Lang Resources & Evaluation (2023) 57:155–188
https://doi.org/10.1007/s10579-022-09582-8

ORIGINAL PAPER

# Investigating the role of swear words in abusive language detection tasks

**Endang Wahyu Pamungkas[1,2] · Valerio Basile[1] · Viviana Patti[1]**

Accepted: 11 January 2022 / Published online: 17 February 2022

**Abstract** Swearing plays an ubiquitous role in everyday conversations among humans, both in oral and textual communication, and occurs frequently in social media texts, typically featured by informal language and spontaneous writing. Such occurrences can be linked to an abusive context, when they contribute to the expression of hatred and to the abusive effect, causing harm and offense. However, swearing is multifaceted and is often used in casual contexts, also with positive social functions. In this study, we explore the phenomenon of swearing in Twitter conversations, by automatically predicting the *abusiveness* of a swear word in a tweet as the main investigation perspective. We developed the Twitter English corpus SWAD (Swear Words Abusiveness Dataset), where abusive swearing is manually annotated at the word level. Our collection consists of 2577 instances in total from two phases of manual annotation. We developed models to automatically predict abusive swearing, to provide an intrinsic evaluation of SWAD and confirm the robustness of the resource. We model this prediction task as three different tasks, namely sequence labeling, text classification, and target-based swear word abusiveness prediction. We experimentally found that our intention to model the task similarly to aspect-based sentiment analysis leads to promising results. Subsequently, we employ the classifier to improve the prediction of abusive language in

✉ Endang Wahyu Pamungkas
pamungka@di.unito.it; ewp123@ums.ac.id

Valerio Basile
valerio.basile@unito.it

Viviana Patti
viviana.patti@unito.it

[1] Dipartimento di Informatica, University of Turin, Turin, Italy

[2] Department of Informatics Engineering, Universitas Muhammadiyah Surakarta, Surakarta, Central Java 57162, Indonesia

several standard benchmarks. The results of our experiments show that additional abusiveness feature of the swear words is able to improve the performance of abusive language detection models in several benchmark datasets.

**Keywords** Abusive language detection · Content moderation · Hate speech detection · Social media · Swear words abusiveness

## 1 Introduction

Swearing is the use of taboo language (also referred to as bad language, swear words, offensive language, curse words, or vulgar words) to express the speaker's emotional state to their listeners (Jay, 1992, 1999). Not limited to face to face conversation, swearing also occurs in online conversations, across different languages, including social media and online forums, such as Twitter, typically featured by informal language and spontaneous writing. Twitter is considered a particularly interesting data source for investigations related to swearing. According to the study in Wang et al. (2014) the rate of swear word use in English Twitter is 1.15%, almost double compared to its use in daily conversation (0.5–0.7%) as observed in previous work (Jay, 1992; Mehl & Pennebaker, 2003). The work by Wang et al. (2014) also reports that a portion of 7.73% tweets in their random sampling collection is containing swear words, which means that one tweet out of thirteen includes at least one swear word. Interestingly, they also observed that a list of only seven words covers about 90% of all the swear words occurrences in their Twitter sample: *f*ck*, *sh*t*, **ss*, *b*tch*, *n*gga*, *h*ll*, and *wh*re*.

Swearing in social media can be linked to an abusive context, when it is intended to offend, intimidate or cause emotional or psychological harm, contributing to the expression of hatred, in its various forms. In such contexts, indeed, swear words are often used to insult, such as in case of sexual harassment, hate speech, obscene telephone calls (OTCs), and verbal abuse (Jay et al., 2006; Jay & Janschewitz, 2008). However, swearing is a multifaceted phenomenon. The use of swear words does not always result in harm, and the harm depends on the context where the swear word occurs (Jay, 2009a). Consider for instance the two following tweets containing swearing from the *StackOverflow Offensive Comments* dataset (Fišer et al., 2018):

> If you don't have the answer, move on to the next **f*cking** question and mind your own **f*cking** business
>
> Sh_Khan: **f*cking** genius. Thank you

In the first example, it is obvious that the swear word is used to insult, thus this is an instance of abusive language. However, the second example shows the use of the same swear word in a casual setting, to emphasize an emotion of gratitude without intention to be offensive (Pinker 2007, *emphatic swearing*).

Some studies even found that the use of swear words has also several upsides. Using swear words in communication with friends could promote some advantageous social effects, including strengthen the social bonds and improve conversation harmony, when swear word is used in ironic or sarcastic contexts (Jay, 2009a). Another study by Stephens and Umland (2011) found that swearing in cathartic ways is able to increase pain tolerance. Furthermore, Johnson (2012) has shown that the use of swear words can improve the effectiveness and persuasiveness of a message, especially when used to express an emotion of positive surprise. Also accounts of appropriated uses of slurs should not be neglected (Bianchi, 2014), that is those uses by targeted groups of their own slurs for non-derogatory purposes (e.g., the appropriation of ‘nigger’ by the African–American community, or the appropriation of ‘queer’ by the homosexual community).

In recent years, more and more studies focused on abusive language detection which covers hate speech, cyberbullying, trolling, and offensive language (Waseem et al., 2017; Schmidt & Wiegand, 2017; Michal et al., 2010). Swear words play an important role in these tasks, providing a signal to spot an offensive utterance (Malmasi & Zampieri, 2018). However, the presence of swear words could also lead to false positives when they occur in a casual context (Chen et al., 2012; Nobata et al., 2016; Van Hee et al., 2018; Malmasi & Zampieri, 2018). Distinguishing between abusive and not-abusive swearing contexts seems to be crucial to support and implement better content moderation practices. Indeed, on the one hand, there is a considerable urgency for most popular social media, such as Twitter and Facebook, to develop robust approaches for abusive language detection, also for guaranteeing a better compliance to governments demands for counteracting the phenomenon (see, e.g., the recently issued EU commission *Code of Conduct on countering illegal hate speech online* (EU Commission, 2016)). On the other hand, as reflected in statements from the Twitter Safety and Security[1] users should be allowed to post potentially inflammatory content, as long as they are not-abusive.[2] The idea is that, as long as swear words are used but do not contain abuse/harassment, hateful conduct, sensitive content, and so on, they should not be censored.

In this work, we conduct an in-depth investigation on the role of swear words and their context in abusive language detection tasks. We explore the phenomenon of swearing in online conversation, taking the possibility of predicting the abusiveness of a swear word in a tweet context as the main investigation perspective. In this direction, the main goal is to automatically differentiate between abusive swearing, which should be regulated and countered in online communication, and not abusive one, that should be allowed as part of freedom of speech, also recognizing its positive functions, as in the case of reclaimed uses of slurs. To achieve this objective, we conduct several contributions. First, we develop a new benchmark Twitter corpus, called SWAD (Swear Words Abusiveness Dataset), where abusive swearing is manually annotated at the word level. Based on several previous studies (Jay, 2009a; Dinakar et al., 2011; Golbeck et al., 2017), we define abusive swearing as *the use of swear word or profanity in several cases such as name-calling,*

[1] https://help.twitter.com/en/safety-and-security/offensive-tweets-and-content.

[2] See for instance the Twitter Rules trying to determining what an abusive and hateful conduct is: https://help.twitter.com/en/rules-and-policies/twitter-rules.

*harassment, hate speech, and bullying involving several sensitive topics including physical appearance, sexuality, race & culture, and intelligence, with intention from the author to insult or abuse a target (person or group)*. The other uses, such as reclaimed uses, catharsis, humor, or conversational uses, are considered as not-abusive swearing. Second, we develop and experiment with supervised models to automatically predicting abusive swearing within the tweet context. Such models are trained on the novel SWAD corpus to predict the *abusiveness* of a swear word within a tweet. Finally, we investigate the impact of swear word abusiveness on downstream abusive language detection tasks.

In this paper, we address the following research questions.

- **RQ1** How to model the swear word context in social media text as either abusive or not abusive? The abusiveness of a swear word strongly depends on its context. Therefore, we propose to explore the possibility of building a novel corpus that consists of tweets where swear words are annotated at the word level as either abusive or not abusive based on their use within their context.
- **RQ2** Is it possible to automatically predict the abusiveness of a swear word within the tweet context? To answer this question, we propose three different tasks, namely sequence labeling, text classification, and target-based swear word abusiveness prediction.
- **RQ3** Is the additional information about swear words abusiveness helpful for detecting abusive language? As part of the extrinsic evaluation of our corpus, we explore the impact of swear word context prediction in the downstream task of abusive language detection. We do so by infusing the swear word context prediction as an additional feature to the baseline models.

The contribution of this paper can be summarized as following:

1. We propose a novel corpus which focuses on studying the swear words context as either abusive or not abusive.
2. We propose a new task to predict the abusiveness of swear words within tweet context, taking some inspiration from the target-based sentiment analysis task.
3. We develop a novel architecture to predict the abusiveness of swear words within their context by adopting a similar idea to the previous study.
4. We leverage the swear word abusiveness feature to improve the baseline model in several downstream abusive language detection tasks.

This study is extended version of our previous work on predicting abusive swearing in social media (Pamungkas et al., 2020a), by providing an extensive literature study, corpus extension, and additional experiment to get a better insight for investigating the role of swear word context in abusive language detection tasks.

The paper is organized as follows. Section 2 introduces related work on swear word use and its context in online conversation. In addition, we also review some studies which investigate the relation between swear word use and abusive language detection task. Section 3 reports on the various steps of development of the SWAD Twitter corpus. Section 4 presents the experimental setting of predicting abusiveness of swear words and discusses the result. Then, Sect. 5 presents our experiment in investigating the impact of swear word abusiveness feature in several abusive language detection tasks. Finally, Sect. 6 includes conclusive remarks and ideas for future work.

## 2 Related works

### 2.1 Swearing in online content

Wang et al. (2014) examines the cursing activity on the social media platform Twitter.[3] They explore several research questions including the ubiquity, utility, and also contextual dependency of textual swearing in Twitter. On the same platform, Bak et al. (2012) found that swearing is used frequently between people who have a stronger social relationship, as a part of their study on self-disclosure in Twitter conversation. Furthermore, Gauthier et al. (2015) provide an analysis of swearing on Twitter from several sociolinguistic aspects including age and gender. This study presents a deep exploration of the way British men and women use swear words. A gender- and age-based study of swearing was also conducted by Thelwall (2008), using the social network MySpace[4] to build their corpus. Recently, Cachola et al. (2018) studied vulgar words use in Twitter, by analyzing socio-cultural and pragmatic aspects of vulgarity based on users demographic data. Furthermore, they explored the impact of vulgar words use to the sentiment analysis task, which found that explicitly modeling vulgar words can boost sentiment analysis performance.

Besides social media, the study of swearing is also carried out on online communities. The study by Sood et al. (2012) focused on the use of profanity in an online community called Yahoo! Buzz[5]. They explored several research questions including what are the pitfalls of current profanity detection systems, how profanity differs between different communities, and how different communities receive the swearing in various contexts. Recently, Rojas-Galeano (2017) aimed at tackling the difficulties in detecting obfuscated obscenities on Spanish and Portuguese online news sites. Kwon and Gruzd (2017) studied the contagious diffusion of offensive comments in the Donald Trump's campaign video on Youtube[6]. They examined two kinds of swearing including: public swearing (when swearing has no specific target) and interpersonal swearing (the use of taboo words with a specific target).

### 2.2 Contextual swearing

Swearing is not always abusive—its abusiveness is context-dependent. Swearing context is explored by several prior studies. Fägersten (2012) classifies swearing context into two types, following the dichotomy introduced by Ross (1969):

[3] https://www.twitter.com.

[4] https://www.myspace.com.

[5] A social news commenting site that is no longer active.

[6] https://www.youtube.com.

*annoyance* swearing, "occurring in situations of increased stress", where the use of swear words appears to be "a manifestation of a release of tension", and *social* swearing, "occurring in situations of low stress and intended as a solidarity builder", which is related to a use of swear words in settings that are socially relaxed. Likewise, Allan and Burridge (2006) distinguishes the swearing contexts into *casual* context (when swear words do not cause insult, but are rather cathartic and humorous) and *abusive* context (when swear words are used with an intention to attack or insult).

The work by Jay (2009b) found that the offensiveness of taboo words is very dependent on their context, and postulates the use of taboo words in conversational context (less offensive) and hostile context (very offensive). These findings support prior work by Rieber et al. (1979) who showed that obscenities and swear words used in a *denotative* way are far more offensive than those used in a *connotative* way. Furthermore, Pinker (2007) classified the use of swear words into five categories based on why people swear: *dysphemistic*, exact opposite of euphemistic; *abusive*, using taboo words to abuse or insult someone; *idiomatic*, using taboo words to arouse the interest of listeners without really referring to the matter; *emphatic*, to emphasize another word; *cathartic*, the use of swear words as a response to stress or pain.

### 2.3 The role of swearing in abusive content

In recent years, abusive language detection is gaining interest from the research community. Swear words play a key role in this task, according to several works in the literature. Razavi et al. (2010) developed an automatic system for discriminating between *regular texts* and *flames*. They built a dictionary for this specific purpose called *Insulting and Abusing Language Dictionary* (IALD), which contains words, phrases, and expressions with several degrees of abuse and insult. Several swear words can be found among IALD entries, which are used as features in the automatic classification. Similarly, Chen et al. (2012) built a dictionary containing pejoratives, obscenities and profanities extracted from Urban Dictionary.[7] By combining both lexical features from their dictionary and syntactic features from dependency relations, their models were able to achieve high precision and recall in detecting both offensive content and offensive users. Mubarak et al. (2017) built a list of Arabic obscene words and hashtags by extracting patterns that are frequently used in offensive Twitter posts. This wordlist is used to classify a tweet into three classes: *obscene*, *offensive*, and *clean*. Recent studies also found that swear words are relevant to several related tasks including abusive language detection (Nobata et al., 2016), cyberbullying detection (Van Hee et al., 2018; Michal et al., 2010), and hate speech detection (Malmasi & Zampieri, 2018). The most recent study by Holgate et al. (2018) introduced six vulgar word use functions, and built a novel dataset based on them. They filtered their dataset based on presence of swear words from a list taken from the *noswearing* website.[8] Their results show that classifying

[7] https://www.urbandictionary.com/.

[8] http://www.noswearing.com.

vulgar word use by its function improves the system performance in detecting hate speech content.

### 2.4 Swear words corpora

The development of the swear word usage corpus was started by Holgate et al. (2018). They proposed a novel corpus, consisting of tweets containing swear words, where every swear word is annotated by six different labels based on its function. These vulgar function are including “express aggression”, “express emotion”, “emphasize”, “auxiliary”, “signal group identity”, and “non-vulgar”. The annotation process was done by using the crowd-sourced scenario. Furthermore, they built a model based on logistic regression coupled with several handcrafted features to classify the vulgar words function automatically. Pamungkas et al. (2020a) also introduced SWAD (Swear Words Abusiveness Dataset) corpus by filtering tweets from the OLID dataset (Zampieri et al., 2019a) based on swear word presence and annotating them with a binary label including “abusive” and “not-abusive”. They conducted the intrinsic evaluation of SWAD by predicting swear words’ abusiveness within a tweet as a context in two different models of prediction task, including sequence labeling task and text classification. Recently, Kurrek et al. (2020) also proposed a novel corpus that captures the online slur usage. The corpus consists of 39.8k human-annotated comments gathered from Reddit.[9] The annotation guideline outlines four main categories of online slur usage, divided into 12 sub-categories.

In this work, we follow a similar line as Holgate et al. (2018), which tries to model the pragmatic use of swear words to improve hate speech detection task. However, their work focuses on classifying the swear word used by its function and using it as an additional feature to detect hate speech utterances. In this work, we focus instead on the abusiveness prediction of swear words, rather than their function, to discover the context of a given swear word, whether abusive (should be eventually considered for content moderation, as it hurts) or not-abusive. Furthermore, we also adopt a similar task setting as target-based sentiment analysis to focus only on classifying the swear word’s context at the word-level. We also test the additional feature of swear word abusiveness information into four downstream hate speech detection tasks.

## 3 Corpus creation and analysis

### 3.1 Corpus collection

Our starting point was a corpus of tweets selected from the training set of the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a), which

[9] https://www.reddit.com/.

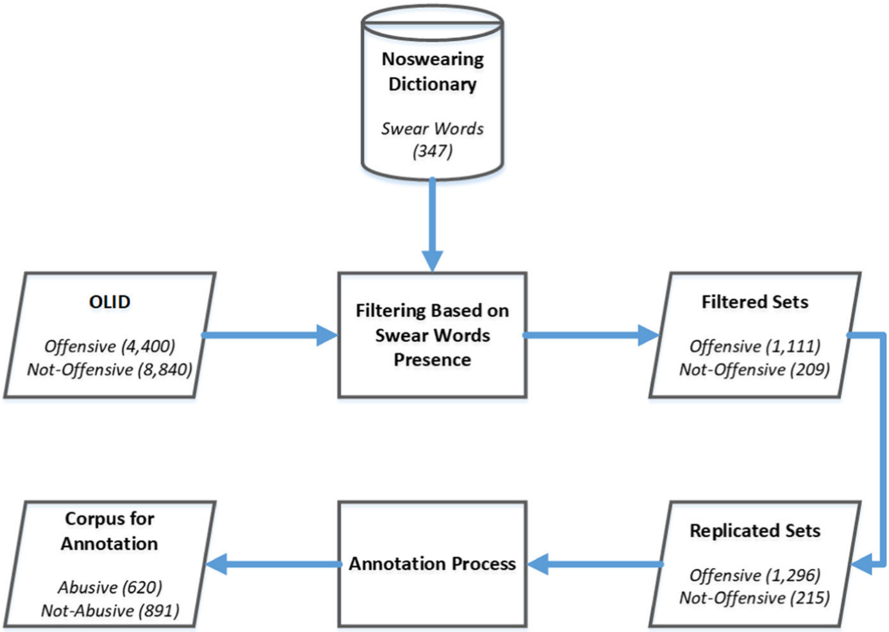

**Fig. 1** Corpus development process

**Table 1** Corpus statistic after filtering process

| | Original | After Filtering | After Replication |
|---|---|---|---|
| Offensive | 4400 | 1111 | 1296 |
| Not | 8840 | 209 | 215 |
| Total | 13,240 | 1381 | 1511 |

was proposed in the context of the shared task OffensEval (Zampieri et al., 2019b) at SemEval 2019.[10] This task is aimed at detecting offensive messages as well as their targets. In OLID, Twitter messages were labelled by applying a multi-layer hierarchical annotation scheme, which encompasses three dimensions, including tags for marking the presence of offensive language (*offensive* vs *not offensive*), tags for categorizing the offensive language (*targeted* vs *untargeted*), and tags for the offensive target identification (*individual*, *group*, or *other*). The broader coverage of the concept and definition of offensive language are the main reasons we choose this dataset as starting point for our finer grained annotation concerning swearing, rather than other datasets developed around more specific typologies of offensive language, such as hate speech, cyberbullying or misogyny, which we think could introduce a bias in our corpus, undermining the generality of its possible future exploitation.

[10] https://sites.google.com/site/offensevalsharedtask/offenseval2019.

Some preprocessing has been applied to the OLID data, such as mention and URL normalization. Since our focus is on analyzing swear words in the tweet context, we first filtered out a subset of tweets from OLID based on the presence of swear words, in order to obtain a collection of tweets that include at least one swear word. At this stage we exploited the list of swear words published on the noswearing website,[11] an online dictionary site which includes a list of swear words. This dictionary includes 349 swear words covering general vulgarities, slurs, and sex-related terms. We manually checked the list to exclude highly ambiguous words, namely swear words like "ho" and "hard on".[12] Table 1 shows the full statistics of our corpus after the filtering process. We identified 1,320 tweets that contain at least one swear word. Since this annotation task is at the (swear) word level, tweets which have more than one swear word were replicated. We generated as many new instances of the same tweet as the number of swear words occurring in the message, and marked each single swear word with special tags $<b>$ and $</b>$ (e.g. $<b>$f*ck$</b>$, $<b>$sh*t$</b>$, and etc.) so that the abusiveness label on each instance records the context of the marked swear word in the tweet (abusive or not). For instance, given the message *@USER This sh*t gon keep me in the crib lol f*ck it*, two instances will be generated: *@USER This* $<b>$ *sh*t* $</b>$ *gon keep me in the crib lol f*ck it* and *@USER This sh*t gon keep me in the crib lol* $<b>$*f*ck* $</b>$ *it.*

We found 154 tweets having more than one swear word, with a range of occurrences from 2 to 6 swear words. As a result, we have 1511 instances to be annotated. Figure 1 shows the overall process of our corpus development.

## 3.2 Annotation task and process

The annotation of 1511 instances involved three expert annotators (the authors), with different gender and ages. All instances were annotated by two independent annotators (A1 and A2). The resulting disagreement was resolved by involving a third annotator (A3), labeling those instances where a disagreement between A1 and A2 was detected. All annotators use English as a second language, with a minimum level of B2. The annotators involved conducted the process with particular care by adopting a cautious attitude, carefully discussing the disagreement, and consulting native speakers when in doubt.

### *3.2.1 Annotation task*

Annotators were asked to annotate (with a binary option) whether the highlighted swear word (tagged with the $<b>$ and $</b>$ tags) can be considered *abusive swearing*, contributing to the construction of an abusive context (by using the tag "yes") or whether the swear word does not contribute to the construction of an

[11] https://www.noswearing.com/.

[12] In the noswearing site "ho" is a short form of "hoe", but in the dataset we found that word "ho" is mostly used as a short form of "how". Similarly, "hard on" is a slang word of "erection" in the noswearing site, but this word is frequently used to express hard effort, as in "...I'm working *hard on* this task right now,..."

abusive context (by using the tag “no”). We first started a trial annotation on a portion of 100 tweets from the collection, to test our annotation guidelines and improve the understanding between annotators. During this trial annotation we also deepened our understanding of the *offensiveness* notion, which underlies the definition of offensive language driving the whole OLID annotation process. There is a crucial difference between the coarse notion of offensive language as defined in OLID and the concept of abusive language we are interested in, given our main goal to reason about abusive swearing. Indeed, according to the OLID definition a tweet can be considered offensive only because of the presence of profanities, even if no occurrence of abusive swearing can be detected.

Such considerations have driven our decision to annotate the abusiveness of swear words on tweets belonging to both classes (offensive and not-offensive) of the OLID data. Another issue discovered during the trial annotation consisted in some cases where the swear word is used for indirect insult: the swear word itself is used to insult, but the overall context of the tweet is not abusive. This mostly happened in the reported speech such as in the Example 3.1 below, where we determined this tweet as not abusive:

> [Example 3.1. Indirect insult.] *@USER Everyone saying* ***f*ck*** *Russ dont know a damn thing about him or watched the interview*🤷‍♀️🤷‍♀️🤷‍♀️

Therefore, in the final annotation guidelines, we decided to include the author *intention* to resolve the swear word context, especially to deal with this kind of swear word use. We consider abusive swearing those uses where *swearing contributes to the construction of an abusive context such as name-calling, harassment, hate speech, and bullying, involving several sensitive topics including physical appearance, sexuality, race and culture, and intelligence, with intention from the author of tweet to insult or abuse a target (person or group of persons).* Let us notice that one tweet can have more than one swear word, but for every tweet, only one swear word will be highlighted as relevant for the annotation in each row (see the replication process explained above). Therefore, the annotator only needs to focus on the marked swear words (e.g., $<b>$ f*ck$</b>$ ). We remark again that abusive swearing can be found on both offensive and not-offensive tweets, therefore during the application of our annotation layer, we decided to ignore the original message-level layer of annotation from the original OLID (offensive vs not-offensive), in order to avoid confusing the annotators during the annotation process. Indeed, we observed four possible cases, when we consider the OLID original labels on the offensiveness of a tweet, namely: (i) the message is offensive and the swear word is abusive, (ii) the message is offensive but the swear word is not abusive, (iii) the message is not offensive but the swear word is abusive, and (iv) the message is not offensive and the swear word is not abusive. Let us provide an example for each case to get a better understanding on such circumstances:

> [Example 3.2. Offensive tweet & abusive swearing] *@USER You are an absolute* ***d*ck***😡

> [Example 3.3. Offensive tweet & not abusive swearing] *@USER I was definitely drunk as* ***sh*t***

[Example 3.4. Not offensive tweet & abusive swearing] *@USER* ***b*llshit*** *there's rich liberals too so what are you saying ???*

[Example 3.5. Not offensive tweet & not abusive swearing] *@USER Haley thanx! you know how to brighten up my* ***sh*tty*** *day*😘

### 3.3 Annotation results and disagreement analysis

Referring to the application of two independent annotations on the whole dataset of tweets (A1 and A2), we can say that annotators achieved a good agreement, selecting the same value in a large portion of the annotated tweets being only 216 out of 1511 the messages where they disagreed by marking in a different way the presence of abusive swearing. The average pairwise agreement percentage amounts to 85.70%. The inter-annotator agreement is 0.652 (Cohen's kappa coefficient), which corresponds to a substantial agreement. The final SWAD annotated corpus consists of 1511 unique swear words immersed in the context of 1320 tweets, where 620 swear words are marked as abusive and 891 are rated as not-abusive.[13] Table 2 shows the detailed distribution of our annotation result. Interestingly, we found more not-abusive swearing than abusive ones in tweets belonging to the offensive class of OLID (728 versus 568). In addition, we also found 52 cases of abusive swearing in tweets belonging to the OLID not-offensive class.

In the following we list and share some interesting findings and elements of discussion related to the annotation task and outcome.

#### *3.3.1 Most of the non-abusive contexts of swearing are dominated by emphatic and cathartic swearing function*

Cathartic swearing is a swear word function when it is used as a response to pain or misfortune (see Example 3.6), while emphatic swearing is another swear word function when a swear word is used to emphasize another word in order to draw more attention (see Example 3.7).

[Example 3.6. Cathartic function] *@USER* ***d*mn*** *I felt this shit Why you so loud lol*

[Example 3.7. Emphatic function] *@USER I AM F*CKING SO* ***F*CKING*** *HAPPY*

#### *3.3.2 Emojis could become an important signal to resolve the context of a swear word within the tweet*

In some tweets when the context of swear word use is difficult to be resolved, the presence of emojis could give key information. As shown in Example 3.8, without the presence of the emoji, the swear word *fucking* seems to contribute to the

[13] The corpus is available for research purpose at the following URL: https://github.com/dadangewp/SWAD-Repository.

**Table 2** Label distribution in the SWAD dataset

| | Original OLID | Abusive | Not-abusive |
|---|---|---|---|
| Offensive | 1296 | 568 | 728 |
| Not | 215 | 52 | 163 |
| Total | 1511 | 620 | 891 |

construction of an abusive context, but the presence of the *Face with Tears of Joy* emoji helped annotators to understand the real context of the whole tweet.

> [Example 3.8. Use of emojis] *@USER ur a* ***f*cking*** *dumbass fr. there's no way she is anyone else's*😂

### *3.3.3 Irony and sarcasm could provide an issue for automatic prediction based on machine learning approach*

We found some tweets which contain sarcasm and irony, most of the times in not-abusive context. As in other related tasks such as sentiment analysis, irony and sarcasm could contribute to the difficulties of this task. An example of tweet where these phenomena are expressed can be seen in Example 3.9.

> [Example 3.9. Irony and sarcasm issues] *@USER Yeah we need some more made up* ***b*llshit*** *protestors and antifa lol time for an epic beatdown*😉

Furthermore, we analyzed cases of disagreement between annotators. We conducted a manual analysis of 216 disagreement cases with the aim to extract the most common patterns, which contribute to the difficulty of the annotation task. As a result, we found several difficult cases:

### *3.3.4 Missing context*

We found that some tweets are very short, resulting in the context missing (see Example 3.10). Other instances are also challenging to understand due to the presence of grammatical errors (see Example 3.11). These issues are very dominant in the annotator disagreement cases.

> [Example 3.10. Very short tweet] *@USER Lmfaoo!*😭

> [Example 3.11. Noisy text with grammatical errors] *@USER* ***d*mn*** *that headgear is lit sucks im not on pc ubi plz for console to*

### *3.3.5 Need of world knowledge to understand the context*

Some tweets are also very difficult to understand due to the lack of world knowledge, as shown in Example 3.12. Sometimes annotators need to gather more information by using search engine to understand the context. The presence of hashtags usually becomes the key to understand the nature of the context.

**Table 3** Interaction between the original Holgate's label with our annotation

| | AGG | EMO | EMPH | AUX | SGI | NV |
|---|---|---|---|---|---|---|
| Abusive | 66 | 62 | 21 | 33 | 18 | 5 |
| Not abusive | 61 | 253 | 142 | 230 | 59 | 50 |

> [Example 3.12. Difficult to understand] *@USER @USER It's probably better to have an❌❌next to my name than a pink* ***p*ssy*** *hat on my head🤔🤔🤔🤔 #MAGA #MakeAmericaGreatAgain*

## 3.4 Corpus extension

After completing the full annotation process, SWAD consists of 1511 instances. We realize that this collection is still relatively small to obtain reliable performance for machine learning models. Therefore, we extended SWAD by conducting another round of annotation. We also included in the collection the test set of the OLID dataset, which contains 860 tweets, and we re-annotated tweets from Holgate's dataset (Holgate et al., 2018) according to the SWAD guidelines. Similar to SWAD, tweets in Holgate's dataset were filtered based on the presence of vulgar words. Then, all instances of vulgar words were annotated with one of the six categories of vulgar word use by using a crowdsourcing approach. They introduced six mutually exclusive labels, namely `express aggression` (AGG), `express emotion` (EMO), `emphasis` (EMPH), `auxiliary` (AUX), `signal group identity` (SGI), and `non-vulgar` (NV) use. The idea of including the Holgate's dataset in our collection, by applying to the data the SWAD annotation scheme, was stimulated by the possibility of investigating the interaction of our swear word abusiveness label with the swear word function, as introduced in the Holgate's study.

We annotated the new data by following the same annotation guidelines described in Pamungkas et al. (2020a) and involving the same pool of three expert annotators. In the case of the OLID test set, we got 66 instances after the filtering and replicating process. For Holgate's dataset, we only selected the first 1000 tweets to be re-annotated. We re-annotated all tweets regardless of their original labels. To avoid bias in the annotation process, we hide the original label based on Holgate's study from our annotators' view. Our effort was therefore towards adding another layer of annotation on the swear word. After the annotation process, we obtained 204 instances annotated as abusive and 796 as not abusive for Holgate's data. Meanwhile, for the OLID test, we obtained 18 instances annotated as abusive and 48 instances as not abusive. The inter-annotator agreement on this corpus extension is 0.516 based on the Cohen Kappa coefficient from the annotation of the first and second annotators on 1066 instances. Therefore, we have 2577 tweets in total after this extension process. Table 3 shows the interaction between the original label from Holgate's work and our new label addition.

Before the annotation process, we expected that most tweets with AGG label will be classified into abusive class. However, we found that these AGG tweets were distributed in both abusive and not abusive classes in a similar proportion. We were interested in AGG tweets, which are categorized into not abusive class. Some examples of these instances are reported in the following. Based on our annotation guidelines, the first example of swear word use (Example 3.13) is labeled as not abusive because the insulted target does not exist. Similarly, the second example (Example 3.14) shows an expression of humor and catharsis, which is not classified as abusive based on our annotation guidelines.

> [Example 3.13. Not abusive based on our guidelines] *My* ***b*llshit*** *radar is on full force today*

> [Example 3.14. Humor and catharsis] *I gained ten pounds this summer.* ***D*mn...*** *L0L*

## 4 Swear words abusiveness prediction

In this section, we provide an intrinsic evaluation of the corpus by conducting cross-validation experiments. We built supervised machine learning models to predict the abusiveness of swear words in SWAD. We model this prediction task in three different tasks, namely sequence labeling, simple text classification, and target-based swear word abusiveness prediction. The main objective of the sequence labeling experiment is to test the consistency of the annotation of the corpus. Meanwhile, we devise the classification experiment to shed some light on the most predictive feature to differentiate between abusive and not-abusive swearing. We also propose to adopt a target-based sentiment analysis task, a more well-explored task in the sentiment analysis area, into our experiment, as presented in the following subsection (see Sect. 4.3).

### 4.1 Sequence labeling task

In order to test the robustness of the annotation of swear words in SWAD, we devise a cross-validation test based on a sequence labeling task. Given a sequence of words (i.e., a tweet from our dataset), the task consists in correctly labeling each word with one of three possible labels: abusive swear word (`SWA`), non-abusive swear word (`SWNA`) or not a swear word (`NSW`). The task is carried out in a supervised fashion, by splitting the dataset in a training set (90% of the instances) and a test set (the remaining 10%).

#### *4.1.1 Model description and evaluation*

For this experiment, we adapt the BERT Transformer-based architecture (Devlin et al., 2019) with the pre-trained model for English `bert-base-cased`. We train the model for 5 epochs, with learning rate $10^{-5}$ and a batch size of 32.

**Table 4** Sequence labeling task: confusion matrix

| Predicted ground truth | SWNA | SWA | NSW |
|---|---|---|---|
| SWNA | 1366 | 217 | 68 |
| SWA | 455 | 322 | 35 |
| NSW | 139 | 52 | 38,950 |

**Table 5** Sequence labeling task: results broken down by label

| | Precision | Recall | $F_1$-score |
|---|---|---|---|
| SWNA | 0.705 | 0.829 | 0.753 |
| SWA | 0.532 | 0.389 | 0.421 |
| NSW | 0.997 | 0.994 | 0.995 |
| Macro avg | 0.745 | 0.737 | 0.723 |

### *4.1.2 Results*

Table 4 shows the confusion matrix resulting from the cross-validation. Unsurprisingly, the majority of classification errors are due to SWA/SWNA confusion, while the distinction between swear words and non-swear words is basically trivial. The classifier is slightly biased towards abusive swear words (217 SWA→SWNA misclassifications) than non-abusive swear words (455 SWNA→SWA misclassifications). These results are confirmed by the performance measured in terms of per-class precision, recall and $F_1$-score, shown in Table 5, where the SWA class has a higher recall than precision, while the opposite is true for the SWNA class. In absolute terms, the per-class and macro $F_1$-score confirms that our annotation is stable when tested in a supervised learning setting. In our test, only one abusive swear word was misclassified as NSW. Interestingly, the word is *sk*nk*, which is semantically ambiguous, conveying the offensive sense as well as the animal sense. Even more interestingly, the few NSW instances misclassified as SWA are all borderline cases of abusive language: *sh*tcago* (an offensive slang for Chicago), *messed*, *c*mming*, and *c*mslave*.

## 4.2 Simple text classification task

In this setting, we explicitly predict the abusiveness of swear words (as the target word) in given tweets as context. We employ several machine learning models including a linear support classifier (LSVC), logistic regression (LR), and random forest (RF) classifier. We use different features, at the word level (focusing on the target word) and at the tweet level (identifying the context).

**Table 6** Ablation test on several feature sets

| Feature | LSVC | | | | LR | | | | RF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature set | P | R | $F_1$ | Acc | P | R | $F_1$ | Acc | P | R | $F_1$ | Acc |
| ALL | 0.660 | 0.618 | 0.638 | 0.694 | 0.724 | 0.641 | **0.680** | 0.727 | 0.657 | 0.582 | 0.617 | 0.684 |
| ALL - Unigram SW | 0.576 | 0.503 | 0.537 | 0.651 | 0.586 | 0.539 | 0.561 | 0.651 | 0.581 | 0.537 | 0.558 | 0.649 |
| ALL - Bigram | 0.652 | 0.626 | 0.639 | 0.690 | 0.732 | 0.635 | **0.680** | 0.727 | 0.667 | 0.590 | 0.626 | 0.690 |
| ALL - Twitter | 0.723 | 0.578 | 0.642 | 0.696 | 0.723 | 0.608 | 0.661 | 0.711 | 0.702 | 0.580 | 0.635 | 0.694 |
| ALL - Sentiment | 0.576 | 0.501 | 0.536 | 0.651 | 0.723 | 0.628 | 0.672 | 0.721 | 0.670 | 0.588 | 0.626 | 0.690 |
| ALL - Stylistic | 0.667 | 0.585 | 0.623 | 0.688 | 0.719 | 0.635 | 0.674 | 0.723 | 0.642 | 0.572 | 0.605 | 0.676 |
| ALL - Syntactic | 0.710 | 0.623 | 0.664 | 0.715 | 0.722 | 0.624 | 0.670 | 0.719 | 0.667 | 0.597 | 0.630 | 0.692 |

### 4.2.1 Features

*Lexical features* In this feature set, we focus on the word-level features. We include the **Swear Word** feature, that is, the unigram of the marked swear word, as we aim to investigate whether the abusiveness of a swear word could be predicted only from the word choice. We also use the **Bigrams** feature, obtained from bigrams of the target word with its next and previous words.

*Twitter features* Since our corpus consists of tweets, we also employ several features which are particular to the Twitter data. This feature set include **Hashtag Presence**, **Emoji Presence**, **Mention Presence**, and **Link Presence**. We use regular expressions to extract hashtags, mentions and URLs, and a specialized library[14] for emoji extraction.

*Sentiment features* This feature is proposed in order to resolve the context of the tweet. We use two features: **Text Sentiment**, to model the polarity of the text, and **Emoji Sentiment** to model the overall sentiment of the emojis in the tweet. We use the VADER dictionary (Hutto et al., 2014) to extract the polarity score of the text and *emoji sentiment ranking*[15] to get the sentiment value for emojis.

*Stylistic features* In this feature set, we consider several common stylistic features for text classification task such as **Capital Word Count**,[16] **Exclamation Mark Count**, **Question Mark Count**, **Text Length**. In addition, we also exploit another word-level feature, namely **Swear Word Position**, indicating the index position of the marked swear word in the tweet.

*Syntactic features* In this feature set, we focus on the word-level features, including **Part of Speech** and the **Dependency Relation** of the target word with its next and previous words. We extract part-of-speech tags with the NLTK library,[17] while dependency relations are extracted with SpaCy.[18]

### 4.2.2 System description and evaluation

We build our models by using the Scikit-learn library.[19] We split the dataset into 80% and 20% for the training and testing respectively. We use several evaluation metrics, including accuracy, macro average precision, macro average recall, and macro average *F*-score. An ablation test is performed to investigate the role of each feature set in the classification result. The swear word unigram feature is used as a baseline in this experimental setting.

[14] https://pypi.org/project/emoji/.

[15] http://kt.ijs.si/data/Emoji_sentiment_ranking/.

[16] This feature consider all capital words on the tweet.

[17] https://www.nltk.org/.

[18] https://spacy.io/.

[19] https://scikit-learn.org/stable/.

### *4.2.3 Results*

Table 6 shows the full results of the text classification experiment by using LSVC, LR, and RF models. We start the experiment by using all feature groups altogether. Then, we remove one feature at a time to see the importance of each feature group in the model performance. Overall, RF is under-performing compared to the two other classifiers. The results also show that LR performed the best compared to two other models. Based on the macro average *F*-score, the best performance is achieved using all the features coupled with LR. With the same model and by removing Bigrams feature also obtained similar performance, but a lower macro average recall. Our goal is to investigate the most predictive feature set in the ablation experiment by removing one feature set at a time. We found that the unigram of a swear word is the most informative feature in this classification task. Bigrams, sentiment, emotion, stylistic and syntactic features all contribute to the classification performance, while the Twitter features have a detrimental effect on the LSVC and RF models. The main issue of this task is the lower recall compared to the precision, which is consistent across all models. It denotes that such models struggle to deal with false-negatives. We argue that this happens due to the dataset imbalance, where the swear words percentage over both classes is dominated by not-abusive class (negative class).

## 4.3 Target-based abusiveness prediction of swear words

This setting is similar to the text classification task presented in Sect. 4.2. However, here we explicitly model the task by adopting a similar setting as the target-dependent sentiment analysis task (Vo & Zhang, 2015; Saeidi et al., 2016). The main objective of this task is to identify the sentiment polarity of a given target in an utterance. This task is also related to aspect-based sentiment analysis. However, in target-dependent sentiment analysis, the target word is known and mentioned explicitly in the given utterance. Meanwhile in aspect-based sentiment analysis, the target aspect could be expressed implicitly, where the aspect detection is also part of the task. Adopting a similar idea of target-dependent sentiment analysis, we use the swear word as the target word, with the main objective to predict its abusiveness in a given utterance as a context.

> [Example 4.1. Tweet from Davidson's dataset] *@USER **d*mn** I hate a **b*tch** that like to argue and **sh*t***

In Example 4.1, we can find three swear words in the tweet, i.e., "d*mn", "b*tch", and "sh*t". Therefore, there are three target words, and the task is to predict the abusiveness of each swear word in the tweet, individually. Based on our manual investigation, the first swear word is not abusive, the second one is abusive, while the third one is more difficult to assess. The first swear word is used to express catharsis, which is not abusive in most of the cases. The second swear is abusive because it can insult some targets. The last swear word is a bit problematic since the swear word is used as an idiomatic expression. The abusiveness of a given swear word is highly dependent on its context in the tweet, which is identical to the target-dependent sentiment analysis task.

### *4.3.1 System description and evaluation*

In this experiment, we adopt several state-of-the-art models from the target-dependent sentiment analysis task as baseline models. In addition, we also implement a BERT model by applying a simple masking approach to mark the target words. We evaluate the model's performance by using several standard evaluation metrics, including precision, recall, *F*-score and accuracy. We present precision, recall, and *F*-score on both positive and negative classes. We split our extended SWAD corpus into training (70%), development (10%), and testing (20%) sets for the experiment. Following is a short description of each model we use in our experiment:

- **TD-LSTM (Target-dependent LSTM)** The basic idea of this architecture is to model the preceding and following context surrounding the target word so that the feature representation consists of the left part (preceding the target word) and the right part (following the target word) (Tang et al., 2016). Specifically, this architecture consists of two LSTMs (LSTM left and LSTM right), which model the preceding and following target word context, respectively. The output of these LSTMs is then concatenated to the softmax layer to predict the label.
- **TC-LSTM (Target-connection LSTM)** This architecture is a further development of TD-LSTM, which tries to incorporate a target connection component. The additional component explicitly models the connection between the target word and each context of the word when building the sentence representation (Tang et al., 2016). This component was implemented as a target word vector obtained by averaging the vectors of context work of words it contains. This vector is then concatenated to the word representation before feeding it to the LSTM network. The rest of the architecture is similar to the TD-LSTM.
- **AE-LSTM (Aspect Embedding LSTM)** This architecture (Wang et al., 2016) tries to learn the embedding vector of each aspect, or in our study, is the target word. This vector is then concatenated to the sentence embedding representation, which is followed by the LSTM network. The additional vector representation of the target word gives vital information to the model to learn the sentiment for each target word.
- **AT-LSTM (Attention-based LSTM)** The standard LSTM is not able to model the important part of aspect-based sentiment classification. This particular model (AT-LSTM) (Wang et al. 2016) has an attention mechanism which captures the important part of a sentence by focusing on the given aspect. This attention mechanism took input from the hidden layer produced by LSTM and aspect embedding vector and produce an attention weight vector and a weighted hidden representation.

- **ATAE-LSTM (Attention-based LSTM with Aspect Embedding)** Basically, this architecture (Wang et al., 2016) is AT-LSTM model which is concatenated with aspect embedding vector as implemented in AE-LSTM.
- **CABASC (Content Attention Based Aspect Based Sentiment Classification)** This architecture consists of two enhanced attention mechanisms (Liu et al., 2018), including sentence-level content attention mechanism which captures the important information about given aspects from a global perspective and the context attention mechanism, which simultaneously takes the order of the words and their correlations into account, by embedding them into a series of customized memories.
- **IAN** This architecture uses two LSTM networks to model the sentences and the target words (Ma et al., 2017). Then, the target word's hidden state and the hidden state of context sentence are placed in parallel to generate an attention vector interactively. Finally, these attention vectors provide a sentence representation and target representation.
- **RAM (Recurrent Attention on Memory)** This framework implements a multiple-attention mechanism that captures sentiment features separated by a long distance so that it is more robust against irrelevant information (Chen et al., 2017). The outputs of these multiple attentions are non-linearly combined with the LSTM network, strengthening the model for handling more complications.
- **TD-BERT (Target-dependent BERT)** We also propose to adopt the idea of TD-LSTM and exploit the state-of-the-art pre-trained model BERT as language representation. Therefore, our model consists of two BERT layers (BERT left and BERT right) to represent the context of preceding and following target words, respectively. The output of these BERT layers is passed into a fully connected dense layer with RELU activation before going into the last sigmoid layer to produce the final prediction. This model is optimized using Adam Optimizer with a learning rate of 1e-5 and trained with three epochs and batch size at 32.[20]
- **TM-BERT (Target-masked BERT)** BERT model has an attention mechanism to model many downstream tasks which involve single text or even text pairs. BERT encodes multiple text segments using two special tokens ([SEP] and [CLS]). [SEP] token is used to separate two or more text segments in case of multiple text segment processing. In the single text, the encoded text is started by [CLS] token and ended by [SEP] token. In this model, we add a special token [SW] and [SW] to mark the swear word in the sentence. The intuition for doing so is to inform the important part (target word) of the text for the model. We expect the BERT model able to construct the representation by focusing on this

[20] These settings were obtained based on the results of our preliminary experiments. We tried to optimize several hyperparameters including varying the optimizers (Adam and RMSProp), learning rate ($1^{-5}$, $2^{-5}$, and $3^{-5}$), batch size (16, 32 and 64), and the number of epochs (1 - 10)

**Table 7** Result of target-based abusiveness prediction of swear words

| Model | $P_0$ | $P_1$ | $R_0$ | $R_1$ | $F_0$ | $F_1$ | $F_{avg}$ | Acc |
|---|---|---|---|---|---|---|---|---|
| TD-LSTM | 0.610 | 0.793 | 0.617 | 0.789 | 0.613 | 0.791 | 0.702 | 0.729 |
| TC-LSTM | 0.628 | 0.801 | 0.628 | 0.801 | 0.628 | 0.801 | 0.714 | 0.740 |
| AT-LSTM | 0.611 | 0.661 | 0.061 | 0.979 | 0.111 | 0.789 | 0.450 | 0.659 |
| AE-LSTM | 0.721 | 0.708 | 0.272 | 0.943 | 0.395 | 0.809 | 0.602 | 0.709 |
| ATAE-LSTM | 0.603 | 0.786 | 0.600 | 0.789 | 0.602 | 0.788 | 0.695 | 0.723 |
| IAN | 0.661 | 0.735 | 0.400 | 0.890 | 0.498 | 0.805 | 0.652 | 0.719 |
| CABASC | 0.747 | 0.723 | 0.328 | 0.940 | 0.456 | 0.818 | 0.637 | 0.727 |
| RAM | 0.628 | 0.744 | 0.450 | 0.857 | 0.524 | 0.797 | 0.660 | 0.715 |
| TD-BERT | 0.719 | 0.814 | 0.580 | 0.848 | 0.636 | 0.827 | 0.731 | 0.784 |
| TM-BERT | 0.708 | 0.825 | 0.625 | 0.859 | **0.665** | **0.843** | **0.754** | 0.806 |

special token. We use an open-source implementation of BERT by HuggingFace,[21] which provides a special method to add a new special token in the BERT masking process.[22]

### *4.3.2 Results*

As shown in Table 7, the **TM-BERT** obtained the best result with .665 in *F*-score in positive class, .843 in *F*-score in negative class, and .754 in macro *F*-score. Overall, the BERT-based models achieved a better result than other models, where **TD-BERT** also obtained a competitive result in all evaluation metrics. We also notice that **TD-LSTM** and **TC-LSTM** get better results than the rest of non-BERT models, including **CABASC** and **RAM**, which achieved better performance in several benchmarks for the aspect-based sentiment analysis task (Liu et al., 2018). We also compare our result in this experiment with the results in our previous experiment, as presented in Table 6. The overall result shows that our models presented in this experiment, which are based on neural architecture, outperformed the traditional models. We also notice that most of the models exploited in this experiment are able to cope with the dataset imbalance issue, as we discovered in previous experiments with traditional models, where we obtained lower recall than precision.

[21] https://huggingface.co.

[22] https://huggingface.co/transformers/model_doc/bert.html.

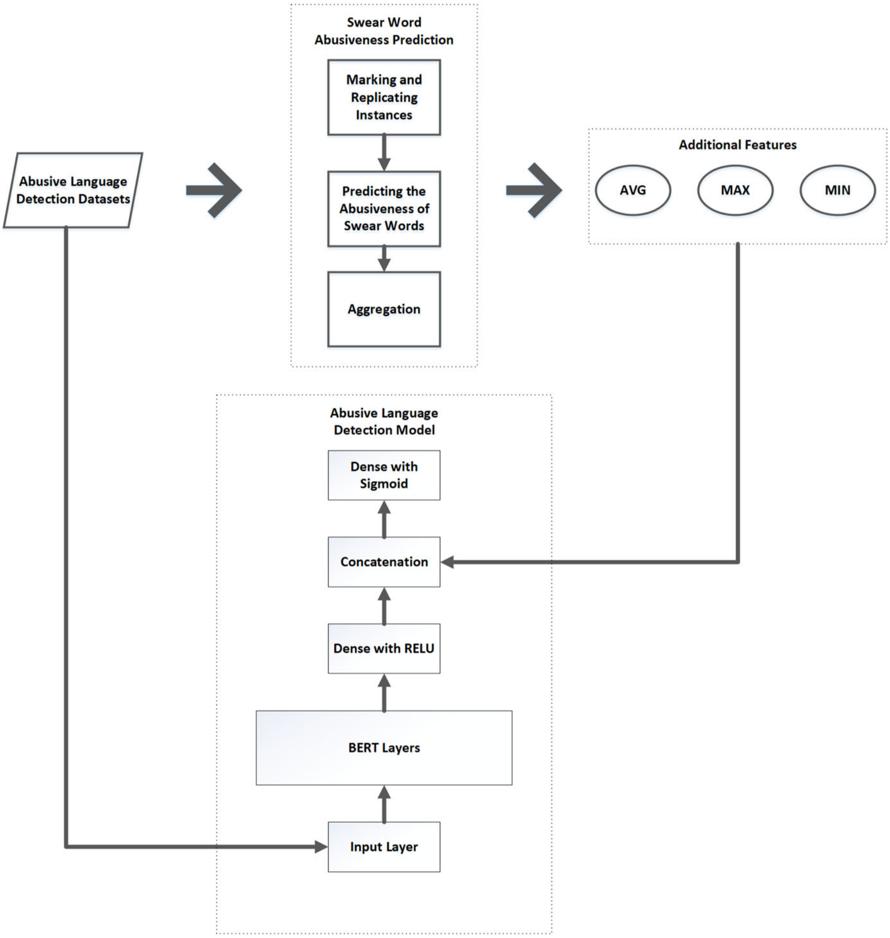
Swear Word Abusiveness Prediction
Marking and Replicating Instances
Predicting the Abusiveness of Swear Words
Aggregation
Abusive Language Detection Datasets
Additional Features
AVG
MAX
MIN
Abusive Language Detection Model
Dense with Sigmoid
Concatenation
Dense with RELU
BERT Layers
Input Layer

**Fig. 2** Process to infuse additional features

# 5 Swear words in abusive language detection

## 5.1 Task description and experimental settings

In order to answer the third research question (RQ3), we explore the usefulness of the swear word abusiveness information feature on several downstream abusive language detection tasks. We reiterate that our assumption is that knowing the swear word context as either abusive or not could help the system resolve the abusiveness of the whole utterance. Therefore, our idea is to explicitly infuse the swear word abusiveness prediction into the abusive language detection model to help the model dealing with swear word ambiguity. The overall experimental scenario is illustrated in Fig. 2.

First, we need to select some abusive language benchmarks which contain a high frequency of swear words. We found four dataset collections, including **HatEval** (Basile et al., 2019), **AMI@IberEval** (Fersini et al., 2018), **AMI@Evalita** (Fersini et al., 2018), and **Davidson** (Davidson et al., 2017) datasets. These datasets contain a fairly high frequency of swear words. Around half or their instances are containing swear words, specifically 42.26%, 56.74%, 62.79%, and 69.2% for HatEval, AMI@Evalita, AMI@IberEval, and Davidson dataset respectively. Following is a short description of each dataset.

### *5.1.1 HatEval dataset*

The dataset focuses on the detection of hate speech in Twitter on two specific targets, namely immigrants and women, in a multilingual perspective (Basile et al., 2019). The HatEval shared task introduced a dataset in two languages, English and Spanish. However, we will only focus on the English collection. The HatEval collection was gathered by using several keywords, including neutral keywords, pejorative words towards targets, and highly polarized hashtags. This dataset was annotated by relying on judges from a crowdsourcing platform, which applied an annotation scheme including three binary labels: hate speech (hate speech or not), target range (generic or individual), and aggressiveness (aggressive or not). The final dataset used for the English HatEval shared task contains 13,000 (about 10,000 for training and for 3000 testing).[23]

### *5.1.2 AMI datasets*

Basically, datasets for AMI@IberEval (Fersini et al., 2018) and AMI@Evalita (Fersini et al., 2018) were selected from the same collection of tweets, which were filtered using three approaches including querying from Twitter streaming API based on some keywords, monitoring account of online harassment victims, and downloading tweets from misogynist accounts. This dataset was annotated with three annotation layers including misogyny identification (misogyny or not), misogynistic behaviours (stereotype, dominance, derailing, sexual harassment, and discredit), and target of misogyny (active or passive). In this task we focus only on the misogyny identification task, where models need to predict whether a given tweet as either misogynous or not. AMI@IberEval dataset contains 3977 tweets (3251 training and 831 testing), while AMI@Evalita collection contains 5000 tweets (4000 training and 1000 testing). Originally, AMI dataset is available in three languages including English, Italian, and Spanish, but here we only focus on English.

[23] Upon manual investigation, organizers decided to exclude 1000 tweets from the English training set, 29 tweets from the English test set due to duplicated instances.

### *5.1.3 Davidson dataset*

The dataset has been built by Davidson et al. (2017) and contains 24,783 tweets[24] manually annotated with crowdsourcing scenario. Differently from the other datasets considered, in this dataset a multilabel annotation is applied, with three labels including *hate speech*, *offensive*, and *neither*. These tweets were sampled from a collection of 85.4 million tweets gathered using the Twitter search API, focusing on tweets containing keywords from HateBase.[25] Only 5.8% of the total tweets were labeled as *hate speech* and 77.4% as *offensive*, while the remaining 16.8% were labelled as *not offensive*.

The second process is to predict the abusiveness of swear words in each instance of these datasets. We pre-processed all instances of these datasets similarly as we did to the SWAD (see Fig. 1), including marking the swear word and replicating instance when more than one swear words are found. After a preprocessing step, we immediately predict all preprocessed instances by employing our best performing system based on results presented in the previous section, which is TM-BERT. We aggregate the prediction score for instances which contain more than one swear words by taking *minimum (MIN)*, *maximum (MAX)*, and *average (AVG)* score. In case of instances which do not containing swear word, we set the prediction score to 0.

The final process is to infuse the swear word abusiveness prediction score into the base model for detecting abusive language in these respective tasks. However, note that this work aims not to produce the best possible system for these shared tasks but rather to test our hypothesis on the usefulness of predicting the pragmatics of swear word use. For this experiment, we employ a straightforward BERT model with a minimum hyperparameter tuning. We use (`bert-base-cased`) model available on TensorFlow-hub[26], which allows us to integrate BERT with the Keras functional layer[27]. Our network starts with the BERT layer, which takes three inputs consisting of id, mask, and segment before passing into a dense layer with RELU activation (256 units) on top and an output layer with sigmoid activation. We train the network with the Adam optimizer with a learning rate of $2^{-5}$.[28] We tune this model by trying several combinations of batch size (32, 64, 128) and the number of epochs (1–5). We infuse the additional feature by simply concatenating the swear word abusiveness probability into the dense layer after the BERT embedding layer.

[24] Although in the original paper the authors mention that the dataset consists of 24,802 annotated tweets, we only found this number of instances in the shared GitHub repository: https://github.com/t-davidson/hate-speech-and-offensive-language.

[25] A multilingual repository, which allows for the identification of HS terms by region: https://hatebase.org.

[26] https://www.tensorflow.org/hub.

[27] https://keras.io/.

[28] In a preliminary stage, we also tried to use another optimizer (RMSProp) and different number of learning rate ($1^{-5}$, $2^{-5}$, and $3^{-5}$).

## 5.2 Results

We apply standard evaluation metrics in this experiment, including a wide coverage of evaluation metrics such as precision, recall, $F$-score, and accuracy. We present precision, recall, and $F$-score on both positive and negative classes to picture the system performance better. Tables 8, 9, 10, and 11 present the result of the experiments on HatEval task, AMI@Evalita task, AMI@IberEval, and Davidson dataset, respectively. As mentioned before that, we experiment with three additional features, namely MIN, MAX, and AVG. These additional features depict the approach in aggregating the abusiveness score when more than one swear words exists in the tweet. We marked with superscript (*) results where the performance improvement is statistically significant compared to baseline models ($F_{avg}$ and Acc columns).[29]

On the HatEval task, the additional feature was able to improve the model performance. The best result is obtained using the MIN score with .482 in the macro average $F$-score with a statistical significance compared to the baseline model. A similar result is observed on both AMI@Evalita and AMI@IberEval task datasets. All models infused by additional features are experiencing performance improvement significantly, where the best result was obtained by using the MIN aggregation score. The performance improvement is consistent in both classes, as observed from the $F$-score in the positive ($F_1$) and negative ($F_0$) class. However, a different result was observed in the experiment of the Davidson dataset as presented in Table 11. We found that the additional features were not able to augment the model performance.

It was an interesting finding that the MIN aggregation is recognized as the most effective approach on most datasets. Based on our further investigation, we found two possible reasons which lead to this result. First, we found several examples where two or more swear words were used in different abusiveness degrees within one tweet. As shown in the Example below, which is taken from AMI Evalita collection. Our model predicted the first swear word with a high abusiveness degree, while the second one with a low degree of abusiveness. With the MIN aggregation, the additional feature informs the model that there is an not-abusive swear word, which could become an important signal to resolve the context of the whole message. On the contrary, if we use MAX aggregation, the additional feature could also deceive the model. In this case, MIN aggregation would provide a better knowledge for the model. Second, there are many instances of HatEval, AMI@Evalita, and AMI@IberEval, which contain more than one swear word. Therefore, aggregating the score in a better way would heavily influence the prediction result, where in this case MIN aggregation provides better information for the models.

> [Example 5.1. Not Misogyny tweet from AMI Evalita dataset] *everytime i reach the highlights of smut im reading me. ok* ***ho**** *calm down calm down sit your* ****ss*** *relax its just a smut*

[29] We used bootstrap sampling significance test tools publicly available at https://github.com/fornaciari/boostsa.

Regarding to the peculiar result on Davidson dataset, we conducted a deeper investigation. We notice that our models struggle to detect the hate speech class as observed in Table 11, where the micro $F$-score in hate speech class was very low. Furthermore, our additional feature also failed to improve the model performance in determining the hate speech instances. Our manual inspection of the dataset highlights that our swear word abusiveness prediction model struggles to differentiate between the swear word in the offensive class and the hate speech class. For example, as shown in the examples below (Example 5.2 and Example 5.3), we can see that our model predicts the swear word use in both classes with a high abusiveness degree. Even with human reasoning, we also could not differentiate the abusiveness degree of the swear words in both messages. We argue that this issue is the main reason for the less impact of our additional features in the Davidson dataset.

> [Example 5.2. Offensive tweet from Davidson's dataset] *@USER @USER so you was in a female DMs talking to another* ***n*gga****... You're a* ***f*ggot****...*

> [Example 5.3. Hate Speech tweet from Davidson's dataset] *Vanessa is such a* ***f*ckin f*ggot****.*

## 6 Conclusion and future work

The research presented in this paper investigates the automatic classification of abusive swearing. We developed a new benchmark corpus called SWAD, consisting of English tweets, where abusive swearing is manually annotated at the word-level. Our initial corpus consists of 1511 instances of swearing from 1320 tweets, where 620 swear words were annotated as abusive and 891 marked as not-abusive. The inter-annotator agreement is 0.708, based on Cohen's Kappa coefficient, which denotes a substantial agreement. Furthermore, we extended our corpus to improve the coverage when it is used by statistical models. We added 66 tweets from the OLID test set, which were missing from the first round of annotation, and 1000 instances from Holgate's dataset. Our second annotation process labeled 204 instances annotated as abusive and 796 instances as not abusive. The annotator agreement for the second annotation process is 0.516, which moderate agreement is achieved. Our final collection consists of 2577 instances from 2282 tweets.

We built models trained on the SWAD corpus to automatically classify abusive and not-abusive swear words and provide an intrinsic evaluation of SWAD. We experimented by modeling this task into three different settings, namely, sequence labeling, simple text classification, and target-based swear word abusiveness prediction. We used BERT for sequence labeling, simpler but more transparent models for text classification, and wide coverage of models, including several state-of-the-art models in aspect-based sentiment analysis for the target-based task. Our results confirm that our annotation is robust as shown by the sequence labeling performance. On the other hand, text classification results provided new insights on

**Table 8** Result of investigating swear words role in HatEval task

| Model | $P_0$ | $P_1$ | $R_0$ | $R_1$ | $F_0$ | $F_1$ | $F_{avg}$ | Acc |
|---|---|---|---|---|---|---|---|---|
| BERT | 0.695 | 0.442 | 0.225 | 0.838 | 0.340 | 0.579 | 0.459 | 0.502 |
| BERT + Features (MAX) | 0.708 | 0.443 | 0.227 | 0.846 | 0.343 | 0.582 | 0.462 | 0.491 |
| BERT + Features (MIN) | 0.690 | 0.454 | 0.261 | 0.822 | **0.379** | **0.585** | **0.482*** | 0.513 |
| BERT + Features (AVG) | 0.720 | 0.436 | 0.184 | 0.885 | 0.294 | 0.584 | 0.439 | 0.485 |

**Table 9** Result of investigating swear words role in AMI Evalita task

| Model | $P_0$ | $P_1$ | $R_0$ | $R_1$ | $F_0$ | $F_1$ | $F_{avg}$ | Acc |
|---|---|---|---|---|---|---|---|---|
| BERT | 0.604 | 0.635 | 0.761 | 0.458 | 0.674 | 0.532 | 0.603 | 0.606 |
| BERT + Features (MAX) | 0.617 | 0.664 | 0.756 | 0.478 | 0.680 | 0.555 | 0.618 | 0.637* |
| BERT + Features (MIN) | 0.627 | 0.676 | 0.762 | 0.486 | **0.688** | **0.565** | **0.627*** | 0.636* |
| BERT + Features (AVG) | 0.587 | 0.647 | 0.764 | 0.469 | 0.664 | 0.544 | 0.604 | 0.616 |

**Table 10** Result of investigating swear words role in AMI IberEval task

| Model | $P_0$ | $P_1$ | $R_0$ | $R_1$ | $F_0$ | $F_1$ | $F_{avg}$ | Acc |
|---|---|---|---|---|---|---|---|---|
| BERT | 0.701 | 0.746 | 0.869 | 0.540 | 0.776 | 0.627 | 0.701 | 0.740 |
| BERT + Features (MAX) | 0.734 | 0.765 | 0.904 | 0.543 | **0.810** | 0.636 | 0.723* | 0.747 |
| BERT + Features (MIN) | 0.715 | 0.785 | 0.931 | 0.555 | 0.809 | **0.650** | **0.730*** | 0.766* |
| BERT + Features (AVG) | 0.718 | 0.765 | 0.912 | 0.542 | 0.803 | 0.634 | 0.719 | 0.773* |

the most predictive features for distinguishing abusive and not-abusive swear words. In particular, we found that a wide range of features can actually improve the models' performance. Meanwhile, our intention to model the task similarly to aspect-based sentiment analysis leads to promising result. Our BERT-based models obtained the best result in this setting, significantly better than simple text classification settings where we implemented more traditional models.

Finally, we explore the usefulness of predicting swear words' abusiveness on several downstream abusive language detection tasks. Based on models built for swear word abusiveness prediction (RQ2), we introduce a novel feature, namely the swear word abusiveness feature, and infuse it to improve current abusive language detection models. We test our approach to several abusive language detection tasks, including HatEval, AMI@Evalita, AMI@IberEval, and Davidson dataset, showing consistent and significant performance improvement across topics, except the Davidson dataset. Our further investigation discovered that the different notion of

**Table 11** Result of investigating swear words role in Davidson dataset

| Model | $P_0$ | $P_1$ | $P_2$ | $R_0$ | $R_1$ | $R_2$ | $F_0$ | $F_1$ | $F_2$ | $F_{avg}$ | Acc |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 0.428 | 0.926 | 0.719 | 0.250 | 0.925 | 0.830 | **0.316** | **0.925** | 0.771 | **0.671** | 0.869 |
| BERT + Features (MAX) | 0.375 | 0.911 | 0.757 | 0.172 | 0.940 | 0.792 | 0.236 | **0.925** | **0.774** | 0.645 | 0.870 |
| BERT + Features (MIN) | 0.392 | 0.923 | 0.719 | 0.203 | 0.927 | 0.832 | 0.267 | **0.925** | 0.771 | 0.655 | 0.868 |
| BERT + Features (AVG) | 0.288 | 0.912 | 0.755 | 0.193 | 0.928 | 0.782 | 0.231 | 0.920 | 0.768 | 0.640 | 0.860 |

annotation in the Davidson dataset was the main reason why our feature was not impactful.

While these results are encouraging, we believe that there is still room for improvement for both the corpus and the automatic classification of swearing. Furthermore, we aim to improve the dataset by proposing a fine-grained categorization of swear words such as the ones introduced by Pinker (2007) and McEnery (2006)). We also plan to apply our swear word abusiveness feature into more tasks and datasets (Poletto et al., 2021), to obtain the full picture of its impact in abusive language detection tasks. We also observe the possibility to use the additional swear word abusiveness features as domain independent features which proven as important features to transfer knowledge in cross domain abusive language detection (Pamungkas et al., 2020b; Pamungkas & Patti, 2019; Chiril et al., 2021). We are also aware that the results of different aggregation approaches are heavily depended on the dataset, and this will deserve further investigation.

Applying our methodology to other languages is not trivial, as it depends on the availability of language resources and robust NLP tools for them (Pamungkas et al., 2021). Fortunately, full-fledged NLP pipelines do exist for many languages, thanks for instance to large-scale initiatives such as Universal Dependencies, which provides among its deliverables the UDpipe software library and a broad set of trained models in more than 70 languages (Nivre et al., 2016; Straka et al., 2016). Deep learning models, including transformer-based networks are also surfacing for languages less resources than English—see for instance the Italian BERT model AlBERTo (Polignano et al., 2019). Moreover, the multilingual lexicon of offensive words HurtLex (Bassignana et al., 2018) could provide a solid basis to compile lists of swear words in its 53 covered languages.

Finally, let us mention the issue related to the implicit constructions denoting abusive content or multi-word swear constructions possibly present in tweets. The detection of implicitly abusive language, i.e. abusive language that is not conveyed by abusive words, has been recently recognised as one of the most prominent challenges in the field (Wiegand et al., 2021; Caselli et al., 2020). Implicit abuse is in rather direct contrast with lexicon-based methods, based on the use of lists of swear words. Therefore, an interesting direction for future work could be extending our method to determine the abusiveness of non-swearing words, to improve our systems by addressing the possibility to detect toxicity without swear words, where the abusive load is masked by the use of euphemistic constructions and figurative devices, i.e. rhetorical questions, comparisons, metaphors, irony and sarcasm.

**Acknowledgements** The work of Endang W. Pamungkas, Valerio Basile and Viviana Patti is partially funded by the project “Be Positive!” (under the 2019 “Google.org Impact Challenge on Safety” call).

## Declarations

**Conflict of interest** All authors state that there are no conflicts of interest.

**Ethical approval** This article does not contain any studies with human participants or animals carried out by any of the authors. In addition, the data that was used is composed of textual content from the public domain taken from datasets publicly available to the research community. These datasets also conform to the Twitter Developer Agreement and Policy that allows unlimited distribution of the numeric identification number of each tweet.

## References

Allan, K., & Burridge, K. (2006). *Forbidden words: Taboo and the censoring of language*. Cambridge University Press.

Bak, J.Y., Kim, S., & Oh, A. (2012). Self-disclosure and relationship strength in twitter conversations. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, (pp. 60–64). Association for Computational Linguistics

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P., & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, (pp. 54–63)

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In: *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, December 10–12, 2018. http://ceur-ws.org/Vol-2253/paper49.pdf

Bianchi, C. (2014). Slurs and appropriation: An echoic account. *Journal of Pragmatics, 66,* 35–44. https://doi.org/10.1016/j.pragma.2014.02.009.

Cachola, I., Holgate, E., Preoţiuc-Pietro, D., & Li, J.J. (2018). Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In: *Proceedings of the 27th International Conference on Computational Linguistics*, (pp. 2927–2938)

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 6193–6202). European Language Resources Association, Marseille, France. https://aclanthology.org/2020.lrec-1.760

Chen, P., Sun, Z., Bing, L., & Yang, W. (2017). Recurrent attention network on memory for aspect sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 452–461. Association for Computational Linguistics, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1047. https://www.aclweb.org/anthology/D17-1047

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), (pp. 71–80). IEEE

Chiril, P., Pamungkas, E.W., Benamara, F., Moriceau, V., & Patti, V. (2021). Emotionally informed hate speech detection: a multi-target perspective. Cognitive Computation pp. 1–31. https://link.springer.com/article/10.1007/s12559-021-09862-5.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. arXiv:1703.04009

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In: The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011, *AAAI Workshops*, vol. WS-11-02. AAAI. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841

EU Commission: Code of conduct on countering illegal hate speech online. (2016). https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en#theeucodeofconduct

Fägersten, K. B. (2012). *Who's swearing now? The social aspects of conversational swearing*. Cambridge: Cambridge Scholars Publishing.

Fersini, E., Anzovino, M., & Rosso, P. (2018a). Overview of the task on automatic misogyny identification at ibereval. In: *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018).* CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain

Fersini, E., Nozza, D., & Rosso, P. (2018b). Overview of the evalita 2018 task on automatic misogyny identification (ami). In: *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org

Fišer, D., Huang, R., Prabhakaran, V., Voigt, R., Waseem, Z., & Wernimont, J. (2018). Proceedings of the 2nd workshop on abusive language online (alw2). In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Association for Computational Linguistics. http://aclweb.org/anthology/W18-5100

Gauthier, M., Guille, A., Deseille, A., & Rico, F. (2015). Text mining and twitter to analyze British swearing habits. Handbook of Twitter for Research

Golbeck, J., Ashktorab, Z., Banjo, R.O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A.A., Gergory, Q., Gnanasekaran, R.K., Gunasekaran, R.R., Hoffman, K.M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M.J., Naik, S., Nixon, H.L., Ramachandran, P., Rogers, K.M., Rogers, L., Sarin, M.S., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z., & Wu, D.M. (2017). A large labeled corpus for online harassment research. In: P. Fox, D.L. McGuinness, L. Poirier, P. Boldi, K. Kinder-Kurlanda (eds.) Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25–28, 2017, pp. 229–233. ACM. https://doi.org/10.1145/3091478.3091509.

Holgate, E., Cachola, I., Preoţiuc-Pietro, D., & Li, J.J. (2018). Why swear? analyzing and inferring the intentions of vulgar expressions. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (pp. 4405–4414). Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1471. https://www.aclweb.org/anthology/D18-1471

Hutto, C.J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: E. Adar, P. Resnick, M.D. Choudhury, B. Hogan, A.H. Oh (eds.) Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014. The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

Jay, T. (1992). *Cursing in America: A psycholinguistic study of dirty language in the courts, in the movies, in the schoolyards, and on the streets*. John Benjamins Publishing.

Jay, T. (1999). *Why we curse: A neuro-psycho-social theory of speech*. John Benjamins Publishing.

Jay, T. (2009a). Do offensive words harm people? *Psychology, Public Policy, and Law, 15*(2), 81.

Jay, T. (2009b). The utility and ubiquity of taboo words. *Perspectives on Psychological Science, 4*(2), 153–161.

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture, 4*(2), 267–288.

Jay, T., King, K., & Duncan, T. (2006). Memories of punishment for cursing. *Sex Roles, 55*(1–2), 123–133.

Johnson, D. I. (2012). Swearing by peers in the work setting: Expectancy violation valence, perceptions of message, and perceptions of speaker. *Communication Studies, 63*(2), 136–151.

Kurrek, J., Saleem, H.M., & Ruths, D. (2020). Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 138–149. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.alw-1.17. https://www.aclweb.org/anthology/2020.alw-1.17

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald trump's youtube campaign videos. *Internet Research, 27*(4), 991–1010.

Liu, Q., Zhang, H., Zeng, Y., Huang, Z., & Wu, Z. (2018). Content attention model for aspect based sentiment analysis. In: P. Champin, F. Gandon, M. Lalmas, P.G. Ipeirotis (eds.) Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23–27, 2018, pp. 1023–1032. ACM. https://doi.org/10.1145/3178876.3186001.

Ma, D., Li, S., Zhang, X., & Wang, H. (2017). Interactive attention networks for aspect-level sentiment classification. In: C. Sierra (ed.) *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19–25, 2017*, (pp. 4068–4074). ijcai.org. https://doi.org/10.24963/ijcai.2017/568.

Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence, 30*(2), 187–202.

McEnery, A. (2006). *Swearing in English: Blasphemy, purity and power from 1586 to the present*. Routledge.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*(4), 857.

Michal, P., Pawel, D., Tatsuaki, M., Fumito, M., Rafal, R., Kenji, A., & Yoshio, M. (2010). In the service of online order: Tackling cyber-bullying with machine learning and affect analysis. *International Journal of Computational Linguistics Research, 1*(3), 135–154.

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on arabic social media. In: *Proceedings of the First Workshop on Abusive Language Online*, (pp. 52–56)

Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (pp. 1659–1666). European Language Resources Association (ELRA), Portorož, Slovenia. https://www.aclweb.org/anthology/L16-1262

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In: *Proc. of the 25th International Conference on World Wide Web*, (pp. 145–153)

Pamungkas, E.W., Basile, V., & Patti, V. (2020a). Do you really want to hurt me? predicting abusive swearing in social media. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6237–6246. European Language Resources Association, Marseille, France. https://www.aclweb.org/anthology/2020.lrec-1.765

Pamungkas, E.W., Basile, V., & Patti, V. (2020b). Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. Information Processing & Management **57**(6), 102360. https://www.sciencedirect.com/science/article/pii/S0306457320308554

Pamungkas, E.W., Basile, V., & Patti, V. (2021). Towards multidomain and multilingual abusive language detection: a survey. Personal and Ubiquitous Computing pp. 1–27. https://link.springer.com/article/10.1007/s00779-021-01609-1

Pamungkas, E.W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In: F. Alva-Manchego, E. Choi, D. Khashabi (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop, pp. 363–370. Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-2051/

Pinker, S. (2007). The stuff of thought: Language as a window into human nature. Penguin

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and Benchmark Corpora for Hate Speech Detection: a Systematic Review. Language Resources and Evaluation **55**(2), 477–523. https://link.springer.com/article/10.1007/s10579-020-09502-8

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019). Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In: Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019. http://ceur-ws.org/Vol-2481/paper57.pdf

Razavi, A.H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In: Canadian Conference on Artificial Intelligence, pp. 16–27. Springer

Rieber, R. W., Wiedemann, C., & D'Amato, J. (1979). Obscenity: Its frequency and context of usage as compared in males, nonfeminist females, and feminist females. *Journal of Psycholinguistic Research, 8*(3), 201–223.

Rojas-Galeano, S. (2017). On obstructing obscenity obfuscation. *ACM Transactions on the Web (TWEB), 11*(2), 12.

Ross, H. (1969). Patterns of swearing. Discovery: The Popular Journal of Knowledge pp. 479–481

Saeidi, M., Bouchard, G., Liakata, M., & Riedel, S. (2016). SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1546–1556. The COLING 2016 Organizing Committee, Osaka, Japan. https://aclanthology.org/C16-1146

Schmidt, A., & Wiegand, M. (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, pp. 1–10

Sood, S., Antin, J., & Churchill, E. (2012). Profanity use in online communities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1481–1490. ACM

Stephens, R., & Umland, C. (2011). Swearing as a response to pain-effect of daily swearing frequency. *The Journal of Pain, 12*(12), 1274–1281.

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4290–4297. European Language Resources Association (ELRA), Portorož, Slovenia. https://www.aclweb.org/anthology/L16-1680

Tang, D., Qin, B., Feng, X., & Liu, T. (2016). Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 3298–3307. The COLING 2016 Organizing Committee, Osaka, Japan. https://www.aclweb.org/anthology/C16-1311

Thelwall, M. (2008). Fk yea i swear: cursing and gender in myspace. *Corpora, 3*(1), 83–107.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., et al. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE, 13*(10), e0203794.

Vo, D., & Zhang, Y. (2015). Target-dependent twitter sentiment classification with rich automatic features. In: Q. Yang, M.J. Wooldridge (eds.) Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, pp. 1347–1353. AAAI Press. http://ijcai.org/Abstract/15/194

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A.P. (2014a). Cursing in English on twitter. In: S.R. Fussell, W.G. Lutters, M.R. Morris, M. Reddy (eds.) Computer Supported Cooperative Work, CSCW '14, Baltimore, MD, USA, February 15-19, 2014, pp. 415–425. ACM. https://doi.org/10.1145/2531602.2531734

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A.P. (2014b). Cursing in english on twitter. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, pp. 415–425. ACM

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 606–615. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1058. https://www.aclweb.org/anthology/D16-1058

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In: Proceedings of the First Workshop on Abusive Language Online, pp. 78–84

Wiegand, M., Ruppenhofer, J., & Eder, E. (2021). Implicitly abusive language – what does it actually look like and why are we not getting there? In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 576–587. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.naacl-main.48. https://aclanthology.org/2021.naacl-main.48

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. 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), pp. 1415–1420. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1144. https://www.aclweb.org/anthology/N19-1144

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 75–86. Association for Computational Linguistics, Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/S19-2010. https://www.aclweb.org/anthology/S19-2010

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.