Toxic Comment Classification
Machine Learning for Natural Language Processing 2020

Eléonore Blanchard
ENSAE Paris
eleonoreblanchard@gmail.com

Yannick Ly
ENSAE Paris
yannick.ly@ensae.fr

Abstract

Our work aims to identify the different types of toxicity amongst comments. [1]

1 Problem Framing

Due to the increase in the use of social media and the anonymity of internet, we observe a surge in the toxicity of comments and cyber-bullying. Some of the comments are event illegal because they are racists, homophobic, pedophile. This is why it is important that online platforms moderate this kind of comments and apply sanctions such as bans or the transmission of the personal data to authorities.

However, those platforms still have issues to correctly identify those comments, and more specifically their types (toxic, severe toxic, obscene, threat, insult, identity hate).

Being able to identify the type of toxicity would make it easier to apply appropriate measures.

2 Experiments Protocol

The data set that we are going to use come from Wikipedia comments. It is made up of 159 171 comments with 6 different labels (toxic, severe toxic, obscene, threat, insult, identity hate). They have been labeled by humans, therefore labelisation mistakes are unavoidable. Some of the comments have multiple labels.

3 Results

4 Discussion/Conclusion

.

[1] https://github.com/YannickLy/NLP-Project-ENSAE-2020

References