An exploration into the use of Natural Language Processing in Stock Market Prediction on the JSE

Marcus Gawronsky Christpher Kleyweg Robert Brink

28 March 2018

1 Introduction

According to the Efficiency Market Hypothesis, (Fama, 1972)

Random Walk Theory (Magdon-Ismail, Nicholson & Abu-Mostafa, 1998)

2 Brief Literature Review

Not just a summmary, criticize and look for questions this research raises

(Gidófalvi, 2001)

3 Problem statement and analysis

Key Research Questions

Using the state-of-the-art techniques in Natural Language Processing, this paper aims to investigate the use of popular news sources in share price prediction on the Johannesburg Stock Exchange.

Using state-of-the-art techniques in Natural Language Processing, do publically availible news articles serve as price signals on the Johannesburg Stock Exchange?

Importance of Research

Use this chapter to present a clear outline of the problem or issue that you will address, including:
-Who has responsibility for the problem?
-What has already been done to try to solve it?
-What will happen if the problem is not solved?
To-date, little research has been used conducted on JSE We aim to extedn the literature int eh use of newly developed continous word vector repsentations ot analyze market sentiment

4 Objective and final outcomes

This research aims to estimate a predictive models that can be used in the real-world to achieve above average market returns.

5 Approach

Datasources

Data for this project will comprise of articles scraped from popular news services. This data is publically availible and contains information on the source, time of publishing, a title and the article itself.

Methodology

The projects aims to compare the use of various vector representations and embeddings for Natural Langauge Processing, these include the use of Bag-of-Words, N-gram and continuous vector representations (Mikolov et al., 2013). This research will take an event-based approach and compare the use of Random Forest, Naive Bayes and Kernel Support Machine Models.

6 Conditions and risk analysis

Resource Requirements

This project will require access to the University of Cape Town's High Performance Computing Fascilities, this is required due to the size and complexity of the models estimated.

Research Planning

Referrences

Fama, E. 1972. American Finance Association, Wiley. 27(3):551–567.

Gidófalvi, G. 2001. Using news articles to predict stock price movements. *Department of Computer Science and Engineering University of California San Diego*. (December 2004):9. DOI: 10.1111/j.1540-6261.1985.tb05004.x.

Magdon-Ismail, M., Nicholson, A. & Abu-Mostafa, Y.S. 1998. Financial markets: Very noisy information processing. *Proceedings of the IEEE*. 86(11):2184–2195. DOI: 10.1109/5.726786.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. 1–12. DOI: 10.1162/153244303322533223.