KI2 Lab 2: Supervised learning

Y. S. Antille, M. E. Solèr, J. Wichser
{*antilyas, solerma1, wichsjoe*}*@students.zhaw.ch*

1 Introduction

This hand-in is organised as follows. First, we evaluate several classification models and approaches and analyse their performance, mainly with the F1-score. Then, we utilise the most successful model in an online hatespeech-evaluator.

All of the models listed here use stemming and removal of non-specific words (stopwords). The performance of the models is summarised in Table 1.

2 TF-IDF-based methods

In the first model, TF-IDF for feature extraction in conjunction with a linear support vector machine is used. Although its overall performance is acceptable for many applications (the F1-score is 89%) and it performs well with non-hate comments, it misclassifies almost a third of hate-comments.

Our first attempt at providing a remedy to this problem is by using AUTO-SKLEARN[1] leading to our second approach. AUTO-SKLEARN automatically selects the model that performs best from a range of algorithms, preprocessors and parameters[2]. The selected model also consists of a linear SVC, and can thus be understood as improvement of the default approach. Its F1-score is however significantly better (95%). In particular, this model detects hate-speech more reliably, scoring 11 F1 points better than the default model.

Comparison of our methods

Method	F1-Score	F1-Neg.	F1-Pos.
TF IDF with linear SVM	89%	94%	60%
auto-sklearn	95%	97%	71%
Word2Vec with ETC	89%	84%	4%
Word embeddings	96%	98%	77%

Table 1: The methods are shown with three scores: **F1-Score** describes the overall score for correctly identified sentences, **F1-Neg.** quantifies the score of non-hate-speech sentences and **F1-Pos.** describes the score of hate-speech sentences.

3 Word-embedding methods

Due to the relative success of the WORD2VEC model in the Lab 1, we evaluate a similar approach. Using a word-embedding technique instead TF-IDF, semantical similarites among words are represented.

In our third model, we use the well-known WORD2VEC method for feature extraction and classify the words using a ExtraTreesClassifier (ETC). The ETC belongs to the class of Random Forests, but is less susceptible to overfitting because the individual tree's prediction is averaged instead of used as one vote for a class. The results are however poor. Although the overall F1-score is similar to the first model, it misclassifies most hate-speech sentences. Due to the more promising approaches, we do not pursue it further. We assume that the poor performance stems from a false application of the WORD2VEC method.

Finally, we introduce the fourth model. It consists of a neural network, which is composed in TENSORFLOW. In contrast to the prior models, the Neural Network takes care of all analysis steps: vectorisation, embedding and classification. The model's flow of data for a single sentence is is described in Figure 1.

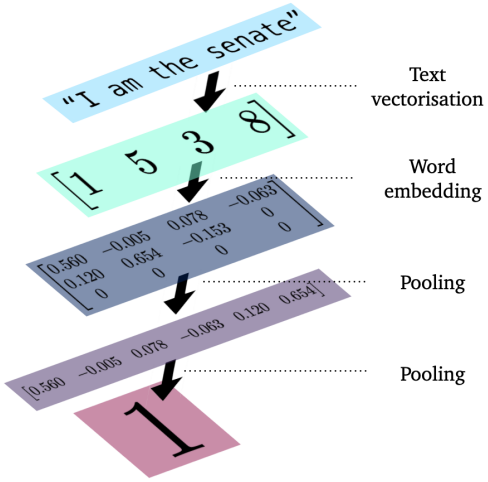

Figure 1: The structure of word embedding model. First, the sentence is vectorised using a bag-of-words model. Then, the vectors are transformed into a larger matrix, where semantic similarity is considered. Empty cells are padded with zeros. Over two pooling layers, the embeddings are transformed into a binary output.

The neural network is trained over 15 epochs, with the learning curve rendered in Figure 2. As the learning rate of the training data surpasses the one of training data, the model suffers from overfitting with more epochs. This is however not a great concern, because it still generalises well after few epochs (and thus, before overfitting).

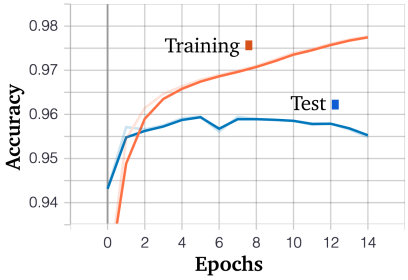

Figure 2: The accuracy of the model (y-axis) vs. epochs (x-axis). With more epochs, the model appears to overfit, and generalise less.

4 Application

To make practical use of the model, we introduce DETOXER, a web-based application that evaluates a sentence and decides whether it is considered as hate speech or not. Figure 3 shows its main interface. DETOXER is available under `https://detoxer.herokuapp.com`

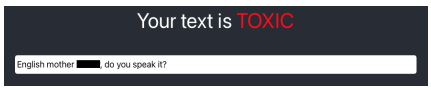

Figure 3: The interface of DETOXER. In this case, hate speech was detected.

[1] *Efficient and Robust Automated Machine Learning* (2015), Matthias Feurer et al.
[2] We evaluated 65 algorithms, with an ensemble size of 1