# **Targets and Aspects in Social Media Hate Speech**

Alexander Shvets<sup>1</sup>, Paula Fortuna<sup>1</sup>, Juan Soler-Company<sup>1</sup>, Leo Wanner<sup>2.1</sup>

<sup>1</sup>NLP Group, Pompeu Fabra University, Barcelona, Spain <sup>2</sup>Catalan Institute for Research and Advanced Studies

alexander.shvets|paula.fortuna|juan.soler|leo.wanner@upf.edu

### Abstract

Mainstream research on hate speech focused so far predominantly on the task of classifying mainly social media posts with respect to predefined typologies of rather coarse-grained hate speech categories. This may be sufficient if the goal is to detect and delete abusive language posts. However, removal is not always possible due to the legislation of a country. Also, there is evidence that hate speech cannot be successfully combated by merely removing hate speech posts; they should be countered by education and counter-narratives. For this purpose, we need to identify (i) who is the target in a given hate speech post, and (ii) what aspects (or characteristics) of the target are attributed to the target in the post. As the first approximation, we propose to adapt a generic state-of-the-art concept extraction model to the hate speech domain. The outcome of the experiments is promising and can serve as inspiration for further work on the task.

# 1 Introduction

Online hate speech and, in particular, hate speech in social media, is the cause for growing concern. Already six years ago, 73% of adult internet users have seen someone harassed online, and 40% have personally experienced it (Duggan, 2014). Therefore, research on hate speech identification is of increasing importance. A significant body of work has been conducted over the last decade; cf., e.g., (Waseem and Hovy, 2016a; Schmidt and Wiegand, 2017; Davidson et al., 2017a; Fortuna and Nunes, 2018; Kennedy et al., 2020). Most of this work focused on the task of classifying, for instance, social media posts, with respect to predefined typologies of rather coarse-grained hate speech categories, such as 'hate speech', 'racism', 'sexism', 'offense', etc. This may be sufficient if the task is to detect and remove abusive language posts. However, for

instance, in the US, hate speech has been repeatedly judged as being covered by the First Amendment.<sup>1</sup> Furthermore, a number of studies suggest that hate speech cannot be successfully combated by merely removing identified hate speech posts<sup>2</sup> and should be countered by education and counternarratives (Tekiroğlu et al., 2020; Mathew et al., 2019). But to provide a basis for education and counter-narratives, we need a more detailed analysis of hate speech. In particular, we need to identify (i) who is the target in the identified hate speech post, and (ii) what aspects of the target are referred to or what aspects are attributed to the target in the post. For instance, we need to be able to determine that post (1) below targets Muslims of Palestine and that it attributes to them to be terrorists. Similarly, for post (2), we need to determine that it targets female sports reporters and that they "should come to an end" (i.e., that they should be removed from their jobs).<sup>3</sup> The analogy to aspect-oriented sentiment analysis (Schouten and Frasincar, 2016) is evident.

- (1) I'm standing outside and looking in and there isn't a shadow of doubt that the Muslims of Palestine are the terrorists.
- (2) I'm not sexist but female sports reporters need to come to an end.

Some recent works on hate speech go beyond the mere classification task and, actually, some of them also use the term *aspect*, but, again, with a

<sup>&</sup>lt;sup>1</sup>See, among others, Brandenburg vs. Ohio (1969), Snyder vs. Phelps (2011), Matal vs. Tam (2017), etc.; https://en.wikipedia.org/wiki/Hate\_ speech\_in\_the\_United\_States provides further details and references.

<sup>&</sup>lt;sup>2</sup>See, e.g., https://unesdoc.unesco.org/ark: /48223/pf0000233231.

<sup>&</sup>lt;sup>3</sup>Both posts are from the (Waseem and Hovy, 2016a) dataset.

different interpretation. In this paper, we present an approach that is different from these works and that aims to identify (i) the entity (most often, a group of individuals or an individual) who is *targeted* in the post, without drawing upon a predefined range of categories (which will necessarily be always limited and coarse-grained and will not cover new or intersecting categories, like 'black women'), (ii) the *aspect* (or *characteristics*) assigned to the targeted entity. We use an open-domain neural network-based concept extraction model for the identification of target and aspect candidates in each post. The obtained candidates are then further processed taking into account the idiosyncrasies of the codification of both targets and aspects in the domain.

The remainder of the paper is structured as follows. In the next section, we introduce the notions of *target* and *aspect* we are working with. Section 3 summarizes the work that is related to ours, including aspect-oriented sentiment analysis, to which our proposal shows some clear analogies. Section 4 outlines the generic concept extraction model from which we start and presents its adaptation to the problem of target and aspect extraction from hate speech data, while Section 5 describes the experiments we carried out and discusses their outcome. Section 6, finally, draws some conclusions and outlines several lines of research that we aim to address in the future.

# 2 Targets and Aspects

Let us define more precisely what we mean by 'target' and 'aspect' in the context of our work.

**Definition 1** (Target). A **target** is the entity that is in the focus of a hate post, i.e., the entity that incurs the hate of the author.

Very often, the target is an individual or a group of individuals, e.g., women, people of color, refugees, Muslims, Jews, etc.:

# (3) Bruh im tired of **niggas** retweetin Miley Cyrus naked that bitch aint no types of bad.

However, the target can also be a specific political conviction, a religion, an object related to an individual or a group of individuals, etc.; see, e.g., *feminist novels* in (4):<sup>4</sup>

# (4) I'm not sexist, but nothing bores me more than *feminist novels*.

**Definition 2** (Aspect). Aspect is a characteristic, attitude, or behavior or the lack of it (as a rule, with a pejorative connotation) that the author attributes to the target.

The aspect is often expressed as a modifier(e.g., *boring, stupid, lazy, not funny*, etc.) of the target in the focus of the author, as in:

- (5) I'm not sexist, but female comedians just aren't funny
- (6) I'm not sexist but \*most girls are fucking stupid.

where *not funny* is the aspect of *female comedians* (5) and *fucking stupid* of (*most*) *girls* (6). It can also be a verbal group, as *can't cook* in (7):

(7) Scoring like a Cunt because you can't cook for shit isn't fighting hard Kat.

In some posts, no targets and/or aspects can be identified; see, e.g., (8).

(8) I asked that question recently and actually got an answer http://t.co/oD98sptcGT.

We discard such posts in our current experiments.

## **3 Related Work**

As mentioned in Section 1, most of the works on online hate speech focused on the task of classifying social media posts with respect to predefined typologies of rather coarse-grained hate speech categories, such as 'hate speech', 'racism', 'sexism', 'offense', etc. (Schmidt and Wiegand, 2017; Davidson et al., 2017a; Fortuna and Nunes, 2018; Swamy et al., 2019; Arango et al., 2019; Salminen et al., 2020; Kennedy et al., 2020; Rajamanickam et al., 2020).<sup>5</sup> Vidgen and Derczynski (2020) distinguish between binary classification (as in (Alfina et al., 2017; Ross et al., 2017)), multi-class classification into several hate speech categories (e.g., 'racism', 'sexism', and 'none' in (Waseem and Hovy, 2016b)), different strengths of abuse classification (e.g., 'hateful', 'offensive' and 'neutral' contents as in (Davidson et al., 2017b)), classification into different types of statements (e.g., 'denouncing', 'facts', 'humor', 'hypocrisy' and others) and themes (e.g., 'crimes', 'culture', 'islamization',

<sup>&</sup>lt;sup>4</sup>In (4), *feminist* is a classifying attribute of *novels* (see also Section 4) and should thus be part of the target.

<sup>&</sup>lt;sup>5</sup>Cf. (Fortuna et al., 2020) for a list of categories used in the most common hate speech datasets.

'rapism' and others) as in (Chung et al., 2019)), and classification of different focuses of abuse (e.g., 'stereotype & objectification', 'dominance', 'derailing', 'sexual harassment', 'threats of violence', and 'discredit' as in (Fersini et al., 2018)). All these works do not aim to identify the specific targeted group of individuals or the individual and neither do they aim to identify characteristics of the targets that provoked hate. Rather, they identify posts related to hate speech in general or to one of its more specific categories – which is a step prior to detection of targets and aspects, where we start.

Some previous works use a similar terminology as we do, but with a different meaning. For instance, Zainuddin et al. (2017, 2018, 2019) aim to identify the sentiment (positive or negative) of the author of a given post towards a range of specific hate speech categories (e.g., 'race' and 'gender'), which they call "aspect". In (Gautam et al., 2020), tweets related to the MeToo movement are annotated manually with respect to five different linguistic "aspects": relevance, stance, hate speech, sarcasm, and dialogue acts. In this case, too, the interpretation of the notion of aspect is different from ours. Ousidhoum et al. (2019) define five different "aspects" that include specific targets, among others: (i) whether the text is direct or indirect; (ii) whether it is offensive, disrespectful, hateful, fearful out of ignorance, abusive, or normal; (iii) whether it is against an individual or a group of people; (iv) the name of the targeted group (16 common target groups are identified); and (v) the annotators' sentiment. Fersini et al. (2018) are also concerned with target detection in that they determine whether the messages were purposely sent to a specific target or to many potential receivers (e.g., groups of women). In (Silva et al., 2016), targets are identified using a short list of offensive words built drawing upon Hatebase<sup>6</sup> and a single template "<one word> people" to capture "black people", "stupid people", "rude people", etc.

Our work also aligns with Mathew et al. (2020) and Sap et al. (2020) in the sense that Mathew et al. (2020) annotate a hate speech dataset at the word and phrase level, capturing human rationales for the labelling (which is similar to the target–aspect labelling), while Sap et al. (2020) propose to understand and fight hate speech prejudices with accurate underlying explanations. However, Mathew et al. (2020) take into account only three labels ('hate', 'offensive', and 'normal') and ten target communities performing supervised classification, while we aim at retrieving and distinguishing open-class targets and aspects in a semi-supervised manner. Sap et al. (2020) perform supervised training of a conditional language generation model that often results in generic stereotypes about the targeted groups rather than in implications meant in the post, while we use a language generation model only to produce candidates and further expand, rank, and select them such that a connection of a target and an aspect to the text is guaranteed.

To summarize, although the identification of the targets and characteristics of hate speech in the above works are significant advancements compared to the more traditional hate speech classification, all of these works still assume predefined target categories and do not identify which characteristics of the targets are concerned. In contrast, open-class target and aspect extraction may allow for modeling of the particular forms of discrimination and hate experienced by individuals or groups of individuals covered or not covered by previously identified target categories.

As already mentioned in Section 1, our work is also related to aspect-oriented sentiment analysis, in which "targets" are specific entities (e.g., products, sights, celebrities) and "aspects" are characteristics or components of a given entity (Kobayashi et al., 2007; Nikolić et al., 2020). For each identified aspect, the "sentiment value" aligned with it is extracted; see, e.g., (Nazir et al., 2020) for a recent comprehensive survey of aspect-oriented sentiment analysis. In some (more traditional) works, aspects and their values are identified in separate stages (Hu and Liu, 2004; Hai et al., 2011). In more recent works, both tasks are addressed by one model, with aspects being partially identified by attention mechanisms realized, e.g., in an LSTM (Wang et al., 2016), CNN (Liu and Shen, 2020) or an alternative common deep NN model. The targets are, as a rule, predefined, such that the challenge consists in analysing the sentiment of tweets towards these predefined targets; cf., e.g., (Tang et al., 2016; Dong et al., 2014). The problem of open-class target identification has not been broadly investigated and sometimes solved simply as a named entity recognition problem due to the nature of the data in wihch the targets are often represented by proper names (Mitchell et al., 2013; Ma et al., 2018). However, targets in hate speech texts go far beyond named

<sup>&</sup>lt;sup>6</sup>http://www.hatebase.org/

entities, and the overall task is inverse to targetoriented sentiment classification: given a known category (hate speech of negative sentiment as a rule), we have to identify the hate target and its corresponding "opinioned" aspect. Still, our proposal is similar to the modern approaches to aspectoriented sentiment analysis in the sense that we also use an NN model (in our case, LSTM-based encoder) with attention mechanisms for initial hate speech target and aspect candidates identification, before a domain-adaptation post-processing stage.

# 4 Outline of the Model

The study of social media hate speech posts reveals that targets are entities that are, as a rule, verbalized in terms of *classifying nominal groups* (Halliday, 2013). Aspects may also be expressed by classifying nominal groups, but adjectival (attributive) and participle groups (actions) are also common. In other words, overall, targets can be considered *concepts* (Waldis et al., 2018). Therefore, we envision the detection of surface forms of targets in the posts primarily as a concept extraction (CE) task. For aspects, it is often not sufficient to apply concept extraction if we want to also capture the adjectival and verbal group aspects.

Given that hate speech datasets are, in general, too small to serve for training neural networks for reliable concept extraction, we opt for applying an open-domain-oriented concept extraction model with a follow-up algorithmic domain adaptation.

# 4.1 Generic Concept Extraction

As an open-domain concept extraction model, we use an open-source state-of-the-art model that comprises two pointer-generator networks pretrained on different concept-annotated datasets within distant supervision (Shvets and Wanner, 2020). Given a sentence, each network generates a list of concepts which are then merged and aligned with the sequence of tokens of a sentence. In case of ambiguity due to the overlap of surface forms of concepts, the first detected and the longest spans are selected as the resulting positions; see the implementation in the original publicly available code published along with the released models.<sup>7</sup>

The model is a sequence-to-sequence model; cf. Figure 1. The *pointer* mechanism makes it possible to copy out-of-vocabulary words directly to the outcome, which is especially relevant to our work, as the hate speech dataset includes specific words unseen during generic training, such as proper names, hashtags, and Twitter names. The *generator* implies the ability to adjust internal vocabulary distribution for selecting the next word (which might be a termination token "\*") based on weights of global attention  $a^t$  (Luong et al., 2015), which are updated at each generation step t. The probability of generating the next word instead of copying one is defined as follows:

$$p_{gen} = \sigma(w_{h*}^T h_t^* + w_x^T x_t + w_s^T s_t + b_{ptr}) \quad (1)$$

where  $h_t^*$  is the sum of the hidden states of the encoder weighted with the attention distribution  $a^t$ ,  $x_t$  is the decoder input,  $s_t$  is the decoder state,  $w_{h*}$ ,  $w_x$ ,  $w_s$ ,  $b_{ptr}$  are learnable parameters, and  $\sigma$ is the sigmoid function. The encoder is a stacked bidirectional LSTM, while the decoder is a stacked unidirectional LSTM (Hochreiter and Schmidhuber, 1997).

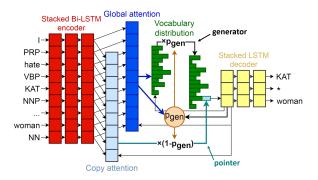


Figure 1: The neural architecture for generic concept extraction

#### 4.2 Domain Adaptation

The goal of the domain adaptation with respect to target and aspect determination in hate speech posts is to take into account the most relevant idiosyncrasies of the genre into account. In the case of targets, the following observations can be made with respect to such idiosyncrasies:

(i) While in generic discourse, targets can be assumed to be classifying nominal groups (see above), in hate speech, we observe also adjectival and participle targets that need to be captured.

(ii) Some targets form part of compounds and would thus be skipped by Shvets and Wanner (2020)'s generic concept extraction algorithm since it was trained to generate tokens from the input sentence without compound decomposition; cf.,

<sup>&</sup>lt;sup>7</sup>https://github.com/TalnUPF/ ConceptExtraction/

e.g.,  $Daeshbags \equiv Daesh+bags$ . The consideration of "subwords" instead of entire words has already proved to be beneficial for many NLP applications, including, e.g., machine translation (Sennrich et al., 2016). We thus consider also subwords of tokens. (iii) As a rule, a single post contains one target only; multiple targets are very seldom in short posts.<sup>8</sup> This means that all target candidates in a post must be ranked in terms of their probability to be a target. A high term frequency of a candidate across the posts implies a higher probability that this candidate is a common target, such that we favour candidates with a higher term frequency in a reference corpus. However, this is not the only criterion as this would introduce a strong bias towards frequent terms and contradict the idea of having unseen open-class targets. If no nominal candidates have been identified, we favour the adjectival/participle candidate with the highest tf\*idf, with the term frequency (tf) being calculated over a reference corpus and the inverse document frequency (idf) being calculated over the English Gigaword v.5 corpus (Parker et al., 2011). The same idea applies to aspects: aspect candidates should be ranked with respect to their likeliness to be a real aspect. To determine aspect candidates, we take into account the PoS and their position with respect to the previously determined target. Candidate aspects are: (i) concepts, which are detected by Shvets and Wanner (2020)'s generic concept extraction algorithm and which precede or follow the target; (ii) adjectival or participle modifiers either preceding or following the target.

Similarly to non-nominal target candidates, we favour aspect candidates with the highest *tf\*idf*, but regardless of their PoS. In addition to frequency terms, we chose several variables that give priority to different target and aspect candidates, depending on the weight assigned to them. They are listed in Table 1. Learning the weights within the domain adaptation stage using target-aspect expert annotated posts results in ranking criteria that are further used for the selection of target and aspect candidates in other (unseen) posts.

Three algorithms carry out the target and aspect identification. Algorithm 1 fine-tunes the weight variables from Table 1 for domain-specific target– aspect identification. An exhaustive weight variable fine-tuning procedure is run over all variable weight combinations. Algorithm 1 takes as input a domain reference dataset from which nominal concepts are extracted using the generic concept extraction model (reference target candidates  $T_{ref}$ ), and a development dataset from which new domainspecific targets and aspects are extracted (not necessarily nominal) using Algorithms 2 and 3. Expert annotation of the development dataset  $TA_d^{TRUE}$ serves as a reference during the weight variable tuning procedure.

Algorithm 2 outputs the target-aspect pair of a given post, extracted using the weight variables. It calls Algorithm 3 for the first stage target identification by a ranking based on variables  $a_1-a_5$ , then refines the delivered target and identifies the aspect by a ranking based on variables  $v_1-v_5$ .

Var	Weight of
$a_1$	nominal target candidate
$a_2$	proper name target candidate
$a_3$	target candidate comprises entire words
$a_4$	position of the candidate in p
$a_5$	expansion of the detected target
$v_1$	temporal expansion of the detected target
	within aspect detection
$v_2$	expansion of the detected aspect
$v_3$	nominal concept aspect candidate located in
	a span following the target
$v_4$	adjectival/participle aspect candidates
	regardless of their location in a post
$v_5$	nominal concept aspect candidate located
	in a span prior to the target

Table 1: Weight variables used in Algorithms 2 and 3

# **5** Experiments

## 5.1 Data

For our experiments, we use the 'sexism' and 'racism' partitions of the (Waseem and Hovy, 2016a) dataset, with 5,355 positive instances in total. The 5,355 instances are split into disjoint reference (90% of the 5,355 instances), development (2%) and test (8%) datasets. The reference dataset is used for the identification of domain-specific nominal group target candidates. The development set serves for fine-tuning the discrete variables used in Algorithms 2 and 3.

The development set (of 100 posts) and test set (of 440 posts) are annotated in terms of targets and aspects by three annotators. For this purpose, the annotators were provided with the definitions of the notions of 'target' and 'aspect' (see Section 2) and the instruction to first identify the target (which

<sup>&</sup>lt;sup>8</sup>Only about 2% of the posts in our dataset contain two targets. In order to expand the coverage of our algorithm, we plan to consider in our future work also datasets with longer texts; see the discussion of Figure 2 for details.

#### Algorithm 1: GetSettings: Domain adaptation

**Input:**  $S_{ref}$ : reference set,  $S_{dev}$ : development set,  $TA_d^{TRUE}$ : expert annotation of  $S_{dev}$ ,  $a_1, \ldots, a_5, v_1, \ldots, v_5$ : sets of possible discrete values for the weight variables **Output:**  $(T_{ref}, \vec{a}_{best}, \vec{v}_{best})$ **Dependencies:** *GetTA\_Pair* // Algorithm 2;  $C_{ref} \leftarrow ExtractConcepts(S_{ref}) // Apply concept extraction to S_{ref};$  $T_{ref} \leftarrow DetectSubjects(S_{ref}, C_{ref}) // Detect target candidates as concepts in the grammatical subject position;$  $TA_d \leftarrow \emptyset;$  $R^* \leftarrow \emptyset;$ foreach  $(\vec{a} \in \{a_1 \times a_2 \times a_3 \times a_4 \times a_5\}, \vec{v} \in \{v_1 \times v_2 \times v_3 \times v_4 \times v_5\})$  do // Select discrete values for components of  $\vec{a}$  and  $\vec{v}$  iteratively on a grid to extract target-aspect pairs from  $S_{dev}$ ; foreach  $p_d : post \in S_{dev}$  do  $TA_d \leftarrow TA_d \cup GetTA_Pair(p_d, T_{ref}, \vec{a}, \vec{v}) //$  Get target-aspect pairs using Algorithm 2: GetTA\_Pair; end  $r_{av} \leftarrow Score(TA_d, TA_d^{TRUE}) // Score resulting pairs TA_d;$  $R^* \leftarrow R^* \cup (\vec{a}, \vec{v}, r_{av})$ end

 $(\vec{a}_{d_{best}}, \vec{v}_{d_{best}}) \leftarrow (\vec{a}, \vec{v}) \in R^* \mid max(r_{av_{best}}) //$  variable values that give the best targets and aspects on dev set;  $(T_{ref}, \vec{a}_{best}, \vec{v}_{best}) \leftarrow (T_{ref}, \vec{a}_{d_{best}}, \vec{v}_{d_{best}}) //$  Output tuned settings for Algorithm 2: *GetTA\_Pair* for using them at all subsequent extractions (including extractions on test set)

#### Algorithm 2: GetTA\_Pair: Target and aspect extraction

**Input:** p: post,  $T_{ref}$ : target candidates in reference data,  $\vec{a}, \vec{v}$ : fine-tuned weight variables

**Output:**  $(t_{out}, a_{out})$ 

**Dependencies:** *GetTarget* // Algorithm 3;

 $C \leftarrow ExtractConcepts(p)$  // Apply concept extraction to p;

 $APM \leftarrow AdjectivalMod(p) \cup ParticipleMod(p)$  // Obtain the adjectival and participle modifiers in p;

 $t_{in} \leftarrow GetTarget(p, T_{ref}, C, APM, \vec{a})$  // Apply Algorithm 3: GetTarget;

 $t_{in} \leftarrow SelectIF(t_{in}, Expand(t_{in}), \vec{v})$  // Select  $t_{in}$  or  $t_{in}$  expanded to a complete group depending on  $\vec{v}$ ; if  $t_{in} \equiv modifier \in APM + concept$  then

 $t_{in} \leftarrow concept; a_{best} \leftarrow modifier //$ Select concept in  $t_{in}$  as updated target  $t_{in}$  and its modifier as  $a_{best}$ ; else

 $A_c \leftarrow \{c \mid \forall c_s : c_s \text{ IS } subword(c), c \in C \text{ OR } c \in APM \land \nexists t_s : t_s \text{ IS } subword(t_{in}) \land c_s = t_s\};$ // Identify concepts and modifiers in p which do not have common subwords with the extracted target  $t_{in};$  $A^* \leftarrow Order(Weight(A_c, \vec{v}));$ 

// Weight concepts and modifiers in p according to  $\vec{v}$  and order them in descending weight order  $a_{best} \leftarrow FirstElement(A^*)$  // the top-ranked aspect candidate;

 $t_{out} \leftarrow t_{in};$ 

 $a_{out} \leftarrow SelectIF(a_{best}, Expand(a_{best}), \vec{v}) //$ Output  $a_{best}$  or  $a_{best}$  expanded to a complete group depending on  $\vec{v}$ .

#### Algorithm 3: GetTarget: Target determination

**Input:** *p*: post,  $T_{ref}$ : target candidates in reference data, C: concepts in *p*, APM: adj/participle modifiers in *p*,  $\vec{a}$ : fine-tuned weight variables

**Output:**  $t_{out}$ : identified target

 $T_p \leftarrow \{t \mid t \in C \land t \in T_{ref}\}$  // Identify concepts in p already seen as target candidates in the reference data;  $T_c \leftarrow \{c \mid (c \in C \land \nexists t_p \in T_p : t_p = c)\}$  // Identify other concepts in p;

 $T_{sub} \leftarrow SubwordConcepts(C) \cup SubwordConcepts(APM);$ 

// Identify concepts in p which are subwords in nominal compounds or adjectival/participle modifiers;  $T_{overlap} \leftarrow \{c \mid c \in T_{sub} \land c \in T_{ref}\};$ 

// Collect subword concepts in p that overlap with the target candidates seen in the reference data;  $T_{disj} \leftarrow \{c \mid c \in T_{sub} \land c \notin T_{ref}\};$ 

// Collect subword concepts in p that do not overlap with the target candidates seen in the reference data;  $T_1^* \leftarrow Order(Weight(T_p \cup T_{overlap}, \vec{a}));$ 

// Weight concepts + subword concepts in p seen as target candidates in the reference data according to  $\vec{a}$  and order // them in descending weight order

 $T_2^* \leftarrow Order(Weight(T_c \cup T_{disj} \cup APM, \vec{a}));$ 

// Weight other concepts + subword concepts in p according to  $\vec{a}$  and order them in descending weight order  $T^* \leftarrow APPEND(T_1^*, T_2^*)$ ;

 $t_{best} \leftarrow FirstElement(T^*) //$  the top-ranked target candidate;

 $t_{out} \leftarrow SelectIF(t_{best}, Expand(t_{best}), \vec{a}) //$ Output  $t_{best}$  or  $t_{best}$  expanded to a complete group depending on  $\vec{a}$ .

Sexism	Racism				
women (143), girls (102), woman (43),	islam (97), muslims (89), mohammed (84),				
men (35), kat (33), feminists (23), peo-	isis (34), prophet (22), quran (20),				
ple (20), girl (14), andre (13), man (12), fe-	people (19), jews (15), muslim (14),				
males (11), nikki (8), guy (8), bitches (7),	religion (12), women (11), world (10),				
annie (6), feminism (5), bitch (5), produc-	hamas (10), salon (9), jesus (9), hadith (8),				
ers (4), football (4), female comedians (4),	woman (7), prophet mohammed (7),				
guys (4), gender (4)	men (6), christians (6)				

Table 2: Concepts with the highest TF over the reference set, which appear in the grammatical subject position in the reference set

should be explicitly mentioned in the text and not inferred) and then the (potentially multiple) aspects, keeping in mind that the target and the aspect can be the same. The annotation was carried out in several iterations. After each iteration, a consensus among the annotators with respect to the annotation of each post was reached, such that the annotated 540 posts can be considered a solid ground truth.<sup>9</sup>

#### 5.2 Experiments and Their Results

## 5.2.1 Domain adaptation

Our domain adaptation consists in applying Algorithms 1–3 to the reference dataset  $(S_{ref})$  of 5205 posts and the development set  $(S_{develop})$  of 100 posts from 'racism' and 'sexism' categories of (Waseem and Hovy, 2016a). Shvets and Wanner (2020)'s concept extraction detects in  $S_{ref}$  about 7K concepts in the 'sexism' subset (e.g., 'dinner', 'iq', 'wings', 'abortion', 'female commentator', 'women', 'girls', etc.), and about 4K concepts in the 'racism' subset (e.g., 'hypocrite', 'armies', 'death cult', 'countries', 'honor killings'). Already at the first glance, we reckon that not all of them can be targets in the sense defined in Section 2. This shows the importance of the proposed domain adaptation. The concepts with the highest tf in the  $S_{ref}$ (and thus the candidates to be targets) are shown in Table 2. Note that for the *tf* figures, we used only concepts from the  $S_{ref}$  that appear in the subject position in  $S_{ref}$ , as we observed that 94% of the targets in  $S_{develop}$  are subjects in  $S_{ref}$ . It is also worth noting that this list of generic targets provides only candidates that are further dynamically extended by other concepts for each new post, such that generic candidates may appear in a compound target or can even be dropped altogether.

The fine-tuning procedure of Algorithm 1 provides  $a_1 = 10^6 \gg a_2 = 10^3 \gg max(tf_{T_{ref}}) > min(tf_{T_{ref}}) \gg a_3 = 10^{-3} \gg a_4 = 10^{-6} \neq$   $0; a_5 = 0$ , and  $v_1 = 1, v_2 = 1; v_3 = 10^9 \gg v_4 = 10^6 \gg v_5 = 10^3 \gg Length(p)$  (p being the post under consideration). Thus, the importance of variables for target detection is the following: nominal target candidate > proper name target candidate > target candidate comprises entire words > position of the candidate in p. For aspects, this procedure results in: nominal concept candidate following target > adjectival/participle candidate > nominal concept candidate preceding the target.

## 5.2.2 Target and Aspect Extraction

After the adaptation, we identify the targets and aspects using the fine-tuned weight variables  $\vec{a}_{best}$ and  $\vec{v}_{best}$  (specified in Section 4.2) in the test set  $(S_{test})$  of 440 posts. Consider a few examples, with the identified targets and aspects marked in bold.

- (9) The Muslims (Target) conquered 2/3 of the Christian world (Aspect) before it attacked back. So again, what are you crying about.
- (10) There's something wrong when a girl (Target) wins Wayne Rooney street striker (Aspect) #NotSexist.
- (11) *Feminism* (Target) *is a snoring issue* (Aspect).
- (12) But why propagandize your bigotry when Pakistani Muslims (Target) are murdering Christians and Hindus for blasphemy (Aspect)?
- (13) Why haven't you stopped the sick Muslims (Target) from trying to exterminate Israel (Aspect)?
- (14) Kat (Target) is a sociopath (Aspect) #mkr

We can observe that the identified targets are nominal entities, while the aspects are mainly verbal groups that have been obtained by the expansion of an initial nominal aspect candidate (*Christian world, women*, etc.) to a full verbal group. However, as (11) and (14) show, we cannot reduce aspect identification to verbal group extraction: an aspect can readily be also a nominal group.

We evaluated the performance of the proposed model along with several baselines for target identification on  $S_{develop}$  (dev) and  $S_{test}$  (test) with respect to accuracy in terms of the Jaccard index, partial and exact match and with respect to precision, recall and F1 for ROUGE-L (Lin, 2004); cf. Table 3. The first baseline takes the first noun as a target. This baseline already provides many correct matches due to the reduced lengths of the posts in our dataset. The second baseline identifies a noun

<sup>&</sup>lt;sup>9</sup>This makes the calculation of the inter-annotator agreement obsolete; it will, obviously, become of relevance in the case of the annotation of larger datasets.

		Accuracy		ROUGE-		·L			
Algorithm	Jaccard	Partial	Exact	P	R	$F_1$			
	index	match	match						
Targets (dev)									
Baseline 1 - first noun as a 0.1 0.08 0.08 0.11 0.1 0.									
target									
Baseline 2 - noun with a hyper-	0.28	0.34	0.24	0.34	0.29	0.3			
nym "person"/"group"									
$GetTA\_Pair(T_{ref}, p, \vec{a}_{best}, \vec{v}_{best})$	0.68	0.79	0.65	0.74	0.7	0.7			
Targets (test)									
BERT - fine-tuned on the dev set	0.58	0.76	0.45	0.65	0.67	0.63			
$GetTA\_Pair(T_{ref}, p, \vec{a}_{best}, \vec{v}_{best})$	0.63	0.74	0.57	0.7	0.66	0.66			
$GetTA\_Pair(T_{ref}, p, \vec{a}_{best}, \vec{v}_{best})$	0.63	0.82	0.49	0.69	0.74	0.68			
+BERT									
Aspects (dev)									
$GetTA\_Pair(T_{ref}, p, \vec{a}_{best}, \vec{v}_{best})$	0.39	0.64	0.18	0.51	0.54	0.45			
Aspects (test)									
BERT - fine-tuned on the dev set	0.34	0.67	0.11	0.5	0.45	0.42			
$GetTA\_Pair(T_{ref}, p, \vec{a}_{best}, \vec{v}_{best})$	0.29	0.62	0.11	0.44	0.41	0.36			
$GetTA\_Pair(T_{ref}, p, \vec{a}_{best}, \vec{v}_{best})$	0.36	0.74	0.12	0.48	0.55	0.45			
+BERT									

Table 3: Evaluation of the quality of the detected targets and aspects on the development and test set

with a hypernym *person* or *group* that is a relevant candidate entity according to the definition of a target. We also fine-tuned a BERT model (Devlin et al., 2019) on the development set for target recognition in order to compare our pointer-generator-based model to transformer-based models.

We can observe that target identification as invoked by the GetTA\_Pair (Algorithm 2) achieves a rather good performance. Thus, the accuracy for the exact match between the ground truth targets and predicted targets is 0.65 for the development set and 0.57 for the test set. With BERT, we achieve somewhat lower accuracy. It is interesting to observe that combining GetTA\_Pair with BERT results in lower accuracy for the exact match, but in considerably higher accuracy (of 0.82) for a partial match, i.e., the match between the semantic head of the predicted target and the semantic head of the ground truth target. This is likely due to the limited amount of material in the development set, which seems to be sufficient to learn the essence of what an aspect is, but is not sufficient to learn well the composition of the aspect in terms of lexico-syntactic patterns.<sup>10</sup> The performance for aspect recognition is, in general, lower, which can be explained by the higher complexity of the task. However, for the partial aspect match, the accuracy is still 0.74, and the ROUGE-L F1 score is 0.45.

Table 4 shows the performance of our target detection algorithm with different variable settings. As can be observed, just the use of the  $tf^*idf$  feature

		Accuracy		ROUGE-L		
Algorithm setup	Jaccard	Partial	Exact	P	R	$F_1$
	index	match	match			
w/o learning targets with refer-	0.16	0.21	0.14	0.18	0.16	0.16
ence set and w/o tf*idf for nomi-						
nals and $\alpha_1=0$						
w/o learning targets with refer-	0.38	0.49	0.36	0.43	0.39	0.39
ence set and w/o tf*idf for adjec-						
tival/past participle groups						
w/o learning targets with refer-	0.38	0.49	0.36	0.44	0.39	0.4
ence set						
w/o using subject position in ref-	0.55	0.67	0.53	0.61	0.57	0.57
erence set for tf						
w/ target expanding ( $\alpha_5=1$ ) and	0.59	0.76	0.52	0.64	0.69	0.63
w/o tf and $\alpha_2=0$						
w/ target expanding ( $\alpha_5=1$ ) and	0.6	0.76	0.53	0.65	0.69	0.64
w/o tf and $\alpha_4=0$						
w/ target expanding ( $\alpha_5=1$ ) and	0.61	0.75	0.55	0.65	0.69	0.64
w/o tf and $\alpha_3=0$						
w/o all subwords	0.63	0.73	0.6	0.69	0.64	0.64
w/o nominal subwords	0.63	0.74	0.61	0.7	0.65	0.65
w/ target expanding ( $\alpha_5=1$ ) and	0.63	0.79	0.57	0.68	0.71	0.66
w/o tf						
w/ target expanding ( $\alpha_5=1$ )	0.63	0.79	0.56	0.68	0.73	0.67
$GetTA\_Pair(T_{ref}, p, \vec{a}_{best}, \vec{v}_{best})$	0.68	0.79	0.65	0.74	0.7	0.7

 Table 4: Evaluation of the quality of the detected targets during fine-tuning on the development set

already improves the performance considerably. When only concepts in the subject position are taken into account as target candidates, the Jaccard index improves significantly; the best performance is achieved when all variables are set as indicated in the description of the Algorithms 2 and 3.

In addition, we assessed the performance of the model when Algorithm 3 is applied successively several times, excluding targets predicted at previous steps from consideration. Similarly, for each detected target we ran several times Algorithm 2. The improvement in ROUGE-L score with each run is shown in Figure 2, when the best of the predicted top n targets and the best corresponding top n aspects are scored. Figures provided for aspects correspond to the second run of the Algorithm 3, but this does not distort the overall picture since they are at the same scale for any number of predicted targets. We can observe a steady increase in performance already for small values of n, which shows the potential of our model. This strategy of selecting top n targets can also be used for detecting multiple targets in longer texts.

To verify that the proposed fine-tuning procedure of the weight variables is not dataset-specific, we ran it also on the negative sentiment subset of (Dong et al., 2014) as  $T_{ref}$ , with the targets originally obtained through dictionary search as test set targets.<sup>11</sup> To avoid a bias in the evaluation by "seen" targets, we ensured that 50% of the targets in

<sup>&</sup>lt;sup>10</sup>Pretraining BERT on concept annotated datasets may improve the figures for the exact match. If this proves to be the case, transformer-based models are likely to outperform other models on the overall target identification task.

<sup>&</sup>lt;sup>11</sup>Recall that no aspects in our sense are annotated in this sentiment dataset.

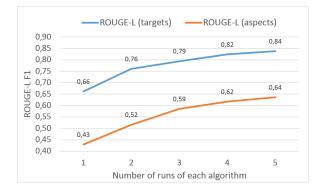


Figure 2: Mean ROUGE values over the test set for different number of algorithm runs

		Accuracy		R	-L	
Part of the test set	Jaccard	Partial	Exact	P	R	$F_1$
	index	match	match			
Only posts with targets from $T_{ref}$	0.88	0.92	0.85	0.88	0.91	0.89
Only posts with unseen targets	0.53	0.57	0.51	0.54	0.55	0.54
All posts	0.73	0.77	0.71	0.73	0.76	0.74

Table 5: Evaluation of target detection on Dong et al.(2014)'s negative sentiment sub-dataset

the test set are unseen by removing a number of examples with targets appearing in both the reference set and the test set from the reference set. Table 5 shows the scores obtained in this experiment for targets. We can observe that the evaluation figures are even considerably higher than those in Tables 3 and 4. This is likely because of the high percentage of named entities in this dataset, which facilitates an accurate detection of concepts.

# 6 Conclusions and Future Work

Classification of hate speech in terms of broad categories is not sufficient; in order to effectively combat hate speech, a detailed target–aspect analysis is necessary. We presented a model that adapts a generic concept extraction model and showed that it is able to reach a reasonable quality for target and aspect identification in the 'sexism' and 'racism' categories of the (Waseem and Hovy, 2016a) hate speech dataset. The model is semi-supervised and works already with a small annotated dataset. This is an advantage in view of the absence of large hate speech datasets annotated with the target–aspect information.

Despite the promising figures, our model still has some limitations. Thus, aspect identification quality should be further improved. Furthermore, we plan to use distance learning in order to make the model language-independent, which will be an advantage compared to the presented implementation, which is to a certain extent language-specific. In addition, experiments on other hate speech datasets should be carried out in order to demonstrate that the proposed variable tuning and implemented syntactic target and aspect patterns generalize well across datasets. Finally, although the vast majority of posts indeed contains just one target, to capture multiple targets would be desirable.

The annotated development and test sets and the code are available in the following GitHub repository: https://github.com/ TalnUPF/HateSpeechTargetsAspects/.

#### Acknowledgements

We would like to thank the three anonymous reviewers for their insightful comments, which helped to improve the final version of the paper considerably.

Paula Fortuna is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of the Operational Program *Human Capital* (POCH), supported by the European Social Fund and by national funds from MCTES. The work of Alexander Shvets, Juan Soler-Company, and Leo Wanner has been supported by the European Commission in the context of the H2020 Research Program under the contract numbers 700024, 786731, and 825079.

#### References

- Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pages 233–238. IEEE.
- Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 2819–2829. Association for Computational Linguistics.

- Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017a. Automated hate speech detection and the problem of offensive language. In Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017, pages 512–515. AAAI Press.
- Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017b. Automated hate speech detection and the problem of offensive language. In Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017, pages 512–515. AAAI Press.
- 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.
- Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers, pages 49–54. The Association for Computer Linguistics.
- Maeve Duggan. 2014. Online harassment. Technical report, Pew Research Center, Washington, USA. Available at https://radimrehurek. com/gensim/summarization/keywords.html.
- Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA* @ *CLiC-it*.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. ACM Computer Surveys, 51(4):85:1–85:30.
- Paula Fortuna, Juan Soler Company, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pages 6786–6794. European Language Resources Association.
- Akash Kumar Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. #metooma: Multi-aspect annotations of tweets related to the metoo movement. In Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020, pages 209–216. AAAI Press.

- Zhen Hai, Kuiyu Chang, and Jung-jae Kim. 2011. Implicit feature identification via co-occurrence association rule mining. In Computational Linguistics and Intelligent Text Processing 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I, volume 6608 of Lecture Notes in Computer Science, pages 393–404. Springer.
- M.A.K. Halliday. 2013. *Halliday's Introduction to Functional Grammar*. Routledge, London & New York.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004, pages 168–177. ACM.
- Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5435–5442, Online. Association for Computational Linguistics.
- Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1065–1074.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Ning Liu and Bo Shen. 2020. Aspect-based sentiment analysis with gated alternate neural network. *Knowl. Based Syst.*, 188.
- T. Luong, H. Pham, and C.D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of the EMNLP*, pages 1412– 1421.
- Dehong Ma, Sujian Li, and Houfeng Wang. 2018. Joint learning for targeted sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4737–4742.
- Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019, pages 369–380. AAAI Press.

- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *CoRR*, abs/2012.10289.
- Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.
- A. Nazir, Y. Rao, L. Wu, and L. Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing; doi:* 10.1109/TAFFC.2020.2970399.
- N. Nikolić, O. Grljević, and A. Kovačević. 2020. Aspect-based sentiment analysis of reviews in the domain of higher education. *The Electronic Library*, 38(1):44–64.
- Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4675– 4684, Hong Kong, China. Association for Computational Linguistics.
- Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition ldc2011t07, 2011. URL https://catalog. ldc. upenn. edu/LDC2011T07.[Online].
- Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118.
- Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10,

2020, pages 5477–5490. Association for Computational Linguistics.

- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Trans. Knowl. Data Eng.*, 28(3):813–830.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715– 1725, Berlin, Germany. Association for Computational Linguistics.
- Alexander V. Shvets and Leo Wanner. 2020. Concept extraction using pointer-generator networks and distant supervision for data augmentation. In *Knowledge Engineering and Knowledge Management - 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16-20, 2020, Proceedings,* volume 12387 of *Lecture Notes in Computer Science,* pages 120–135. Springer.
- Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016, pages 687–690. AAAI Press.
- Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.
- Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, pages 3298– 3307. ACL.
- Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1177– 1190, Online. Association for Computational Linguistics.
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

- A. Waldis, L. Mazzola, and M.A. Kaufmann. 2018. Concept extraction with convolutional neural networks. In *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)*, Porto, Portugal.
- Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspectlevel sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pages 606–615. The Association for Computational Linguistics.
- Zeerak Waseem and Dirk Hovy. 2016a. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Zeerak Waseem and Dirk Hovy. 2016b. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 88–93. The Association for Computational Linguistics.
- Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2017. Twitter hate aspect extraction using association analysis and dictionary-based approach. In New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 16th International Conference, SoMeT\_17, Kitakyushu City, Japan, September 26-28, 2017, volume 297 of Frontiers in Artificial Intelligence and Applications, pages 641–651. IOS Press.
- Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2018. Evaluating aspect-based sentiment classification on twitter hate speech using neural networks and word embedding features. In New Trends in Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 17th International Conference SoMeT\_18, Granada, Spain, 26-28 September 2018, volume 303 of Frontiers in Artificial Intelligence and Applications, pages 723–734. IOS Press.
- Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2019. Hate crime on twitter: Aspect-based sentiment analysis approach. In Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques - Proceedings of the 18th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT\_19), Kuching, Malaysia, 23-25 September 2019, volume 318 of Frontiers in Artificial Intelligence and Applications, pages 284–297. IOS Press.