

# Report on Bachelor Thesis

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

<b>Student:</b>	<b>Jan Hynek</b>
<b>Advisor:</b>	<b>doc. PhDr. Ladislav Křištofek, Ph.D.</b>
<b>Title of the thesis:</b>	<b>Stock market prediction using Twitter</b>

## **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.

### **Contribution**

The main motivation of the thesis is to find influential words used in Tweets for a given company hashtag and possible use this word's occurrence as an indicator of upcoming significant positive or negative change in price. To do so, the author applied a combination of econometrics and data science, specifically text mining and semantical analysis combined with logistic regressions. However, the outcomes are not as clear as probably hoped for. The contribution from this perspective is thus rather weak. However, the author provides quite a detailed discussion about possible issues. An important one goes with the dataset collection which yielded maybe unreliable time series. My personal view is maybe a bit different – without a selection bias, many studies using Twitter would yield insignificant results due to huge amounts of noise (i.e. useless, meaningless Tweets) in the data. It is thus not completely surprising that the results are not strong or “miraculous”.

### **Methods**

The utilized methods combine the data mining techniques (specifically text mining and sentiment analysis) and traditional econometrics methods (logit and multinomial logistic regression). Therefore, the used methodology goes beyond the standard bachelor's curriculum. Apart from mining and analyzing the tweets, these needed to be collected. This has turned out to be quite complicated thing as Twitted restricts these downloads and the collection is frequently disturbed by (maybe intentional) outages of the system. This has affected the dataset size.

The methodological approach towards analyzing the dataset is quite novel and unique so that the author had to describe the steps carefully. I think this has been done nicely

### **Literature**

The topical literature is a rather rapidly expanding collection and it is not easy to collect a reasonable set of works with narrowly defined focus. Nevertheless, I believe the author has covered useful references. However, there are some other important “classical” references on the topic of utilizing online data that could have been mentioned, such as the works of Da et al. (2011, Journal of Finance) and Choi & Varian (2012, Economic Record) (even though they do not focus on Twitter, they are still essential).

### **Manuscript form**

From the formal perspective, the thesis is nicely written, well-structured and there are only minor issues such as rather short abstract and the Conclusion section not being on a separate page.

# Report on Bachelor Thesis

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

<b>Student:</b>	<b>Jan Hynek</b>
<b>Advisor:</b>	<b>doc. PhDr. Ladislav Krištofuk, Ph.D.</b>
<b>Title of the thesis:</b>	<b>Stock market prediction using Twitter</b>

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

<b>CATEGORY</b>	<b>POINTS</b>
<i>Contribution</i> (max. 30 points)	18
<i>Methods</i> (max. 30 points)	28
<i>Literature</i> (max. 20 points)	16
<i>Manuscript Form</i> (max. 20 points)	19
<b>TOTAL POINTS</b> (max. 100 points)	<b>81</b>
<b>GRADE</b> (1 – 2 – 3 – 4)	<b>1</b>

**NAME OF THE REFEREE:** doc. PhDr. Ladislav Krištofuk, Ph.D.

**DATE OF EVALUATION:** 5.6.2017

Digitally signed by Ladislav Krištofuk  
DN: cn=Ladislav Krištofuk, o=Faculty of  
Social Sciences, Charles University in  
Prague, ou=Institute of Economic  
Studies,  
email=ladislav.kristoufek@fsv.cuni.cz,  
c=CZ  
Date: 2017.06.05 20:10:18 +02'00'

---

**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	<b>1</b>	= excellent	= výborně
61 – 80	<b>2</b>	= good	= velmi dobře
41 – 60	<b>3</b>	= satisfactory	= dobře
0 – 40	<b>4</b>	= fail	= nedoporučuji k obhajobě