**DEPARTMENT OF ECONOMICS**

UNIVERSITY OF STELLENBOSCH

Correlating Factors of U.S. Presidential Speeches with Stock Market Movements – a Deep Learning Approach

Experiment design and results analysis

By Pablo Rees

[19119461]

Results analysis formally presented with intention of fulfilment of the requirements for completion of the MCom Economics Programme (full thesis only) at Stellenbosch University.

With supervision from Dawie van Lill

[08 June 2022]

# Experiment Design

Two sets of experiments were run on two categories of data design. The first is a speech centred design. For every speech in this dataset there is a row of data attached. Thus there is no data for dates when no speeches occurred and there are duplicate dates because there are days when more than one speech occurred. This dataset is the larger of the two datasets and contains 35251 rows of data – one for each unique speech since 1998-01-01. The second dataset uses a date centred design. For every date (weekday) there is one row in the dataset. Thus on dates when more than one speech occurred the speech data has been aggregated into a single vector (see the Doc2Vec methods section for an explanation of this method). The date centred dataset contains 25 383 rows of data – one for each trading day between 1950-01-01 and 2022-03-22. The speech centred and date centred datasets are exemplified in Table 1 and 2, respectively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Date** | **Speech** | **LD\_Date\_Resid\_1** | **V1** | **V2** | **V3** |
| 1950-01-02 | Good morning… | 0,02 | 3,6 | 2,6 | 1 |
| 1950-01-06 | Ladies and… | -0,05 | 2,1 | 2 | 0,36 |
| 1950-01-06 | Congress… | -0,05 | 0,12 | 3,4 | 2,8 |
| … |  |  |  |  |  |

Table 1:Example of the speech centred data design – Note that there is no data for dates that no speeches occurred and that there are duplicate dates. Note also that on dates where more than one speech occurred there is duplicate data in the LD\_Date\_Resid\_1 column which is representative of the autoregressive data in this example. Finally, note that the V1 and V2 columns (representative of the vectorized speech data) is different for every row of data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Date** | **Speech** | **LD\_Date\_Resid\_1** | **Days since last speech** | **V1** | **V2** |
| 1950-01-02 | Good morning… | 0,02 | 0 | 1,2 | 1,8 |
| 1950-01-03 | N/A | 0,025 | 1 | 1,2 | 1,8 |
| 1950-01-04 | N/A | -0,03 | 2 | 1,2 | 1,8 |
| 1950-01-05 | (Multiple speeches) | 0,05 | 0 | 2,1 | -0,6 |
| 1950-01-06 | Today marks… | -0,03 | 0 | 5,3 | -0,86 |
| 1950-01-09 | N/A | 0,04 | 3 | 5,3 | -0,86 |
| ... |  |  |  |  |  |

Table 2:Example of the date centred data design – Note that every consecutive trading day is included in the data. Note also that on days when no speeches occurred the V1 and V2 (representing vectorized speech data) from the last date that a speech occurred are duplicated and the number of days since the last speech is recorded. Finally, note that on days when multiple speeches occurred there is only one row of data – i.e. the speeches are amalgamated into a single vector.

For the speech centred dataset a total of 1152 models were tested. These consisted of 384 regression tasks and 768 classification tasks. While for the date centred data 336 models were tested. These consisted of 112 regression models and 224 classification models. The models were selected by incrementally altering a single hyperparameter across 5 fields for regression and 6 fields for classification. These fields are detailed in the following sections.

### Regression fields

The 5 fields of hyperparameters for regression were StartDates, Remove\_duplicates, Reg\_Types, Reg\_algos and Datasets. StartDates refers to the date from which the dataset began: 1998-01-01, 2000-01-01 and 2010-01-01. The 1998 dataset runs from 1998-01-01 until 2022-03-22, the 2000 dataset runs from 2000-01-01 until 2022-03-22 etc.

The Remove\_duplicates field divides the speech centred data into two subsets – one with all the duplicate dates removed and one including the duplicate dates. Duplicate dates occurred because some speeches occurred on the same date. This hyperparameter is necessary because the Meta and Auto datasets only contain one value in each field – thus duplicating dates duplicates Meta and Auto data which creates the risk of data contamination between the train and test sets. This field is not valid for the dates centred dataset and is set to ‘False’ for all the models using it.

The Reg\_Types field contains two options, TS\_Regressor and CS\_Regressor. These refer to the manner in which the data is split for cross validation. The TS\_Regressor cross validation method splits the data into 5 time consecutive subsets using the ‘TimeSeriesSplit’ out of sample (OOS) method from Sci Kit Learn. Model training is done on the first subset, then the first and second subset etc. up until the fourth subset. Training scores are calculated for each of the four training subsets and the maximum score is passed forward as the final representation of the training score. Test scores are calculated for only the fifth (and latest) split. This method is employed to force the algorithm to perform prediction instead of interpolation. It avoids giving the models access to the future of a data trend when making predictions for a data point. However, the method is not strictly necessary in this case because the dependent variable has been centred and is close to normally distributed – i.e. it does not have a trend (Bergmeir, Hyndman & Koo, 2018). The CS\_Regressor option performs regular 5-fold cross validation.

The Reg\_algos field refers to the four different ML algorithms available for training. These are stochastic gradient descent, a 3 hidden layer neural network, multiple linear regression and gradient boosting. The hyper-parameters for each of these fields can be found in Appendix B.1.

Datasets is a complicated field. There are three groups of prediction sub-datasets in the speech centred dataset – the control subset: X\_control ; the meta subset: X\_meta ; and the test subset: X\_test. X\_control contains the autoregressive variables described in section \_\_\_\_\_, X\_meta contains the meta variables (other possibly relevant financial variables, such as the dollar strength index) and X\_test contains the variables of focus – the NLP variables. However, in total there are 458 variables across these three subsets. The X\_test subset makes up the majority of this. There are 4 variables derived from the VADER sentiment analysis, 2 variables from the textBlob sentiment analysis, 200 variables from Word2Vec, 200 variables from Doc2Vec\_200 and 20 variables from Doc2Vec\_20. In order to minimize training times and avoid multi-collinearity the X\_control dataset was reduced to contain only 8 variables, X\_meta was reduced to contain 5 variables and X\_test was reduced to contain 26 variables. (Run an elastic net or lasso to deduce the datasets selected for this section).

The date centred dataset also contains three groups of prediction subsets. These are the same X\_control dataset as in the speech centred data, and two versions of the vectorised speech data – the paragraph vector distributed bag of words (PVDBOW) dataset and the paragraph vector distributed memory (PVDM) dataset – both detailed in section \_\_\_\_\_\_. A detailed description of the variables available and tested is available in Appendix B.3.

The final X\_control variables selected for both the speech centred and the date centred datasets were DlogDif\_1, DlogDif\_2, absDlogDif\_1, blackSwan\_SD3\_1, blackSwan\_SD4\_1, blackSwan\_SD5\_1, stdVol\_1DateResid and pos\_neg\_transform. The final X\_meta variables selected were Nasdaq\_ld\_1, Oil\_ld\_1, SSE\_ld\_1, USDX\_ld\_1 and VIX\_ld\_1. The final X\_test variables selected were the VADER scores, the TextBlob scores and the set of 20 Doc2Vec variables. While the PVDBOW and PVDM datasets each contained all 20 of their variables and the number of days since the last speech.

The datasets field provided the option for any combination of each data designs’ three subsets as training variables for a sum of 7 combinations each. Additionally a PossibleBest set of variables was selected for the speech centred dataset using an Elastic Net algorithm from SciKitLearn (hyperparameters available in appendix B). This PossibleBest dataset consisted of DlogDif\_1, DlogDif\_2, pos\_neg\_transform, Nasdaq\_ld\_1, Oil\_ld\_1, VIX\_ld\_1, DV\_20\_6, DV\_20\_8, DV\_20\_13, DV\_20\_15.

### Classification fields

The 5 fields of hyperparameters for classification were StartDates, Remove\_duplicates, Binary, Clf\_Types, Clf\_algos and Datasets. StartDates, Remove\_duplicates and Datasets are identical to their equivalents in the regression section above.

Binary refers to the Y-variable (classification fields) being predicted. The options for Binary are True and False. If False is selected the continuous Y-variable is split into 8 categories denoted by the numbers 1-8. These categories represent a number of standard deviations from the mean of the input Y-variable. See Equation 1.1.2.1. If Binary is set to True then the Y-variable is split into 2 categories denoted by 1 and 0 which indicate whether the entry is above or below the mean of the continuous Y variable. See Equation 1.1.2.2.

*Y*

Equation 1.1.2.1: Conversion of continuous Y variable to a non-binary categorical variable

Equation 1.1.2.2: Conversion of continuous Y variable to a binary categorical variable

Clf\_Types refers to the Time Series cross validation and Cross-section cross validation methods as described in the regression section. The options in this category are CS\_Classifier and TS\_Classifier. Clf\_algos is very similar to the Reg\_algos field described in the previous section. The algorithms available for classification are stochastic gradient descent, a 3 hidden layer neural network, logistic regression and gradient boosting. The hyper-parameters for each of these fields can be found in Appendix B.2.

# Results and discussion

The results of the models indicated that US Presidential Speeches hold at least a small amount of predictive power over movements in the S&P 500 stock index. This conclusion was deduced from the fact that the best performing datasets in terms of test accuracies or test mean absolute errors across both data design types, and classification and regression tasks all included vectorized speech data.

The date centred dataset outperformed the speech centred dataset by about 5 percentage points (~0,55 vs ~0,6) on the test set accuracy score for binary classification analysis. Notably, across both datasets the time series binary classification outperformed the other classification tasks in terms of maximum accuracies achieved while predicting the unseen test data.

The date centred dataset also outperformed the speech centred dataset in regression tasks. Notably \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_...

Interestingly, an implication of the superior performance of the (smaller) date centred dataset is that data with high quality has outperformed data with high quantity. The section below explains the results in more detail.

## Speech centred regression analysis

The average Mean Absolute Error (MAE) recorded across all 384 regression models run was 0,87 for the training data, 0,95 for the test data and 0,8 for the validation data. Comparing these with a standard deviation of 1,26 for the *logDif\_date\_resid* variable after 1998-01-01 shows that the average absolute error across all the regressions run lies within one standard deviation of the dependent variable indicating that the regression prediction is a better predictor than the mean of the series. This holds across each of the three date subsets (from 1998, 2000, and 2010, respectively) as well.

The top ten performing models in regression tasks were all trained on the cross-sectional 2010 dataset and achieved MAE’s between 0,68 and 0,7. Interestingly, there was an even split between datasets with duplicates removed and without. The best performer did not have duplicates removed whilst the second, third and fourth best performers did. Analysis of only data with duplicates not removed gave the top four places to gradient boosting models trained on the AutoMeta, All, Meta and Auto datasets (in that order). The MAEs for the top four performers ranged from 0,68 to 0,7. NLP inclusive datasets ranked in five of the top ten places. Analysis on the duplicates removed top ten performers shows the NN taking 8 of the top ten places – including the top seven spots. NLP inclusive datasets ranked well taking the second, third, fourth, fifth and seventh spots but not beating the purely autoregressive data.

It is difficult to draw a solid conjecture about the predictive power of the NLP data from these mixed results. However, data contamination in the duplicate inclusive dataset is likely and the results should probably be discarded. Given the fair performance of the NLP inclusive datasets in the duplicates removed dataset, at this point, it remains possible that the NLP data holds predictive power.

A final analysis of all the regression models trained on only 1998 NLP data from the speech centred dataset shows that the top ten performing models all achieved a MAE below 1,06. This is still below the standard deviation of 1,26 for the 1998 *logDif\_date\_resid* variable and thus indicates that regression models trained only on the NLP data outperform the mean of the series as a predictor. This evidence points towards US Presidential Speeches having predictive power over S&P 500 movements.

## Speech centred classification analysis

Analysing the best performing classification models, in terms of test set accuracy, across all categories of the speech centred data reveals a very high likelihood of data contamination in the data containing duplicate dates. All ten of the top performers are on data containing duplicates and the accuracies range from 0,72 to 0,80. It is highly unlikely that these models have managed to achieve 80% accuracy in prediction of the S&P 500 on unseen data. Thus results of models trained and tested on duplicate date containing data is discarded henceforth.

Of the duplicate removed results the top ten performers contain seven models trained and tested on the 2010 dataset – including the top four. Further, eight of the top ten models performed binary classification. Finally seven of the ten were trained on cross-sectional cross validation. These initial mixed results re depicted in Figure 2. Thus further analysis will take place within the binary, cross-sectional and 2010 categories. The test set accuracy ranges from 0,54 to 0,58. The mean of the target variable is ~ 0,5 and thus every model in the top ten models outperforms chance (recall that the target variable is binary). NLP data is included in six of the ten best models, while autoregressive data is included in five and meta data is included in eight of them. Notably an NLP only (gradient boosting) model ranks fourth with a test set accuracy of 0,56. This constitutes evidence of the predictive power of US Presidential speeches over S&P 500 movements and strengthens the evidence presented in section 2.1.

![Chart, line chart

Description automatically generated]()

Figure 1: Test set accuracy for cross-section classification across starting dates for binary and non-binary NLP datasets across all four classification algorithms.

![Chart, line chart

Description automatically generated]()

Figure 2: : Test set accuracy for time series classification across all starting dates for binary and non-binary NLP datasets across all four classification algorithms. Note that the binary Gradient Boosting (blue dotted) line lies under the non-binary NN (yellow solid) line between 2000 and 2010.

## Date centred classification analysis

The means of both the full set of 1950-01-01 binary Y’s and the test set of 1950-01-01 Y’s are maintained at ~ 0,5. This implies that any accuracy above 0,5 beats chance.

In the Date centred classification results analysis the top ten models in terms of test set accuracies contain ten 1950 subsets, ten TS subsets and eight binary subsets. All ten positions datasets contained autoregressive data, five positions included the PVDBOW vector set and two included the PVDM vector set. The range of the top ten models test set accuracies is 0,57 to 0,59. Thus the date centred dataset initially outperforms the (duplicate dates removed) speech centred dataset by 2 percentage points at the bottom of the range and 1 percentage point at the top of the range in terms of test set classification accuracy. Because of this the models trained on the date centred dataset were optimised for maximum performance. Log-regression is initially the best performing classifier in the TS section generally taking the top two and the seventh positions. However, AutoPVDBOW outperforms Auto in the Binary NN and the Binary Gradient Boosting categories. Thus, these three models were selected for optimisation.

### Optimization of classification models for TS classification of date centred data

The training set accuracy on the 1950 binary AutoPVDBOW for the gradient boosting classifier was initially 77,5% indicating that it may have been slightly overfitting the training data. For the NN the training accuracy is only 59% indicating that it may have underfit the data. There is no validation score for the time series tests. Grid search should be performed to optimize NN and Gradient Boosting for the 1950 AutoPVDBOW binary dataset.

After closer optimization of hyper-parameters it becomes clear that the AutoPVDBOW dataset outperforms the Auto dataset on test score both overall and across both the Stochastic Gradient Descent and Log Regression algorithms. The best performing algorithm for the AutoPVDBOW dataset was the LogReg\_4 algorithm (please see appendix B for hyperparameters) which achieved a test set accuracy of 0,601 while the best performing algorithm for the Auto dataset was the NN\_7 algorithm which achieved a test set accuracy of 0,599. This difference indicates that there is at least a slight benefit to including the PVDBOW dataset in predictive data and strengthens the evidence for the predictive ability of the speech data.

## Date centred regression analysis

The initial ten lowest test set MAE’s across all the regression categories were in the 1950 and cross section categories. The standard deviation for the 1950 test set is 0.96. All ten of these MAEs clustered around 0,66 so the regressions are better predictors than the mean. Again the date centred dataset has outperformed the speech centred dataset. The best score was achieved for the PVDBOW dataset by the SGD algorithm. The Subsets including NLP data were included in seven of the top ten performing models and three of them only contained NLP data. Notably, the AutoPVDM subset outperformed the Auto subset in the NN model undeniably indicating the predictive power of the NLP data.

NN algorithms were responsible for six of the lowest ten MAEs and SGD models were responsible for a further 3. The PVDBOW dataset appeared twice, the Auto dataset 3 times, the PVDM dataset once, AutoPVDBOW twice, AutoPVDM once and AutoBoth once. In total the Auto dataset appeared in 7 of the top ten performers, the PVDBOW appeared in 5 and the PVDM dataset appeared 3 times. Given this and the predictive ability of the PVDBOW shown in the classification section above, PVDBOW and Auto data sets were compared across attempted optimizations of the NN and SGD algorithms. However, no significant improvement could be engineered and almost all models performed worse than the initial models. Given that the best performing dataset in terms of test MAE was the PVDBOW dataset it must be said that speech data has predictive power over the target variable.

# Conclusion