Machine Learning – Mini Project 2
Marie-Helen Lindbäck

1. Introduction

This report presents using machine learning techniques for sentiment analysis on tweets. Sentiment analysis aims to understand the emotional tone of text data categorizing it for example as positive or negative. Two machine learning models was used for the task, recurrent neural network (RNN) and Naive Bayes.

2. Data processing

The data consists of 160 000 tweets with sentiment labels positive or negative. The tweets were preprocessed by removing usernames, punctuation and numbers. They were converted to lowercase and by using lemmatization the words were converted to their base form. Additionally stopwords were removed. For the RNN the data was further preprocessed. The text was converted to sequences of integers using a tokenizer. Sequences were then padded to a fixed length. Pre-trained word embeddings, GloVe, were incorporated to represent words numerically. For the Naïve Beyes TF-IDF (Term Frequency-Inverse Document Frequency) was used to convert text into a numerical feature matrix.

3. Modelling

For the Recurrent Neural Network (RNN) a sequential model was created using Keras. An LSTM layer with 128 units processed the data. A final dense layer with sigmoid activation provided the probability of a tweet being positive. Adam optimizer and binary cross-entropy loss were used for training. The model was trained on 80% of the data with a validation split of 20%. Training lasted for 6 epochs with a batch size of 128. Training and validation accuracy/loss plots were generated to monitor the model's learning progress.

For the Naïve Bayes classifier a TF-IDF vectorizer extracted features representing the relative importance of words in each tweet, and a multinomial Naïve Bayes classifier was trained on the features.

Both the RNN model and the Naïve Beyes achieved an accuracy of 75 %. Figure 1 displays the accuracy of the RNN after each epoch.

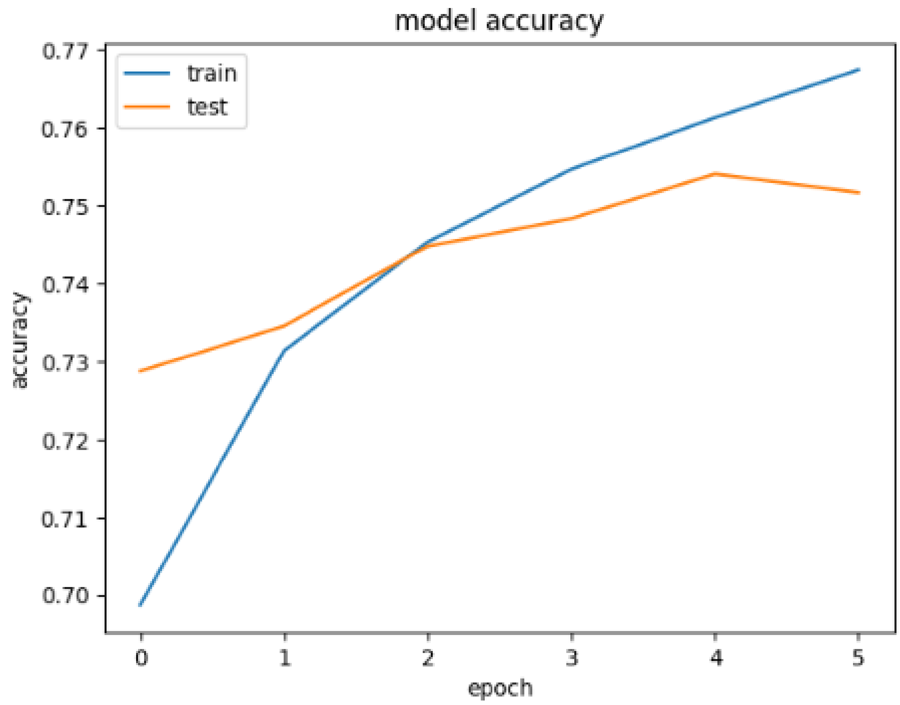

Figure 1. The accuracy of the RNN model

Potential improvements could be exploring hyperparameter configurations such as learning rate, number of LSTM units could potentially lead to better results. Other preprocessing techniques could also be explored.

7. Conclusion

In this report the use of two differnet machine learning models for sentiment analysis was explored. The two models used was Recurrent Neural Networks and Naïve Bayes. One bottleneck that was encountered was the many options available for preprocessing the natural language data. Some of the techniques also revealed to be computationally expensive, such as lemmatization, which lead to the choice of a simpler lemmatization technique. Additionally, the RNN model was computationally expensive to train, which also made the tuning of the model challenging.