Annotation guideline

Classe Target :

La classe cible est "Offensive".

Nous étiquetons un message comme offensant s'il contient une forme quelconque de langage inacceptable (grossièreté) ou une offense ciblée, qu'elle soit voilée ou directe. Les tweets de cette classe contiennent des insultes, un langage offensant, un contenu explicite, des discours de haine, des menaces, des jurons ou toute autre forme de communication nuisible ou irrespectueuse.

Les annotateurs doivent se concentrer sur l'identification des tweets qui relèvent clairement de cette catégorie.

Classe Non Offensive :

Les tweets ne contenant aucun langage offensant ou contenu nuisible doivent être classés comme "Non-offensant". Ces tweets peuvent être neutres, informatifs, positifs ou sans rapport avec des sujets offensants.

Classe not annotable :

Cette classe est utilisée pour les tweets qui ne peuvent pas être classés de manière définitive comme offensants ou non-offensants en raison de facteurs tels que l'ambiguïté du langage, un contexte incomplet ou un contenu difficile à interpréter.

Les annotateurs doivent attribuer cette classe seulement lorsqu'il est impossible de déterminer la nature offensive ou non-offensive du tweet.

Exemples : Pour illustrer le processus de classification, voici quelques exemples de tweets avec leurs classes respectives :

Example 1:

Tweet ID: 1410492618778157059 / Text: Fuckoff / Class: Offensive

Example 2:

Tweet ID: 1410492622972588038 / Text: 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? / Class: Non-Offensive

Guidelines :

1. Lisez attentivement le tweet en tenant compte du sens général et de l'intention.
2. Déterminez si le tweet contient un langage offensant, des discours de haine, de la discrimination, des menaces ou un contenu explicite.
3. Classez le tweet comme Offensant s'il répond à l'un des critères mentionnés à l'étape 2.
4. Si le tweet ne contient pas de langage offensant ni de contenu nuisible, classez-le comme Non-offensive.
5. En cas de cas ambigus, prenez en compte le sarcasme, l'ironie, les nuances contextuelles et l'interprétation subjective. Considérez comment le tweet pourrait être perçu par différentes personnes et, si nécessaire.
6. Si vous ne pouvez pas déterminer de manière définitive si le tweet est offensant ou non-offensant en raison d'un langage ambigu, d'un contexte incomplet ou d'autres facteurs, attribuez-lui la classe Impossible à déterminer / non annotable.
7. Évitez les préjugés personnels et les jugements. Concentrez-vous sur le contenu du tweet plutôt que sur l'identité de l'utilisateur.
8. Si un tweet contient un contenu mixte, où une partie est offensante et une autre est non-offensante, classez-le par rapport à sa signification globale contextualisée.
9. Ignorez les URL, les hashtags ou autres éléments non textuels qui ne sont pas directement liés à un contenu offensant et concentrez-vous sur le texte du tweet lui-même.
10. Ne pas utiliser trop souvent la classe Impossible à déterminer / non annotable.
11. Assurez-vous de maintenir la cohérence et de suivre précisément les directives pour garantir des annotations de haute qualité.

.

BONUS : On a pu voir par la pratique, que certaines utilisations de mots peuvent gêner les personnes par rapport à leur pudeur. Par exemple, les mots parlant de l’anatomie peuvent être considérés comme offensants dans des contextes se voulant neutres.

Par exemple dans le sample sélectionné on a ce commentaire :

“She is beautiful her boobs are perfect! Don't project your insecurities on her... pleaseðŸ¤Œ”

Ce commentaire tente d’être neutre mais reste déplacé dans le regard sur le corps et peut être une atteinte à la pudeur des personnes concernés par le commentaire.