**DEPARTMENT OF ECONOMICS**

UNIVERSITY OF STELLENBOSCH

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

Appendix A

By Pablo Rees

[19119461]

Appendix A 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

September 1 2020

# Appendix A: Data collection and feature extraction

## A.1. Speech (NLP) data at collection and cleaning intervals

Table A 1: Speech data after webscrape and cleaning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | Name | Date | Title | Transcript | No Stops Transcript |
| elections-and-transitions/convention-speeches | /people/president/joseph-r-biden | 2020-08-20 22:48:00+00:00 | Address Accepting the Democratic Presidential Nomination in Wilmington, Delaware | ['good', 'evening', '.', 'ella', 'baker', 'a', 'giant', 'of', 'the', 'civil', 'rights', 'movement',…] | ['good', 'evening', '.', 'ella', 'baker', 'giant', 'civil', 'rights', 'movement',…] |
|  |  |  |  |  |  |

## A.2. Feature extraction parameters

Word2Vec – the [Gensim Word2Vec module](https://radimrehurek.com/gensim/models/word2vec.html) was used to create the Word2Vec NLP variables. The parameters specified during model specification were:

* min\_count = 5
* vector\_size = 200
* workers = cores
* sg = 1
* compute\_loss = True

Where ‘cores’ is equal to the number of central processing units available as determined by the multiprocessing.cpu\_count() function available in the Python multiprocessing package.

Doc2Vec– the [Gensim Doc2Vec module](https://radimrehurek.com/gensim/models/doc2vec.html) was used to create 4 sets of Doc2Vec NLP variables. The parameters specified during model specification were:

* dm = 0
* vecto\_size = vecSize
* negative = 5
* hs = 0
* min\_count = 2
* sample = 0
* epochs = 20
* workers = cores

Where ‘cores’ is determined in the same manner as in the Word2Vec description above and where vecSize is set to either 20 for the two 20 variable sets and 200 for the two 200 variable sets.