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

Robert Brink Marcus Gawronsky Christopher Kleyweg

16 March 2018

1 Introduction

According to the Efficienct Market Hypothesis markets take into account all relevent information in efficiently pricing securities (Samuelson, 1965). While many studies have explored this supposition under its strong, semi-strong and weak form, the growing volume, velocity and variety of market data has forced financiers to invest more-and-more in technology as a tool for decision making.

Investors form a mosaic of information pulling from financial reports, news articles and price data. With the growing trend towards automated trading a requirement to explore new forms of unstructured and semi-structured data in order to remain competitive has emerged. This research aims to explore the use of text data in quantitative stock price prediction.

2 Brief Literature Review

Natural Language Processing (NLP) is a complex challenge in feature extraction and model building. Many techniques represent a trade-off between computability and complexity, sacrificing elements of speech such as word order, conjugation and meaning. Even small corpuses can contain hundreds-of-thousands of unique words. Reducing this dimensionality whilst capturing the sentiment and meaning of documents is a broad and long-standing research area in the fields of computer science and statistical research. Dominant in this literature is document vector representations such as the bag-of-words approach, in which documents are stripped of features such as punctation, capitallization and word conjugates and represented as a vector counting the occurance of each word in the document.

Recent conference preceedings from the 11th International Workshop on Semantic Evaluation Cortis2017 summarise the work of 31 research teams with the task of performing fine-grained sentiment analysis on financial micro-blogs and news. In this work one can observe a blossoming of new techniques from works published only a few years earlier. The emergence of new tools such as Latent Sentiment Analysis, Convolutional Neural Networks and Doc2Vec for document vector extraction demonstrates the potential in revisiting such research in this area. These tools can be used in conjunction with new models from the field of Machine

Learning such Naive Bayes, Random Forest and Artificial Neural Networks (Le & Mikolov, n.d., Blei et al. (2003), Johnson & Zhang (2014)).

While many papers focus on the predictive power of natural language in stock market prediction questions remain on the temporal impact of this data on stock prices. Gidófalvi (2001) demonstrates a clear 20-minute lead and lag window around news articles where price response is observed - indicating an important dimension to the problem.

In a review of some 262 natural language and readility studies in the field of Accounting, Auditing and Finance, Fisher et al. (2009) identified narrative disclosures as untapped repositories of qualitative data key in stock price prediction.

Currently little research has been done on narrative disclosures and South African data in the field of NLP stock prediction. This research aims to bridge this gap using a corpus of news articles and JSE SENS (Stock Exchange News Service) data implementing modern tachniques in the analysis.

3 Approach

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

This project 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.

4 Conditions and risk analysis

This project will require access to the University of Cape Town's High Performance Computing facilities due to the size and complexity of the models estimated.

Referrences

Blei, D.M., Edu, B.B., Ng, A.Y., Edu, A.S., Jordan, M.I. & Edu, J.B. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3:993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

Fisher, I., Garnsey, M., Hughes, M., Fischer, I., Garnsey, M. & Hughes, M. 2009. Natural Language Processing in Accoutning, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management.* 16(1-2):21–31. DOI: 10.1002/isaf.

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.

Johnson, R. & Zhang, T. 2014. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. (2011). Available: http://arxiv.org/abs/1412.1058.

Le, Q. & Mikolov, T. n.d. Distributed Representations of Sentences and Documents.

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

Samuelson, P. 1965. Proof that Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review.* 6(2):41–49.