I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language

Tommaso Caselli♣, Valerio Basile⋄, Jelena Mitrović‡, Inga Kartoziya♣, Michael Granitzer‡
♣University of Groningen, ⋄University of Turin, ‡University of Passau
Groningen The Netherlands, Turin Italy, Passau Germany
⋄ valerio.basile@unito.it, ♣t.caselli@rug.nl, ♣i.kartoziya.1@student.rug.nl
‡{jelena.mitrovic|michael.granitzer}@uni-passau.de

Abstract

Abusive language detection is an unsolved and challenging problem for the NLP community. Recent literature suggests various approaches to distinguish between different language phenomena (e.g., hate speech vs. cyberbullying vs. offensive language) and factors (degree of explicitness and target) that may help to classify different abusive language phenomena. There are data sets that annotate the target of abusive messages (i.e.OLID/OffensEval (Zampieri et al., 2019a)). However, there is a lack of data sets that take into account the degree of explicitness. In this paper, we propose annotation guidelines to distinguish between explicit and implicit abuse in English and apply them to OLID/OffensEval. The outcome is a newly created resource, AbuseEval v1.0, which aims to address some of the existing issues in the annotation of offensive and abusive language (e.g., explicitness of the message, presence of a target, need of context, and interaction across different phenomena).

Keywords: Corpus Annotation, Social Media Processing, Statistical and Machine Learning Methods

1. Introduction

Social media platforms have been promoted as (virtual) places where many people from different parts of the world can engage in productive discussions and share opinions (Jurgens et al., 2019). At the same time, quoting Umberto Eco, "everyone who inhabits the planet, including crazy people and idiots, has the right to the public word [and], on Internet, your message has the same authority as the Nobel laureate [...]." [1] Alongside such flattening in the process of production, consumption and sharing of information, toxic and abusive behavior online has surged.

Recently, there has been an increasing effort from the Natural Language Processing (NLP) community to develop methods for the automatic detection of abusive language and related phenomena, as the volume of data is such that it has become impossible to manually track and monitor it (Nobata et al., 2016; Kennedy et al., 2017). This has taken different forms such as the creation of data sets or corpora in multiple languages (Waseem and Hovy, 2016a; Ross et al., 2017; Poletto et al., 2017; Founta et al., 2018; Ibrohim and Budi, 2019), [2] the promotion of evaluation campaigns (Kumar et al., 2018; Wiegand et al., 2018b; Bosco et al., 2018; Zampieri et al., 2019b; Basile et al., 2019), the organization of thematic workshops and conferences, [3] and the compilation of surveys (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018).

However, research in this area is still fragmented and the field has been flooded with different terminologies, perspectives and understandings of this phenomenon, somehow lacking of consensus on shared definitions. Waseem et al. (2017) clearly point out how such lack of consensus in the definitions has resulted in contradictory guidelines and annotations. To address these issues and differentiate across subtypes, they propose a typology that is built upon two factors: (i.) whether the abusive language is directed towards a specific target (an individual, a group, or an entity); and (ii.) degree of explicitness of the abusive language, i.e., the extent to which the abusive message is unambiguously perceived as abusive, without the need to decipher hidden meaning such as from rhetorical devices. Developing annotation schemes on the basis of these two factors is suggested as a way to better distinguish different phenomena (e.g., hate speech vs. cyberbullying vs. offensive language) and it may help to reconcile definitions and reach a better consensus in the community.

In this contribution, we investigate a recent data set for offensive language in English, namely OLID/OffensEval (Zampieri et al., 2019a; Zampieri et al., 2019b), in the light of the two factors previously illustrated. Although OLID/OffensEval is among few data sets that explicitly take the information about targets into account, we will show that the data suffer from a lack of attention on the explicitness parameters resulting in debatable, though coherent, annotations. In particular, we annotate the distinction between explicit and implicit messages in OLID/OffensEval, enriching the data with a complementary annotation layer. In addition to this, we also propose newly developed annotation guidelines and apply them on OLID/OffensEval. The outcome is a newly created resource for English, called AbuseEval v1.0, that targets some of the pending issues in the annotation of offensive/abusive language (e.g., explicitness of the message, presence of a target, need of context, and interaction across different phenomena). [4]

[1] https://www.elmundo.es/cultura/2015/03/26/551385fc22601dfd398b456b.html

[2] For a more comprehensive overview see http://hatespeechdata.com

[3] https://sites.google.com/view/alw3/

[4] The enriched OLID/OffensEval and AbuseEval v1.0 are available https://github.com/tommasoc80/AbuseEval

2. OLID: An In-Depth Analysis

OLID (Zampieri et al., 2019a), Offensive Language Identification Dataset, has been introduced in the context of the SemEval 2019 shared task on offensive language detection (OffensEval, Zampieri et al. (2019b)). This data set is a collection of English tweets. The most innovative aspect of it is the annotation of the target of an offensive message. As it was mentioned above, Waseem et al. (2017) introduced target as one of the factors in their proposed abusive language typology. OLID/OffensEval has been created by applying a hierarchical annotation scheme distinguishing three different levels (subtasks A, B and C):

- whether a message is offensive or not (A);
- whether the offensive message has a target or not (B);
- whether the target of the offensive message is an individual, a group, or other (i.e., an organization, an event, an issue, a situation) (C).

The annotation has been conducted via Figure Eight, a crowdsourcing platform[5]. Data quality was ensured by selecting only experienced annotators and using test questions to discard individuals not reaching a minimum reliability threshold. Messages were retrieved using a keyword approach through Twitter API. As reported in Zampieri et al. (2019a), the authors selected political and non-political keywords, refining their list after a trial annotation. The largest amount of offensive material (58%) resulted from the Twitter "safe" filter (i.e., messages already flagged by twitter as unsafe).

We focus on sub-task A, i.e, whether a message is offensive or not, and we break our analysis into two blocks, namely (i.) basic properties (§ 2.1.); and (ii.) keywords (§ 2.2.).

2.1. Basic Properties

Our analysis of the basic properties of OLID consists of two parameters (besides the raw number of messages per class) – average message length and *offensive prior* (described later in this section). To compute these statistics, we pre-processed the data by removing hashtag symbols and tokenizing the messages using the NLTK Tweet tokenizer. We did not split hashtags composed by multiple words in separate tokens. In Table 1 we report the results for both training and test distributions.

Table 1: OLID statistics per class: number of messages, average message length in tokens, average Offensive Prior. Asterisks mark statistical significance differences ($p < 0.05$). OFF = offensive; NOT = not offensive.

Class	Stats	Train	Test
OFF	# messages	4,400	240
	Avg. Length (token)	24.88*	25.91
	Offensive Prior (avg.)	0.2547*	0.2306*
NOT	# messages	8,840	620
	Avg. Length (token)	21.90	28.10
	Offensive Prior (avg.)	0.0614	0.0370

[5] https://www.figure-eight.com/

The distribution of the messages in the two classes reflects a *de facto* standard in the creation of offensive or abusive language data sets, whereby the negative class, i.e., the non offensive messages, actually represent the majority of the data (in this case, 67%). Such a distribution is interpreted as an attempt to mirror the distribution of offensive/abusive messages on social media platforms.

The average message length gives a quick overview of the distribution of the data in the two classes. We can observe that both in training and test, offensive messages (OFF) are actually quite long (24.88 tokens for training and 25.91 tokens for test, respectively), and comparable with the non offensive ones (NOT). However, standard deviations suggest that the distribution of the messages is quite skewed in both training and test for both classes, having values between 15.73 (NOT) and 16.52 (OFF) for training, and 15.31 (NOT) and 15.94 (OFF) for test. Furthermore, we observe that the most frequent message length in the training distribution is comparable between the two classes, namely 7 tokens (138 cases) for OFF and 6 tokens (358 cases) for NOT. On the contrary, we find that in the test distribution the most frequent message length is actually higher: 9 tokens for OFF (11 cases) and 19 tokens for NOT (18 cases) [6]. Finally, we observe that length difference between offensive and not offensive messages in training is statistically significant ($p<0.05$; Mann-Whitney test) while this does not occur in the test data.

The second parameter, on the other hand, assesses the offensive prior, or offensiveness prior score, of messages per class. The score is calculated by means of a weighted offensive dictionary (Wiegand et al., 2018a), [7] and is inspired by previous work on bias identification in abusive language data sets (Wiegand et al., 2019). The score has been obtained as follows: first, we stemmed the messages and the entries in the dictionary to increase the possibility of finding a token; secondly, we calculate the offensive prior by averaging the dictionary scores of all matched items in the dictionary. In this case, differences between offensive and not offensive messages, both in the training and in the test distribution, clearly emerge. Furthermore, the difference in offensive prior is statistically significant ($p<0.05$; Mann-Whitney test) in both distributions.

2.2. Keywords

We extracted the top 50 keywords per class per distribution by applying a tf-idf approach. In Table 2 we report the top 10 keywords, due to space limitations. [8] As the table illustrates, and as the manual investigation confirms, offensive messages show a higher number of profanities, racial and sexist slurs than the not offensive ones. However, the distinction appears to be slightly less clear-cut than what was expected, as some slurs and profanities also appear in messages labelled as not offensive.

The trend emerging from this analysis is that the keyword

[6] Notice that the top five most frequent length messages in test for the OFF are: 9 tokens (11), 16 tokens (10), 12 tokens (9), 13 tokens (9) and 15 tokens (9)

[7] We have used the extended version of the dictionary

[8] The whole list is available at https://github.com/tommasoc80/AbuseEval

Table 2: OLID top 10 keywords per class

Class	Train	Test
OFF	fuckbucket	davidhogg
	pornhub	bitch
	ostrich	female
	dickmatized	fuck
	bunk	clown
	hungery	oh
	batting	potus
	dickhead	extremely
	dillusional	racist
	fuckass	5k
NOT	austria	nickidagoat
	follback	fucking
	fluffy	revolting
	kingggg	literally
	dd	titty
	eggman	irish
	burger	sam
	lmfaoooooo	muslim
	darling	ripmacmiller
	razzinfrazzinmaggle	pink

approach used to retrieve potentially offensive messages seems to be somehow biased towards *explicit* expression of offense, i.e., words or phrases that are unambiguous in their potential to be offensive (Waseem et al., 2017).

3. Explicit or Implicit?

Among the participants to SemEval 2019 Task 6: OffensEval, the Duluth system (Pedersen, 2019) is particularly interesting, being a very competitive dictionary-based approach to distinguish whether a message is offensive or not (sub-task A). In particular, a list of 563 profanity words (or black-listed words) has been created by merging together different sources of information, such as terms in the offensive messages of the OffensEval training data that occur five times or more, terms in the hateful messages of the HatEval training data that occur five times or more, and blacklists found online. The final approach is very simple: if the message contains one or more of the words in the 563 word list, it is considered as offensive.

We re-implemented the Duluth's approach using again Wiegand et al. (2018a)'s extended dictionary. In this case, we decided not to use all words in the dictionary but rather a subset of highly potentially offensive and abusive words by empirically setting an offensiveness threshold at 0.75, for a total of 861 terms. Following Pedersen (2019)'s approach, we have marked any message that contains one or more offensive words above the threshold as offensive. Similarly to the computation of the offensive prior, we stemmed the tokenized data both in the messages and in the dictionary. Table 3 reports the results of our dictionary-based approach (Dictionary) in terms of precision (P), recall (R) and macro-averaged F1 as in the OffensEval shared task. It is compared to the best system at OffensEval — NULI (Liu et al., 2019), a model based on fine-tuning BERT (Devlin et al., 2019) using OLID; Duluth and a linear SVM baseline introduced by the task organizers (Zampieri et al., 2019a).

The results of our dictionary-based approach are very competitive and point in the same direction as the Duluth system, showing that OLID/OffensEval is very rich in messages that contain explicit markers of offensiveness. To have a better overview of the actual amount and instances

Table 3: OffensEval - Test: Evaluation of dictionary-based system and comparison against NULI, Duluth, and SVM baseline.

Approach	Class	P	R	F1 (macro)
Dictionary	NOT	.836	.872	.722
	OFF	.629	.558	
NULI	NOT	.902	.908	**.828**
	OFF	.758	.745	
Duluth	NOT	.832	.900	.735
	OFF	.673	.533	
SVM	NOT	.800	.920	.690
	OFF	.660	.430	

of explicit offensive messages in the data set, we performed a manual annotation.

We considered only the messages originally labelled as offensive in both the training and test data, and we applied a straightforward annotation rule: if a message contains a profanity or a slur, we consider it as an explicit instance; otherwise, messages are annotated as implicit. Given that explicit markers of offensiveness are in part culturally dependent, and that none of the authors is a native English speaker nor is based in any of the Anglo-Saxon speaking countries, cases of doubts were addressed by making use of an online resource, namely Urban Dictionary [9], a crowd-sourced online dictionary for slang words and phrases. For any word or expression the annotators did not know or were not familiar with, all associated meanings were checked.

The use of Urban Dictionary has helped to clarify doubts and to take non-ambiguous decisions, i.e., decisions not dependent on the personal annotator background, on how to treat special constructions such as abbreviations. Abbreviations in the context of instant messaging or microblogging can be used either to save space, thus substituting long phrases: for instance, "*as far as I know*" is commonly abbreviated with "afaik"; or to mask negative expressions, profanities, or slurs. In this annotation task, whenever an abbreviation is found in Urban Dictionary and any of its meaning denotes an expression of an offense, abuse, or hate, the whole message is marked as explicit. To better illustrate our decisions, consider the following examples where we have marked abbreviations in bold:

1. @USER @USER **F** Hope! Since he's 'NOT smarter than a 5th grader' maybe she is! [OffensEval:train id: 36226]

2. @USER @USER I'm sorry but damn she is hot **af** [OffensEval:train id: 26674]

In example 1 the use of *F* is not considered as an abbreviation that may trigger an explicit offense. Although *F* exists as an entry in Urban Dictionary, its meaning is highly positive as it is used in the context of online gaming to pay respect to gamers who got killed[10]. On the other hand, the abbreviation *af* is attested as the abbreviation of a well-known phrase that is offensive and potentially abusive, as

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

[10] https://knowyourmeme.com/memes/press-f-to-pay-respects

in this case. However, in this case, even without the expression *af*, the message would still be considered offensive due to the presence of *damn*.
Table 4 illustrates the results of our annotation on the offensive messages in the training and test distributions.

Table 4: OffensEval: Explicit vs. Implicit offensive messages. EXP = EXPLICIT; IMP = IMPLICIT.

Data distribution	Class	Messages
Train	EXP	2,901
	IMP	1,499
Test	EXP	154
	IMP	86

Table 4 shows that more than 65% of the messages in the training distribution contain at least one token that explicitly encodes an offense, while in the test distribution this occurs in 59% of the cases. Furthermore, by assuming a "perfect" dictionary for markers of explicit offense/abuse, i.e., a dictionary with all profanities and slurs with all possible spelling variations, the results of a dictionary-based system would reach an F1 of .783 for the offensive messages, which would correspond to an increase of 3.17 points when compared to the results of the best participating system at OffensEval (i.e., NULI).

4. Abusive or Offensive?

The annotation of the explicit-implicit dimension of the offensive data raised questions on the phenomenon represented in OLID/OffensEval, in the lights of previous annotation initiatives, namely Founta et al. (2018), existing definitions of offensive language and recent discussions in the NLP community (Jurgens et al., 2019; Vidgen et al., 2019).
Offensive language is a phenomenon closely interconnected with a number of other linguistic and societal phenomena, including: abusive and aggressive language, cyberbullying, racism, extremism, radicalization, toxicity, profanity, flaming, discrimination, hate and hate speech. We want to focus on the distinction between **abusive** and **offensive** language, in order to better understand the relationship between the two phenomena and create better language resources for both of them.
Abusive language is defined in popular English dictionaries as "extremely offensive and insulting; engaging in or characterized by habitual violence and cruelty." (Oxford English Dictionary, 2019) and "using harsh, insulting language; harsh and insulting abusive language; using or involving physical violence or emotional cruelty"[11] (Merriam-Webster Online, 2009). Founta et al. (2018) define abusive language as "used to refer to hurtful language, including hate speech, derogatory language and also profanity", while Fortuna and Nunes (2018) summarize the previous definitions by Papegnies et al. (2017), Park and Fung (2017), and Nobata et al. (2016) into the following: "any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion." The use of the words *insulting* and hurtful across these definitions points out at a strong component of intentionality intrinsic in this phenomenon.
By contrast, offensive language is defined as "causing someone to feel resentful, upset, or annoyed; actively aggressive; attacking" (Oxford English Dictionary) and "causing displeasure or resentment."[12] (Merriam-Webster Online, 2009). Fortuna and Nunes (2018) provide a synthesis of definitions by Chen et al. (2012) and Razavi et al. (2010) as "profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group." In the context of OLID and the OffensEval shared task, offensive language is defined as containing "any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct." (Zampieri et al., 2019b). We note that these definitions emphasize more the lexical content (profanity) and the emotional response (feeling resentful, upset, or annoyed) of the receiver, rather than the intentionality of the producer.
The annotation study reported in Founta et al. (2018) has shown that categories such as *aggressive*, *abusive*, and *offensive* tend to be significantly correlated, highly coexisting, and very similar (where similarity is defined by means of cosine distance across vectorizations of the messages). As a result, in the final annotation of their data set, the authors merge these categories and select **abusive** as the final label. [13] Although the decision has been justified on the basis of experimental data, it indicates that different phenomena are collapsed and confounded.
Considering the definitions described above, we further notice that the definition of **offensive** language in OLID/OffensEval covers two cases: (i.) containing non-acceptable language; and (ii.) targeted offense. While the messages from the second case, are closely related to abusive language ("targeted" can be interpreted as the *intention* to offend/debase someone or a group), messages pertaining to the first case might be not necessarily abusive. Additional information on this distinction is conveyed by the sub-task B, in which participants are asked to categorize the offense type as either *targeted* (TIN) or *untargeted* (UNT). According to our interpretation, messages that contain abusive language should all be labeled TIN, while messages containing non-acceptable language, or perceived as being offensive, could either be abusive and TIN, or not abusive, and therefore UNT.
Recently Vidgen et al. (2019) have pointed out that the sensibility of (online) audiences when it comes to offensive language may affect definitions (and consequently annotated data) and may "inevitably" (Vidgen et al., 2019, 83) lead to some mischaracterization. Indeed in some contexts of occurrence the same message may be perceived as offensive, while in others it may be perfectly safe. This may apply to abusiveness as well, and, particularly, to the distinction between messages whose content is inherently abu-

[11] https://www.merriam-webster.com/dictionary/abusiveness

[12] https://www.merriam-webster.com/dictionary/offensive

[13] The choice of the label is done on the basis of a further annotation study.

sive (and supposedly mainly explicit) and messages whose effect is perceived as abusive (and supposedly mainly implicit). Taking into account the ideas from Waseem et al. (2017), we believe that the analysis of Vidgen et al. (2019) represents a strong motivation to undergo a systematic, empirical study of the explicit/implicit distinction in offensive and abusive language. Finally, we focus on abusive language because it is a hub of other phenomena that may have a very strong negative impact in society.

4.1. The AbuseEval Data Set

Our annotation experiment is an attempt to take into account current reflections and pending issues in the creation of data sets for abusive language detection into a unified framework. In particular, we focus on the following aspects:

- a comprehensive definition of abusive language;
- a distinction between the explicit and implicit levels of expressions of abusive language;
- context of occurrence as a necessary factor for the correct understanding of the abusive content of a message.

Furthermore, we have applied our proposal to an existing and popular data set (OLID/OffensEval) in order to test which suggested parameters are reflected and present in the data. Our aim is not to criticize or reproach the effort and the creators of the data set.

We define abusive language as *hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions*. This might include hate speech, derogatory language, profanity, toxic comments, racist and sexist statements. In this definition, the speaker targets other people, but not himself/herself. Furthermore, they do not quote anyone or present potential abusive content as a statement.

Our definition of abusive language is more comprehensive with respect to those reported in literature, and it attempts to characterize relationships among the various phenomena. We have envisioned each phenomenon as a potential class, and then, we have structured our definition in terms of relationships among classes by taking into account subsumption, intersections, and disjoint relations. In our definition, for instance, the class of offensive language partially overlaps (i.e. non-empty intersection) with abusive language. In particular, we consider only directed offense to targets, being either individuals or groups. As a consequence, we explicitly exclude generic expressions (e.g., use of a profanity in isolation) or untargeted messages from our positive class as their use is not easily interpretable.

We propose a three-way annotation rather than the more common binary one. We are not simply interested in distinguishing between abusive or not abusive messages, but we want to address the degree of explicitness of the abusive message. The access to such information is highly important to reduce bias and dependency of systems on lexical cues in the data. Furthermore, we intend to make available such distinctions, to strengthen the generalisation ability and portability of systems, across different distributions of the same domain, if not also across different platforms. We define our labels as follows:

(a.) **Explicit** (abuse): it always has a surface evidence of abuse with respect to a target by means of profanity, performative constructions, imperatives, idioms, adjectives or nouns with a clear negative connotation (for instance, examples 5, 6, 8);

(b.) **Implicit** (abuse): it does not have any surface evidence, abuse can only be suggested or inferred. It can be hidden with sarcasm, metonymy, irony, litotes, euphemism, and inside jokes among other linguistic devices (for instance, examples 12 13, 14);

(c.) **Not** (abusive): this class is used for messages that are not interpretable without additional information about their context of occurrence, or that are actually not abusive.

The definition of the negative class (i.e., **Not** (abusive)) highlights a difference with respect to literature: we make an explicit reference to interpretability with respect to the context of occurrence. We see this requirement as a strategy to reduce biases in the data, especially when related to a community-accepted jargon or dialect (Sap et al., 2019; Davidson et al., 2019). For instance, consider the following messages:

3. @USER I miss you bitch!! [OffensEval:train id:85858]
4. @USER Nigga we're going next week [OffensEval:train id:72880]

Both examples have been marked as **Not** (abusive) because of lack of the context of occurrence. In particular, in both cases it is not possible to decide whether the messages are threats to individuals (both signaled by "@USER") or acceptable expressions in specific communities.

4.2. Annotation Guidelines

The annotation guidelines have been formulated in terms of a decision tree. [14] In this way, we can reduce subjective interpretation from the annotators and, at the same time, help to reconstruct how the annotation decisions have been made, facilitating a more transparent annotation process. However, the decision tree is not exclusive and multiple markers of abusive language could be present in the same message.

The first distinction we make is whether a message is an *utterance* or not. We consider utterances uninterrupted sequences of spoken or written language, grammatical, and fully intelligible, i.e., expressing a meaning and an intention (Grice et al., 1975; Strawson, 1964). In other cases,

[14] Available at `http://bit.ly/2PsWRJJ`

and especially in social media platforms, messages are actually composed by unconnected words, hashtags, or unfinished thoughts. These latter cases are not considered utterances and thus are excluded from the annotations. To clarify the difference, consider the following examples, where the first two messages are considered utterances while example 7 is not. Although, this message may express a meaning, the intention (i.e., being abusive with respect to a target) is not clear, leaving the message ambiguous. [15]

5. @USER @USER @USER He. Is. A. Sociopath. They are incapable of feeling empathy. Period. [OffensEval:train id:10414]

6. #Spanish #unjustice vs. #FreedomOfExpression and #HumanRights #Spain is a #fakedemocracy [OffensEval:train id:49138]

7. @USER you fucking - [OffensEval:train id:31853]

Notice that even messages that do not qualify as utterances can still be carriers of abusive language. In the context of Twitter messages, as it is in our case, annotators should focus their attention on the presence of hashtags that encode abusive language, otherwise the message should be labelled as not abusive. On the basis of our experience in the annotation of explicit and implicit messages in the original OffensEval data set, messages that carry an abusive content only on the basis of their hashtags must be considered as explicit.

A further distinction concerns the presence of quotes in a message. Although the content of a quote can be abusive, this does not make the message automatically abusive. Quotes can be used to report someone's opinion and the author of the message can use it either to show support or express disagreement. In cases when the abusive content is only inside a quote and there is no other cue of an agreement of the author with respect to that content, we consider the message as not abusive. In all other cases, additional checks must be conducted in order to determine whether the message is ether explicit, implicit, or not abusive.

Profanities are considered as markers of explicit abuse when used in a negative context to debase or target an individual or a group, as illustrated in the following example:

8. #ThursdayThoughts- FUCK liberals. Forever. [OffensEval:train id:77089]

Further ways of expressing abuse in an explicit way have been identified in the use of performative construction (example 9) or use of imperatives (example 10) in negative contexts. In both examples, we have underlined the target construction.

9. @USER @USER SHE IS A FUCKING MESS!! <u>I HATE HER SO MUCH</u> [OffensEval:train id:94169]

10. @USER <u>Go to hell</u>! This is NOT Queen for a Day. I believe you less and less with every bit of bullsh*t you pull. You're nothing but a lying Demonrat! #MAGA #Trump2020 [OffensEval:train id:77392]

As far as implicit abuse is concerned, we have identified mainly linguistic constructions involving sarcasm, irony and rhetorical questions:

11. 4 out of 10 British people are basically full-on racists. 4 out of 10 voters vote for the Conservatives. Coincidence!???!??? [OffensEval:train id:54991]

12. @USER @USER Oh you are in England? Your views on gun control stopped mattering in 1776. [OffensEval:train id:54991]

13. @USER @USER Wonder how many children he molested [OffensEval:train id:97580]

14. @USER Isn't the coalition for gun control headed up by the lady who was turned down for a job because she was a bully? [OffensEval:train id:17971]

4.3. Inter-Annotator Agreement and Data

We tested the reliability of the annotation guidelines and definition of abusive language through an inter-annotator agreement study. 100 random messages from the training set of OffensEval were selected and annotated by three annotators [16] by labelling each message as implicit, explicit or not abusive.

After the first round of annotation, we measured a Fleiss' Kappa = 0.61 indicating substantial agreement (Fleiss, 1971). Complete agreement was reached on 63 cases out of 100. However, looking at pairwise agreement, asymmetries emerge. Two annotators agreed with each other on 88 cases, while the third agreed with them 67 and 70 times respectively. Following an analysis of the disagreements, three major areas emerged. The first area concerns a specific characteristic of offensive messages, namely their inextricably subjective nature when it comes to the sensibilities of audiences. In lots of cases, we have found the distinction between negative stance and implicitly expressed abuse to be blurred. In example 15, the author of the message clearly expresses a negative stance towards gun control using a form of sarcasm. This utterance was labeled **implicit** by one of the annotators and **not** (abusive) by the others.

15. @USER @USER I believe gun control should consist of guarding your firearms from thivery and kids. [OffensEval:train id:42611]

The second substantial amount of disagreement comes from different perceptions of "borderline" profanity and slurs, such as *weirdo* or *blathering*. The abusiveness degree of swear words is contextual (Pamungkas et al., 2019) and its perception, or effect (Vidgen et al., 2019), may depend on the annotator's background. Finally, the last aspect concerns rhetorical devices such as irony and sarcasm. In these cases, the lack of context of occurrence (i.e., the conversational context) and the need to "unpack" the rhetorical

[15] Both endings are perfectly fine, but completely different in their intentions and effects: (a.) "@USER you fucking genius" *vs*. "@USER you fucking idiot".

[16] All of the annotators are authors of this paper.

expressions to decode their pragmatic meaning introduce noise in the annotation process. After discussion, the annotators performed a second round of annotation on the same set of tweets, achieving a Fleiss' Kappa = 0.86.[17] Table 5 illustrates the statistics of the AbuseEval v1.0 data set. We also report figures for offensive messages (sub-task A) and the target of the offense (sub-task B) from OLID to better compare similarities and difference of the two annotations.

Table 5: AbuseEval v1.0: annotated data and annotation overlap with OLID/OffensEval. OLID/OffensEval labels: OFF = offensive; TIN = target; UTN = not targeted; NOT = not offensive. AbuseEval v1.0 labels: EXP = explicitly abusive; IMP = implicitly abusive; NOTABU = not abusive.

Data Distribution		OFF	TIN	UTN	NOT
Train	EXP	2,023	1,887	136	0
	IMP	726	668	58	0
	NOTABU	1,651	1,321	330	8,840
Test	EXP	106	103	3	0
	IMP	72	70	2	0
	NOTABU	62	40	22	620

As expected, there is a large overlap between the OFF annotation of OffensEval and the EXP annotation of AbuseEval v1.0. However, a surprising amount of instances considered offensive are marked NOTABU in the new resource. Finally, during the AbuseEval v1.0 annotation we found a small but not negligible portion of instances (about 7% in the training set and 2.8% in the test set) that are marked as abusive (EXP or IMP) but also as untargeted. By our definition, instances of abusive language are always targeted.

5. AbuseEval: Experiments

In this section, we report on a series of experiments conducted in order to empirically test the content value of the resources introduced in the present work as well as their applications in supervised learning settings.

We used the pre-trained BERT (Devlin et al., 2019) model for English (`uncased_L-12_H-768_A-12`), fine-tuned it on different training sets, and applied the fine-tuned models to predict the labels in different classification tasks. In all the experiments we used a standard learning rate of 10^{-5}, a batch size of 16, and a variable number of epochs between 5 and 10. We used the models implemented in the `keras_bert` library.[18] All results are averaged over multiple runs (i.e., 5 different runs).

5.1. Abusive vs. Not Abusive Classification

We first investigated whether the new annotation layer concerning the distinction of abusive vs. non-abusive messages can be effectively learned, thus collapsing the **explicit** and **implicit** labels into one abusive class (ABU). We train the model on the AbuseEval training set, comprising the same instances as the OffensEval training set, and test it on the AbuseEval test set. In Table 6 we report the evaluation in terms of precision (P), recall (R), and macro-average F1 (F1-macro). We further compare the performance of the classifier to the Dictionary system, introduced in § 3.

Table 6: AbuseEval: Evaluation on the Test set

Approach	Class	P	R	F1 (macro)
Dictionary	NOT	.867	.822	.657
	ABU	.431	.516	
BERT	NOT	$.868 \pm .017$	$.772 \pm .678$	**.716**$\pm .034$
	ABU	$.659 \pm .031$	$.446 \pm .096$	

Although the results are not directly comparable with those reported in Table 3 on the OffensEval data, we see a drop of the Dictionary approach. This indicates that our annotation actually captures a different phenomenon rather than the mere presence of profanities or slurs that may be perceived (or not) as offensive. In other words, we find that abusive language is captured to a lesser extent by modeling lexical items only, compared to offensive language. Indeed, a more complex system such as our implementation of BERT, is supposedly able to capture higher-level linguistic features. As can be seen, even in its "vanilla" version, it performs better in this task.

5.2. Implicit vs. Explicit Classification

In the second experiment, we test the prediction capability of the BERT model trained on the Implicit/Explicit annotation of OffensEval (§ 3.) and on the same layer as we revised it for AbuseEval (§ 4.). The task is therefore a three-label classification. We keep the training/test splits from OffensEval, and train the model for 10 epochs.

Table 7: Results of the experiments on the Implicit vs. Explicit distinction.

Data set	Class	P	R	F1 (macro)
OffensEval	NOT	$.868 \pm .023$	$.867 \pm .035$	$.614 \pm .157$
	IMP	$.240 \pm .059$	$.225 \pm .156$	
	EXP	$.637 \pm .029$	$.671 \pm .028$	
AbuseEval	NOTABU	$.864 \pm .019$	$.936 \pm .013$	$.535 \pm .023$
	IMP	$.234 \pm .086$	$.098 \pm .092$	
	EXP	$.640 \pm .060$	$.509 \pm .135$	

The results, reported in Table 7, show that the prediction of the Implicit class (IMP) is challenging. The limited amount of training data in both data sets is certainly a reason. However we also think that this is an indications of the complexity of the phenomenon. On the contrary, the prediction of explicit instances (EXP) is significantly more stable, despite the lower amount of instances of this class compared to the neutral ones (i.e., both NOT and NOTABU).

5.3. Cross-domain Hate Speech Detection

In this experiment, we explore the usefulness of the AbuseEval annotation in a downstream task, namely hate speech (HS) detection, a particular kind of abusive language. The HatEval shared task at SemEval 2019 provides a benchmark for hate speech (HS) detection systems on English and Spanish Twitter data (Basile et al., 2019). We train the BERT model for 5 epochs on the original OffensEval training set, as well as the AbuseEval training set,

[17] During the harmonization process, sometimes the initial majority vote was overturned, as in the case of example 15, which was labeled **implicit** in the final version.

[18] `https://pypi.org/project/keras-bert/`

and test the model against the official HatEval English test set. We further compare the performance of the same model trained on the official HatEval training set. In order to provide a meaningful comparison, the **offensive** label of OffensEval and the **abusive** label of AbuseEval (either IMP or EXP) are mapped to the HS class of HatEval. The model is trained for 5 epochs.

Table 8: Results of the cross-domain experiments.

Training set	Class	P	R	F1 (macro)
HatEval	NOT	$.877 \pm .021$	$.254 \pm .053$	$.514 \pm .033$
	HS	$.479 \pm .012$	$.950 \pm .022$	
OffensEval	NOT	$.665 \pm .068$	$.402 \pm .091$	$.528 \pm .016$
	HS	$.462 \pm .025$	$.712 \pm .170$	
AbuseEval	NOT	$.661 \pm .047$	$.672 \pm .134$	**.591**$\pm .023$
	HS	$.531 \pm .031$	$.510 \pm .182$	

From the results presented in Table 8, we draw several considerations. Firstly, the "vanilla" BERT model (i.e., without ad-hoc adjustments to suit the task) is quite competitive: with a macro F1-score of .514, this system would have been ranked 7 out of 71 in the official competition. Training on OffensEval achieves an even higher results. We ascribe this result to the larger amount of training data in OffensEval, which counterbalance the focus being on offensive language rather than hate speech. This also proves the high overlap between the two phenomena as discussed in § 4.

Finally, training on AbuseEval yields a significantly higher prediction performance on HatEval than training on OffensEval, both in relative terms (+.063 F1-score) and absolute terms (such system would have been ranked second in the competition). We consider this result as a clear indication that the annotation of AbuseEval indeed captures abusive phenomena (including hate speech) that are missing from the OffensEval dataset, due to its different focus.

6. Conclusion and Future Work

This contribution has presented an in-depth analysis of an existing and popular data set for offensive language detection, namely OLID/OffensEval. Following Wiegand et al. (2019), we have computed the offensive prior of the messages using a dictionary and observed how messages labelled as offensive are highly skewed towards the presence of explicit markers of offensiveness. As a follow-up step on this aspect, we have enriched OLID/OffensEval by manually annotating offensive messages with explicit and implicit labels. This has shown that $\sim$65% of the messages in training and $\sim$ 59% in test are explicit. Such a results calls for a reflection and development of new strategies on how data sets for offensive messages are actually generated. Applying the proposed annotation scheme using crowdsourcing could provide with more varied judgments about what is seen as explicit, implicit, or not abusive language by taking into account native English speakers with different demographics (e.g., educational level, ethnicity, age) (Mladenović et al., 2017). Investigation on a sample of 1,000 random messages originally annotated as racist/sexist or hateful from two additional data sets, such as Waseem and Hovy (2016b) and HatEval, we found that the majority of them is realised by explicit messages (38% in Waseem and Hovy (2016b) [19], and up to 98% in HatEval).

Inspired by current pending issues and on-going debates in the NLP community on data sets and definitions of offensive/abusive language, we propose a newly developed annotation scheme for abusive language classification. In our proposal, the aim is to clearly label the abusive potential of a message (i.e., the intention and/or effect) and, at the same time, the way the abuse is realised, i.e., whether in an explicit or implicit way. One aspect we make clear in our annotation is the need of contextual information: in all cases where the surface content of a message in isolation is not enough to take a decision, either the annotator is able to retrieve the context of occurrence or it should have a more conservative approach and mark the message as not abusive. We have applied our guidelines to the OLID/OffensEval data by refining the existing annotation of (perceived) offensiveness by identifying abusive messages. When compared to the original OLID/OffensEval annotation, we have identified a large portion of messages marked as offensive that do not qualify as abusive. We claim that one shortcoming of the OLID/OffensEval data set is the annotation of every message containing a profanity as offensive. This is particularly dangerous as it may lead to misrepresentations of communities.

We have further validated the annotated data with a series of experiments that have shown how the detection of abusive language, though strictly related to offensive languages, requires more flexible and "language-aware" methods than a simple look-up in a dictionary of profanities or slurs.

We want to advocate three directions for the future. First, data sets for the detection of abusive language phenomena must be contextually grounded: the annotation of a message and of its abusiveness must be conducted with respect to the context of occurrence rather than in isolation. Second, the use of keywords to retrieve potentially abusive messages should be deprecated, and we should collect data from "hateful" users. Some work in this direction has already been done (Ribeiro et al., 2018; Mishra et al., 2018; Wiegand et al., 2018b). The idea is that by directly collecting messages from users that are potentially more prone to use abusive/offensive language, we may reduce the bias with respect to explicit expressions of abuse/offense and increase the identification of more complex and implicit expressions of abuse/offense. Third, the use of figurative language and its relationship with abusive/offensive language needs to be further explored and may help in creating a data set that can help to address messages with a strong abusive effect but weak surface forms, as seen in the rhetorical figure litotes (Mitrović et al., 2017), e.g. "He is not the smartest pea in the pod".

Acknowledgments

The work of V. Basile is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

[19] Following our annotation guidelines on abusive languages, $\sim$38% of the sample may actually be interpreted as not abusive/racist/sexist.

Appendix: Data Statement - #BenderRule

Language of AbuseEval is English. Medium: Twitter. AbuseEval is created on top of the OLID/OffensEval data set. The complete data statement is available at `http://bit.ly/3ah4Ql8`

7. Bibliographical References

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and 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*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Bosco, C., Dell'Orletta, Felice, F. P., Sanuguinetti, M., and Tesconi, M. (2018). Overview of the EVALITA Hate Speech Detection (HaSpeeDe) Task. In Tommaso Caselli, et al., editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 71–80, Washington, DC, USA. IEEE Computer Society.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August. Association for Computational Linguistics.

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.

Fleiss, Joseph., L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Grice, H. P., Cole, P., and Morgan, J. (1975). Speech acts: Syntax and semantics.

Ibrohim, M. O. and Budi, I. (2019). Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy, July. Association for Computational Linguistics.

Kennedy, G., McCollough, A., Dixon, E., Bastidas, A., Ryan, J., Loo, C., and Sahay, S. (2017). Technology solutions to combat online harassment. In *Proceedings of the First Workshop on Abusive Language Online*, pages 73–77.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Liu, P., Li, W., and Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Merriam-Webster Online. (2009). Merriam-Webster Online Dictionary.

Mishra, P., Del Tredici, M., Yannakoudakis, H., and Shutova, E. (2018). Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098.

Mitrović, J., O'Reilly, C., Mladenović, M., and Handschuh, S. (2017). Ontological representations of rhetorical figures for argument mining. *Argument & Computation*, 8:267–287.

Mladenović, M., Krstev, C., Mitrović, J., and Stanković, R. (2017). Using lexical resources for irony and sarcasm classification. In *Proceedings of the 8th Balkan Conference in Informatics, BCI 2017, Skopje, Macedonia, September 20 - 23, 2017*, pages 13:1–13:8. ACM.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Oxford English Dictionary. (2019). Oxford English Dictionary Online.

Pamungkas, E. W., Basile, V., and Patti, V. (2019). Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure. *CoRR*, abs/1901.01911.

Papegnies, E., Labatut, V., Dufour, R., and Linarès, G. (2017). Detection of abusive messages in an on-line community. In *CORIA*.

Park, J. H. and Fung, P. (2017). One-step and two-step classification for abusive language detection on twitter. *CoRR*, abs/1706.01206.

Pedersen, T. (2019). Duluth at SemEval-2019 task 6: Lexical approaches to identify and categorize offensive

tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 593–599, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., and Bosco, C. (2017). Hate speech annotation: Analysis of an italian twitter corpus. In *CEUR WORKSHOP PROCEEDINGS*, volume 2006, pages 1–6. CEUR-WS.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, AI'10, pages 16–27, Berlin, Heidelberg. Springer-Verlag.

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., and Meira Jr, W. (2018). Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July. Association for Computational Linguistics.

Schmidt, A. and 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. Association for Computational Linguistics, Valencia, Spain*, pages 1–10.

Strawson, P. F. (1964). Intention and convention in speech acts. *The philosophical review*, 73(4):439–460.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016a). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016b). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

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

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018a). Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018b). Overview of the germeval 2018 shared task on the identification of offensive language. *Austrian Academy of Sciences, Vienna September 21, 2018*.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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 NAACL*.

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.

8. Language Resource References

Basile, Valerio and Bosco, Cristina and Fersini, Elisabetta and Nozza, Debora and Patti, Viviana and Rangel Pardo, Francisco Manuel and Rosso, Paolo and Sanguinetti, Manuela. (2019). *HatEval - Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*.

Zampieri, Marcos and Malmasi, Shervin and Nakov, Preslav and Rosenthal, Sara and Farra, Noura and Kumar, Ritesh. (2019). *Offensive Language Identification Dataset - OLID*.