COMP 550: Programming Assignment 1

Jan Tiegges

September 28, 2023

1 Introduction

In this assignment, the use of different linear classifiers as well as the impact of different choices regarding preprocessing, feature extraction and hyperparameters was investigated based on a toy experiment. The task was to classify short statements about animals whether it is a true or an imaginary one. The primary research question was to find out whether linear classifiers are suitable for such tasks and which effect the choice of a classifier has.

2 Methodology

The experiments used data generated by ChatGPT by providing clear guidelines and examples to maintain quality. 230 statements were obtained for 23 animals (10 facts and fakes each) and the combination with additional data from students resulted in 1030 statements for each category. For preprocessing, the data was split into individual tokens and then optionally lemmatization and porter stemming were applied and the stop words were removed. For feature extraction, the data was split into an 80/20 train/test split and processed using a count vectorizer with unigrams. Classification was performed using Naive Bayes (NB),Logistic Regression (LR) and Linear Support Vector Machines (SVM). A 5-fold cross-validation optimized the regularization parameter C for LR and SVM and the smoothing parameter α for NB as well as the choice of preprocessing options for each model. The LR solver was "Saga" with a maximum number of 5000 iterations.

3 Results

4 Discussion

not general. The strongest assumption is that the type of language/words used indicates wether it is a fact or not. This might be true for very easy cases, but if facts would sound more realistic, then the model does not work anymore. Also, it does not take into account the context of the animal, which also is a strong assumption. In the case where the made-up facts are actually very creative and non-sense like, while the facts are more neutral scientifc ones, then these assumptions actually work out reasonable and make the model perform well.

5 Conclusion

gave some good hints on the interplay between choosing the right pre-processing, feature extraction and classification models for such a task. Generally one should first conduct some data analysis to get a good understanding of the data which can be used as prior knowledge to make some model decisions upfront. This task however could already show how strong simple mechanisms and linear classification already work on data sets like this.

References