NLP(DEEP) - Lab 08 - Annotation Guideline

Eliot LECLAIR, Alex POIRON, Tom THIL, Aurelien VISENTIN

1 Target class

We chose to work with the "offensive" dataset for this lab. Has we said before, when we load it in our notebook, we think that select the offensive one will contains more generalized sentences and be, we hope, more interesting to look at it. Nevertheless, we need to define correctly the word **offensive** with generic rules. Define properly these rules will allow us to annotate more efficiently and especially to have the same directive.

According to the definition of the Larousse dictionary, an offense is defined as follows : "A word or action that injures someone's dignity or honor." How we are gonna translate this definition and applies it on our tweets ? We thought about some basics detection rules as follows :

— violent terms from insults
— any form of discrimination such as racism, homophobia, ...
— belittling, attacking the person and/or his values

As we all know, these rules can be interpreted differently following the person who want to annotate. This is why we give just below some examples for a better understanding.

2 Some examples

We select here from the dataset some examples of tweets that may concern what we said earlier. Each example corresponds to a specific level of "how are we certain about the offensive content ?" Let's get started with a simple one :

Tweet 1 : "YOU BETTER SUCK HIS DICK KOZY I SEE YOU WITH KNUCKLES GET EM GYAAAAL"

Here, clearly we can say that this tweet must be annotated as **offensive**. This is an example where the doubt doesn't exist. The easiest case.

Tweet 2 : "Do you even know what a proper midfielder is ? Plz go to sleep."

With this tweet, the offensive part is more subtile but we can still find clearly that this tweet can be categorized as **offensive**.

Tweet 3 : "A boy died without seeing his family you say ? I say 'vaccines, vaccines, vaccines'. GET BORIS OUT The people have had enough. @user @user SackBorisJohnson SackJohnson JohnsonOut'"

In this case, we thought that this example is a tricky/good one due to its content. Here, a person is just giving is opinion about a political subject. The freedom of expression comes into play especially in this example and it is difficult to decide whether we should prohibit and "condemn" this kind of speech by classifying it as **offensive** or not.

Tweet 4 : "Who cares"

This final example is a perfect transition for the last part of this guideline. We can't annotate this kind of tweet due to one missing parameter : **the context**.

3 Can't tell / not annotable

At the end, each annotator will be faced to a tweet with absolutely no **context**. In this particular case, it is then impossible to class this example either in the non-offensive class either in the offensive one. Due essentially to the lack of context, we must don't imagine or over-interpret the possible continuation of the tweet.