Ethics for NLP: Spring 2022
Homework 2

TECHNISCHE UNIVERSITÄT DARMSTADT

Due until Thursday, 09.06.2022 at 11:30am

Submission Guidelines for Homework

1) Use our given LaTex template to write your report in any LaTex editor of your choice and save it as: hw02.<matriculation_number>.lastname.firstname.pdf.
2) Write your code in a Jupyter Notebook and save it with its outputs as a PDF-file, so that we can understand your results better.

- Your submission in Moodle should be 1 ZIP-file named "hw02.<matriculation_number>.lastname.firstname.zip" and include the report and Jupyter Notebook as a PDF-file.
- This homework worth 20 Points (extra credit shall be given to well-written submissions).
- In case of questions or remarks post on Moodle or contact:
  - Aniket Pramanick, pramanick@ukp.informatik.tu-darmstadt.de

1 Identifying and Categorizing Offensive Language in Social Media (20 Points)

As we all know, "Abusive Language" is one of the major concerns on online platforms nowadays. In this assignment you will analyze the performance of a toxicity classifier and additionally build your own classifier.

Dataset: For this assignment the required data originates here and is available from two sources **(please be aware that it contains offensive or sensitive content including profanity and racial slurs)**.

The files train.tsv and dev.tsv consist of annotated tweets for offensiveness, published at OffensEval 2019. Their first column text contains the text of a tweet and their second column label contains the following labels for each tweet:

- Not Offensive (NOT): The post does not contain offensiveness or profanity,
- Offensive (OFF): The post contains offensive language or targeted (veiled or direct) offense.

Please refer to offenseval-annotation.txt for details on the annotation scheme.

Additionally, we provide a dataset of tweets proxy-labelled for race in the file mini_demographic_dev.tsv. This data has been sampled from TwitterAAE dataset and uses posterior proportions of demographic topics as a proxy for racial dialect. The first column text contains the tweet and the second column demographic consists of one of the following labels: "AA" (African American), "White", "Hispanic" or "Other". Assume that no tweet in the TwitterAAE dataset contains toxic language.

Finally, both (dev.tsv & mini_demographic_dev.tsv) contain a column perspective_score which contains the toxicity score. These scores were obtained using PerspectiveAPI tool. In all datasets, user mentions have been

replaced with the token @user .

Evaluation: You will evaluate your models using two criteria:

1. Performance on hate speech detection (Accuracy & F1 Score where "NOT" is considered as positive).
2. False Positive Rate (FPR), how often the model misclassifies non-toxic speech as toxic (specifically for comments associated with different demographic dialects).

Poor performance over hate speech classification suggests that the model is not accurate enough to be useful, while poor or imbalanced FPR indicates that the model may impose racial biases.

1.1 Classifier Performance Analysis (9 Points)

a) Use the provided perspecitve_score values to classify each tweet in dev.tsv and mini_demographic_dev.tsv as toxic or non-toxic. As a starting point, assume that a tweet is considered offensive if it contains a toxicity score > 0.8 (you may optionally explore other thresholds). [2P]

b) Use the dev.tsv to report the Accuracy and F1 Scores of PerspectiveAPI for OFF classification. [3P]

c) Use the mini_demographic_dev.tsv to separately report the FPR for each demographic group (assuming no tweet in mini_demographic_dev.tsv is actually offensive). [2P]

d) Briefly explain your results. [2P]

1.2 Custom Classifier (11 Points)

a) Build your own classifier to distinguish offensive (OFF) tweets from non-offensive (NOT) tweets. Your model should be trained on train.tsv and you should achieve an accuracy of at least 70% as well as a F1 score of at least 80% on the dev.tsv (this should be easy to obtain with surface-level features). Then report the Accuracy and F1 score of your model on the dev.tsv . [6P]

b) Report FPR on the mini_demographic_dev.tsv . [2P]

c) Briefly explain your results and how your model compares to PerspectiveAPI. [3P]