1_Intro

Nicholas Mitchell

March 30, 2016

Contents

1 Introduction

1.1 Acknowledgements

I would first like to thank my two magnificent champions, Rupert Hughes and Nikolay Robinzonov. They have both invested their own time, spending many a late evening with me discussing ideas and helping make the hardest of the decisions. I would also like to thank Michael Scherer for invaluable feedback and an eagle eye, helping bring everything together at the end. The help of all three would not have been possible without the blessing of DEVnet GmbH, to whom I express my gratitude for all the support they have provided.

From the academic side, I would like to thank Professor Yarema Okhrin of the Chair for Statistics at the University of Augsburg. I greatly appreciated his feedback, ideas, and patience with me from day one.

Lastly, I must thank my brother and two sisters - not least for their proof-reading, but for their lasting support; giving me the inspiration and courage to live abroad and fulfil a dream.
I dedicate this work to our amazing and loving parents, *Mum and Dad*.

1.2 Abstract

The emergence and ensuing explosion in popularity of social media platforms since the beginning of the twenty-first century, facilitated and catalysed by technological advancements in global connectivity, has created an abundance of data in completely new dimensions to traditionally available data. This has lead to rise in popularity of terms such as *big data*, *sentiment analysis* and *machine learning*, to name but just a few. The task of the data scientist is to extract information from masses of information in a way that allows new insights to be made.

This study utilises a combination of these developments with the aim of enhancing data sets commonly used to analyse and predict the movements of stock markets; with particularly interest given to the Dow Jones Industrial Average (DJIA). Social media data from the Twitter platform has been used to add greater breadth to our data set; the results of multi-model sentiment analysis performed on individual tweets being incorporated as additional covariates into a forecasting model.

Component-wise gradient boosting was selected as the methodology for its inherent features, such as variable selection and scalability to high-dimensional data sets. These are major advantages, considering both the large number of covariates produced by sentiment analysis and the scarcity of prior knowledge regarding their individual influence levels.

The aim of this study was to determine whether or not the addition of social media data, in the form of sentiment analysis, to traditional financial market data would enhance the predictive accuracy of a set of forecasting models. Over numerous modelling conditions, we were able to show that the predictive accuracy was indeed increased through the inclusion of social media data

Expectations/Results:

1.3 Thesis overview

There are three main parts to this study, with each phase feeding directly into the next. Each of the three phases required considerable independent effort, which contained little overlap; only the flow of raw data linked them. They were all, however, completely necessary in order to accomplish the aim of this study. The steps may be summarised as follows:

Twitter mining	→ obtaining social media data associated with certain topics
↪ **Sentiment analysis**	→ quantifying the sentiment contained of the collected data
↪ **Market forecasting**	→ using sentiment data to enhance traditional forecasting models

The structure of this report follows the path laid out above. Each of the three phases has a dedicated chapter, including: a brief background, the methodologies employed for this study, and lastly the results. An overview of the steps involved for *twitter mining* and *sentiment analysis* are additionally visualised using flowchart diagrams, which can be found in the Appendices. The flowchart describing the workflow for Twitter mining and scraping is given by Figure `flowchart-scraping` and may be found in Section `flowchart-twitter-mining`. The entire modelling process is put into perspective by Figures `flowchart-modelling1` and `flowchart-modelling2`, both found in Section `flowchart-mod`. A summary of all results and some ideas for future work are given in a final