TDDE09 Project status report

What has been done

Starting with what we are supposed to have done.

- Gather and label data (this refers to our own collected news), DONE.
- Clean up, label and tokenize collected news, DONE
- Translate financial dataset and change both datasets to a uniform labelling system DONE
- Import a Swedish bert and set up for document classification (SEMI-DONE)

We have set up a bertForSequenceClassification in a nearly identical way to lab3x, and have been able to fine-tune this model on a dataset of financial phrases. This model achieved 88.6% accuracy on a held out dataset of similar manner.

The same model preformed with an accuracy of 39% on a small sample of the translated amazon-reviews from our review dataset, this was a bit underwhelming, but also understandable given the dissimilarity of the subjects.

Current problems

1. No setup for docbert currently. A problem we are facing is that the amazon-reviews (max length 1232 words) and some of our own collected news, are way longer than the financial phrases (max length 78 words). In this article https://arxiv.org/pdf/1910.10781.pdf the authors suggest splitting long pieces of tokens into segments of a fixed size with overlap and feeding these into an LSTM. This might prove to be difficult.
2. Our Swedish dataset of reviews turned out to be labelled sub-optimally. We have addressed this by switching review dataset to a set of amazon reviews which we have translated.
3. Training is taking long. Fine tuning a BERT on the set of financial phrases, 4000 sentences between 2 and 78 words long, took about 40 minutes. We believe that this will take even longer for the long amazon-reviews. We have tried to mitigate this by saving this model so that we can load it without training it again, and by sorting sentences by length before batchifying. The sorting improved performance slightly.

What will be done this week

This week we will preform the experiments, and hopefully address our current problems in an efficient manner. We will start by contacting the examiner of the course in hope of receiving some guidance.