NAME: Axel Bremer
STUDENTID: 11023325
MAIL: axel.bremer@student.uva.nl

NAME:Rochelle Choenni
STUDENTID:10999949
MAIL:rochelle.choenni@student.uva.nl

Sentiment Classification
Project Proposal
2019-09-13

NAME:Tim de Haan
STUDENTID:11029668
MAIL:tim.dehaan@student.uva.nl

NAME: Shane Koppers
STUDENTID:12233188
MAIL: shanekoppers@hotmail.com

1 Dataset

We will use the IMDB Large Movie Review Dataset provided with the task suggestion.

2 Pre-processing

- Lowercase data
- Removing punctuation
- Encode words and labels
- Removing outliers (extremely long or short reveiws)
- Padding/truncating data: For reviews shorter than seq_length, we will pad with 0s. For reviews longer than seq_length we will truncate them to the first seq_length words.

3 Initial model

Use LSTM neural network with pre-trained word embeddings. We will implement the model in PyTorch. Since we have determined that the average length of the reviews is 230 words, we're not sure yet whether we want to feed the review word-by-word or sentence-by-sentence.

4 Unsupervised component

Check for explanation with TAs.

5 Relevant literature

[1] implements a sentiment classification algorithm using neural networks, that is, a combination of CNN and LSTM networks. The performance of the presented algorithm outperforms other families of algorithms that were considered to be the state of the art at the time. Hence, this paper can give a useful overview of how to structure a neural network such that it can be successfully used for sentiment classification.
[2] performs a comparative analysis of several sentiment analysis algorithms, including CNN and LSTM networks. The paper concludes that the LSTM network shows the best performance, thus supporting our choice for a LSTM network.
https://medium.com/@pi19404/using-pre-trained-word-vector-embeddings-for-sequence-classification-using-lstm-277dee188348

References

[1] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

[2] Wlodek Zadrozny. A comparison of neural network methods for accurate sentiment analysis of stock market tweets. In *ECML PKDD 2018 Workshops: MIDAS 2018 and PAP 2018, Dublin, Ireland, September 10-14, 2018, Proceedings*, volume 11054, page 51. Springer, 2019.