Report Structure

Nicholas Mitchell

January 8, 2016

Contents

This file: A rough outline to the thesis. Each heading below may be separated into its own .tex file (in folder "content")

1 Introduction

1. Overview of research area
2. Motivation for thesis An introduction: what are we setting out to better understand? Where can we shed light within the current state of prediction making in a financial context using social media data? On top of incorporating social media data, which specific/special statistical methodology are we implementing, and why?
3. Literature review (combined with the motivations?) An overview of similar projects using social media data to effect (Google work from Okhrin's chair). Work using our statistical methodology, a jsutification why it is fitting for our model (boosting allows the inclusion of many factors without necessarily watering down the model at the same time)
4. Thesis breakdown
   (a) The order of the thesis -> try to tell a story that can be read from start to finish, each step easily comprehendable.
   (b) List hypotheses?

2 Data

1. Data Overview
   (a) What data do we aim to use? Here a brief overview, but a section later on explaining how the data was collected and prepared for modelling.
   (b) What is our justification for this data?
2. Twitter Mining
   (a) Overview [Background for reader but also for future reference for DEVnet]
      i. The end goal We want data that looks like **this** which can be edited like **this** and is reliable, reproducible, relevant, ...
      ii. The challenges There are several sources of Twitter data, each with their own strengths and shortcomings. Below is a summary of each.
   (b) Twitter API for Developers
      i. What is it?

ii. How does it work?
iii. Advantages
iv. Limitations

(c) Third party companies
i. What is it?
ii. How does it work?
iii. Advantages
iv. Limitations

(d) Twitter advanced search
i. What is it?
ii. How does it work?
iii. Advantages
iv. Limitations

(e) How we have used the advanced search
i. Advantages vs. Disadvantages

3. [Optional] Scraping with Python

(a) Overview Explain general methodology, difficulties and their solutions This may be better as a larger appendix

4. [Include in appendix?] Data Preprocessing

(a) Features of the Twitter data Possibly talk about any differences between our scraped data and the data that is available from the API

(b) Our final version of Twitter data A simple example table of the final version that gets imported into R

5. Inspection of Entire Data Set

(a) Starting point for modelling
Having collected and cleaned the Twitter data, the next step is to look at it in context, alongside the common and more directly related financial market data It is important that some level of correlation is present. Here we could explain our hypotheses (mentioned in first section):

- Frequency of tweets is telling of near future market movements?
- Or rather it may give us a measure of momentum? A lot of tweets don't tell us how the direction will change, but rather hopw long it may stay on its present course.
- Can we measure a general delay between market movements and the respose on Twitter? (Maybe it does indeed run in the opposite direction?)

3 Sentiment Analysis

1. Introduction
   (a) What is sentiment analysis and why can it help us to model the markets
2. Models to be applied
   (a) SentiStrength, Emolex, Sentinet140, Vader Afinn, Vader
   (b) Short explanation of each of the five models used:
      i. the underlying philosophy
      ii. the algorithm
      iii. understanding the output

4 Boosting

1. Theoretical background
   (a) Friedman - sequential regression, using residuals to fit next learner
   (b) Parameters: number of iterations, shrinkage (learning rate), tree depth. For each, explain:
      i. importance - how can it affects/helps refine results
      ii. limitations - what happens if we get parameters wrong or, for example, had infinite time to compute things? Where are the bottle necks? Why are other models better in certain situations?
2. Strengths & Weaknesses
   (a) Number of covariates is no longer an issue (assuming we hae enough iterations)
   (b) (Multi)collinearity isn't such a worry, as the most important predictors are used - in addition we can 'prune' early on
   (c) Collinearity isn't a problem perhaps anyway in a strict sense as there is a lot of noise in our data sets - see the comments on: `http://stats.stackexchange.com/questions/30903/what-does-that-mean-that-two-time-series-a`
   (d) Can be optimised according to any given loss function (Least squares, absolute error, Huber error, ...) These can be tailor-fitted to data. If we believe the data set to be non-Gaussian, a different loss-function can be used.
   (e) The sequential learning steps can be performed stochastically to increase model performance AND computation times

(f)

3. Why does it suit the requirements of this research?

   (a) We have many predictors, meaning the dimensions of the data [e.g. 670 x 400] are not typically great time-series/predictive analysis

   (b) ... Compare to other models that are not suited to this data?

5 Description of our model

1. Define our boosting model:

   (a) How are the model parameters optimised

   (b) Cross validation

   (c) Loss function plots

   (d) AUC/ROC plots?

2. Principal Component Analysis / (one other common and robust techniques to analyse data?)

   (a) Can we show that the addition of sentiment analysis data improves the ability of the model to explain variance? Is it possible that Sentiment analysis would add another dimension to the PCA (or SCA)

6 Data subsets

Name	Components ($\log_{ret}$ ~ ...)	Reasoning
dow_{only}	lagged $\log_{ret}$	Most basic example for comparison
dow_{trad}	gold, oil, sp500, int_{rates}	traditional model factors
dow_{macro}	all macro data	Many macro factors handled well
dow_{SAavg}	average sentiment scores	Sentiment analysis explains var.
dow_{SAall}	all individual SA results	SA from certain models might perform better
dow_{best}	trad + macro + best of SA	All data to showcase component-wise boosting

7 For each model:

1. Explanation/Reasoning behind the model

2. Analysis of Results

8 Discussion of all results / comparison to literature

9 Further work

1. Other sources of social media data
2. Extensions to the model:
   (a) Use of PCA to capture the variance of model with fewer predictors
   (b) This may allow boosting to run for longer and so in sum produce better results