Annotation Guidelines - Swear Words Abusiveness Dataset

September, 2019

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 [Jay92, Jay99]. 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. 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 [JKD06, JJ08]. 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 [Jay09]. Some studies even found that the use of swear words has also several upsides [Jay09, SU11, Joh12]. In this work we will build a new benchmark Twitter corpus, called SWAD (Swear Words Abusiveness Dataset), where abusive swearing is manually annotated at the word level, whether a given swear word is abusive or not-abusive. The final goal of this corpus development is to automatically classify 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.

Abusive Swearing

In this annotation task, we ask annotators to annotate (with a binary option) whether the highlighted swear word (tagged with the and 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").

For the annotation purpose, tweets which have more than one swear words were replicated. We generated as many new instances of the same tweet, as the number of swear words occurring in the message, and marked each single swear word with special tags and (i.e. fuck, shit, and etc.). For instance, given the message :

@USER This shit gon keep me in the crib lol fuck it

then two instances will be generated:

@USER This shit gon keep me in the crib lol fuck it

@USER This shit gon keep me in the crib lol fuck it.

In this case, annotators just need to focus on the marked swear words in every instance.

Here we give some examples that could help annotator to understand the task. We observe three possibilities of case, can be categorized as follows :

– Abusive swear word in a abusive instance. This is categorized as a normal case, when an abusive swearing is a part of an abusive sentence, and the swear word contribute to the abusive context of the sentence. See at the example 1 and 2. You can see that in the example 1 and 2, swear words fuck and bullshit have the main role to the abusive context of the full sentence. In this case these swear words should be annotated as abusive.

@USER fuck you. It shouldn't be happened if you do it correctly.
(example 1)

@USER shut up! the words from your mouth is such a bullshit!.
(example 2)

– Not-abusive swear word in a not-abusive instance. Similar to previous case, this case is also a trivial case, when a not-abusive swear word appear in a not-abusive instance. Look at the example 3 and 4. In example 3 and 4, swear words fucking and fucking do not contain intention to be abusive, and they also does not change the context of the full text to become abusive. In this case these swear words should be marked as not-abusive.

@USER the party last night was just fucking crazy. I love it!.
(example 3)

@USER what the fuck is going on! I just knew that he will marry her.
(example 4)

– Not-abusive swear word in an abusive instance. This case is more challenging compared to two previous cases, when an abusive instance contain a not-abusive swear word. There is two possibilities, there is another swear word in the sentence which contributes to the abusiveness of the sentence (see example 5), or the abusiveness is caused by other words (not a swear word) as shown in example 6. In example 5, swear word damn does not have abusive intention, while the abusiveness of the sentence is caused by another swear word, *whore*. Meanwhile in the example 6, swear word fuck also does not have abusive context, but the abusiveness context is mainly resulted by discrimination intention (black man). In these both cases, the marked swear word should be rated as not-abusive.

@USER damn look at this picture. @USER girlfriend looked like a whore
(example 5)

@USER what the fuck! I just don't understand. @USER is just a black man who only do nothing. Useless.
(example 6)

References

[Jay92] Timothy Jay. *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, 1992.

[Jay99] Timothy Jay. *Why we curse: A neuro-psycho-social theory of speech*. John Benjamins Publishing, 1999.

[Jay09] Timothy Jay. Do offensive words harm people? *Psychology, public policy, and law*, 15(2):81, 2009.

[JJ08] Timothy Jay and Kristin Janschewitz. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288, 2008.

[JKD06] Timothy Jay, Krista King, and Tim Duncan. Memories of punishment for cursing. *Sex Roles*, 55(1-2):123–133, 2006.

[Joh12] Danette Ifert Johnson. Swearing by peers in the work setting: Expectancy violation valence, perceptions of message, and perceptions of speaker. *Communication Studies*, 63(2):136–151, 2012.

[SU11] Richard Stephens and Claudia Umland. Swearing as a response to pain-effect of daily swearing frequency. *The Journal of Pain*, 12(12):1274–1281, 2011.