# Report on **Bachelor / Master Thesis**

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

| | |
|---|---|
| **Student:** | Bc. Vojtěch Čermák |
| **Advisor:** | doc. PhDr. Jozef Barunik, Ph.D. |
| **Title of the thesis:** | Analysis of stock market sentiment with social media |

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

*Please provide your assessment of each of the following four categories, summary and suggested questions for the discussion. The minimum length of the report is 300 words.*

Before providing my assessment of the thesis, I need to say the following: The thesis by Vojtěch Čermák is a study that uses techniques from the field of computational linguistics. Computational linguistics is not within the area of my expertise. Since I am not an expert in this field, I am not familiar with the existing literature and methods used in the field. Therefore, I cannot assess what the contribution of the thesis is and I cannot evaluate the correctness of used methodology either. Thus, I will just comment on a few points that came across to me when reading the thesis.

**Contribution**

I am persuaded that the strenght of the thesis lies in the overlap between computational linguistics (CL) and economics. The author uses techniques from the CL to shed more light on economic phenomena. Specifically, Vojtěch Čermák studies whether sentiment contained in and extracted from Twitter messages can serve as a predictor of the behaviour of financial market.

I find this line of research extremely important. Sentiment, a mental feeling and human emotions that are independent of reason influence behaviour of each individual. It is not possible to do business or run political campaign successfully without taking into account human emotions.

The author claims (page 60) that he has made both methodological and empirical contributions to the existing literature. I feel it is important for the author to explain his value added in a more clear way than he did in his thesis – this should be an important task for the defense. The author describes his contribution on page 2 (copied below), but it is not clear enough.

*In this thesis, we would like to explore the possibility of using methods of text representation for direct sentiment modelling. Instead of mapping text to sentiment vector and with it subsequently modelling some economic phenomena, we want to find an efficient representation of text usable for direct modelling of some desired variables. High dimensional text representations allow us to encode lot more information and use such information to model an economic variables directly. That allows us to skip one layer of estimation, presumably reducing total estimation error. However, this approach comes with challenges related to the large dimensionality of the data. Standard frequency based vector space models, commonly used in information retrieval to model text, have usually tens of thousands dimensions. Consequently, the data are very noisy, prone to overfitting and requires proceeding with caution. Also, we are limited to statistical and machine learning methods that are prone to over-specification. There have been recent advances in natural language processing, which allows us to overcome many of those problems. For example, using word embeddings as a method of text representations will enable us to reduce the dimensionality of text data substantially. We aim to connect state of*

# Report on Bachelor / Master Thesis

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

| | |
|---|---|
| **Student:** | Bc. Vojtěch Čermák |
| **Advisor:** | doc. PhDr. Jozef Barunik, Ph.D. |
| **Title of the thesis:** | Analysis of stock market sentiment with social media |

*the art information retrieval and computational linguistics with the goals of behavioural finance and show the effect of Twitter sentiment in the light of recent advances in natural language processing.*

I appreciate that author's empirical findings are not limited to the Twitter feed and can generally be applied to arbitrary text documents. The thesis can be used as a base for future researchers doing text regression.

**Methods**

The author uses advanced methods of text processing and text vectorization. As explained before, I am not in a position to comment on the correctness of used methodology.

**Literature**

Although the literature review in chapter 2 provides a well-written review of publications examining the relationships between sentiment and markets, I am lacking a comparison of author's work with the existing literature. The purpose of the literature review part in the thesis should be an explanation what has already been accomplished in the relevant line of research, what the state of the art is, how one's research fits into the existing literature and how it differs from the work done by others. In this regard, the chapter 2 does not do the job.

**Manuscript form**

Manuscript form is very good. Despite of it, it would be beneficial if the author used a spell checker to eliminate typos (e.g. p. 2 – asses instead of assess, etc.).

**Summary and suggested questions for the discussion during the defense**

- Could the author compare his work in a more transparent way to the existing literature?

- Could the author explain again the cause of his spurious results and to what extent it is a problem for applications in real-world scenarios?

I consider the thesis to be of a very high quality. I appreciate the sophisticated techniques used, ability of the author to conduct such a difficult analysis and the selection of the topic that I find extremely important and . Vojtěch Čermák should try to convert his thesis into a journal article and try to publish it.

# Report on **Bachelor / Master Thesis**

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

| | |
|---|---|
| **Student:** | Bc. Vojtěch Čermák |
| **Advisor:** | doc. PhDr. Jozef Barunik, Ph.D. |
| **Title of the thesis:** | Analysis of stock market sentiment with social media |

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

| CATEGORY | | POINTS |
|---|---|---|
| *Contribution* | *(max. 30 points)* | 28 |
| *Methods* | *(max. 30 points)* | 28 |
| *Literature* | *(max. 20 points)* | 17 |
| *Manuscript Form* | *(max. 20 points)* | 19 |
| **TOTAL POINTS** | *(max. 100 points)* | **92** |
| **GRADE** | **(A – B – C – D – E – F)** | **A** |

***NAME OF THE REFEREE: PhDr. Pavel Vacek, Ph.D.***

***DATE OF EVALUATION: 10 June 2018***

Vacek P.

***Referee Signature***