 2.4% could be observed in the type and target classification tasks, respectively. These results indicate that hateful metaphor features contribute to type and target classification of hate speech comments in the current experimental setting.

5.2 Qualitative results

In this section, individual instances that were classified correctly only after adding hateful metaphor features are discussed. We focus on two comparisons, namely between the model that showed the highest increase in performance after adding metaphor information and the same model without additional features (per task). For the target prediction task, these are SVM and SVM to which all generic features have been added. For the type prediction task, on the other hand, these are the

[5] The optimal SVM parameter settings for the target prediction task were {"C": 1, "loss": "squared_hinge", "penalty": "l2"} and {"C": 0.5, "loss": "hinge", "penalty": "l2", "class_weight": "balanced"} for the type prediction task.

baseline SVM classifier and the SVM classifier enriched with numerical features based on the number of hateful metaphorical tokens (regardless of their source domains). The confusion matrices of these models are provided in Figures 1, 2, 3 and 4, respectively.

Target prediction task For this task, it can be observed that the additional features improved the classification accuracy for all classes. The only exception was the "journalist/medium" class, which is the most accurately predicted class using the SVM baseline and is predicted equally accurately when using additional features. On a deeper level, we observed that 52.8% of all instances in the target prediction task that were classified correctly only after adding metaphor features to the SVM baseline contained at least one hateful metaphor. These metaphors were often implicit cases of hate speech, such as "nep Belgen" (*fake Belgians*), "soortgenoten" (*conspecifics*), and "die leven nog in de middeleeuwen" (*they still live in the Middle Ages*). Still, we also found less subtle hateful metaphors, e.g., "strontvretende kakkerlakken" (*shit eating cockroaches*).

Type prediction task As evidenced by Figures 3 and 4, adding hateful metaphor features to the SVM model drastically decreases the number of cases where violent comments are confused with offensive comments, while retaining high classification accuracy for the “offensive” class. More specifically, 36.4% of all instances that were classified correctly only after adding hateful metaphor features contained at least one hateful metaphor. Similar to the improvements in the target prediction task, these metaphors were often implicit forms of hate speech, such as "op [ANONIEM]’s gezicht kan je pannenkoeken bakken" (*"you could cook pancakes on* [ANONYMOUS]*’s face"*) and afschaffen da klubke (*abolish that little club*, referring to the Catholic Church).

6 Conclusion

In this paper, we investigated the usefulness of hateful metaphors as predictive features for two less studied hate speech detection subtasks (namely type and target prediction) and analyzed the annotated hateful metaphors in our corpus in line with Conceptual Metaphor Theory and Critical Metaphor Analysis.

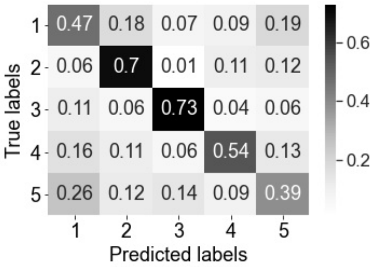

Figure 1: Confusion matrix for the **target** classification **SVM baseline** (1=“migrants/LGBT”, 2=“related to migrants/LGBT”, 3=“journalist/medium”, 4=“commenter”, 5=“other”).

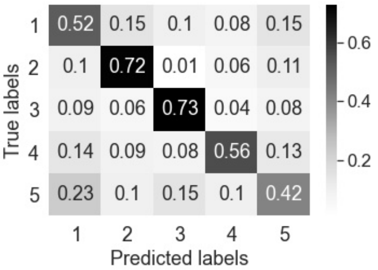

Figure 2: Confusion matrix for the **target** classification **SVM enriched with all generic features** (1=“migrants/LGBT”, 2=“related to migrants/LGBT”, 3=“journalist/medium”, 4=“commenter”, 5=“other”).

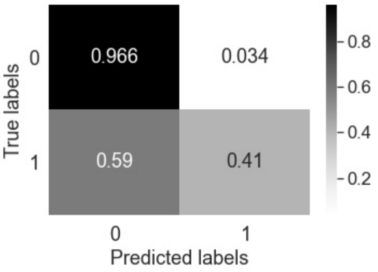

Figure 3: Confusion matrix for the **type** classification **SVM baseline** (1=“violence”, 0=“offensive”).

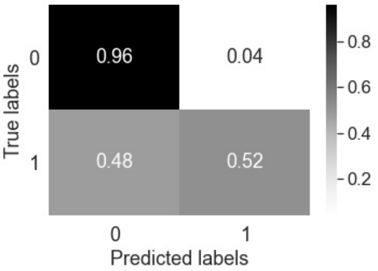

Figure 4: Confusion matrix for the **type** classification **SVM enriched with generic n tokens feature** (1=“violence”, 0=“offensive”).

Performances of SVM, BERTje and RobBERT were provided for both type and target prediction tasks and these models were then enriched with the hateful metaphor features in various ways to show their usefulness. The results show that the target SVM baseline improved by 2.4%. Conversely, BERTje and RobBERT could not be improved with additional features for this task. Regarding the type prediction task, an improvement up to 4.6% was observed for the SVM baseline, whereas the already high-performing BERTje and RobBERT baselines improved by 1.9% F1-score each. From the qualitative analysis that was conducted, it was observed that these improvements contained a large number of implicit forms of hate speech, which is considered to be one of the main challenges of hate speech detection at the moment.

This paper is a starting point for further research into the new area of (hateful) metaphors as predictive features for the hate speech classification tasks. Further research may include investigating whether the same results achieved with an upper-bound baseline in this paper (provided by our manually annotated features) can also be obtained when using labels predicted by models that have been trained to detect hateful metaphors. Other future research directions could include investigating more feature encoding methods and conducting ablation studies when combining multiple ways to encode hateful metaphors. In addition, it was observed that the SVM model cen be improved more strongly than BERTje and RobBERT, which suggest that the latter models already contain metaphorical information due to pretraining. Whether this is indeed the case is yet another subject worth investigating in future studies.

Acknowledgments

This research was supported by an IOF SEP project of the University of Antwerp. It also received funding from the Flemish Government (AI Research Program) and the Flemish Research Foundation through the bilateral research project FWO G070619N “The linguistic landscape of hate speech on social media”.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis (MN), USA. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don’t be abusive! Implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Jonathan Charteris-Black. 2004. *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

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 (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

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.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a dutch roBERTa-based language model.

Zsofia Demjen and Claire Hardaker. 2017. Metaphor, impoliteness, and offence in online communication. In *The Routledge Handbook of Metaphor and Language*, pages 353–367. Routledge.

Inga Dervinytė. 2009. Conceptual emigration and immigration metaphors in the language of the press: a contrastive analysis. *Studies about languages*, 14:49–55.

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, Minneapolis (MN), USA. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe (NM), USA. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago (IL) USA.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans (LA), USA. Association for Computational Linguistics.

Zhongwei Li, Xuancong Wang, Ai Ti Aw, Eng Siong Chng, and Haizhou Li. 2018. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in slovene and english. In *Text, Speech, and Dialogue*, pages 103–114, Cham. Springer International Publishing.

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. The LiLaH emotion lexicon of Croatian, Dutch and Slovene. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 153–157, Barcelona, Spain (Online). Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

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). Association for Computational Linguistics.

John T. Nockleby. 2000. Hate speech. In *Encyclopedia of the American Constitution*, pages 1277–1279. Macmillan, New York, USA.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

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.

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).

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 (offenseval 2020).