Improving Hate Speech Detection with Self-Attention Mechanism and Multi-Task Learning

Nicolas Zampieri, Irina Illina, Dominique Fohr

Improving Hate Speech Detection with Self-Attention Mechanism and Multi-Task Learning

Nicolas Zampieri, Irina Illina, Dominique Fohr

Lorraine University, CNRS, Inria, Loria, F-54000 Nancy, France

Abstract

Hate speech detection is a challenging task of natural language processing. Recently, some works have focused on the use of multiword expressions for hate speech detection. In this paper, we propose to use an auxiliary task to improve hate speech detection: multiword expression identification. Our proposed system, based on multi-task with self-attention, outperforms an MWE-based features state-of-the-art system on four hate speech corpora.

Keywords: hate speech, detection, neural networks, self-attention, multi-task

1. Introduction

Social media have an important place in today's society, in particular thanks to their forms of communication which are intended to be instantaneous and uncensored. Social networks make possible to communicate an idea, a thought, or any other form of the message whether it is harmful or not. Millions of messages are posted every day: e.g. Twitter with around 500 million posts every day (Bendler et al., 2014). Each social media has its own definitions of unwanted content. Hate speech is part of the unwanted content of social media and is punished by several countries. Automatic detection of hateful contents is essential due to the huge amount of posts on social media.

According to the Committee of Ministers of the Council of Europe, hate speech is "*any types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as race, color, religion etc.*"[1].

Hate Speech Detection (HSD) is a challenging task in the field of natural language processing. The nature of social media posts makes it difficult to detect hate speech, especially in Twitter posts (tweets). Indeed, tweets consist of short texts (maximum of 280 characters) that often employ non-standard syntax, and can contain misspellings, abbreviations, or even non-texts (e.g., emojis, images, and URLs). Annotate hate speech corpus is time consuming and expensive, so there are only a few annotated corpora.

Nowadays, state-of-the-art systems in this field are based on Deep Neural Networks (DNN). Chakrabarty et al. (2019) studied the impact of self-attention and contextual-attention. Kapil et al. (2020) explored multi-task learning, in parallel on five hate speech corpora. Awal et al. (2021) developed the AngryBERT system, which was trained on HSD and sentiment classification tasks.

In this article, we propose to incorporate syntactic and semantic information in a DNN-based system to improve HSD. Syntactic and semantic information will be learned from the Multiword Expression (MWE) identification task. MWE is a group of words (more than two lexemes) that express some form of idiosyncrasy: lexical, morphological, syntactic, semantic, and/or statistic (Baldwin and Kim, 2010) (e.g., *shut up*, *break a leg*, *black and white*). An MWE can be idiomatic or noun compound, and can have several meanings if we consider the word-by-word meaning of the MWE or if we take the meaning of the words composing an MWE as a single lexical unit. For example, *break a leg* could mean *good luck* when it is an idiom MWE and depends on the context of the sentence.

Multiword expressions and HSD have been the subject of some recent studies. Ptaszynski et al. (2017) proposed the use of morphosemantic patterns, such as part-of-speech and semantic role. Stankovic et al. (2020) extended a Serbian lexicon of abusive language with special attention to MWEs and proposed to exploit it to create an abusive corpus for the Serbian language. Zampieri et al. (2021) developed a DNN-based system that uses MWE features. MWE features have been integrated into a DNN-based system that utilizes MWE categories. Zampieri et al. (2022) compared the impact of two MWE identification systems: the first is based on a lexicon and the second is based on DNN. These works have shown that MWEs are helpful for the HSD task.

In this article, we propose a new HSD-system based on MWE which outperforms the system proposed by Zampieri et al. (2022). Our system uses self-attention mechanism and multi-task learning. The advantage of our system is to learn MWE and hate speech features thanks to a self-attention layer. Compared to Chakrabarty et al. (2019), we use multi-head self-attention. In contrast with Kapil et al. (2020), where several corpora were used for the same task, we use two different tasks.

2. Methodology

Zampieri et al. (2021) and Zampieri et al. (2022) showed that system using MWE features outperforms system without MWE features on the HSD task. In this current work, we pursue this idea in the framework of multi-task learning.

[1] https://www.coe.int/en/web/freedom-expression/hate-speech

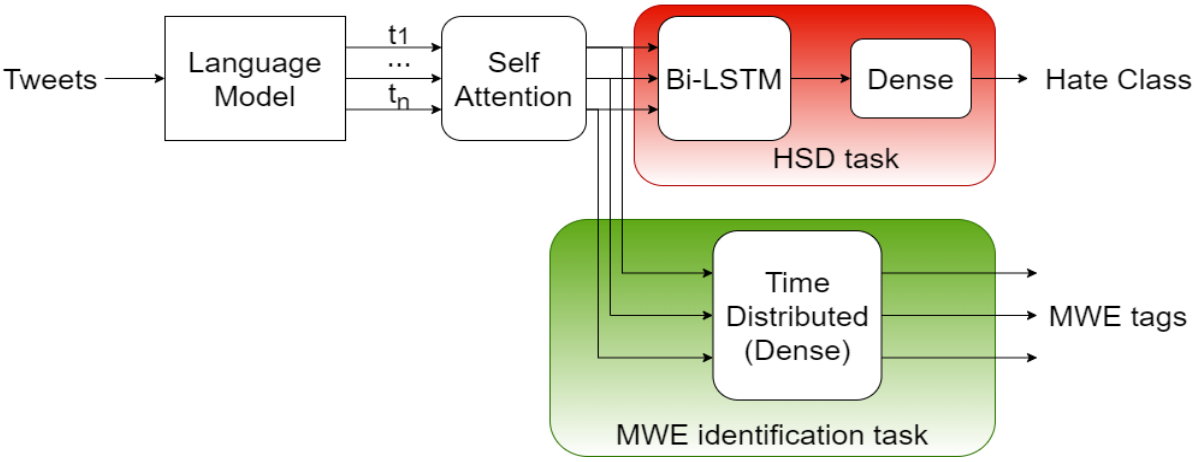


Fig. 1: Proposed HSD system based on multitask learning with self-attention.

The two tasks are MWE identification and HSD. Moreover, we propose to share, between the two tasks, a multi-head self-attention layer proposed by Vaswani et al. (2017) in order to learn attention simultaneously from both tasks. We believe that attention to MWEs could help the system distinguish hate from non-hate speech.

Figure 1 shows the architecture of our proposed HSD system based on multi-task learning and self-attention. The self-attention layer is there to learn representations considering the two tasks. Our system uses the contextual token embeddings provided by the outputs of a pre-trained language model like BERT-based models (Delvin et al., 2019). These embeddings are given as input to a self-attention layer. For the HSD task, we utilize a bidirectional long short-term memory layer followed by a dense layer. The final prediction is made from the output of specialized dense HSD task layer. For the MWE identification task, a dense time-distributed layer is used, and the outputs are formatted as "BIOo": each lexical unit is tagged "B" if it is at the start of an MWE, "I" if it is inside an MWE, "O" if it does not belong to an MWE. The "o" tag has the same meaning as the "O" tags, but the word is nested in an enclosing MWE.

We compare our approach with a baseline system trained only on the HSD task without self-attention. To perform the training of the multi-task system, we need a HSD corpus annotated in terms of MWEs. Since no such corpus exists, we utilize the predictions provided by the deep neural network MWE identification system of Zampieri et al. (2022). Compared to the work of Zampieri et al. (2022), where MWE features are used at the input of the DNN-based system, we design a multi-task approach that consists of hate speech detection and MWEs identification by using a self-attention mechanism.

3. Experimental Setup

In this section, we describe hate speech corpora and system configuration.

3.1. Datasets

Waseem and Hovy (2016) corpus (**Waseem**) contains 16,919 tweets annotated in three classes: sexist, racist, and neither. We recovered 10,807 tweets because some tweets have been removed from social media. We focus on HSD task, so we combine the sexist and the racist classes into single class: hate class. Tweets labeled as ''neither'' are considered to belong to the non-hate class. The corpus contains 73% and 27% of non-hateful and hateful tweets, respectively.

Davidson et al. (2017) corpus (**Davidson**) is a corpus annotated in terms of hate speech, offensive speech, or neither. The corpus contains 24,802 tweets: 76% are offensive, 7.4% hateful, and 16.6% neither. We do not merge offensive and hate speech classes, as the corpus is designed to distinguish offensive content from hateful content.

Founta et al. (2018) corpus (**Founta**) contains 100k tweets, annotated in four classes: hateful, abusive, normal, and spam. Our experiments focus on HSD, so we remove spam tweets, and we keep around 86k tweets. The corpus contains 63% normal, 31% abusive, and 6% hateful tweets. As in the Davidson dataset, we do not aggregate abusive and hateful tweets under the same label.

Basile et al. (2019) corpus (**HatEval**) is a balanced corpus annotated in hate and non-hate speech: 42% hateful and 58% non-hateful tweets. It contains 13k tweets and is partitioned into training, development, and test sets with 9k, 1k, and 3k tweets, respectively. It is provided by the SemEval2019 shared task 5.

For Waseem, Davidson, and Founta datasets, we utilize 60%, 20%, and 20% for **training**, **validation**, and **testing sets**, respectively. For the HatEval corpus, we use the standard corpus partition of the SemEval shared task 5. We apply the same preprocessing as in Zampieri et al. (2022): we remove mentions, hashtags, URLs, and we replace emojis with readable text (e.g., ♥ → *:heart:*).

3.2. MWE identification system

To annotate the MWEs on the four tweet corpora, we use the transformer-based system proposed by Liu *et al.* (2021). Indeed, Zampieri et al. (2022) showed that this MWE identification system achieves better performance than a lexicon-based approach. In our study, we apply the same configuration of the MWE identification system as in Zampieri et al. (2022). We use this MWE identification system to automatically annotate hate speech corpora in terms of MWEs. The MWE identification system tagged about 4k, 9k, 10k and 46k MWEs in Waseem, HatEval, Davidson, and Founta training sets, respectively.

HSD Systems	#Head	Binary classification		Ternary classification		Average
		Waseem	HatEval	Davidson	Founta	
Zampieri et al. (2022)	-	81.9 (±0.6)	64.6 (±1.1)	73.9 (±1.4)	74.0 (±0.7)	73.5
BERTweet embeddings						
Single task (baseline)	-	82.8 (±0.5)	61.1 (±4.8)	71.0 (±0.6)	73.7 (±1.0)	72.2
Single task	2	84.5 (±0.8)	61.5 (±2.3)	72.0 (±3.2)	74.0 (±0.9)	73.0
	4	84.3 (±2.0)	**64.4** (±3.7)	73.2 (±2.3)	**74.1** (±1.1)	74.0
Multi-task	2	**85.5** (±2.2)	64.2 (±1.8)	**74.2** (±1.0)	73.5 (±0.4)	**74.5**
	4	85.1 (±0.7)	63.3 (±4.6)	73.6 (±2.2)	**74.1** (±1.1)	74.0
HateBERT embeddings						
Single task (baseline)	-	81.6 (±1.2)	63.1 (±3.6)	71.9 (±3.5)	74.3 (±0.6)	72.7
Single task	2	82.4 (±0.6)	61.0 (±2.4)	72.8 (±3.1)	73.8 (±1.7)	72.5
	4	82.7 (±1.6)	60.4 (±2.9)	73.2 (±1.8)	74.4 (±0.7)	72.7
Multi-task	2	83.2 (±0.8)	63.7 (±2.6)	**75.1** (±2.0)	**74.6** (±0.3)	**74.2**
	4	**83.5** (±2.3)	**65.0** (±1.7)	73.3 (±3.1)	73.4 (±1.2)	73.8

Table 1: Median macro-F1 of HSD and standard deviation of 5 runs. The column #Head represents the number of heads for the attention layer. The results that are significantly better than the ''baseline'' systems are underlined. The *Average* column represents the average of median macro-F1 on the four corpora and the significant improvement is computed by merging all predictions.

Note that in this article, we do not evaluate our proposed multi-task system on the MWE identification task because the corpora used are not labeled in terms of MWE.

3.3. Hyperparameters of HSD system

To generate contextual token embeddings, we use state-of-the-art transformers-based models trained on tweets or hateful data: the BERTweet-base model (Nguyen et al., 2020), the HateBERT model (Caselli et al., 2021), and the fBERT model (Sarkar et al., 2021). The BERTweet model is trained on tweets. The HateBERT model is trained on Reddit comments that potentially contain abusive or hateful speech. The fBERT model is a BERT-based model fine-tuned on offensive tweets. Sarkar et al. (2021) showed that the fBERT model outperforms the HateBERT model. However, in our preliminary experiments, we found that the fBERT embeddings achieves lower performance compared to the two other models. So, in this article, we are experimenting with the BERTweet and the HateBERT embeddings.

For the MWE identification task, we use a ''IO'' tagging scheme with two labels: if a word belongs to an MWE, then it is tagged by ''I'', otherwise it is tagged by ''O''. Concerning the HSD task, we use a bidirectional long short-term memory layer with 128 neurons and followed by a dense layer. The output size of the bidirectional long short-term memory is 256.

3.4. Evaluation Metrics

We evaluate our models in terms of macro-average F1. It is the average of the F1 scores of all classes. We compute the median macro-F1 score over 5 runs. We use a matched pairs test with a 5% risk (Gillick and Cox, 1989) to determine if there is a significant improvement compared to the baseline system.

4. Results

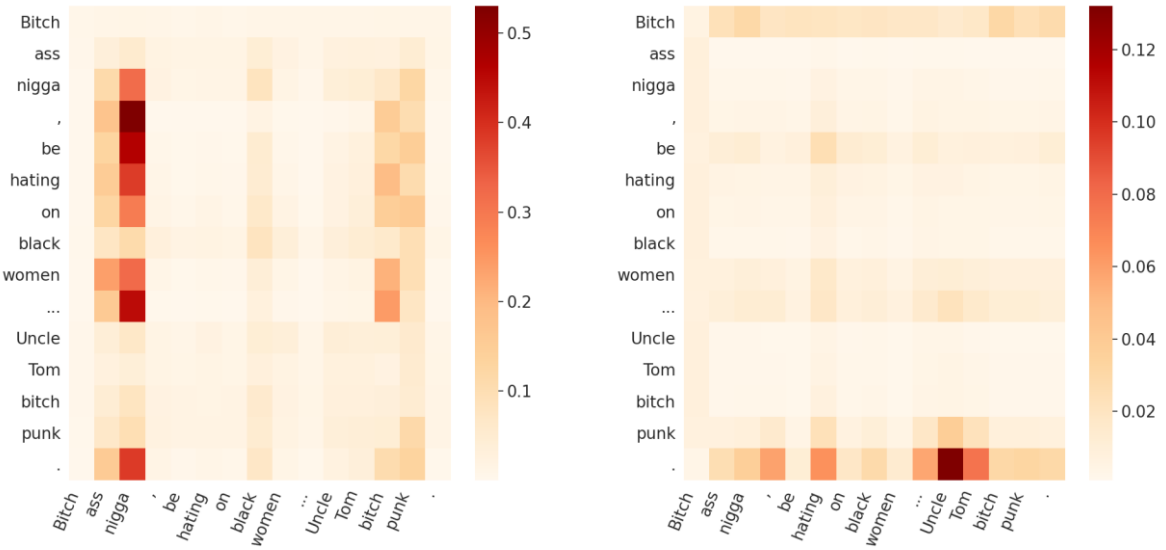
The goal of our experiments is to improve the performance of the HSD task using the MWE identification task. We study the effect of the self-attention mechanism on the HSD task. Moreover, we assess the multi-task approach using two different contextual token embeddings. We compare our new approach with the approach proposed by Zampieri et al. (2022) as they obtained good performance for the hate detection task and they also used MWEs.

Table 1 shows that our approach based on self-attention with multi-task learning outperforms the Zampieri et al. (2022) system for all datasets. Our best configuration of HSD system with multi-task learning, two heads of self-attention and BERTweet embeddings improves the average score by 1% relative compared to the Zampieri et al. (2022) HSD system (74.5% versus 73.5%). The best improvement is achieved for Waseem test corpus with an increase of 3.6% relative (85.5% versus 81.9%).

4.1. Impact of the self-attention mechanism

For the BERTweet embeddings and the single task approach, we find that using 2 attention heads does not significantly improve the average score compared to Zampieri et al. (2022) system. Using 4 attention heads, the average macro-F1 score is significantly better than baseline: 74.0% versus 72.2%. This improvement is observed in three corpora: Waseem, HatEval and Davidson. Using more than 4 self-attention heads does not provide any further improvement and it is not shown here.

Regarding the use of HateBERT embeddings, we do not observe an improvement using the self-attention mechanism compared to the baseline: the baseline achieves 72.7% versus 72.5% and 72.7% using 2 and 4 attention heads, respectively. It can be due to the fact that there is a mismatch between the training HateBERT embeddings (on Reddit) and testing on tweets.



(a) Weights of the first head of self-attention (b) Weights of the second head of self-attention

Fig. 2: Attention weights of the multi-task system with two heads (a and b) and the BERTweet embeddings. Example from the Davidson development set: ‘‘*Bitch ass nigga, be hating on black women... Uncle Tom bitch punk.*’’. Two MWEs are detected: *Bitch ass nigga* and *Uncle Tom*.

4.2. Impact of the multi-task learning

Table 1 shows that in the case of BERTweet embeddings, the multi-task system significantly outperforms the baseline system: the baseline reaches 72.7% of the average macro-F1 score, compared to 74.5% and 74.0% using 2 and 4 attention heads with multi-task learning, respectively. This is the case for 3 corpora. However, multi-task systems do not outperform single task with 4 attention heads. Using HateBERT embeddings, all multi-task configurations significantly outperform single-task systems: multi-task systems obtained 74.2% and 73.8% of the average macro-F1 scores compared to 72.5% and 72.7% obtained by single-task systems using 2 and 4 self-attention heads, respectively.

It is important to note that multi-task performance is obtained using an automatic MWE tagging system (used only during training). As Zampieri et al. (2022), this confirms that MWEs are helpful for the HSD task. For the two studied embeddings, the best performance is achieved by the multi-task system with 2 attention heads: 74.5% and 74.2% using BERTweet and HateBERT, respectively.

For further analysis, Figure 2 provides an example of the weights of the multi-task system with two self-attention heads. The example is extracted from the Davidson development set. We observe in some samples that each self-attention head often focuses on one task: the first head (2a) tends to specialize on harmful words (*ass*, *nigga* and *bitch*) and the second (2b) on MWEs (*Uncle Tom*).

4.4. Limitations

One limitation of our approach is the fact that it requires both MWE and hate/non-hate annotations of the data. To the best of our knowledge, such corpus does not exist. Therefore, in this work, we used an automatic MWE annotation system. The performance of the multi-task system may depend on the accuracy of this automatic MWE annotation system. As no corpus annotated in terms of both MWE and hate speech is available, we cannot fine-tune the BERTweet or HateBERT models for multi-task learning.

5. Conclusions

In this work, we investigated the impact of the self-attention mechanism and the multi-task learning for the hate speech detection. The two tasks that we want to investigated are the MWE identification task and the hate speech detection task. We carried out our experiments on four corpora and using two contextual embeddings: BERTweet and HateBERT. We observed that multi-task system significantly outperforms the baseline single task system. The best performance is obtained using the multi-task system with two attention heads.

For future work, we would like to take advantage of multi-task and multi-corpus approaches: MWE-annotated corpus and hate-speech-annotated corpus can be used simultaneously to train the system.

6. Acknowledgements

Experiments presented in this article were carried out using the Grid’5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations.

References

Awal, M.R., Cao, R., Lee, R.KW., Mitrović, S. (2021). AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection. In *Advances in Knowledge Discovery and Data Mining*. PAKDD Lecture Notes in Computer Science, Vol. 12712. Springer International Publishing.

Baldwin, T., and Kim, S. N. (2010). Multiword expressions. *Handbook of Natural Language Processing*, pp. 267–292. CRC Press, Taylor and Francis Group, Boca Raton, 2nd edition.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., and Sanguinetti, M. (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*, pp. 54–63. Association for Computational Linguistics.

Bendler, J.T., Brandt, T. L., Wagner, S. and Neumann, D. (2014). Investigating Crime-to-Twitter Relationships in Urban Environments - Facilitating a Virtual Neighborhood Watch. *European Conference on Information Systems*.

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 17–25. Association for Computational Linguistics.

Chakrabarty, T., Gupta, T., and Muresan, S. (2019). Pay "Attention" to your Context when Classifying Abusive Language. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 70–79. Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *CoRR, abs/1703.04009*. http://arxiv.org/abs/1703.04009

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pp. 512–515.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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*, Vol. 1, pp. 4171–4186.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., and Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

Gillick, L., and Cox, S. J. (1989). Some statistical issues in the comparison of speech recognition algorithms. *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 532–535.

Kapil, P., and Ekbal, A., (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, Vol. 210.

Liu, N. F., Hershcovich, D., Kranzlein, M., and Schneider, N. (2021). Lexical Semantic Recognition. *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pp. 49–56. Association for Computational Linguistics.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 9–14. Association for Computational Linguistics.

Ptaszynski, M., Masui, F., Nakajima, Y., Kimura, Y., Rzepka, R., and Araki, K. (2017). A Method for Detecting Harmful Entries on Informal School Websites Using Morphosemantic Patterns. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 21, No. 7, pp. 1189–1201.

Sakar, D., Zampieri, M., Ranasinghe, T., and Ororbia, A. (2021). fBERT: A Neural Transformer for Identifying Offensive Content. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pp. 1792–1798. Association for Computational Linguistics.

Savary, A., Cordeiro, S., and Ramisch, C. (2019, August). Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pp. 79–91. Association for Computational Linguistics.

Stanković, R., Mitrović, J., Jokić, D., and Krstev, C. (2020). Multi-word Expressions for Abusive Speech Detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pp. 74–84. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems.* NeurIPS, Vol. 30, pp. 5998–6008.

Waseem, Z., and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pp. 88–93. Association for Computational Linguistics.

Zampieri, N., Illina, I., and Fohr, D. (2021). Multiword Expression Features for Automatic Hate Speech Detection. Dans E. Métais, F. Meziane, H. Horacek, and E. Kapetanios (Éd.), *Natural Language Processing and Information Systems*, pp. 156–164. Springer International Publishing.

Zampieri, N., Ramisch, C., Illina, I., and Fohr, D. (2022). Identification of Multiword Expressions in Tweets for Hate Speech Detection. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pp. 202–210. European Language Resources Association.

Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. Dans A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, M. Alam (Éd.), *The Semantic Web*, pp. 745–760. Springer International Publishing.