Big Data and Language Technologies - Project Exposé

Nico Schmidt, Marvin Müller, Klaus Schmidt

July 2, 2022

1 Main Research Question

Our main focus is to create a robust method to extract statements about the future from website data and to improve the accuracy of this process. If this method performs well, we want to analyse the extracted statements for sentiment.

2 Basic Research Plan

To answer those questions we need to dig deeper on how to define such a statement. We have access to several petabytes of web-archive data, which we are able to extract with the given web-archive-keras pipeline [1]. Using regular expressions, we want to extract about 2000 sentences that could include statements about the future. From those candidates we want to create a gold standard dataset by manually labelling those sentences whether they contain statements about the future or not. After labelling, this dataset will be shuffled and afterwards split into training, test and validation data in a ratio of 60 - 20 - 20. Depending on the distribution of classes we need to perform data-balancing on the training set to prevent a class from being underrepresented in the training set [2]. With those datasets and the use of active learning [3] we want to fine-tune a pre-trained transformer model such as the BERT derivation RoBERTa [4]. To find the best parameters and finally evaluate the model metrics like accuracy, precision, recall and the F-score may be relevant.
If we are able to train a well performing model we can extract another 2000 sentences from the web-archive data that now definitely (or primarily) contain statements about the future. Now we want to label these sentences depending on their sentiment value, e.g. whether the statement is positively or negatively intended. Afterwards we want to train a transformer model as described above that categorises the given statements.

3 Group Roles

The role distribution for the upcoming tasks will be as follows:
Nico Schmidt will be responsible for extracting and pre-filtering the WARC-files from the cluster with regular expressions. With the filtered data we will then be able to annotate and create our gold standard dataset. This task will be split equally between us. The deployment process as well as the preparation of our used data will be done by Marvin Müller. This includes among others the creation of the train-test-validation split and the precautions to prevent train-test leakage, such as shuffling and balancing the annotated data. To improve the performance of our chosen model, we want to use an active learning approach. This task will be implemented by Klaus Schmidt. All general tasks, like documentation, will again be conducted by all of us in an equal manner.

References

[1] `https://github.com/webis-de/web-archive-keras`

[2] C. D. Larose and D. T. Larose. *Data Science Using Python and R*. 2019. doi: 10.1002/9781119526865

[3] C. Schröder, L. Müller, A. Niekler and M. Potthast. *Small-text: Active Learning for Text Classification in Python*. 2021. arXiv preprint arXiv:2107.10314.

[4] `https://huggingface.co/docs/transformers/model_doc/roberta`