



Thesis Evaluation Report

Student:	Tomáš Juchelka
Advisor:	Ladislav Křišťoufek
Title:	Prediction of Stock Return Volatility Using Internet Data

Assessment

Overall

I believe the thesis examines an interesting research question of stock price volatility prediction using modern data on Twitter feeds, which contributes to its novelty. At the same time, I am quite disappointed with its overall quality. I believe the author does not motivate his research question well and he does not explain his original contribution relative to the existing studies. The thesis makes an impression of being a pure data mining exercise that lacks a compelling argument about the envisaged underlying mechanism that brings about the documented phenomena. It is not clear to me how the author ensures that the documented correlations are not driven by confounding events (see below for details). It is not clear to me whether horse racing the alternative empirical approaches is reasonable because the empirical approaches only estimate the dependent variable rather than the predictor variable (see below for details). I find the literature review rather superficial and sometimes ill-focused on the (easy to discuss) technical aspects rather than on the (relevant) prior research on the impact on internet based activity on stock prices (see below for details). The writing style is somewhat negligent. Hence, I believe the thesis significantly lags behind its potential and in the current form the level of its sophistication tends to be appropriate for a good Bachelor's rather than a Master's thesis.

Motivation

I believe the author does not provide a good motivation for his research question. The author should provide a *prima facie* explanation of why we should care about the precision in predicting stock market volatility? In fact, I think the argument is fairly easy to make as the stock price volatility has direct implications (besides others) for option pricing. However, the author simply states "In general, stock return volatility is a very important variable that enters many asset pricing and risk assessment models that are broadly used by either general public or professionals of financial markets." (p. 2) without providing any specific examples or indeed references (!). I consider such a statement very vague and quite insufficient to motivate the author's work.

Contribution

I believe the author does not properly explain how his work contributes to the existing research on stock price volatility prediction. The author acknowledges that much research has already been done in the field "Although stock return volatility is quite extensively researched topic, the interaction of users on the Internet and social media generates data of unprecedented scale and content

of information." (p. 68). However, he does not explain why we should a priori expect that using Twitter data is likely to yield better predictions than e.g. Google company name searches, Yahoo Finance ticker searches, etc. He simply states "We assume that such employment of these [Twitter] data enhances the quality of results obtained from currently available analytics, either in finance or in other sciences." (p. 4). I think such a sentence is an expression of an author's faith or preference rather than an objective argument for why considering Twitter data is likely to provide some new insights.

Causality

The author provides very limited (if any) discussion of the direction of causality or of the underlying mechanism he envisages. The title uses the word "prediction", which seems to imply that the author believes that the Twitter posts proxy for early investor behavior that later on "leads to"/"implies"/"causes". I could easily imagine that the volume of Twitter posts and stock price volatility are co-determined by other factors. I find no results inconsistent with this alternative view.

Confounding Events

If I understand the author's method discussion correctly, he does not control for important information-rich events, such as earnings announcements, investor conference calls, new product launch announcements, etc. These can plausibly drive the results. If the author does not control for these events the results can be considered almost trivial. It is hardly surprising that these events attract significant investor attention (which is well established in the past literature) and at the same time they increase stock price volatility as there is some uncertainty about the impact these events have on company value. Documenting the correlation gives little additional insight and it can hardly be labeled as a study of the predictive power Twitter posts have as the underlying volatility hike is due to the confounding events rather than due to the Twitter posts. If so, the finding could be reformulates as: "Earnings announcements are associated with increased stock price volatility as well as with heightened investor attention"...

Empirical Models

The author presents the horse race of the alternative empirical approaches discussed in section 2.3 and 2.4 as one of his contributions: "The second task was to discover which model is the best among the models we compared." (p. 69). On the fundamental level one can ask whether such a comparison makes sense given that there is no theoretical prediction as to what model(s) should be superior to others. Without such a prediction the results are hardly generalizable. More importantly though if I understand the author's intentions correctly he uses these models as "proxies for realized volatility" (p. 27) (see section 3.3). In other words, these are alternative empirical approaches of estimating the dependent variable (Y) rather than alternative ways of measuring the volatility predictor (i.e. the independent variable X). If so, reporting results from such a horse race makes little sense because these models are likely to measure realized volatility with various level of noise and so whether they are more or less correlated with the volume of Twitter searches says very little (if anything, indeed) about whether Twitter data can be used to predict volatility. Such analysis would address a (completely) different research question related to the alternative methods of estimating stock price volatility.

Literature

I find the literature review rather superficial and ill-focused. That implies that the development of the position argued in the thesis is rather weak. I suspect the author has obtained only limited understanding of the relevant research. I would expect him to review (much) more thoroughly literature (i) on the stock price volatility prediction, and (ii) on the modern investor attention proxies, e.g. Google company name searches, Yahoo Finance ticker searches. The author essentially aims to extend the literature on "bounded rationality" without making reference to the

concept. In contrast, I am not convinced about the usefulness of the textbook-like discussion of the empirical models in section 2.3 and 2.4. I find it rather remarkable that while the introduction contains 7 (!) references (some of which are perhaps not quite relevant) the author cites his own Bachelor's thesis Juchelka (2014) perhaps suggesting that it is one of the key studies in the area... Notwithstanding that the way of referencing defies standard conventions and it is sometimes inconsistent, e.g. "In particular, theories from the perspective of Socioeconomic Theory of Finance written by Robert R. Prechter & Parker (2007) or Behavioral Economics and Behavioral Finance suggests that stock market prices can be, at least to some degree, successfully predicted." (p. 2).

Results

In my opinion, the very formulation of the main result in the conclusion indicates that the author understand that the results are not particularly insightful: "Let us start with a description of the Twitter effects on future volatility. In general, we can say that almost all the variables have positive effects on tomorrow's volatility. The interpretation is straightforward as well, the higher the number of tweets regardless their polarity, the higher the volatility on the other trading day." (p. 69). I would argue that if something is "straightforward" there is little need to analyze it in a thesis...

Style

I find the writing style rather imprecise/vague and sometimes negligent, e.g. "Stock market features, as well as its prediction, have been widely discussed topic either by academic town and gown or business professionals, working in financial services industry." (p. 1). In addition, sometimes the same idea is needlessly repeated several times, e.g. "It also directly implies that it is impossible to overperform the market consistently over a long period of time. ... Such fact would imply, that it is impossible to achieve the abnormal return on a long term basis." (p. 2).

Awarded Points and Grade

Literature (max 20)	14
Methods (max 30)	24
Contribution (max 30)	22
Form (max 20)	14
Total (max 100)	74
Grade (1 – 2 – 3 – 4)	2

Referee's Signature

August 10, 2017

Jiří Novák

Evaluation Date

Referee's Name

Grading 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	= nedoporuji k obhajobě