Review of Master Thesis

Faculty of Mathematics and Physics, Charles Universty

Author	Anastasia Serebryannikova		
Title	Predicting Stock Market Trends from News Articles		
Year	2018		
Study Programme	Informatika	**Field of Study**	Matematická lingvistika

Reviewer	Mgr. Barbora Vidová Hladká, Ph.D.	**Role**	Oponent
Affiliation	ÚFAL MFF UK		

Review:

The main topic of the thesis is an automatic prediction of changes in stock market prices using financial news articles. More specifically, the student conducted a machine learning study of binary classification of movements in the prices of stocks in the S&P 500 index using the articles published during the 2006-2013 period by the Bloomberg and Reuters agencies. The method of Support Vector Machines was used.

The thesis consists of six regular chapters accompanied by Conclusion, References, six appendices and DVD. Chapter 1 provides general information about functioning stock market and predicting stock market changes. Chapter 2 includes a review of the literature focusing on research conducted in the given area. Chapter 3 describes the data, namely the news articles and the stock market data. Chapters 4-6 form the core of the thesis. Chapter 4 documents the initial experiments and their settings. Special attention is paid to the prediction of price changes on Mondays. Chapter 5 includes a review of other related works that inspired the student to perform more experiments, namely with the initial feature set enhanced by sentiment information, various forms of n-grams, and feature extraction techniques. Chapter 6 presents a final model and its evaluation on the evaluation data set. Finally, Conclusion summarizes the main findings. The list of bibliographies consists of 44 items. Appendix A.1 is a short description of DVD, A.2 is an article published by Bloomberg, A.3 is an article published by Reuters, A.4 is a table showing performance of selected models, A.5 is an excerpt from the financial dictionary by Loughran and mcDonald, and A.6 is a table showing correlation between the sentiment scores and the target class.

Overall comment

The thesis is well and clearly structured. I really appreciate a careful and thorough literature review presented in the thesis. Many experiments were carried out. However, they did not show improvement over the baseline. I recommend to split the data into training, development test and evaluation test sets differently to understand both the task and the data better. Then I would expect an improvement.

Comments and questions

Chapter 2

2.1 – Do you have any evidence for the statement presented in the very last sentence of Secion 2.1?

Chapter 3

3.1 – I do respect that you split the data according to the previous works. However, I propose to split the data in a way so that the articles from each year occur in the three subsets. In the current setting, the articles from e.g. 2006 occur only in the training data. It would be useful to view the number of articles for each year in 2006-2013. Also, the cross-validation would certainly help.

Chapter 4

4.1 – I would like to see a clear formulation of your machine learning task(s), e.g., Predict movement of opening stock prices.

4.2 – For each experiment, present the features in a form of feature vectors and indicate their number.

4.3 – I need more explanation on how to understand Table 4.7.

Chapter 5

5.1 Have you studied the financial dictionary and SentiWordNet in details? Have you observed any overlap? For each dictionary, what is its coverage in the data?

5.2 Which lemmatizer and POS tagger did you use?

5.3 Please, see 4.2.

DVD

1. I need more explanation on how to understand the file `code/data/prices/^GSPC.csv`.
2. Which script converts the articles into the feature vectors?

Despite the shortcomings listed above, I **recommend** this thesis for the defense.

Date September 3, 2018 **Signature**