Inhaltsverzeichnis

Nicholas Mitchell

January 18, 2016

Contents

1 Abstract (English and German)

1. How did we get here?
   (a) What does this research contribute to the field, i.e. what is new?
2. How can we use this to create an advantage?
3. What is our toolkit
4. What was our expectation?
5. What was our result?
6. Draft 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 and ever increasing global connectivity, has created an abundance of data in completely new dimensions to traditionally available data. This has lead to new terminology, such as: big data, sentiment analysis and machine learning, to name but just a few.

   This study utilises a combination of these developments with the aim of enhancing commonly used models used to analyse and predict the movements of stock markets. Social media data from the Twitter platform has been used to add greater breadth to our model; the results of multi-model sentiment analysis performed on individual tweets being incorporated as additional covariates.

   Component-wise gradient boosting was selected as the methodology for its inherent capabilities in variable selection and adaptability to high-dimensional data sets (*large n large p*). These are major advantages, considering both the large number of covariates produced by sentiment analysis and the little prior knowledge regarding their individual significance levels.

   Expectations/Results: ...

2 Introduction

1. An overview of the research area
   (a) stock market predictions with social media -> overlap of two research areas
   (b) sentiment analysis -> usage of results (not the theory/background - later!)
2. What are we setting out to show? (Make the aim crystal clear)
3. What are the limits of the thesis?
   (a) What are relevant topics/questions, but not included in this study?
4. Thesis Breakdown
   (a) The story of the thesis from start to finish. Describe the sections, how one leads naturally onto the next.

3 Sentiment Analysis

1. Short intro (summarising what has been presented so far)
2. Models to be applied
   (a) List the models (SentiStrength, Emolex, Sentinet140, Vader Afinn, Vader)
   (b) Examplanation of each:
      i. Underlying philosophy
      ii. The algorithm
      iii. Understanding the results

4 Gradient Boosting

1. Origins and development of the methodology
   (a) Original idea by Friedman and it has been adapted into the version used here
2. Relevant theoretical aspects - why we optimise certain parameters
3. Strengths and weaknesses
4. Why does it suit our data and purpose?

5 Twitter Mining

1. Overview of social media data / big-data sources
2. Twitter APIs
3. Scraping
   (a) Firehose / 3rd Party Company / Advanced Search
      i. What are they? How do they work? Pros and Cons of each?
   (b) How the advanced search was utilised
   (c) Which information did we select from Twitter (i.e. the scraped HTML code)?
4. Summary of the raw data obtained
   (a) Explain the basic table of results we carried forward into modelling

6 Data Preparation and Overview

1. How were SA results aggregated?
2. Which market data did we get?
   (a) The 'how' is not very relevant
   (b) Differences between SA data and market data (weekends, scale, etc.)
   (c) Dealing with missing data
3. Descriptive statistics and breakdowns of the final combined data sets

7 Modelling

1. Data Subsets
   (a) Which subsets were chosen

Name	Components [$DJI_{logRet} \sim$]	Reasoning	Name in R
dow_{only}	lagged log returns	Most basic example for comparison	dow_{only}
dow_{trad}	gold, oil, sp500, int.rates	traditional model factors	dow_{trad}
* dow_{macro}	all macro data	Many macro factors handled well	* glm_{macro}
* dow_{SAall}	all individual SA results	SA from certain models might perform better	* glm_{saavg}
* dow_{SAavg}	average sentiment scores	SA in general explains variance	* glm_{saall}
* dow_{best}	macro + best of SA	All data to showcase component-wise boosting	* glm_{mix}

   * = currently being implemented and modelled (i.e. there are already results)
   (b) Lag values for each subset for comparisons autoregressive models
   (c) Choose smaller subsets of the most promising data and drill further into the modelling
      i. This may include different base learning for SA data versus market data inside of gamboost()
      ii. It might be a topic to add to "Further Work"
2. Inspection/superficial comparison of market and SA data (better in)
   (a) Plot interesting relationships e.g. market returns against number of tweets
3. Boosting Models
   (a) Which models with which parameters were used?
      i. Justify these decisions with comparable works or from initial data inspection above
   (b) How was the cross-validation performed?
   (c) What are the results?
   (d) What is the interpretation of the results?

8 Discussion and Conclusion

1. Summarise all results and compare to expectations
2. Compare results to traditional models (i.e. without social media data)
3. Compare to similar studies (if there are any?) - validates our model/methodology
4. Highlight the limits of empirical work and so these results
   (a) How might we improve the outcomes in the future?

9 Still open...

1. How best to embed references to literature? I think during the first section of each of the chapters, as several large areas are covered. It might be difficult to follow if one literature section covers sentiment analysis, boosting and market hypotheses.