Annotation Instructions

In the following, you are asked to annotate a list of expressions with regard to **offensiveness**. Before we describe the specific task, we will first give you a brief introduction to what we understand by offensiveness.

Definition:
By offensive utterances we understand (unfounded) accusations, threats, general insults, profanity, or utterances generally considered rude.

Below you find a list of words in context that are commonly perceived as offensive. The offensive expressions are highlighted in **bold font and red colour**.

- `You are a` **`coward`**`.`
- `You are so` **`treacherous`**`.`
- `You are an` **`idiot`**`.`
- `You` **`sicken`** `me.`
- `I really can't stand your` **`hypocrisy`**`.`
- `You are so` **`lazy`**`.`
- `Stop` **`boasting`**`.`
- `Don't be so` **`stupid`**`.`
- `You just` **`screwed`** `everything.`
- `You are responsible for this` **`massacre`**`.`
- `Don't be so` **`foolish`**`.`
- `I will` **`kill`** `you.`
- `I really can't stand your` **`meanness`**`.`
- `You are such a` **`useless`** `being.`
- `You are such an` **`obtuse`** `person.`
- `You` **`disgust`** `me.`
- `We don't want your` **`selfishness`** `here.`
- `I really can't stand your` **`stubbornness`**`.`

All expressions to be annotated bear a **negative** connotation. However, **by far this does not mean that every expression is also offensive**. In the following, we give some examples which we consider not offensive and explain why. The relevant negative expressions (which are <u>not</u> offensive) are highlighted in **bold font and green colour**.

- `You are a` **`victim`** `of domestic violence.`
  Explanation: Such a statement suggests that the speaker feels pity for the person (s)he refers to as a victim. Pity and offensiveness are two different things.
- `We don't need your` **`warning`**`.`
  Explanation: A warning is no offensive action. (On the contrary, often you warn someone in order to protect him/her from danger.)
- `We don't want your` **`regret`**`.`
  Explanation: Regret is a frame of mind which has no offensive connotation.
- `You are responsible for this` **`mistake`**`.`

Explanation: Too mild in order to be perceived as offensive.

- `Don't be so` **`irritated`**`.`
  Explanation: This utterance could be made to calm someone down. An offensive intention is less likely.
- `You are so` **`unhappy`**`.`
  Explanation: Unhappy describes the frame of mind of the addressee (i.e. *you*). There is nothing inherently offensive about being sad, upset or disappointed etc.
- `Don't be so` **`reluctant`**`.`
  Explanation: The speaker may want to criticize the person addressed. However, an offense is unlikely to be intended by this remark. Besides, reluctance is not a negative human property per se.
- `You are so` **`inappropriate`**`.`
  Explanation: inappropriate is much too polite in order to be perceived as offensive.
- `I can't stand your` **`criticism`**`.`
  Explanation: Criticism is not offensive. However, semantically-related nouns such as *nagging*, *harping* or *squabbling* are. For instance, if someone says that you are nagging, (s)he challenges your ability to provide an adequate form of criticism -- if your actions are described as criticism there is no such pejorative connotation that would make you feel hurt.
- `I` **`protest`** `against you.`
  Explanation: Protest is not inherently offensive (this is similar to *criticism*).
- `Stop` **`arguing`**`.`
  Explanation: This is an order which typically a parent gives to his/her child. (It may be inappropriate to give such an order to other persons, e.g. your boss -- but the order as such is not offensive.)

The actual annotation task:

You are asked to annotate lists of expressions. Each expression on a list has already been rated in a previous survey. We grouped the list of expressions by semantic similarity. All expressions within a list have originally been assigned the same label (i.e. **offensive** or **not offensive**) except one member. This may be an error in the manual annotation of the previous survey. We will not tell you which member is the odd one out. You are to decide whether the annotation is consistent. If you also identify the odd one out, then this means that the previous annotation is correct. If not, then we found an error in the previous annotation.

Your specific task is:

- to identify the expression within the list you find **inconsistent.** This means: on a list of expressions which are predominantly not offensive you mark the expression which is offensive; on a list of predominantly offensive expressions you mark the expression which is not offensive
- you are only to specify **at most one member** of the given list
- if you think that there is actually more than one member inappropriate in the list, then please only specify the member you think is most inappropriate
- if you think that the given list is consistent, you simply type in: ***NONE***

Examples:

- Imagine you are given the following list of terms which are claimed to be **not offensive:**

  `accuse`, `disappoint`, **`disgust`**, `disapprove`

  The word ***disgust*** is likely to be perceived as offensive (as in *`You`* ***`disgust`*** *`me.`*) and therefore should be marked.
- Imagine you are given the following list of terms which are claimed to be **offensive:**

  `stupid`, `dumb`, **`inappropriate`**, `moronic`

  The word ***inappropriate*** is not likely to be perceived as offensive (as it is a fairly polite word) and therefore should be marked.
- Imagine you are given the following list of terms which are claimed to be **offensive:**

  `moron`, `bastard`, `turd`, `dumbass`

  This is a consistent list. All these expressions are offensive. Therefore, type in *NONE*.
- Imagine you are given the following list of terms which are claimed to be **not offensive:**

  `worry`, `bother`, `afraid`, `doubt`

  This is a consistent list. All these expressions are not offensive. Therefore, type in *NONE*. Also notice that while *worry*, *bother* and *doubt* are verbs, *afraid* is an adjective. This is perfectly fine. It will often be the case that you are given lists of words where the individual words belong to different parts of speech.

Some hints:

- We are not asking to mark different levels of offensiveness. For example, if ***slut*** and ***fool*** appear on a list of offensive words, this is considered "consistent". It does not matter if the given expressions differ in the degree of offensiveness (e.g. ***slut*** is more offensive than ***fool***). "Inconsistent" just means an expression which is not offensive among offensive words or an offensive word among list of words which are not offensive.
- Please do not challenge the label assigned to an entire list of expressions. Assume that for the majority of expressions in that group this label is correct. We just want you to identify **the odd one out** (if there is one**).**
- Please do not feel forced to always identify a inconsistent member from the list. If you think that all members fit in that list, then this is perfectly fine (in this case, type in *NONE*).
- You are to make a decision on the basis of whether you think that *someone* may perceive such an utterance (that you conceive with the given expression) as offensive. We are **not asking whether you personally would feel offended or insulted**. You may be toughened up or thin-skinned. Try to give a judgment that you think would be representative of the general public. Regard yourself as a ***watchdog*** whose job is to identify authors of offensive utterances and thus prevent future attacks from these wrongdoers.

Our recommendation for the annotation is that you store the above guidelines or print them out. You should **use the above examples as a reference** for your annotation.