Preliminary Results

1. Problem statement

As mentioned in deliverable 1, the proposed project is a sentiment analysis classifier for financial news. The Naive Bayes model categorizes sentiment into three categories: negative, neutral, or positive. Further, named entity recognition will allow the detection of the organization/company in question.

2. Data Preprocessing

The data preprocessing methods on the fiancial_Phrasebank dataset were modified to increase model accuracy. Instead of repeating underrepresented classes (oversampling), the dataset was truncated (undersampling). Text cleaning (deletion of punctuation and numbers) was conducted for a more accurate TfidfVectorizer.

3. Machine Learning model

The chosen machine-learning model is a Naive Bayes algorithm implemented through the scikit-learn MultinomialNB library. This algorithm relies on the simple Naive Bayes theorem. The dataset was separated into 90% train and 10% test using scikit-learn's train_test_split library. The decision for a 90/10 split rather than a traditional 80/20 split is due to the small amount of data available. The test set's accuracy score was calculated to validate the model. It was found that the model correctly fitted the model. The named entity recognition was implemented through the spacy library and programmed to select the most-recurring entity (NER).

4. Preliminary results

a) Accuracy score: 81%

b) Confusion Matrix:

57	**8**	**3**
4	**45**	**5**
3	**9**	**45**

The project remains feasible, considering a high accuracy score for financial news sentiment analysis and a high ratio of True Positives for all three classes. Additionally, when tested with large news article sizes (multiple paragraphs), the model's overall performance seems unchanged, as with NER.

5. Next Steps

- Simplify preprocessing pipeline