Profiling Hate Speech Spreaders on Twitter

Introduction

Hate speech (HS) is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. Given the huge amount of user-generated contents on Twitter, the problem of detecting, and therefore possibly contrasting the HS diffusion, is becoming fundamental, for instance for fighting against misogyny and xenophobia. To this end, in this task, we aim at identifying possible hate speech spreaders on Twitter as a first step towards preventing hate speech from being propagated among online users. The project's idea is to develop a model that can determine whether its author spreads hate speech, given a Twitter feed.

Dataset

The dataset contains timelines of users sharing hate speech towards, for instance, immigrants and women, 100 training cases/authors each (with 100 tweets per author). The data directory contains a XML file per author (Twitter user) with 100 tweets. The name of the XML file corresponds to the unique author id. Moreover, a truth.txt file with the list of authors and the ground truth. The format of the XML files and the truth.txt are as follow:

```
<author lang="en">
        <documents>
            <document>Tweet 1 textual contents</document>
            <document>Tweet 2 textual contents</document>
            ...
        </documents>
</author>
```

```
b2d5748083d6fdffec6c2d68d4d4442d:::0
2bed15d46872169dc7deaf8d2b43a56:::0
8234ac5cca1aed3f9029277b2cb851b:::1
5ccd228e21485568016b4ee82deb0d28:::0
60d068f9cafb656431e62a6542de2dc0:::1
...
```

The dataset can be downloaded from [1].

Mission

The final objective is to build a system that can predict if a user in social media is a hate speech spreader or not. The output of the system can be used as a filtering module in online social media.

To do that, the following tasks need to be accomplished:

1. Understand the content that is available in the dataset .
2. Pre-process the data to improve the final results.
3. Baseline model
    a. Follow a Support Vector Machine (SVM) or RF based approach to classify users.
    b. Use tfidf as input features
    c. Analyze the outcomes.

    OPTIONAL:
    d. Test different lengths of N-Grams and report the best N in which the classifier achieved the best performance
    e. Report the top 10 representative (most repeated) 1, 2 and 3-grams for each of the classes.

4. LSTM model
    a. Build a LSTM (Long Short-term Memory) text classifier with the same goal as above.
    b. Analyze the outcomes.
5. Compare the results of both models using the F1 measure.
6. Optional: Think about further steps to improve your performance, such as pre processing, feature embeddings, post processing etc. and try them out if you have time

Additional Hints

You can use whatever data you can find additionally to the data we provided, but don’t spend too much time on searching.
This project is about applying the skills you learned during GDA. It is not about developing the perfect solution for a given problem.
The project might be a little bit harder than what you are used to. Take it as a “bridge” to the practice phase.