FOLT Software Project

Luis Dreher **Jan Buchmann**

1 Project Documentation

The 2019/20 FOLT software project consists of two tasks that work with the same data. This data is a set of comments from “Wikipedia talk page edits” (The FOLT 2019/20 Practice Class Organizers, 2020). In these comments, Wikipedia users discuss edits of Wikipedia articles. The first task is a *shared task*, in which the comments are to be classified as *toxic* or *non-toxic* based on the language used. The classification is evaluated with the *accuracy* measure, which gives the proportion of correctly classified comments.

For the classification, one of the classifiers provided in the natural language toolkit (NLTK) (Bird et al., 2009) is to be used. We decided to use the naive Bayes classifier, as this is a well-established tool for this purpose. The most well-known application of naive Bayes classifiers in text classification is spam e-mail detection (Sahami et al., 1998), which is a somewhat similar task to the one described here.

In the training, the classifier is supposed to “learn” specific properties of the comments that distinguish toxic from non-toxic comments. These properties are often called *features*, and the main challenge of this task was to select

To train the classifier, a part of the data (the train split, consisting of 1800 comments) was provided with a *toxicity* label for each comment. These labels had been obtained by human annotation. Inspection of the train data revealed that 877 comments were labeled as toxic and 923 were labeled as non-toxic. This means that the dataset is fairly balanced in terms of class distribution. A non-balanced class distribution can cause problems, because this might mean that there are not enough examples to learn informative features for some of the classes.

2 Results Analysis

References

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.

M. Sahami, S. Dumais, D. Heckermann, and E. Horvitz. 1998. A bayesian approach to filtering junk e-mail. In *Technical Report WS-98-05*, pages 98–105. The AAAI Press.

The FOLT 2019/20 Practice Class Organizers. 2020. FOLT Software Project Task Description.