Detecting discrimination in French texts

Degree programme: BSc in Computer Science | Specialisation: Data Engineering
Thesis advisor: Prof. Dr. Mascha Kurpicz-Briki
Expert: Pierre-Yves Voirol (Abacus Research SA)

The presence of discriminatory language on social media platforms and its implications are becoming a serious concern in modern society. Preventing such acts requires having a means of detecting them. Fortunately, technologies from the field of natural language processing (NLP) can help to automatically detect such discriminatory content. While most of the existing work is dedicated to English, research on other languages like French is limited.

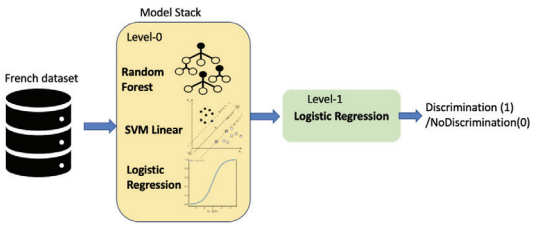
Introduction

Discrimination crimes are, unfortunately, not a new phenomenon in our society. However, the strong network connection provided by the internet has contributed in one way or another to bringing this issue more to the forefront. In this work, we adopt the definition of discrimination given by Swiss law, which considers any act or thought that targets a person or group of people for any reason to be discriminatory, and we use NLP to try to identify discriminatory texts in French data.

Goals

The main objective of this work aligns with previous work in natural language processing (NLP) that has examined hate speech in real-world data. We work with French datasets from Twitter and examine different state-of-the-art supervised and unsupervised discrimination detection techniques. Furthermore, we seek ways to mitigate the imbalance in our datasets using data augmentation methods. Our work is structured as follows:

- Experiment with supervised Machine Learning (ML) models
- Perform data augmentation to improve our model's performance
- Experiment with transformer-based models (French BERT)
- Experiment with Ensemble models
- Quantify discrimination with a so-called Discrimination Score



Stacking ensemble architecture

Methods

As a first step, we conducted a comparative study between supervised ML models (Support Vector Machine, Logistic Regression, and Random Forest) and different feature extraction techniques. In addition, to solve the data imbalance problem, which complicated our mission, we tried different data augmentation techniques. Finally, we used some supervised ML models to create a Stacking ensemble classifier that leverages the capabilities of these powerful single models to make better binary predictions and improve the final performance.

Results

Using the NLPAug library with contextual BERT word embeddings for data augmentation, we managed to improve the F1 score by around 10%. Additionally, we have found that it is most effective to start with supervised ML before moving on to more complex deep learning algorithms (like BERT), as they do not always perform better than traditional ML models. In conclusion, the best overall result on the aggregated dataset was achieved by a Random Forest model yielding an 86.3% F1 score and an accuracy of 83.3%.

Test F1 score comparison of Supervised classifiers with and without Data Augmentation (DA)

Ghofrane Merhbene
Ghofranemer@gmail.com