NLP2

- SCIA 2023 -

Jules DORBEAU - Noé JENN-TREYER - Yacine ANANE - Adrien HOUPERT

1 Guidelines

For a manual classification of our 100 Tweets, we must assign a label to each between "Offensive", "Not offensive" and "Cannot tell". We deliberately take a broad definition of the term offensive as these tweets are intended to be used to train a moderation AI commercial application and we want to avoid controversial topics. We made the following guidelines to process these Tweets by hand:

1.1 Main guideline

Our main guideline is to classify "Offensive" Tweets, complemented by "Cannot tell". The rest is then classified as "Not offensive".

1.2 Label "Offensive"

As we want to mainly classify offensive Tweets in a business situation, we want to encompass as many as possible Tweets that could be seen as offensive by the public. Our rules to recognize offensive Tweets are the following:
It contains any combination of the following rules:

- The subject of the Tweet is a person, community, group or organization
- It contains insults, hate, mockery

1.3 Label "Cannot tell"

This section is for all the situations in which having the complete context would help to do a better classification. This could be the case for many Tweets that would be wrongly classified. This is why we have this type of other classification for those situations:

- The Tweet contains sensible words or is talking about a sensitive subject but is not enough explicit to understand what is exactly the position of the Tweet just by reading it. In this case, the context into which the Tweet was sent would help to classify it more clearly.
- If the Tweet only contains one or a few words/symbols that don't make any sense by themselves. Just like in the previous situation, the context into which the Tweet was sent would help to classify it.

2 Bonus: Changes on the Guidelines

After seeing the results of the individual classification of Tweets for each of the members of our group (Cohen Kappa score ranging from 0.30 to 0.44), we noticed that some changes needed to be done to the guidelines. To illustrate that, we will present 4 examples here to show how different can be the classifications and then explain the changes that we choose to apply to our guidelines.

2.1 Different classifications examples

For this section, we selected 4 examples to show how different can be the answer of many persons on the same Tweet and with the same guidelines :

- Even if they didn't exploit people to acquire their riches, how are you gonna be okay literally wasting thousands and thousands of dollars while there are still people who are homeless? While there are people skipping life saving medical treatments bc of the cost?

We thought that this would be a great example to show here because it has two classifications as "Non-offensive". But as the guidelines say it, as this Tweet is talking in a negative way about a group of persons, it should be classified as "Offensive". After a discussion, we came to the conclusion that, as this was not focusing on a single person directly, some of us thought that it was not that offending just by reading it.

- Good one bro

For this one, this could be easily misunderstood at first look. Indeed, the Tweet looks like really friendly and nice which lead two of us to classify it as "Non-offendent". But, taken in another context, this Tweet could also be really offendent. So, by following the guidelines, here, it should have been a classification to "Cannot tell" as the context is missing.

- ilm out, your eyes tell, Light, stay gold, crystal snow, don't leave me, let go, bnyk sihh

Another one that should have been classified as "Cannot tell" by the majority of the people but was not. Here, at the first look, it does not say anything offendent, we could probably think that this is from a music. But, by following the guidelines, once again, it should have been classified otherwise as we can't tell the meaning of this Tweet.

- I'm usually always on the team of women with facial scars cause they're bad ass but I don't like her right now maybe it will change later but I don't like her

This last example is a really good one as it showed us that we should have separate classifications for an "Offendent" Tweet and a "Negative" Tweet. Indeed, as this one might be negative, it is not classified as an offendent Tweet by following the guidelines. As the difference may be small, it can also be offendent for some people so the difference must be done for those Tweets.

2.2 The changes

After reviewing our respective differences in classification choices and a first experience in labelling the content of the tweets dataset, we agreed on several new rules and guidelines modification planned to minimize human mistakes and encompass more possible outliers and unexpected content in the Tweets. The added rules as are follow:

2.3 Label "Offensive"

It contains any of the following new rules:

- The subject of the Tweet is of sexual nature. (Because the subject is inappropriate for a commercial application).
- The tweet is considered as harassing.
- The tweet is discriminative.
- The tweet is shocking in nature.

2.4 Label "Cannot tell"

It contains any of the following new rules:

- The subject of the Tweet is not explicit and can be easily interpreted differently depending on context.

2.5 The changes results

The Cohen Kappa Scores we got this time were much better, averaging at 0.477 (more than the previous maximum Cohen Kappa Score we had at 0.44) and ranging from 0.40 up to 0.63. This proves the efficiency of reviewing and fixing the existing guidelines.