MUTED 🔇: Multilingual Targeted Offensive Speech Identification and Visualization

Christoph Tillmann, Aashka Trivedi, Sara Rosenthal, Santosh Borse
**Rong Zhang, Avirup Sil, Bishwaranjan Bhattacharjee**
IBM Research AI

Abstract

Offensive language such as hate, abuse, and profanity (HAP) occurs in various content on the web. While previous work has mostly dealt with sentence level annotations, there have been a few recent attempts to identify *offensive spans* as well. We build upon this work and introduce MUTED, a system to identify multilingual HAP content by displaying offensive ARGUMENTS and their TARGETS using heat maps to indicate their intensity. MUTED can leverage any transformer-based HAP-classification model and its attention mechanism out-of-the-box to identify toxic spans, *without further fine-tuning*. In addition, we use the spaCy library to identify the specific TARGETS and ARGUMENTS for the words predicted by the attention heatmaps. We present the model's performance on identifying offensive spans and their targets in existing datasets and present new annotations on German text. Finally, we demonstrate our proposed visualization tool on multilingual inputs.

1 Introduction

Offensive language such as hate, abuse, and profanity (HAP) occurs in various content on the web such as social media sites (e.g. Twitter) and discussion forums (e.g. Reddit). Such content can be hurtful to the reader, and identifying and visualizing HAP speech is necessary to understand and avoid harm. It increases interpretability and can be used to hide and provide a warning for offensive terms, and to avoid generating hate in large language models.

While such visualizations exist, the focus has primarily been on English HAP and on identifying offensive language on the sentence level (McMillan-Major et al., 2022). There are few works that explore spans and other languages (Ranasinghe and

WARNING: This paper contains offensive examples.
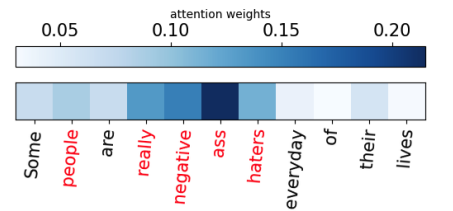
(a) Attention Heatmap

(b) SpaCy: visually identifying Target and Argument

Figure 1: Example system output that shows the intensity of the offensive ARGUMENT and its TARGET,<T,A>: (a), (b): <people, really negative ass haters> .
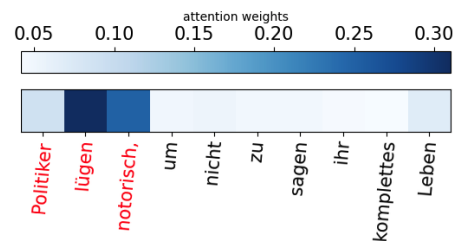
Figure 2: German Input (*"Politicians notoriously lie, not to say their entire lives"*): <Politiker, lügen notorisch> .

Zampieri, 2021; Wright et al., 2021) but these do not identify and visualize the TARGET of the offensive ARGUMENT which is an important indicator regarding whether the offensive argument is harmful or not, as shown in Zampieri et al. (2023).

We propose identifying hate using existing approaches (Caselli et al., 2021) to display multilingual offensive ARGUMENTS and their TARGETS using heat maps as a means of showing their intensity. Moreover, the spaCy library (Honnibal et al., 2020) can also be used to identify the specific target and argument from the predicted words. An example with a <T,A> pair is shown for English
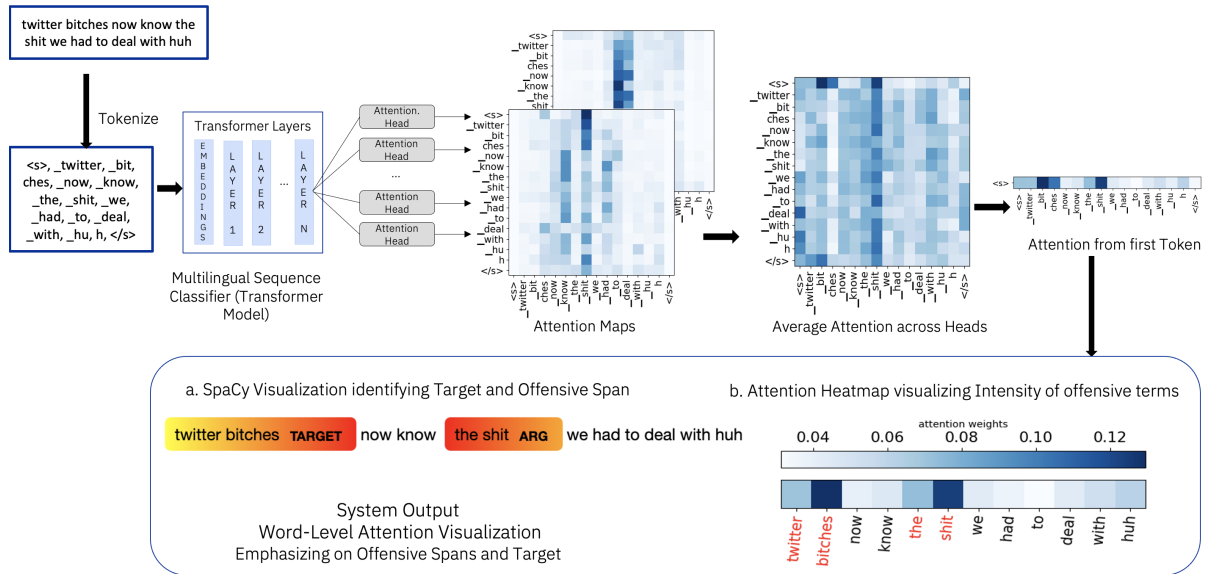
Figure 3: MUTED 🔇: Visualizing offensive spans and targets using Attention Heatmaps. A token-level attention score of a given sentence is obtained using the average attention across all heads of the last layer of the given HAP classifier, and extracting the attention from the first token (often the CLS vector). The score for a word is calculated as the maximum token-level attention score of its constituent tokens. Finally, we display the predicted spans using the attention heatmap, and use spaCy's dependency parser to identify the target and argument in the predictions.

and German inputs in Fig. 1 and Fig. 2, with the resulting visualizations. Our contributions are as follows:

- We present MUTED: A **MU**ltilingual **T**argeted **D**emonstration providing an intuitive way of visualizing existing classifiers by using transformer attention to identify the target of the offensive text as well as the offensive span.
- Unlike similar token classification techniques (Ranasinghe and Zampieri, 2021), our system can be used with off-the-shelf hate/abuse/profanity detectors.
- Our approach is multilingual and we demonstrate it on English (Zampieri et al., 2023) and a new German targeted offensive speech dataset. In the future, we plan to extend to more languages, e.g. a Spanish data set.
- We present easy-to-use Python notebooks and a front-end UI to run our approach on any encoder-only HAP classifier to visualize the offensive <T,A> pair using heat-maps and spaCy [1].

The rest of this paper describes related work, our approach for detecting offensive speech, and our model which outperforms existing sentence classifiers on the TBO (Zampieri et al., 2023) and TSD (Pavlopoulos et al., 2021) datasets. Finally, we present our system demonstration and its efficiency.

[1] https://spacy.io/api/dependencyparser

2 Related Work

Identifying offensive content has been a popular area of research in recent years (Davidson et al., 2017; Jahan and Oussalah, 2023). One popular model that is available is HateBERT (Caselli et al., 2021) which is a Bert-based model finetuned on offensive speech from Reddit comments. Similar models exist in other languages such as deHate-BERT(Aluru et al., 2020) in German. We present our own multilingual model for detecting offensive content which outperforms HateBERT on offensive span selection. However, our notebooks demonstrating our approach for identifying the offensive target and argument can be used with any transformer-based offensive classifier.

Several demos on offensive text exist that perform on the span or sentence level, mostly in English (McMillan-Major et al., 2022; Wright et al., 2021). Perhaps the most relevant demo is MUDES (Ranasinghe and Zampieri, 2021). They identify offensive spans in input text by classifying each token as offensive or not, and support English, Danish and Greek. The UI is token-classification based, and can be used with their trained models and the datasets used in the paper (or any input text) to identify offensive spans which will be displayed in red. In contrast to other prior work, our heat map-

based system can be used to visualize the offensive argument and target for any language for which a sentence level hate classifier is available.

3 Approach

MUTED provides an intuitive visualization of existing HAP classifiers by using attention maps to identify offensive text and their targets, as shown in Fig. 3. Formally, for a transformer model (of L transformer layers and H attention heads) finetuned to classify whether a given input sentence x contains offensive language, we first obtain the attention outputs $A_i^L \in \mathbb{R}^{|x| \times |x|}, i \in [1, H]$ of the last transformer layer. We then compute the average attention across all heads, $A' = \frac{1}{H} \sum_{i=1}^{H} A_i^L$, and extract the attention vector for the first token (e.g., the CLS token for BERT (Devlin et al., 2019) models), $A_0' \in \mathbb{R}^{1 \times |x|}$. Based on a threshold, we obtain the set of tokens T with the highest attention score, which can be intuitively viewed as the tokens that contribute most to the classification decision. We convert the token-level attentions into word-level attentions by assigning a word the maximum attention of any of its constituent tokens. We provide the word-level attention visualization in the form of heat maps, and mark the target and the argument of the offensive span in the sentence (see the System Output in Fig. 3).

Our system can be used to visualize any transformer-based model that is trained to classify if a given sequence has HAP content or not, irrespective of the language. In this work, we present the *Piccolo-HAP classifier*[2], a tiny 4-layer XLM-Roberta (Conneau et al., 2020) model (with 153 Million parameters) finetuned on the HAP detection task for 6 languages (English, German, Japanese, Spanish, French and Portuguese). Specifically, we distil the self-attention relations of an in-house XLM-Roberta Base Model on a task-agnostic (general purpose) manner into a 4-layer architecture, as proposed in Wang et al. (2021). We finetune this general purpose language model on the HAP classification objective, using open-source multilingual annotated datasets (Founta et al., 2018; Davidson et al., 2017; Röttger et al., 2021; de Gibert et al., 2018; Ousidhoum et al., 2019; Jigsaw, 2019; Pereira-Kohatsu et al., 2019; Wiegand et al., 2018; Roß et al., 2016; Leite et al., 2020) originating from social media data, as well as internally an-

notated samples from CC100 (Conneau et al., 2020) and scraped news data from the internet in the six languages mentioned above. For non-English data, we also translate English datasets (Davidson et al., 2017; Founta et al., 2018) to the language required. We finetune the model on a total of 1.7 million sentences, with the majority of data being in English.

4 Experiments

We compare our model to a random baseline, as well as open-source toxicity classifiers (monolingual and multilingual). First, we evaluate a **random** selection of spans as target and arguments in the sentence. Specifically, each span in the sentence is marked as HAP with a probability of 0.50. We also use three off-the-shelf English **Hate-BERT** models (Caselli et al., 2021), each finetuned on either Hateval (Basile et al., 2019), Offenseval (Zampieri et al., 2019b) or Abuseval (Caselli et al., 2020). These models were made available by the HateBERT authors[3], and we have not finetuned them ourselves. We also compare our multilingual model to another open-source multilingual classifier available on HuggingFace, **Multilingual Toxicity Classifier Plus [MTC+]**[4], and two German (monolingual) classifiers, **DeHateBERT-de**[5](Aluru et al., 2020) and **German Toxicity Classifier Plus (V2)**[6].

4.1 Datasets

For experiments, we use the following datasets, all of which contain data that is already known to be offensive. The data is converted into a span-selection task, where the classification model is used to identify the toxic spans (and the target of the span when applicable), using the attention maps.

- Target Based Offensive Language dataset (TBO) (Zampieri et al., 2023): TBO contains around 4500 examples of English twitter data that has been found to be offensive (Zampieri et al., 2019a; Rosenthal et al., 2021), providing token-level annotations and identifying both the offensive spans (ARGUMENT) and its TARGET in the input text. Each tweet can have multiple <T,A> pairs, and may have a "null" target if the target of the offense is not mentioned in the text. For this demonstration we did not explore the Harmful

Model	TSD: F1 Score ↑	English TBO: F1 Score ↑		
	TARGET ONLY	TARGET + ARG.	ARG. ONLY	TARGET ONLY
Random	0.08	0.19	0.16	0.13
HateBERT (AbusEval)	0.15	0.36	0.30	0.24
HateBERT (HatEval)	0.16	0.36	0.30	0.27
HateBERT (OffenseEval)	0.23	0.43	0.37	**0.34**
HF Multilingual Toxicity Classifier Plus	0.29	0.36	0.31	0.22
Piccolo-HAP (Ours)	**0.51**	**0.50**	**0.50**	0.32

Table 1: Results on the TSD and TBO datasets (English). Best results in bold.

Model	German TBO: F1 Score ↑		
	TARGET + ARG.	ARG. ONLY	TARGET ONLY
Random	0.14	0.11	0.08
DeHateBERT (monolingual)	0.17	0.16	0.05
HF German Toxicity Classifier Plus V2	0.19	0.28	**0.21**
HF Multilingual Toxicity Classifier Plus	0.33	0.23	0.15
Piccolo-HAP (Ours)	**0.44**	**0.34**	**0.21**

Table 2: Results on the German TBO dataset . Best results in bold.

label assigned to each tweet. We evaluate on the 475 test examples.

- German TBO: We evaluate our model on another language by annotating a small evaluation set of offensive German tweets from the GermEval corpus (Wiegand et al., 2018). Two skilled German speaking annotators were trained in the English TBO annotation task, excluding the Harmful label. In total, 255 German tweets were annotated.
- Toxic Spans Detection (TSD) (Pavlopoulos et al., 2021): The toxic spans detection task (Sem-Eval 2021 Task-5) annotated English toxic comments at the span level, marking spans of text that contribute to the offensive score. They release code that evaluates predictions at the character level.

For both TBO datasets, we experiment with using our attention-based approach to identify both the TARGET and ARGUMENT (TARGET + ARG.), only the ARGUMENT (ARG. ONLY) and only the TARGET (TARGET ONLY). In the TARGET ONLY setting, we exclude the examples that have no TARGET and only evaluate on the remaining examples; 342 English sentences, and 229 German sentences. We find this to be a fairer evaluation of our attention-based approach, as for sentences without a target the model may still produce argument spans as a prediction (as argument spans will always be attended to heavily). We leave evaluation on the NULL target examples as future work.

Note, our models are not trained on any of the above datasets, we only use them as a tool to evaluate our attention-based span detection approach for available HAP classification models.

4.2 Results

We re-format each dataset as a span identification task, where the output of our system is the character-level spans for the predicted offensive arguments/targets (the spans are computed using attention maps, as described in Section 3). The F1 scores are computed on a character level, following the approach of Pavlopoulos et al. (2021). Here, the training set is used to identify the best attention thresholds to choose the offensive spans, and the test set for evaluating model performance.

Table 1 compares the results of our models to the baseline models on the English datasets. As shown, our model strongly outperforms the Hate-BERT classification models, the MTC+ Classifier, and the random baseline on the TSD task. We evaluate models on the the TBO dataset under three settings, and show that our models significantly outperform all baselines on identifying both the target and argument, and only the argument. On identifying only the target, it is slightly behind HateBERT finetuned on OffensEval.

The results on the German TBO dataset are shown in Table 2. We follow the same experimental setup as for the English results in Table 1 and present separate results for predicting both target and argument individually and jointly. Our Piccolo-HAP model outperforms all other German Huggingface models and also the multi-lingual model with the exception of the target-only score by the
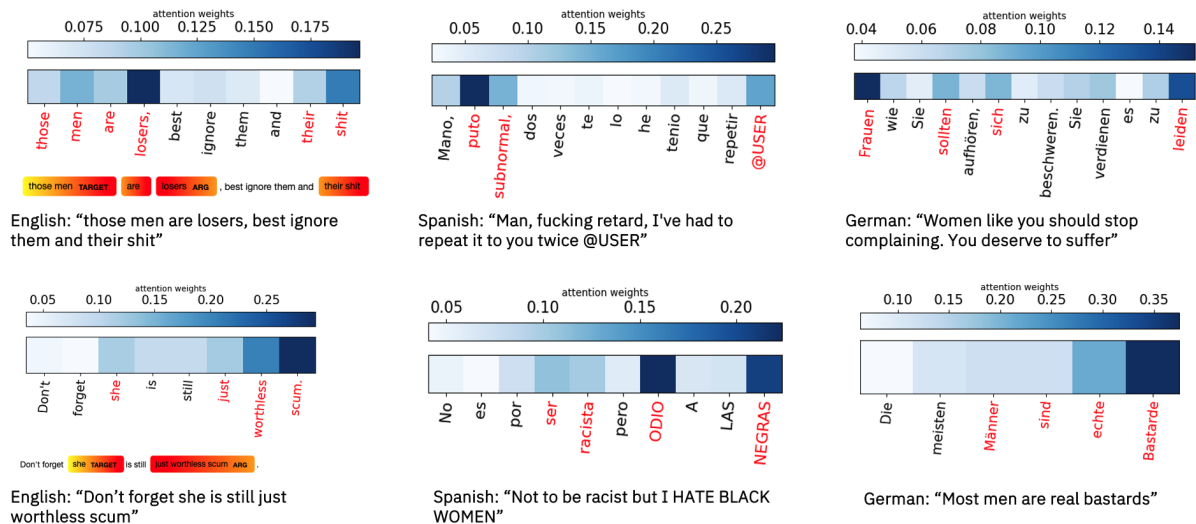
Figure 4: System outputs for examples in English, Spanish, and German. Offensive spans and targets marked in red, and images are captioned with the English translations of the input.

HF German Toxicity Classifier.

As seen for all models, predicting both TARGET and ARGUMENT is an easier task than predicting each individually, with Target-only being the hardest setting. A way to improve the performance on this task is to modify the existing method of using the CLS token's attention to identify targets, and instead use the attention of the argument to identify the target. We leave this as future work.

For further understanding, we analyze a set of sentences from the English TBO dataset for which our model performs poorly in the Target-only setting. We find no clear patterns in this data, however we do find that our approach works very well when the targets themselves are described using offensive or derogatory terms (e.g. "these bitches", "little twats", "clowns", "idiots"). Moreover, our model does not correctly identify targets containing typos (which are common in tweets), such as *yal* instead of *y'all*. As part of future work, a spelling corrector and parser can be built into the HAP prediction system, along with current attention-based thresholds. We also analyze our model's output for some test cases where there is a NULL target annotated in the gold data, and find that our model may predict spans that could be interpreted as the TARGET. For example, the text *"The rich white people don't give a fuck about you unless you affect their bottom line"* marks NULL target in the gold data, but our model outputs *"the rich white people"* as one span, which could be interpreted as the TARGET of offensive ARGUMENT (*"don't give a fuck"*).

5 System Demonstration

We have Jupyter Notebooks and a front-end UI where users can load their models, and obtain visualizations for inputs in any language.

5.1 Jupyter Notebook

We have created a Python Jupyter notebook for displaying the <T,A> offensive pairs in a sentence. The notebook will load any encoder-only sentence level offensive classifier. It can be used on multilingual models trained on any language (e.g. English and German as we presented in our experiments). Given a sentence, we generate a heat map using the attention of the model. Then, we identify the offensive TARGETS and ARGUMENTS using a threshold on the attention. We use the *subj* and *obj* labels from the spaCy dependency parser to identify the TARGET (subject) and ARGUMENT (object) of offense. Finally, we use the spaCy visualization tools to render the sentence with the offensive TARGETS and ARGUMENTS[7]. Example visualizations for inputs in several languages are shown in Fig. 4. We would like to extend the tool to more languages based on multi-lingual parsing models.

5.2 User Interface

The MUTED user interface allows the user to play with the HAP classification model without having to know any technical details. The user interface is implemented in Flask which is a lightweight native

---

[7]In English only as the spaCy German parser did not provide the proper information to identify the target and argument.
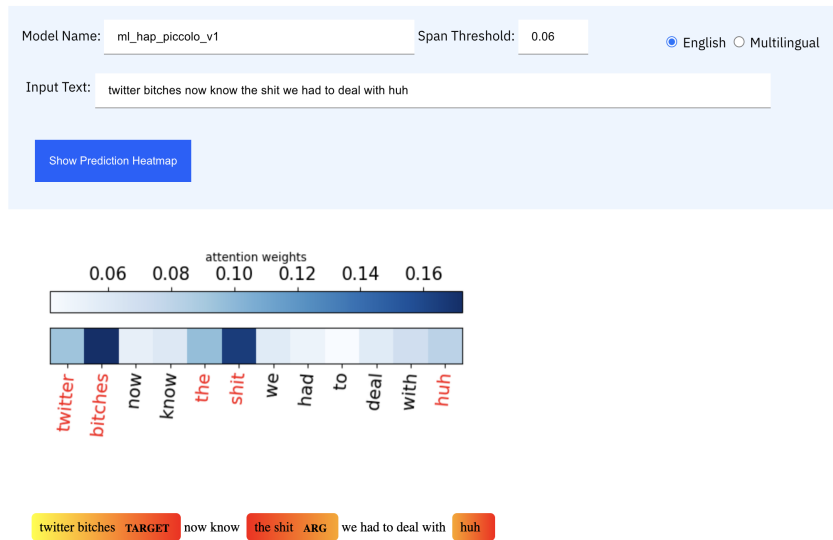
Figure 5: Screen-shot of MUTED User Interface: The user inputs the model name and input text, and selects the language and attention threshold. The system produces the attention heatmap, and (for English inputs) the spaCy visualization marking the target and argument.

python web application framework. We show an example of the user interface in Fig. 5. The UI allows the user to input the sentence, select if the language is English or non-English and select the value of Span Threshold. Upon clicking "Show prediction Heatmap", the UI renders the output visualizations on the same page. Same page rendering allows the user to tune the output with the best possible parameter values.

5.3 System Efficiency

We evaluate the time taken to produce the predictions and visualizations for a single input by averaging the inference time for 100 English texts. Note that the major difference between the CPU and GPU latencies is contributed by the time taken to make a prediction (which happens on the GPU when available). The visualizations always happen on the CPU, and also utilize more time.

We show the results for two multilingual models-our Piccolo HAP classifier (a 4-layer model with 153 million parameters), and the MTC+ Classifier (a 12-layer model with 277 million parameters). It takes 0.65/0.64s on CPU/GPU to run with our small model, and 0.76/0.65s on CPU/GPU for the base size model, for a single input. Table 3 shows the average latency of a single input for the different steps in the process. Thus, the system is quite efficient, and can process 100 examples in about a minute on both CPU and GPU.

		Piccolo Model	MTC+ Model
CPU	Span Prediction	0.02	0.11
	Attention Map	0.22	0.23
	SpaCy Visuals	0.41	0.42
GPU	Span Prediction	0.01	0.02
	Attention Map	0.22	0.22
	SpaCy Visuals	0.41	0.41

Table 3: Time taken (s) for span prediction and visualization of a single input. Avg. metric reported over 100 sentences, using a single core CPU and V100 GPU.

6 Conclusion

We present a method for identifying and visualizing offensive arguments and their targets using the attention of the sentence-based offensive classifier to create a heat map. Our multilingual model outperforms existing popular approaches on multiple datasets in English and German. We provide a notebook and user interface to run any multilingual transformer classifier on sentences and visualize the heat map as well as the <T,A> pair using spaCy visualization. In the future, we would like to add a classifier to indicate harm of the <T,A> pair as described in the TBO paper. We would also like to extend our demo to provide warnings and hide the offensive content to users.

Ethics Statement

Limitations

In this work, we focus primarily on English and German offensive texts. While our Piccolo model supports 6 languages, and there exists open source HAP classification models of many languages, there is a limitation on datasets available for testing. Specifically, we primarily test on datasets that include annotations of the target of offense, which are not widely available. Creating such datasets for multiple languages would be an interesting direction for future research. Moreover, the test sets that we evaluate on are relatively small in size, and consist of shorter text spans such as tweets.

As mentioned, a more robust way to use transformer attentions to identify the target of the offense is to find the words most heavily attended to by the tokens in offensive span (argument), instead of the CLS vector. In this approach, some type of aggregation strategy would be needed to select the correct tokens from the span, and we aim to implement this as part of future work.

Intended Use

Detecting offensive content is an important task that is necessary for avoiding harm. While hateful and harmful content is used to train the models, our intended use is solely for the purpose of avoiding and removing such content and we do not support any malicious or unintended use.

Biases

Due to the subjective nature of the task, our German annotated dataset may have unintended biases. These kind of biases are unintentional and will be prevalent in any subjective task. Anyone that uses the data should be aware that such biases may exist. Our TBO annotations are built on top of the existing GermEval dataset (Wiegand et al., 2018). We also use the TBO (Zampieri et al., 2023) and TSD(Pavlopoulos et al., 2021) dataset. Any biases in those original datasets will exist in ours as well which may impact the trained model.

Acknowledgement

We would like to acknowledge the work of our annotators who worked on the German TBO set: Eva-Maria Wolfe, Joekie Gurski, and Mohamed Nasr.

References

Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

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, Minnesota, USA. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. 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 Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification data.

João A. Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis.

Angelina McMillan-Major, Amandalynne Paullada, and Yacine Jernite. 2022. An interactive exploratory tool for the task of hate speech detection. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 11–20, Seattle, Washington. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21).

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online. Association for Computational Linguistics.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the ACL*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Björn Roß, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. Minilmv2: Multi-head self-attention relation distillation for compressing pre-trained transformers.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng (Polo) Chau, and Diyi Yang. 2021. Recast: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

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 NAACL*.

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. Association for Computational Linguistics.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Toronto, Canada. Association for Computational Linguistics.