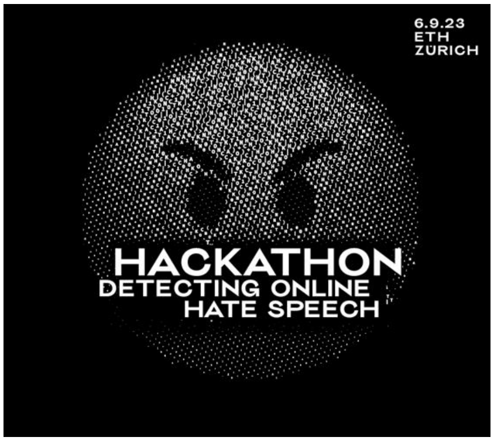
Latsis Hackathon 2023

Instructions for participants

Help us to contribute to safer and more inclusive online spaces!



1. Overview

Your task in the Latsis Hackathon 2023 is to implement an NLP system which is able to detect German hate speech in comment sections of Swiss online news outlets.

Hate speech is any kind of offensive or denigrating speech against humans based on their identity (e.g., gender, nationality, ethnicity, sexual orientation etc.).

2. What the day looks like

Time	Program
08:30-09:00	Registration / Signing of Non-Disclosure Agreement
09:00-09:15	Opening
09:15-09:30	Data access
09:30-12:00	Start hackathon
12:00-13:30	Lunch break
13:30-17:00	Hackathon
17:00	Submission
17:00-19:00	Snacks & apero
19:00	Award ceremony

Location

The hackathon takes place in the ETH Student Project House located at Clausiusstrasse 16, 8006 Zurich.

Food and beverages

Food and beverages will be provided throughout the day.

For the lunch break we will provide lunch bags so you can relax at your own pace. Happy to chat with you then, but in case you are busy hacking – this is completely fine as well.

3. What we provide

3.1 Data set

We provide a proprietary data set of approximately 9k German user comments from leading Swiss newspapers. Each comment is endowed with an annotation indicating whether it is hate speech or not. Prior to receiving the data set you will be required to sign a non-disclosure agreement (NDA). Please follow the instructions you received in a separate e-mail to sign it.

The data set is formatted as *csv (comma-separated values)* and contains two columns—the comment text and its annotations (named “text” and “label”, respectively). You'll find more information on this in Section 4.2 below.

3.2. Computing resources

We will provide GPU-resources via lambdalabs - here’s a first step tutorial on how to use them.

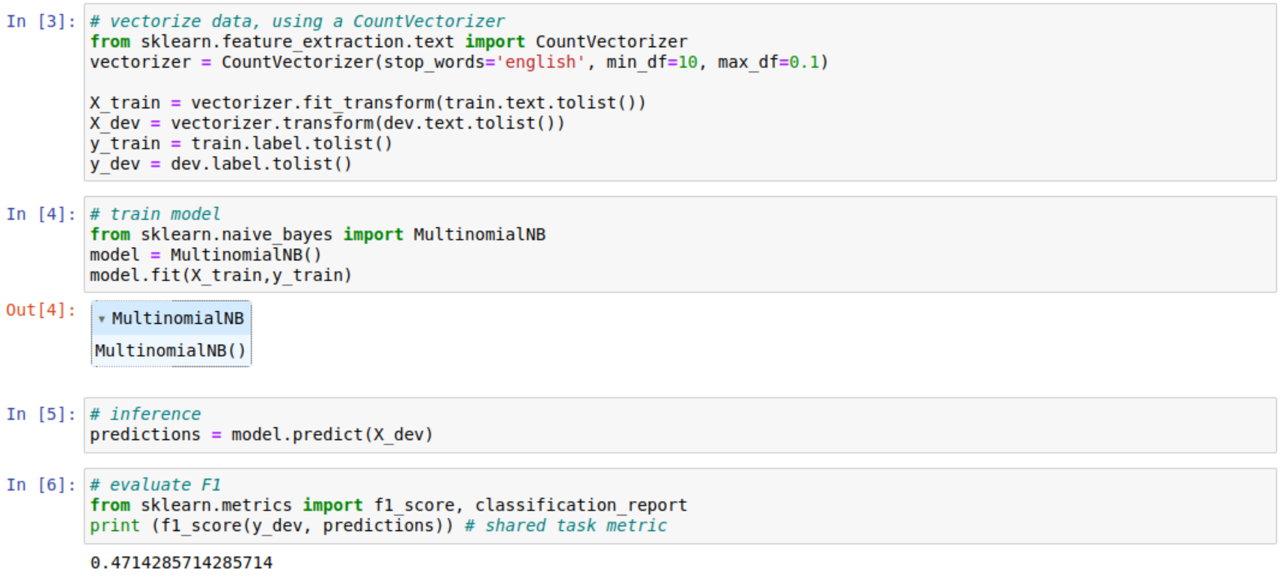
4. How to develop your NLP system

4.1 Rules of the game

You are free to use all methods you think are adequate. Be it rule-based, dictionary-based methods or advanced pre-trained neural networks such as BERT. You will work in groups of 2-4 persons.

4.2 An example

To familiarize yourself with the task before the day of the Hackathon, we encourage you to inspect the 2019 GermEval Offensive Language data (overview, data sets [labels are 0 = not offensive, 1 = offensive]). We transformed this dataset into the same format in which the hackathon data will be provided. However, consider that this exploratory data is from a different context (tweets instead of newspaper comments) and country (Germany instead of Switzerland). Thus, it might be possible that methods which work well on this dataset do not necessarily translate perfectly to the data provided during the hackathon.
We provide a jupyter notebook with sample code to train a Naive Bayes classifier to detect offensive language in the GermEval 2019 dataset here.



```
In [3]: # vectorize data, using a CountVectorizer
        from sklearn.feature_extraction.text import CountVectorizer
        vectorizer = CountVectorizer(stop_words='english', min_df=10, max_df=0.1)

        X_train = vectorizer.fit_transform(train.text.tolist())
        X_dev = vectorizer.transform(dev.text.tolist())
        y_train = train.label.tolist()
        y_dev = dev.label.tolist()

In [4]: # train model
        from sklearn.naive_bayes import MultinomialNB
        model = MultinomialNB()
        model.fit(X_train,y_train)

Out[4]: ▾ MultinomialNB
        MultinomialNB()

In [5]: # inference
        predictions = model.predict(X_dev)

In [6]: # evaluate F1
        from sklearn.metrics import f1_score, classification_report
        print (f1_score(y_dev, predictions)) # shared task metric

        0.4714285714285714
```

4.3. Helpful documents

- The dataset provided during the shared task is a subset of the dataset described in (Kotarcic et al., 2022)
- An overview of hate speech datasets can be found here
- Material from a workshop held in July 2023 on Online Abuse and Harm
- Further resources, tools and literature can be found at the bottom of this page

5. How we evaluate your solutions

At 16:50 sharp, we will distribute a *test data* set containing2,000 comments (without annotations!) and ask each team to use their NLP system to make predictions for each comment in the test set. You will then return these predictions to us.

We will evaluate all predictions by comparing them to the ground-truth annotations (to which only the organizers have access) and rank solutions according to highest F1 scores on detecting hate speech comments.

6. Prices

We distribute the following prizes to the three teams obtaining the highest F1 scores on hate speech comments.

1. 1’000 CHF voucher for Galaxus
2. 800 CHF voucher for Digitec
3. 500 CHF voucher for Zalando

Prizes are for the whole team and can be split up into individual vouchers for each team member if requested. For example, if the winning team consists of 4 members, each member would receive a 250.- voucher.