Report on Master Thesis

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

Student:	**Bc. Ksenia Pogodina**
Advisor:	**PhDr. Boril Šopov, MSc., LL.M.**
Title of the thesis:	**News Feed Classifications to Improve Volatility Predictions**

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.

Contribution

The thesis concerns analysis of stock volatility and improvement of stock volatility models via inclusion of variables of news sentiment. Thus, the author analyzes the impact of news articles on stock return volatility. The author uses three types of methods, Naive Bayes classifier as a machine learning method, lexicon-based sentiment derivation technique as a linguistic method, and hybrid Naive Bayes classifier as a combination of both. Further, two types of lexicons are considered – binary and multiclass. In particular, the author considers three stocks – Apple Inc., Microsoft Corp., Amazon.com – and investigates performance of pure Naive Bayes classifier, hybrid Naive Bayes classifier with binary lexicon and hybrid Naive Bayes classifier with multiclass lexicon. The major contribution of the author is the supervised learning which requires manual labeling of the training dataset. Further, the author examines three types of GARCH models including the news sentiment variable and for each stock compares their performance with respect to „plain" GARCH, both with respect to in-sample period and out-of-sample period.

Methods

The author works with conditional heteroscedastic models of volatility, in particular, for each of the three considered stocks, she identifies GARCH(1,1) model for the log-differences. For obtaining numerical results, the author works with STATA and Gretl. She also performs analysis of several text-based sentiment derivation techniques using programming language Python.

Literature

The literature cited is adequate and chapter 2 provides satisfying overview of the topic. However, I feel as if some topics considered in chapter 2 are not reflected further in the thesis and, namely, the achieved numerical results are not confronted with previously achieved results by other authors supported by precise citations. The list of references itself suffers few shortcomings, e.g. reference items of papers in proceedings seem incomplete (Fernandez et al 2014, Wang and Ho 2016, or Yang and Pedersen 1997). Further, Francq and Zakonian 2010, which is cited on page 28, is not included in the list of references.

Manuscript form

Overall, the manuscript is written with logical structure and presentation of results is executed with adequate detail. However, the text suffers with more than just occasional typos and grammar mistakes. I do not attempt to provide the list of typos here as it would be quite long and hardly complete. I would be happier for more thorough presentation of the theory of GARCH models (e.g. the last paragraph in section 4.2 seems corrupted with mistakes and notational inconsistencies). In (4.11) there seems to be missing the square of the difference. I was mostly dissappointed by lack of graphs and tables for analysis of Microsoft and Amazon timeseries which made it impossible to verify the authors conclusions. Even when the corresponding table or figure is included, the derived conclusions by the author are rather difficult to see, e.g. interpretation of explanatory power using AIC and BIC for

Report on Master Thesis

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

Student:	**Bc. Ksenia Pogodina**
Advisor:	**PhDr. Boril Šopov, MSc., LL.M.**
Title of the thesis:	**News Feed Classifications to Improve Volatility Predictions**

the three considered types of distribution from table 6.1 or difference in magnitude or the direction for lexicons within one stock from Figure 5.7. While the graphical presentation of results up to Section 5 included is pleasing, improvements could be made in Section 6 and the Appendix, which seem rather technical and much less reader-friendly.

Summary and suggested questions for the discussion during the defense

Due to reasons above, subject to successful performance during the thesis defence, I suggest grade B. Below are few suggested questions the author of the thesis could address during the defence.

1) On page 16 below formula (3.8) you mention the so-called underflow and acceletation as something one should avoid during Naive Bayes computations and working in log-space is a suggested remedy. Could you please illuminate what causes underflow and acceleration and why one should attempt to avoid these phenomena?
2) On page 33 you present p-value of KPSS test of 0.187. However, Figure A.5 states the test statistic of KPSS test of this value. From the critical values of KPSS it is clear that the corresponding p-values is larger than 0.1. Is it true that the p-value and value of the test statistic coincide?
3) What information can be derived from comparison of values of AIC or BIC for particular „augmented“ GARCH models of selected stock with respect to different distributions assumed?

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

CATEGORY		**POINTS**
Contribution	*(max. 30 points)*	30
Methods	*(max. 30 points)*	29
Literature	*(max. 20 points)*	15
Manuscript Form	*(max. 20 points)*	10
TOTAL POINTS	*(max. 100 points)*	**84**
GRADE	**(A – B – C – D – E – F)**	**B**

NAME OF THE REFEREE: Michal Červinka

DATE OF EVALUATION: January 20, 2018

Referee Signature

EXPLANATION OF CATEGORIES AND SCALE:

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*

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*

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*

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	**GRADE**
91 – 100	**A**
81 - 90	**B**
71 - 80	**C**
61 – 70	**D**
51 – 60	**E**
0 – 50	**F**