Report on **Bachelor / Master Thesis**

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

Student:	**Jan Hynek**
Advisor:	**doc. Ladislav Krištoufek, Ph.D.**
Title of the thesis:	**Stock market prediction using Twitter**

OVERALL ASSESSMENT *(provided in English, Czech, or Slovak):*

Please provide your assessment of each of the following four categories. The minimum length of the report is 300 words.

In the thesis the author downloads data from Twitter and tries to predict, using penalized multi-/bi-nomial classification, whether the market moves up or down. The author finds no interesting insights apart from the fact that the data are not usable for prediction.

While I understand that the topic might seem *sexy,* it does not have much economic background and the lack of relevant empirical literature connected with the topic should signal that it is relatively a poor choice for investigation.

Apart from my own discontent with the topic, which might arguably be a peeve, there are other significant issues in the thesis.

First of all, the presentation of the research sometimes a false statements on the nature of the analysis. While we often aspire to provide causal relationships in economics, the analysis in this thesis is certainly not a case of causal relationship. Hence statements like: "*Therefore, I can evaluate which words have the biggest influence on the stock market."* is most certainly wrong. To put it blatantly, the tweet does not cause anything, it is just a proxy of an underlying event or some latent sentiment.
Furthermore, the whole thesis ignores the time-series structure of the data and I suspect many of the features could be improved upon, were the time structure taken into account.
At some places in the thesis the author clearly skips important arguments behind the choices he makes (e.g. the choice to use AIC or BIC instead of CV is a valid choice with sound theoretical link going back to Kullback-Leibler divergence) and refers the reader to the *Bible of ML*, the Friedman, Hastie, Tibshirani book. While providing relevant references is a must in academic work, it is author's sole responsibility to make the text self-sufficient and well-explained.
The benchmark classification choice is standard, however, in the finance literature there are other possible choices of benchmark than only using one class. The author could possibly use some kind of autoregressive benchmark.
Lastly, I would like to point to an unfortunate choice of the stock Apple. Downloading the tweets that contain words "apple" will result into an unfortunate amount of results containing something like "I ate an apple, I liked it." that has nothing in common with the company whatsoever, yet would produce a positive sentiment.

To conclude, I find the thesis to be too ambitious in both analytical and programming skills and too little ambitious in the economic domain.

Contribution
The main contribution of the thesis is the author's gained experience with the non-standard methods of the machine learning. The outcomes of the work are negative in terms of the hypothesis.

Methods

Report on **Bachelor / Master Thesis**

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

Student:	**Jan Hynek**
Advisor:	**doc. Ladislav Krištoufek, Ph.D.**
Title of the thesis:	**Stock market prediction using Twitter**

The used methods are samewhat standard in the machine learning literature. Though they ignore the time-series nature of the data and the author sometimes compromises (could not compute CV, hence I used something else).

Literature
Given the fact that there is not much serious literature on predicing the financial markets the survey provides the relevant information. Though, the author could have provided a survey from behavioral finance that could at least partly justify the sentiment analysis and prediction based on that.

Manuscript form
The author's command of English is rather lacking. It is hard to discern whether the mistakes are typos or rather incomplete knowledge of English grammar.
Furthemore, the results in the Appendix have a form that would be excusable in a homework, not in a bachelor thesis.

Hence, as I probably conveyed beforehand, my overall assessment is on the lower bound of acceptability of the thesis.

SUMMARY OF POINTS AWARDED *(for details, see below):*

CATEGORY		**POINTS**
Contribution	*(max. 30 points)*	1
Methods	*(max. 30 points)*	20
Literature	*(max. 20 points)*	10
Manuscript Form	*(max. 20 points)*	10
TOTAL POINTS	*(max. 100 points)*	**41**
GRADE	**(1 – 2 – 3 – 4)**	**3**

NAME OF THE REFEREE: *Mgr. Tomáš Křehlík, M.A.*

DATE OF EVALUATION: *June 2, 2017*

Referee Signature

EXPLANATION OF CATEGORIES AND SCALE:

LITERATURE REVIEW: *The thesis demonstrates author's full understanding and command of recent literature. The author quotes relevant literature in a proper way.*

Strong	*Average*	*Weak*
20	*10*	*0*

METHODS: *The tools used are relevant to the research question being investigated, and adequate to the author's level of studies. The thesis topic is comprehensively analyzed.*

Strong	*Average*	*Weak*
30	*15*	*0*

CONTRIBUTION: *The author presents original ideas on the topic demonstrating critical thinking and ability to draw conclusions based on the knowledge of relevant theory and empirics. There is a distinct value added of the thesis.*

Strong	*Average*	*Weak*
30	*15*	*0*

MANUSCRIPT FORM: *The thesis is well structured. The student uses appropriate language and style, including academic format for graphs and tables. The text effectively refers to graphs and tables and disposes with a complete bibliography.*

Strong	*Average*	*Weak*
20	*10*	*0*

Overall grading:

TOTAL POINTS	**GRADE**		
81 – 100	**1**	= excellent	= výborně
61 – 80	**2**	= good	= velmi dobře
41 – 60	**3**	= satisfactory	= dobře
0 – 40	**4**	= fail	= nedoporučuji k obhajobě