Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6237–6246
Marseille, 11–16 May 2020

Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media

Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti
Dipartimento di Informatica, University of Turin
C.so Svizzera 105, 10149, Turin, Italy
{pamungka, basile, patti}@di.unito.it

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, taking the possibility of predicting the *abusiveness* of a swear word in a tweet context 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 1,511 unique swear words from 1,320 tweets. We developed models to automatically predict abusive swearing, to provide an intrinsic evaluation of SWAD and confirm the robustness of the resource. We also present the results of a *glass box* ablation study in order to investigate which lexical, syntactic, and affective features are more informative towards the automatic prediction of the function of swearing.
Keywords: swearing, social media, abusive language detection

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; Jay, 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 and 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: *fuck*, *shit*, *ass*, *bitch*, *nigga*, *hell*, and *whore*.

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 and 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). 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).

Many studies have been proposed in recent years to deal with online abuse, where swear words have an important role, providing a signal to spot abusive content. However, as we can expect observing the different facets of swearing in social environments, the presence of swear words could also lead to false positives when they occur in a non-abusive context. 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-

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

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.

Our Motivation and Contribution. We explore the phenomenon of swearing in Twitter conversations, taking the possibility of predicting the *abusiveness* of a swear word in a tweet context as the main investigation perspective. The main goal is to automatically differentiate between abusive swearing, which should be regulated and countered in online communications, and not-abusive one, that should be allowed as part of freedom of speech, also recognising its positive functions, as in the case of reclaimed uses of slurs. To achieve our goal, we propose a two-fold contribution. 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, harassment, hate speech, and bullying involving several sensitive topic 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. Such models are trained on the novel SWAD corpus, to predict the *abusiveness* of a swear word within a tweet. The results confirm the robustness of the annotation in the SWAD corpus. We obtained 0.788 in macro F_1-score in sequence labeling setting by using BERT, and explored the role of different features, also related to affect, in a standard text classification setting, with the aim to shed a better light on the properties which allow to distinguish between abusive and not-abusive swearing.

The paper is organized as follows. Section 2 introduces related work on swearing in context. Section 3 reports on the various steps of development of the SWAD Twitter corpus. The annotation scheme applied and the main issues in the annotation process are described in Section 4. Section 5 presents the experimental setting and discusses the result. Finally, Section 6 includes conclusive remarks and ideas for future work.

2. Related Works

Wang et al. (2014) examines the cursing activity in 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 develop the corpus.

Swearing is not always offensive or abusive and its offensiveness or abusiveness is context-dependant. Swearing context is explored by several prior studies. Fägersten (2012), following the dichotomy introduced by Ross (1969), classifies swearing context into two types: *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. The work by Jay (2009b) found that the offensiveness of taboo words is very dependant 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 found that obscenities/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 euphemism; *abusive*, using taboo words to abuse or insult someone; *idiomatic*, using taboo words to arouse 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.

The most similar work to ours is the study by Holgate et al. (2018) that introduced six vulgar word use functions, and built a novel English dataset based on them. The classification of the function of swear words is used to improve the classification of hate speech in social media. In this work, instead we focus on the abusiveness prediction of swear words, rather than their function, with the goal of discovering the context of a given swear word whether abusive (should be censored) or not-abusive.

3. Corpus Creation

Our starting point was a corpus of tweets selected from the training set of Offensive Language Identification Dataset (OLID)(Zampieri et al., 2019a), which was proposed in the context of the shared task OffensEval (Zampieri et al., 2019b) at SemEval 2019[5]. This task is aimed to detect 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

[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

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

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

[5]http://alt.qcri.org/semeval2019/index.php?id=tasks.

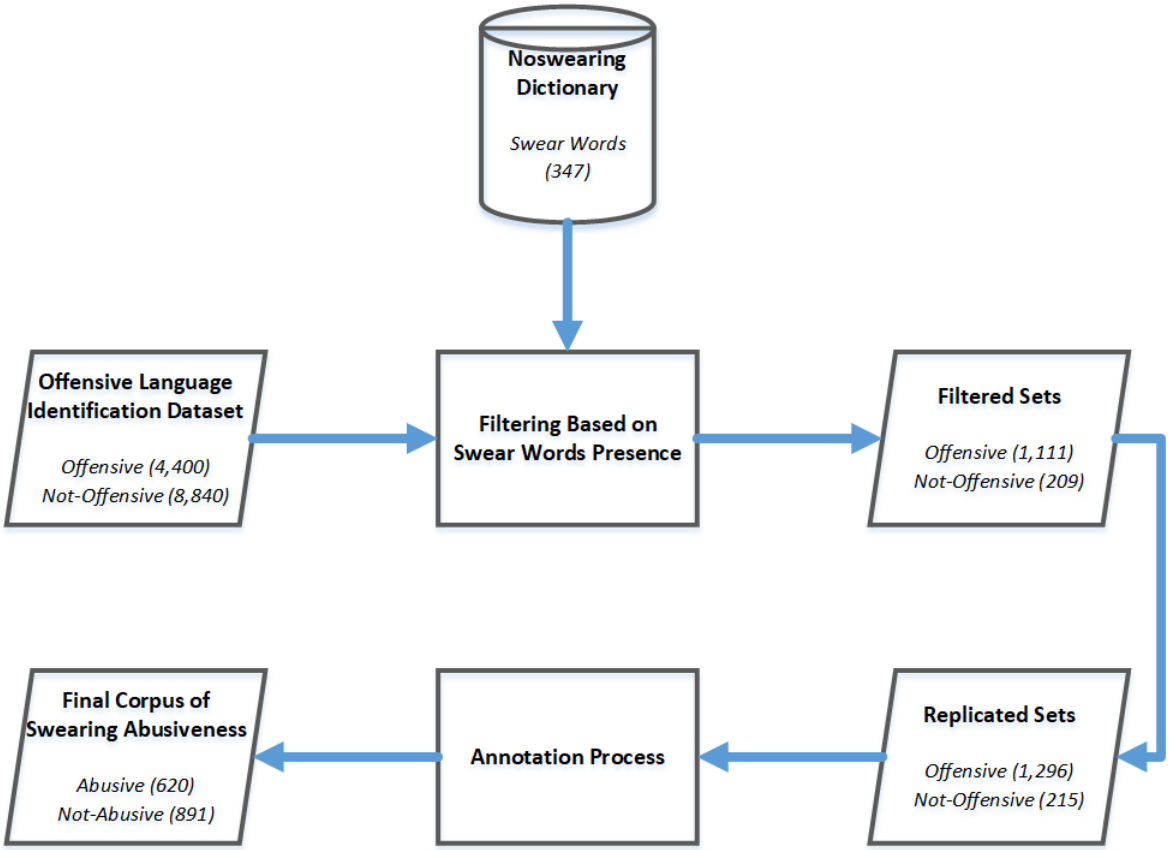

Figure 1: Corpus Development Process.

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. 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[6], 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 word "ho" and "hard on"[7]. Table1 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

	Original	After Filtering	After Replication
Offensive	4,400	1,111	1,296
Not	8,840	209	215
Total	13,240	1,320	1,511

Table 1: Corpus statistic after filtering process.

single swear word with special tags $< b >$ and $< /b >$ (e.g. $< b >$fuck$< /b >$, $< b >$shit$< /b >$, 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 shit gon keep me in the crib lol fuck it*, two instances will be generated: *@USER This* $< b >$*shit*$< /b >$ *gon keep me in the crib lol fuck it* and *@USER This shit gon keep me in the crib lol* $< b >$*fuck*$< /b >$ *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 1,511 instances to be annotated. Figure 1 shows the overall process of our corpus development.

3.1. Annotation Task and Process

The annotation of 1,511 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 the third annotator (A3), labeling those instances

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

[7] 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,..."

where a disagreement between A1 and A2 was detected. All annotators use English as a second language, with a minimum level of B2.

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 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 has 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 below, where we determined this tweet as not abusive:

> [Example of indirect insult.]
> *@USER Everyone saying **fuck** 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., $< b >$fuck$< /b >$). 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:

	Original OLID	**Abusive**	**Not-Abusive**
Offensive	1,296	568	728
Not	215	52	163
Total	1,511	620	891

Table 2: Label distribution in the SWAD dataset.

> [Ex. i): offensive tweet & abusive swearing]
> *@USER You are an absolute **dick*** 😡

> [Ex. ii): offensive tweet & not abusive swearing]
> *@USER I was definitely drunk as **shit***

> [Ex. iii): not offensive tweet & abusive swearing]
> *@USER **bullshit** there's rich liberals too so what are you saying ???*

> [Ex. iv): not offensive tweet & not abusive swearing]
> *@USER Haley thanx! you know how to brighten up my **shitty** day* 😘

4. 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 1,511 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 IAA is 0.708 (Cohen's kappa coefficient), which corresponds to a substantial agreement. The final SWAD annotated corpus consists of 1,511 unique swear words immersed in the context of 1,320 tweets, where 620 swear words are marked as abusive and 891 are rated as not-abusive [8]. 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.

We also extracted the top ten swear words for both classes (abusive and not-abusive) from SWAD, as shown in Table 3. We calculated the percentage of swear word use in each class and the percentage of each swear word used over both classes. We can see that the top 3 words on both classes are the same (not in order), including *fuck*, *ass*, and

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

	Abusive			Not-Abusive		
No.	Swear Words	WP on Class (%)	WP over Classes (%)	Swear Words	WP on Class (%)	WP over Classes (%)
1.	shit	19.68	33.80	shit	26.82	66.20
2.	ass	11.45	39.66	fuck	13.02	63.74
3.	fuck	10.65	36.26	ass	12.12	60.34
4.	bitch	10.00	63.27	fucking	11.56	71.03
5.	bullshit	7.26	81.82	hell	5.95	69.74
6.	fucking	6.77	28.97	damn	5.27	81.03
7.	hell	3.71	30.26	bitch	4.04	36.73
8.	asshole	3.06	90.48	nigga	2.92	83.87
9.	shitty	75.00	2.42	gay	2.36	77.78
10.	pussy	56.00	2.26	fucked	2.24	66.67
...	...	...	...	...	...	...
	cunt	1.13	100	cunt	0	0
	whore	1.13	87.50	whore	0.11	12.50

Table 3: Top ten swear words in each class.

shit. The percentage of these common swear words over both classes are relatively balanced. It means that the abusiveness of common swear words could not be resolved only based on the word choice, but it needs context. However, for some swear words we can observe that their usage is more inclined to abusive contexts such as *cunt*, *whore*, *bullshit*, and *asshole*, despite their presence is not significant in our corpus. Meanwhile, swear words such as *damn*, *nigga*, and *gay* seems to be more inclined to not-abusive uses.

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

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, while emphatic swearing is another swear word function when a swear word is used to emphasize another word in order to draw more attention. Two examples, one for each swearing function mentioned, follow:

> [Cathartic function]
> *@USER* ***damn*** *I felt this shit Why you so loud lol*

> [Emphatic function]
> *@USER I AM FUCKING SO* ***FUCKING*** *HAPPY*

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 the following example, without the presence of the emoji, the swear word *fucking* seems to contribute to the 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.

> *@USER ur a* ***fucking*** *dumbass fr. there's no way she is anyone else's* 😂

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 follows:

> *@USER Yeah we need some more made up* ***bullshit*** *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:

Missing context. We found some tweets are very short, resulting in the context missing. Other instances are also challenging to understand due to the presence of grammatical errors. These issues are very dominant in the annotator disagreement cases. In the following we show two examples where the context is hard to resolve:

> [Very short tweet]
> *@USER Lmfaoo!* 😭 ***bitch***

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

Need of world knowledge to understand the context. Some tweets are also very difficult to understand due to the lack of world knowledge. 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. Let us see an example for this issue:

> *@USER @USER It's probably better to have an* ❌❌ *next to my name than a pink* ***pussy*** *hat on my head* 🤔🤔🤔🤔 *#MAGA #MakeAmericaGreatAgain*

5. Experiments

In this section, we provide an intrinsic evaluation of the corpus by conducting cross-validation experiments. We build supervised machine learning models to predict the abusiveness of swear words in SWAD. We model this prediction task as two different tasks, namely sequence labeling and text classification. The main objective of the sequence labeling experiment is to test the consistency of the annotation of the corpus, while we devised the classification experiment to shed some light on the most predictive feature to differentiate between abusive and not-abusive swearing.

5.1. Sequence Labeling Task

In order to test the robustness of the annotation of swear words in SWAD, we devised 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%).

5.1.1. Model Description

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.

5.1.2. Results

predicted ground truth	SWNA	SWA	NSW
SWNA	59	25	0
SWA	19	45	1
NSW	2	4	2,764

Table 4: Sequence labeling task: confusion matrix.

	precision	recall	F_1-score
SWNA	.737	.702	.719
SWA	.608	.692	.647
NSW	.999	.997	.998
macro avg	.781	.797	.788

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

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 (25 SWA→SWNA misclassifications) than non-abusive swear words (19 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 *skank*, 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: *shitcago* (an offensive slang for Chicago), *messed*, *cumming*, and *cumslave*.

5.2. 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).

5.2.1. 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 [9] 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 and Gilbert, 2014) to extract the polarity score of the text and *emoji sentiment ranking* [10] to get the sentiment value for emojis.

Emotion Features - Similar to the sentiment features, this feature is used to explore the context of the tweet, under the hypothesis that there is a relation between swear word use and the emotional state of the author of the tweet. We use two available affective resources. The first is Emolex (Mohammad and Turney, 2013), a crowdsourced lexicon containing 14,182 words associated with eight primary emotions based on the model by Plutchik (2001): joy, sadness, anger, fear, trust, surprise, disgust, and anticipation. We extracted eight individual features representing the emotion categories of the words. In addition, we also use EmoSenticNet (Poria et al., 2013), an enriched version of SenticNet (Cambria et al., 2014) including 13,189 words labeled by the six basic emotions from Ekman (1992). Therefore, we have 14 emotion features in total, eight from Emolex and six from EmoSenticNet.

[9] `https://pypi.org/project/emoji/`

[10] `http://kt.ijs.si/data/Emoji_sentiment_ranking/`

Feature Set	Feature Set	LSVC P	LSVC R	LSVC F_1	LSVC Acc	LR P	LR R	LR F_1	LR Acc	RF P	RF R	RF F_1	RF Acc
A	Unigram SW	.691	.382	.489	.675	.691	.382	.489	.675	**.708**	.350	.463	.672
B	A + Bigrams	.673	.448	.537	.684	.674	.432	.525	**.680**	.665	.421	.513	**.674**
C	A + Twitter	.695	.374	.483	.674	**.698**	.374	.483	.675	.670	.366	.469	.663
D	A + Sentiment	**.697**	.373	.481	.674	.689	.369	.477	.671	.654	.408	.498	.666
E	A + Emotion	.660	.410	.502	.668	.656	.423	.512	.670	.549	.387	.452	.617
F	A + Stylistic	.675	.379	.482	.669	.676	.379	.483	.670	.515	.360	.423	.598
G	A + Syntactic	.637	.390	.481	.658	.639	.397	.487	.660	.548	.469	.504	.622
H	B + Twitter	.678	.448	.538	**.686**	.666	.406	.503	.671	.652	.374	.472	.659
I	B + Sentiment	.672	.435	.526	.680	.687	.415	.514	**.680**	.595	**.484**	.532	.653
J	B + Emotion	.642	.469	.540	.673	.649	.445	.526	.672	.551	.368	.440	.619
K	B + Stylistic	.659	.439	.525	.676	.650	.389	.484	.662	.584	.418	.486	.639
L	B + Syntactic	.620	.437	.512	.658	.626	.418	.500	.658	.567	.444	.496	.632
M	H + Sentiment	.658	.434	.520	.673	.679	.419	.516	.678	.606	.479	**.534**	.658
N	H + Emotion	.645	.473	.543	.674	.625	.466	.532	.664	.502	.284	.361	.592
O	H + Stylistic	.664	.442	.529	.678	.652	.392	.488	.663	.553	.348	.426	.617
P	H + Syntactic	.617	.429	.505	.655	.627	.415	.497	.658	.526	.418	.464	.607
Q	M + Emotion	.645	.473	.543	.674	.647	.465	.539	.675	.675	.408	.487	.649
R	M + Stylistic	.662	.453	.535	.679	.662	.426	.515	.674	.629	.405	.491	.657
S	M + Syntactic	.627	.453	.525	.664	.637	.424	.507	.664	.559	.397	.462	.623
T	Q + Stylistic	.634	.484	.547	.672	.646	.465	.538	.678	.559	.321	.406	.617
U	Q + Syntactic	.618	.487	.544	.665	.635	.477	.543	.671	.584	.347	.431	.629
V	All Features	.626	.494	.550	.669	.627	.481	.542	.668	.558	.316	.402	.617
	V - Unigram SW	.529	.381	.439	.607	.538	.374	.438	.610	.521	.303	.381	.600
	V - Bigrams	.625	.440	.515	.661	.619	.453	.521	.660	.575	.345	.430	.626
	V - Twitter	.630	**.495**	**.553**	.672	.644	**.485**	**.551**	.677	.572	.321	.410	.623
	V - Sentiment	.605	.460	.521	.654	.625	.463	.531	.664	.523	.287	.369	.601
	V - Emotion	.629	.456	.527	.666	.635	.437	.515	.666	.571	.316	.405	.623
	V - Stylistic	.618	.487	.544	.665	.635	.477	.543	.671	.584	.347	.431	.629
	V - Syntactic	.634	.484	.547	.672	.646	.465	.538	.678	.559	.321	.406	.617

Table 6: Ablation test on several feature sets.

Stylistic Features - In this feature set, we consider several common stylistic features for text classification task such as **Capital Word Count**[11], **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[12], while dependency relations are extracted with SpaCy[13].

5.2.2. System Description and Evaluation

We build our models by using the Scikit-learn library[14]. The performance is evaluated based on 10-fold cross-validation on the whole dataset. We use several evaluation metrics, including accuracy, precision, recall, and F_1-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.

5.2.3. Results

Table 6 shows the full results of the text classification experiment by using LSVC, LR, and RF models. We add each feature incrementally during the experiment, by using the unigrams of swear words as the initial configuration. We notice that bigrams and emotion features provide a significant improvement on the classification performance (feature sets B and E). Overall, RF is under-performing compared to the two other classifiers. LSVC performes slightly better than LR. Based on F_1-score, the best performance is achieved by using all the features except Twitter features[15] with the LSVC model. In the ablation experiment, our goal is to investigate the most predictive feature set by removing one feature set at a time. We found that unigram of 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 LR models. The main issue of this task is the very low recall, which denotes that such models struggle to deal with false-negatives. We argue that this happens due to the dataset imbalance: as shown in Table 3 the swear words percentage over both classes is dominated by not-abusive class (negative class).

[11] This feature consider all capital words on the tweet

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

[13] https://spacy.io/

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

[15] The effect of the Twitter features is however more neutral than detrimental, as shown in lines 1, 3 and 21, 23 of Table 6

6. Conclusion and Future Works

The research presented in this paper investigates the automatic classification of abusive swearing. In this direction, we developed a new benchmark corpus called SWAD, consisting of English tweets, where abusive swearing is manually annotated at the word-level. Our final corpus consists of 1,511 instances of swearing from 1,320 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.

We also built models trained on the SWAD corpus, to automatically classify abusive and not-abusive swear words, and to provide an intrinsic evaluation of SWAD. We experimented by modeling this task into two different settings, namely, sequence labeling and text classification. We used BERT for sequence labeling, and simpler but more transparent models for text classification. Our results confirm that our annotation is robust based on the sequence labeling performance. On the other hand, text classification results provided new insights on 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.

While these results are encouraging, we believe that there is still room for improvement for both the corpus and the automatic classification of swearing. We plan to extend the size of the SWAD corpus, both in its sheer size, in order to be able to train neural-based models, and in breadth, that is, covering different domains and genres, to improve the analysis of rare and domain-specific insults. 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 employ the corpus presented in this work in the context of abusive language detection tasks, to tackle the false positive issue caused by the presence of swear words (Chen et al., 2012; Nobata et al., 2016; Van Hee et al., 2018; Malmasi and Zampieri, 2018), at the same tome providing a further extrinsic evaluation of the SWAD corpus.

Applying our methodology to other languages is not trivial, as it depends on the availability of language resources and robust NLP tools for them. 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). Finally, 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.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

7. Bibliographical References

Bak, J., Kim, S., and Oh, A. H. (2012). Self-disclosure and relationship strength in twitter conversations. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 60–64. The Association for Computer Linguistics.

Bassignana, E., Basile, V., and 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*.

Bianchi, C. (2014). Slurs and appropriation: An echoic account. *Journal of Pragmatics*, 66:35 – 44.

Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In Carla E. Brodley et al., editors, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1515–1521. AAAI Press.

Chen, Y., Zhou, Y., Zhu, S., and 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)*, pages 71–80. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dinakar, K., Reichart, R., and 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*, volume WS-11-02 of *AAAI Workshops*. AAAI.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

EU Commission. (2016). Code of conduct on countering illegal hate speech online.

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

Gauthier, M., Guille, A., Deseille, A., and 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., and Wu,

D. M. (2017). A large labeled corpus for online harassment research. In Peter Fox, et al., editors, *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 229–233. ACM.

Holgate, E., Cachola, I., Preoţiuc-Pietro, D., and 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*, pages 4405–4414, Brussels, Belgium, October-November. Association for Computational Linguistics.

Hutto, C. J. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Eytan Adar, et al., editors, *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.

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

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

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.

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.

Malmasi, S. and 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*. London: Routledge.

Mehl, M. R. and 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.

Mohammad, S. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

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., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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

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

Plutchik, R. (2001). The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and 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*.

Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., and Bandyopadhyay, S. (2013). Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38.

Rieber, R. W., Wiedemann, C., and 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.

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

Stephens, R. and 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., and 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)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2014). Cursing in English on twitter. In Susan R. Fussell, et al., editors, *Computer Supported Cooperative Work, CSCW '14, Baltimore, MD, USA, February 15-19, 2014*, pages 415–425. ACM.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and 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)*, pages 1415–1420, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and 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*, pages 75–86, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.