Guidelines for IGGSA Shared Task on the Identification of Offensive Language

Josef Ruppenhofer, Melanie Siegel, Michael Wiegand

March 12, 2018

1 Definitions for main categories

INSULT is the ascription of negatively evaluated qualities or deficiencies or the labeling of persons as unworthy (in some sense) or unvalued. Insults convey disrespect and contempt.

(1) ein #Tatort mit der Presswurst #Saalfeld geht gar nicht #ARD

(2) @KarlLagerfeld ist in meinen Augen strunzdumm wie ein Knäckebrot.

(3) Wo ist #Kubicki heute? Ist er schon besoffen im Puff?

Whether an utterance is an insult usually depends on the community in which it is made, on the social context (ongoing activity etc.) in which it is made, and on the linguistic means that are used (which have to be found to be *conventional* means whose assessment as insulting are intersubjectively reasonably stable).

Defamation/Slander consists of *claiming things that are unprovable* but which are apt to debase a person or group in the eyes of other people. Claiming here means that the speaker presents a situation as true or correct and that he or she is convinced that this is so. (Legally, the spreading of dematory claims also counts as defamation. In our twitter data, we can try to eliminate such problems by simply getting rid of all retweets that are identifiable as such.)

Slander need not be present linguistically as an assertion / statement. Rhetorical questions as well as the uttering of speculations or suspicions may rise to the level of slander.

Libel A libelous *statement is one that is untrue but which is nevertheless knowingly made or spread* by a person, and which is apt to debase a person or group in the eyes of (some relevant part of) the public.

ABUSE By abuse we define a special type of degradation. This type of degrading consists in ascribing a social identity to a person that is judged

negatively by a (perceived) majority of society. The identity in question is seen as a shameful, unworthy, morally objectionable or marginal identity. In contrast to insults, instances of abusive language require that the target of judgment is seen as a representative of a group and it is ascribed negative qualities that are taken to be universal, omnipresent and unchangeable characteristics of the group. (This part of the definition largely co-incides with what is referred to as *abusive speech* in other research.)

Aside from the cases where people are degraded based on their membership in some group, we also classify it as abusive language when dehumanization is employed even just towards an individual (i.e. describing a person as *scum* or *vermin* etc.).

(4) was mich stört ist wenn am frühen Morgen im #Morgenmagazin schon strunzdumme #Migranten moderieren

(5) ich würde auch nicht mit einer schwarzen #Schwuchtel zusammenarbeiten #Tatort

(6) Ich persönlich scheisse auf die grüne Kinderfickerpartei

Note: mere impoliteness, lack of tact, joking, or teasing/ribbing are not to be treated as insults. The difficulty is, of course, in distinguishing these from genuine insults given the general lack of context about the conversation and its participants.

PROFANITY Unacceptable language may also be encountered in the absence of insults and abuse. This typically concerns the usage of swearwords (*Scheiße, Fuck* etc.) and cursing (Zur Hölle! Verdammt! etc.). This can be often found in youth language:

(7) ob ich sterbe darauf geb ich fick

(8) bring die scheiss kohle vorbei und halt die Fresse haste mich verstanden

Such cases should be labeled as PROFANITY. Swearwords and cursing may also co-occur with insults or abusive speech. If such words are not directed towards a specific person or group of persons, then this is most likely to be profanity.

This label is also used if profane language is attributed to a quoted speaker (potentially a vague group) but it is clear from the context that the author of that utterance does not want to convey any offense.

(9) Sorry , Leute . Aber wenn man mich Fotze nennt , weil ich 'ne eigene Meinung hab , dann werd ich stinkig , ok .

(10) @flohbude gestern auf zdf kultur nen aspekte ausschnitt aus den siebzigern gesehen , in dem das wort negermusik fiel .

In other words, if a sentence contains an abusive word (e.g. *"Fotze"*, *"Neger"* etc.) and it is clear that the sentence is not meant to insult anyone, then this sentence must still be tagged as PROFANITY.

OTHER The category OTHER includes all utterances that have true positive or neutral polarity and all utterances of negative polarity that do not fall under the categories ABUSE, INSULT, or PROFANITY.

Regarding negative polarity: note that an utterance that constitutes an insult or abuse must have negative polarity towards a target of the right type, i.e. towards a person, organization or human social group or against another entity with clear ties to a person, organization or social group. The negative polarity may potentially be conveyed ironically and appear positive on the surface (*"Wie ich mich freue, dass wir die ganzen Neger hier haben!"*). No utterance with truly positive or neutral polarity can be abusive or insulting.

In the realm of negative polarity, the class OTHER can be defined negatively, as the complement of the union of the classes ABUSE, INSULT and PROFANITY. Mainly this will be negative news and/or criticism etc. For instance, on twitter, many racist people disseminate negative news involving foreign people particularly immigrants/muslims:

(11) Hamas-Vertreter bekräftigt Ziel der Zerstörung Israels

(12) Berlin: „Flüchtlingsschule“ legt grandiosen Fehlstart hin

(13) Laut eines Berichts des #NRW-Innenministers unterstützen türkische Behörden die Rockergruppe Osmanen Germania:

(14) 500 Millionen Afrikaner wollen nach Europa: „Daraus ergibt sich zwingend, dass Europa die Immigration regulieren...

(15) 11 amtliche Migrations-Mythen im Bullshitchec

Note that we take all these items at face value: it is not necessary to fact-check such claims as the number “500 million” (14).

Although the intention of most of these tweets is to increase dislike towards these minorities, these tweets should not be considered as abusive language. For our annotation, no ABUSE label should be assigned to such utterances unless there is a clear sign of abuse, insults or profanity. The same holds for criticism against Germany's migration policy.

(16) Ich finde es nicht richtig das Asylanten Freien Eintritt bekommen und hartz 4 Empfänger sollen von dem wenigen Geld was sie bekommen auch noch2,50 pro Person bezahlen. Daß ist gegenüber den deutschen Unrecht

A quite frequent related type of tweets are retweets of news about criminal incidents claiming that no information regarding the ethnic background of the perpetrator is available. The authors of those tweets typically rely

on the ethnic/cultural prejudices of their reader. They want their readers to assume that minorities, such as migrants or muslims, were the perpetrators.

(17) Festnahme an der Frankenwerft in der Altstadt von 2 Verdächtigen (19, 27). Das Duo soll die Nachtstunden damit verbracht, mehrere Touristen zu berauben und zu bestehlen - und ihre Opfer auch noch sexuell zu belästigen. Täterherkunft fehlt.

(18) Unbekanntes Trio überfällt immer wieder Dortmunder und raubt diese nach Bahnfahrten aus. Masche: „Ansprechen, Hilfe anbieten, ins Gesicht schlagen, ausrauben“ Täterherkunft liegt nicht vor. Fahndung.

However, in such tweets, there is typically no explicit accusation, racist remark or insult. For that reason, despite the likely intention of the writers to incite xenophobia, we label those tweets as OTHER.

2 Discussion

It is unclear if there are necessary and sufficient properties for something to be language abuse. E.g. does one have to name/mention a specific individual or group? Is the mention of an entity associated with a person or group sufficient, even in a context where one is discussing a topic unrelated to the group or where one is discussing the issue of discrimination itelf? E.g. can one innocently talk about *“Negermusik”* in any context? It seems these issues cannot be settled and clarified by linguistic analysis alone. Which utterances count as degrading, which kinds of quality ascriptions count as contributing to social stigma is a social and political question that is constantly being renegotiated.

3 Filtering annotations

In section §1, we described the 4 labels that we use to `categorize` the tweets that will be used as part of the shared task.

Here, for completeness sake, we report on a set of labels that we use to `filter` out tweets that are not relevant for annotation with one of our main category labels. Tweets marked with any of these labels will not figure in the shared task.

The idea behind our filtering is that we want the tweets that we label to be mostly self-contained. Accordingly, some of the tweets that we ask annotators to consider must be filtered out by them. For one, since tweets are very short they may not provide enough context to allow the the annotator to understand with certainty what is being talked about. Problems also arise because the language of social media is often ungrammatical, cryptic or full of social media-specific constructs (such as links, emojis, abbreviations) which prevent understanding.

Likewise, we do not consider tweets for which it would be necessary to consult further tweets or linked entities (i.e. follow urls etc).

HUNH is to be used for incomprehensible utterance. We do not require that a sentence is perfectly grammatical well formed and correctly spelled. However, if a sentence is so erroneous that the annotator does not understand its content, then this sentence should be labeled as HUNH. This label also applies if the sentence is formally correct but the annotator still does not understand what is meant by this utterance.

EXEMPT Tweets that are exempt from the subtyping annotation involve:

- Tweets which only contain abuse or insults that represent the view of somebody other than the tweeter. In other words, if a tweet contains an abusive quote by somebody else, we ignore it.
- Utterances which depend on non-textual information (e.g. an attached image or video) in order to understand should be exempt from annotation.
- Utterances that are not proper sentences, that is, just a series of hashtags and/or usernames, even if they indicate abusive speech (e.g. *#crimigrants* or *#rapefugees*).
- This category also applies to utterances that are incomplete. In some cases, the automatic extraction is responsible for that. For instance, some (longer) tweets sometimes end in "...":

  > @LissUncensored tweeted: @Ankhseram_ @ideal2320 Sie hat mich per sprachnachricht hure genannt während er daneben saß.Er hat sich entschuldig...

  Even though this utterance obviously contains a (reported) insult, the tweet has only been incompletely extracted. Therefore, we label the utterance with EXEMPT.
- In many cases, tweeters simply create texts that are broken up across multiple posts. In some cases, the Twitter user has marked this using notations such as [1/2] or (4/6). In other cases such as (19), there is no marking but readers can still recognize that the tweet is a continuation of a previous tweet. If the breaking up of a text across multiple tweets results in incomplete sentences, we mark the affected tweet as exempt, even if it contains some complete sentences. But keep in mind: if ABUSE or INSULT can be determined based on the incomplete text of a tweet alone, we do use those labels on the tweet!

  (19) trifft und selbst danach lebt. Und zwar nur für sich selbst. Nicht für Andere. Damit käme man ganz gut durchs Leben. Und heute abend gäb's

DIALOG is to be used for cases of insufficient context.

- Some tweets contain anaphoras which cannot be resolved without considering context beyond the tweet itself, that is, the thread structure. Unlike HUNH, these tweets are not incomprehensible as such. It is just not clear who the antecedent is. For instance, tweet (20) constitutes a criticism of the proposed office of *Antisemitismusbeauftragter*, who is referred to by the pronoun *der*. Tweet (21) refers to a Muslim refugee featured in a German TV documentation in 2017 that portrayed his relationship with a German girl. If one resolves the reference, it is possible to see the tweet as using the stereotype that all/most minor refugees are in fact adults who lied about their age.

  (20) der sollte sich bei der nächsten Demo mit dazu stellen, "Jude Jude Feiges Schwein"(2012), für IS werben od Zionistenblut fordern (2017) etc

  (21) @berlinerzeitung Ich denke nicht, dass er sich mit 13/14 allein auf den weiten Weg gemacht hat. Er wird älter sein als 17 oder nun 19. Und

  Note, however, that sometimes, even tweets with anaphoras can be judged as ABUSE, INSULT or PROFANITY without the resolution of the antecedent. Such tweets will then not be labeled as DIALOG but as ABUSE, INSULT or PROFANITY.

  (22) @MiriamOzen Weil er eine linke Bazille ist...

  It is not clear to whom er actually refers to. However, it is obvious that it refers to some person. If a person is called a *linke Bazille* then this remark is always abusive speech (it is not just an insult since comparing a human with bacteria is some form of dehumanization, a distinct feature of abusive speech). It is irrelevant who the person being addressed actually is.

3.1 Feature annotations related to the main categories

The following annotations are not part of the official dataset released with the shared task. They were meant as auxiliary labels that facilitates the annotation process.

3.1.1 Epithets

Does the tweet contain any explicit abusive, defamatory, or derogatory words or phrases. In principle there are two types which we however both unite under the overall label of epithets.

1. Evaluative words with negative connotation and some (at least etymologically) descriptive content: ‘ *"idiot", "asshole"; "Idiot", "Arschloch"*.

2. Derogatory words that target persons based on their group membership: *"Spaghetti", "Neger"*.

3.1.2 Typical targets of abusive speech

As pointed out above, one the key characteristic of abusive speech is that the negative qualities attributed to the target 1) arise intrinsically from his or her membership in a group or collective, 2) are shared by the other members of the group, and 3) are unchangeable. So the most prototypical classes are the ones that are defined by birth (sex/gender, nationality, ethnicity, faith/belief). However, there are other classes which, although not defined by birth, we will treat as targets of discrimination on an equal footing with the classes defined by birth, for instance migrants/refugees; political parties/ideologies.

Group identities that are often taken to be targets of abusive speech (this list is not claimed to be exhaustive):

- feminists,
- black people
- muslims
- jews homosexuals (LGBT)
- refugees
- members of political parties ("lefties", "Greens", "Sozis", "Altparteien" etc.)
- public media

3.1.3 Stereotype

Given that abusive language by definition involves reference to universal properties of groups, it is likely that abusive language mentions such properties. Some of these stereotypes are evoked as statements (e.g. 23), while others are evoked presupposed group characteristics, especially in agentive compounds such as *"Ziegenficker"* in 24.

(23) Arbeitskräfte ??? Das ich nicht lache... Die beuten unsere Sozialsysteme aus. Das ist Tatsache

(24) Gehts noch warum schmeißt diese Ziegen Esel Ficker raus ich kann diese Kopftuch Matzen die Islamische verstünde ne Schweine aus unsern Land

3.1.4 Curses

We also record the expression of wishes that misfortune, evil, doom, etc., befall a person or a group.

(25) Reichen diesem kriminellen Sektenführer seine Millionen-Gewinne die er über die Caritas und seine christlichen Schleuser (Merkel / G.Eckardt / Gauck / Steinmeier / Schäuble) macht,noch nicht?! Ihn, seine Schleuser&das Dummvolk der Christen&Katholiken, soll der Teufel holen!!!

(26) So ein süßer Schatz. Menschen, die Tiere entsorgen, denen wünsche ich die Pest an den Hals. Ich wünsche der Fellnase ein langes und glückliches Hundeleben

3.1.5 Threat

The speaker expresses a threat such as

(27) Ich mach Dich fertig

(28) Deshalb sind die ja in der Krise! Agenda 2020 oder es gibt was auf die Mütze, ihr Tzatzikifresser!!

Please be careful to distinguish between @threat in which the speaker announces to take some action himself and @direct-call-for-action in which the speaker calls for others to take some action.

3.1.6 Call for action

In some cases, the reader is directly called upon to attack a person or a group of persons.

(29) Haut den dreckigen Antisemiten in die Fresse. Das ganze linksgrünschwarzgelb versiffte Gesindel ist antisemitisch bis auf die Knochen.

3.1.7 Disguise

If an epithet is present, it may be disguised by the use of asterisks or other characters, e.g. *"Ziegenf***er"*.

3.1.8 Certainty of annotation

We ask annotators to mark cases where they are uncertain of one or more of the labels they chose.

4 General advice

- Annotators should not annotate utterances under the assumption that if they consider something as abusive, then this utterance is so unacceptable that it should automatically be *censored*; previous annotation experiments revealed that this results in a very conservative annotation; annotators should rather consider their rating as a pre-screening of suspicious candidates.

- Abusive language and insults can also be directed against people generally considered bad guys; it does not even matter whether they are dead or alive: e.g. *das Arschloch Hitler* (INSULT); *die NPD-Idioten sind alle paranoid* (ABUSE).

- It does not matter whether an utterance considered an insult or abuse is actually noticed by the person or group of persons targeted.

- Please always carefully consider the given context of a potentially offensive expression. Annotators should always ask themselves whether the utterance is really meant as an offense. For example, the adjective *dumm* may be offensive in the first utterance below (notice that the speaker generally despises the party), but it should only be considered as some criticism in the second utterance (here the speaker just criticizes some particular action of the party):

  - Die dumme SPD, immer von gestern, hat noch nie etwas Sinnvolles zu Stande gebracht.
  - Die SPD ist dumm, wenn sie glaubt einfach mit dem Schlagwort "soziale Gerechtigkeit" die Wahl zu gewinnen.

- Abuse may come in sentence types other than plain assertions, for example, rhetorical questions:

  - Wenn ich diese Ausländer als Asylschmarotzer bezeichne, bin ich dann ein Nazi?

  Annotators should ask themselves what they think the speaker of the utterance is actually thinking. In the above utterance, it is fairly obvious that the speaker opposes asylum seekers.