Target-Based Offensive Language Identification

Marcos Zampieri[1], Skye Morgan[2], Kai North[1], Tharindu Ranasinghe[3]
Austin Simmons[2], Paridhi Khandelwal[2], Sara Rosenthal[4], Preslav Nakov[5]

[1]George Mason University, USA, [2]Rochester Institute of Technology, USA
[3]Aston University, UK, [4]IBM Research, USA
[5]Mohamed bin Zayed University of Artificial Intelligence, UAE

mzampier@gmu.edu

Abstract

We present TBO, a new dataset for Target-based Offensive language identification. TBO contains post-level annotations regarding the harmfulness of an offensive post and token-level annotations comprising of the target and the offensive argument expression. Popular offensive language identification datasets for social media focus on annotation taxonomies only at the post level and more recently, some datasets have been released that feature only token-level annotations. TBO is an important resource that bridges the gap between post-level and token-level annotation datasets by introducing a single comprehensive unified annotation taxonomy. We use the TBO taxonomy to annotate post-level and token-level offensive language on English Twitter posts. We release an initial dataset of over 4,500 instances collected from Twitter and we carry out multiple experiments to compare the performance of different models trained and tested on TBO.

1 Introduction

Confrontational and often offensive behavior is pervasive in social media. Online communities, social media platforms, and tech companies are well aware of the problem and have been investigating ways to cope with the spread of offensive language. This has sparked growing interest in the AI and NLP communities in identifying offensive language, aggression, and hate speech in user-generated content (Davidson et al., 2017; Vidgen and Derczynski, 2020; Mubarak et al., 2020; Aggarwal et al., 2023).

The interest in this topic has motivated the study of offensive language online from different angles. Popular shared tasks organized in the past few years have created benchmark datasets, e.g., OLID (Zampieri et al., 2019a), which are widely used in research on this topic.

Most of these shared tasks and datasets, e.g., HatEval (Basile et al., 2019) and OffensEval (Zampieri et al., 2020), have modeled offensive language at the post-level, where the goal is to predict the label of each post (e.g., offensive vs. not offensive, or hate speech vs. not hate speech). More recently, Pavlopoulos et al. (2021) developed the Toxic Spans Detection (TSD) dataset, which is annotated at the token level to focus on explainability by identifying the token spans that make a post offensive and toxic. One limitation of TSD is that it focuses exclusively on the toxic spans, while the target of the offensive expression is not annotated, for example:

(1) Canadians are very friendly, but their *politicians* are **shit**.

In the TSD dataset, *shit* would be labeled as toxic, but there would be no attempt to identify the actual target, which is *politicians*. Note that both *Canadians* and *politicians* could potentially be the target of the offensive expression. Knowing the target is important for understanding the nature of the offensive post (e.g., hate speech vs. general profanity), an aspect that has been captured by a few annotation taxonomies (Basile et al., 2019; Zampieri et al., 2019a). Another aspect not previously addressed is the issue of harmfulness, which is often related to polarity. All tasks so far have made the assumption that posts containing a curse word would be harmful; yet, consider the following example:

(2) This is one good looking **motherfucker**.

Even though the word *motherfucker* is used, this sentence has positive polarity, and thus annotating this word as offensive or toxic would likely yield incorrect predictions for offensive language detection. Curse words with positive polarity are a relatively common phenomenon, and thus systems should be able to recognize the use of such words in the context of harm.

WARNING: This paper contains offensive examples.

To address these limitations, we introduce a novel annotation taxonomy called Target-Based Offensive language identification (TBO). We use TBO to annotate a new dataset containing over 4,500 posts from Twitter. Our task addresses two important gaps in previous research: detecting (*i*) the offensive span along with its target, and (*ii*) its harmfulness. Furthermore, we derive two post-level labels from the token-level annotation as described in Section 3. We draw inspiration from the popular aspect-based sentiment analysis task (Pontiki et al., 2016), which promoted explainability in sentiment analysis. Here, we apply a similar idea to offensive language identification. The main contributions of our work are as follows:

1. A new target-based taxonomy that will open new avenues for research in offensive language identification with a special focus on explainability.

2. Development and release of the TBO dataset containing 4,500 manually annotated posts from Twitter.

3. An evaluation of multiple models trained on this dataset. To the best of our knowledge, this is the first computational modeling of offensive expressions and targets in offensive language identification. The code and the pre-trained models are made freely available to the community.[1]

2 Related Work

The interest in studying and modeling offensive language in social media continues to grow. This is evidenced by the creation of many widely-used datasets released in the past few years (Founta et al., 2018; Davidson et al., 2017; Zampieri et al., 2019a; Rosenthal et al., 2021) and the organization of multiple popular shared tasks at SemEval and other venues. Along with the aforementioned HatEval (Basile et al., 2019), OffensEval (Zampieri et al., 2019b, 2020), and TSD (Pavlopoulos et al., 2021), some related tasks have recently been organized also at SemEval, namely HaHackathon (Meaney et al., 2021) on humor and offensiveness, and MAMI (Fersini et al., 2022) on multimodal (text and image) offensive content targeted at women.

Popular related tasks organized in other venues include HASOC (Modha et al., 2021; Satapara et al., 2022) at the Forum for Information Retrieval (FIRE) and TRAC (Kumar et al., 2018, 2020) at the TRAC workshop. As discussed in a survey by Poletto et al. (2021), all these competitions have provided participants with important benchmark datasets to evaluate the performance of systems trained to detect multiple different types of offensive content.

With the exception of the aforementioned TSD (Pavlopoulos et al., 2021) and HateXplain (Mathew et al., 2021) datasets, which deal with token spans, all the datasets and competitions discussed in this section target post-level offensive language identification where systems are trained to attribute a label, such as offensive or not offensive, to an instance, typically a post or a comment. The identification of offensive spans has been, so far, mostly unexplored, and our TBO dataset fills this gap. Moreover, the unified taxonomy with target and harmfulness modeled as triples is a new way of reformulating the problem and it is the main new contribution of our work. We believe that our TBO taxonomy opens new avenues for future research.

3 Target-Based Offensive Language Identification

3.1 Annotation Taxonomy

Our TBO taxonomy builds on the annotation framework proposed in OLID (Zampieri et al., 2019a), which was widely replicated in several datasets for English (Rosenthal et al., 2021) and other languages (Pitenis et al., 2020; Sigurbergsson and Derczynski, 2020; Çöltekin, 2020; Gaikwad et al., 2021). As a result, the OLID taxonomy has become a *de facto* standard for general offensive language identification due to the flexibility it provides by representing multiple phenomena, which were treated in isolation in many other studies such as cyberbulling, hate speech, and general profanity (Rosa et al., 2019; Poletto et al., 2021). OLID's hierarchical annotation model comprises of three levels: level A (offensive or not offensive), level B (targeted or untargeted), and level C (group, person, or other). The assumption is that the type and the target of posts is paramount in determining the type of offensive content, e.g., offensive posts targeted at a group are often considered hate speech, while such targeted at an individual are often considered cyberbulling.

[1] https://github.com/LanguageTechnologyLab/TBO

Tweet	TARGET	ARGUMENT	HARMFUL
@USER Liberals are all Kookoo !!!	Liberals	Kookoo	YES
@USER He is a DUMBASS !!!!!	He	DUMBASS	YES
@USER @USER @USER Says the fat Antifa member	Antifa member	fat	YES
@USER Oh shit stay safe!!	NULL	shit	NO
@USER Master of None was so fucking good.	Master of None	fucking	NO

Table 1: Examples of tweets from the **TBO** dataset with corresponding annotations of TARGET, ARGUMENT, HARMFULNESS triples.

In TBO, we consider offensive posts as defined by OLID level A with multiple types and targets. The TBO annotation taxonomy models token-level annotation of these offensive posts in triples: (TARGET, ARGUMENT, HARMFULNESS).

TARGET The target of the offensive argument, such as a person or a group. This can also be NULL when the instance is untargeted, e.g., *Oh shit, stay safe!*

ARGUMENT The span containing the offensive tokens.

HARMFULNESS YES, if the argument is harmful to the target; otherwise, NO. Harmful expressions will often correlate with negative polarity as in the case of sentiment analysis.

Examples of triples are shown in Table 1. Note that the relationship between TARGET and ARGUMENT can be 1:M, M:1, or even M:M. Here is an example of an M:M relationship:

(3) *Peter* is an **idiot** and an **asshole**, and so is *John*.

In this case, four triples can be formed: (*Peter*, *idiot*, YES), (*Peter*, *asshole*, YES), (*John*, *idiot*, YES), and (*John*, *asshole*, YES).

Overall, our two **TBO** subtasks are substantially different from previous tasks on this topic: we address the identification of targets rather than spans, and we further focus on the harmfulness of the offensive arguments. To the best of our knowledge, this is the first work in which these two aspects of offensive language identification have been addressed.

3.2 The TBO Dataset

We sampled data from the SOLID dataset (Rosenthal et al., 2021), the largest English offensive language dataset with over 9 million tweets.

SOLID's semi-supervised annotation strategy follows OLID's three-layer annotation taxonomy, which enabled us to use a sampling strategy geared towards collecting complex targets and a wide variety of offensive arguments. In particular, we sampled social media posts with an aggregate offensive score (OLID/SOLID level A) in the range [0.6–1.0] to ensure that our sampled data was rich in curse words and offensive arguments. We further filtered posts to have at least 11 tokens, ensuring we obtained longer posts, which tended to contain longer arguments and often times, several associated targets. Lastly, we used the SOLID level C score to filter posts that target groups with the goal of obtaining posts that are more likely to contain multiple targets.

To measure the inter-annotator agreement (IAA), we first performed a trial annotation experiment with 350 tweets annotated by seven trained annotators working on the project. Four of them were graduate students in computing based in the USA aged 22-30, while three were researchers in NLP aged 30-45 based in the USA and UK. We report 0.81 Kappa IAA for harmfulness and 0.78 for the target. After the trial experiment, we randomly selected samples for training and testing, which were then annotated by the same annotators.

The final dataset comprises over 4,500 tweets. The training set has a total of 4,000 instances and includes 6,924 triples, 4,863 of which are harmful. Table 2 provides some statistics about the number of tweets per set along with the number of harmful and harmless triples.

Set	Instances	Triples	Harmful	Harmless
Train	4,000	6,924	4,863	2,061
Test	673	1,096	640	456
Total	**4,673**	**8,020**	**5,503**	**2,517**

Table 2: Number of tweets and triples in the **TBO** dataset, and their harmfulness.

Set	Targeted	Harmful
Train	3,167	2,890
Test	505	445
Total	**3,672**	**3,335**

Table 3: Number of targeted and harmful tweets.

We further compute the number of targeted and harmful posts in the dataset and we present this information in Table 3. We considered a post targeted if it contains at least one targeted triple, and harmful if it contains at least one harmful triple.

4 Methods

We experimented with three types of models:

Triple Prediction Models Since the goal of TBO is to predict all elements of an offensive tweet (target, argument, and harmfulness), we are more interested in models that can output triples instead of individual elements. Therefore, we used the following models capable of predicting triples. **Sequence Labeling** (Barnes et al., 2022) where a BiLSTM is used to extract targets and arguments separately and then we train a relation prediction model to predict the harmfulness. **Dependency Graph**, adapted from the head-final approach of Barnes et al. (2021), where the target, the arguments, and the harmfulness are modeled as a dependency graph parsing problem. Finally, two versions in RACL (Chen and Qian, 2020): **RACL-GloVe** and **RACL-BERT**, which use GloVe 840B and BERT-large as input embeddings, respectively.

Token Classification Models We experimented with different token classification architectures, which we trained on two tasks separately: target identification and argument identification. These implementations are largely adopted from the toxic spans detection task (Pavlopoulos et al., 2021). Our **BiLSTM** is a Bi-LSTM-CRF model (Panchendrarajan and Amaresan, 2018). We also experimented with a token classification architecture in transformers, based on BERT-large, to which we refer as **BERT-token**.

Binary Prediction Models Finally, we experimented with a sentence classification architecture in transformers based on BERT-large, referred to as **BERT-post**. The classifier is trained at the post level: if the tweet contained at least one harmful triple, we considered the entire tweet harmful.

4.1 Evaluation Measures

As we are interested in extracting full triples, we propose evaluation measures that capture the relationship between all predicted elements.

(1) Spans - Token-level F1 for TARGET and ARGUMENT This evaluates how well these models are able to identify the elements of a tuple.

(2) Targeted F1 A true positive example requires the combination of exact extraction of the target, and correct harmfulness label.

(3) Target Argument We used two evaluation measures that evaluate the model's capability to extract the target and the argument jointly. The first one is Non-polar Target Argument F1 (NTAF1), where each prediction is considered as a pair of (TARGET, ARGUMENT) and a true positive is defined as an exact match of all elements. We also used Target Argument F1 (TAF1), which uses the same measures as NTAF1, but includes harmfulness as well: (TARGET, ARGUMENT, HARMFULNESS).

(4) Harmful F1 Macro-F1 scores for harmfulness either at the tuple level or at the post level, depending on the model.

4.2 Experimental Setup

We used a GeForce RTX 3090 GPU to train the models. We divided the TBO dataset into a training set and a development set using an 80-20 split.

Transformers We used the configurations presented in Table 4 in all the experiments. We performed *early stopping* if the validation loss did not improve over ten evaluation steps.

Parameter	Value
adam epsilon	1e-8
batch size	64
epochs	3
learning rate	1e-5
warmup ratio	0.1
warmup steps	0
max grad norm	1.0
max seq. length	256
gradient accumulation steps	1

Table 4: Transformer parameter specification.

Parameter	Value
first dense layer units	256
LSTM units	64
voab size	3000

Table 5: BiLSTM parameter specifications.

BiLSTM Model Configurations The configurations for the BiLSTM model are presented in Table 5. The training process was similar to that for the transformer models.

5 Results

Table 6 shows the results for all models from Section 4. We trained each model with five different random seeds, and we report the average evaluation scores. For all models and evaluation measures, the standard deviation was less than 0.0001.

All models performed well at extracting targets scoring more than 0.3 on Target F_1 score. RACL-BERT model performed best with a Target F_1 score of 0.443, and it yielded the best overall result for Targeted F_1. Comparatively, all models struggled with predicting the argument. None of the models we experimented with managed to reach an Argument F_1 score of 0.3. RACL-BERT performed best for predicting arguments as well. All of the triple prediction models performed competitively to token classification architectures. As all models struggled with predicting the arguments, the target argument measures are low for all of them. Among the triple prediction models, RACL-BERT achieved the best NTAF1 and TAF1 scores. Both post-level models and triple-prediction models thrived on harmfulness prediction. RACL-BERT achieved the best result from the triple-prediction models scoring 0.693 Macro F1 score on the triple level.

6 Conclusion and Future Work

We presented our novel target-based Offensive language identification (TBO) taxonomy, which we used to annotate a new English dataset with over 4,500 tweets. We further evaluated the performance of various models on this new dataset and we discussed the evaluation results in detail.

We release all data as well as our code publicly. We believe that the TBO taxonomy and our dataset will be widely used in research on this topic as they have addressed important gaps in previous annotation taxonomies, most notably target identification.

In future work, we plan to annotate more data using the taxonomy proposed above, including other languages. This will allow us to take advantage of existing cross-lingual learning models for making predictions as well as for studying cross-language and cross-cultural aspects of offensive language online. We would also like to create comparable annotated TBO datasets for other languages, which will allow us to take advantage of existing cross-lingual models for offensive language identification (Ranasinghe and Zampieri, 2020; Nozza, 2021). We believe that this will get us closer to what online platforms need (Arora et al., 2023).

Ethics Statement

The dataset we presented in this paper was collected from SOLID (Rosenthal et al., 2021), a freely-available large-scale dataset containing data from Twitter. No new data collection has been carried out as part of this work. We did not collect or process writers'/users' information, nor have we carried out any form of user profiling, thus protecting users' privacy and anonymity. Note also that in SOLID, all Twitter handles are replaced with @USER as a de-identification process. We understand that every dataset is subject to intrinsic biases and that computational models will inevitably learn biased information from any dataset. That being said, we believe that the token-level annotation in TBO will help cope with biases found in models trained on tweet-level annotations by improving the model's interpretability.

Intended Use Our intended use is the same as for SOLID, the dataset we sampled our examples from (Rosenthal et al., 2021). We aim to encourage research in automatically detecting and limiting offensive content towards a target from being disseminated on the web. Using our dataset for its intended use can alleviate the psychological burden for social media moderators who are exposed to extremely offensive content. Improving the performance of offensive content detection systems can decrease the amount of work for human moderators, but some human supervision is still necessary to avoid harm and ensure transparency. We believe that content moderation should be a trustworthy and transparent process applied to clearly harmful content so it does not hinder individual freedom of expression rights. We distribute our dataset under a Creative Commons license, the same as for SOLID. Any biases found in the dataset are unintentional.

Model	(1) Spans		(2) Targeted	(3) Tar. Arg.		(4) Harm
	Target F_1	Arg. F_1	F_1	NTAF_1	TAF_1	F_1
Sequence labeling	0.326	0.193	0.238	0.185	0.178	0.633
Dependency graph	0.368	0.213	0.282	0.206	0.201	0.657
RACL-GloVe	0.335	0.208	0.241	0.202	0.191	0.621
RACL-BERT	0.442	0.256	0.381	0.243	0.233	0.693
BERT-token	0.412	0.236	-	-	-	-
BiLSTM	0.315	0.182	-	-	-	-
BERT-post	-	-	-	-	-	0.745

Table 6: Experiments comparing the different models on the TBO dataset.

Limitations

Biases Human data annotation for a sentiment-related task, e.g., aspect-based sentiment analysis, hate speech detection, etc., involves some degree of subjectivity. While we included important quality control steps in the TBO annotation process, this intrinsic subjectivity will inevitably be present in TBO and learned by the models (see also the Ethics Statement above). That being said, the hierarchical annotations presented in OLID, TBO, and other similar datasets aim to increase the annotation quality by breaking down the decision process, thus providing clearer guidelines to the annotators.

Dataset Collection Another factor that may be considered as a limitation is the dataset size: 4,500 instances and 8,000 triples. We would expect models to perform better when the dataset is expanded in the future. We are addressing this limitation by annotating more data that will be ready for release soon. Finally, another limitation is that this is currently an English-only dataset. We would like to expand TBO to other languages and to take advantage of cross-lingual models (XLM-R, mBERT, etc.) for multilingual predictions.

Risks A dataset containing offensive content is at risk of misuse. The dataset can be maliciously used to build models that unfairly moderate text (e.g., a tweet) that may not be offensive based on biases that may or may not be related to demographic and/or other information present within the text. Due to the nature of the task, this dataset can be also used maliciously to display offensive content. The dataset should not be used for this purpose; our intended use is discussed in the Ethics Statement. Intervention by human moderators would be required to ensure that malicious uses do not occur.

References

Piush Aggarwal, Pranit Chawla, Mithun Das, Punyajoy Saha, Binny Mathew, Torsten Zesch, and Animesh Mukherjee. 2023. HateProof: Are hateful meme detection systems really robust? In *Proceedings of TheWebConf*.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2023. Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Comput. Surv.*

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of ACL*.

Jeremy Barnes, Laura Ana Maria Oberländer, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval-2022 task 10: Structured sentiment analysis. In *Proceedings of SemEval*.

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 SemEval*.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of LREC*.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of ACL*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022

Task 5: multimedia automatic misogyny identification. In *Proceedings of SemEval*.

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. In *Proceedings of ICWSM*.

Saurabh Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher M Homan. 2021. Cross-lingual offensive language identification for low resource languages: The case of Marathi. In *Proceedings of RANLP*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of TRAC*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: a benchmark dataset for explainable hate speech detection. In *Proceedings of AAAI*.

JA Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of SemEval*.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages and conversational hate speech. In *Proceedings of FIRE*.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of OSACT*.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of ACL*.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of PACLIC*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of SemEval*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of LREC*.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of SemEval*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of EMNLP*.

Hugo Rosa, N Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, S Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

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*.

Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2022. Overview of the HASOC subtrack at FIRE 2022: Hate speech and offensive content identification in English and Indo-Aryan languages. In *Proceedings of FIRE*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of LREC*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS One*, 15(12):e0243300.

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 SemEval*.

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). In *Proceedings of SemEval*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 6.3

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2.2, Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 2.2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 6.4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 6.4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data comes from SOLID where the authors replace all twitter handles with @USER. No identifying information is present. See Section 5.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Information regarding the data can be found in the original dataset, SOLID, our dataset is derived from. All content is in English.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2.2 and Tables 2 and 3

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3 and supplemental material

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3 and Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 2.2.

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We had internal documents outlining the annotation tasks. All participants were familiar with the annotation guidelines as they were working in the project. We have not relied on external annotators for this task.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 2,2. Please note that participants were not paid for the annotation because they were collaborators in the project.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 2.2. We have collected data from SOLID which is a dataset that adheres to the Twitter guidelines and the data is anonymized.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The data comes from SOLID, we did not collect new data.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 2.2.