Brief Literature Overview

The use of company information in text form is not a new idea. Many researchers in fields of finance and accounting have tried to extract textual information from all sorts of publications in order to explain or forecast financial phenomena. In the latter class, a lot of effort has been executed in order to predict either returns or return volatility (mostly in equity or FX markets). A broad range of possible research "combinations" arise, depending on what one seeks to forecast, which document serves as predictive feature, and which methods are applied (for both text classification and forecasting). A short list of research papers that will be the most relevant guidance for my thesis are the following:

Huang et al. (2014) use a naive Bayesian algorithm to classify sentences from more than 360,000 analyst reports into three categories (positive, negative, neutral). From the sentence "opinions" an overall report opinion is generated and used to predict future abnormal returns as well as future earnings. This paper will be relevant in both applied methodology as well as using the equal textual underlying (analyst reports).
Luss and d'Aspremont (2009) use support vector machines (SVM) to show how text content from press releases (2000 - 2007) can be used to predict both a binary up- or down movement in the intraday price of equities as well as the magnitude of an abnormal return. Moreover, they apply multiple kernel learning (MKL) to combine quantitative data (i.e. historical returns) with qualitative data (i.e. text from press release).
Following a similar approach with different underlying text source, Rekabsaz et al. (2017) recently used the sentiment of 10-K reports to predict future volatility. By the same token, they apply a Gaussian kernel in a SVM as well as MKL to optimally combine text forecasts with "market" features (such as GARCH predictions).
Kalev et al. (2004) used the number, type and timing of press releases ("news") as an exogenous variable in a GARCH(1,1) framework to forecast the intra-day volatility of the Australian stock market index as well as its five most liquid stocks.

Thesis Title, Goal and Research Question

Based on the existing literature, my research contribution will be to combine the reference papers of Huang et al. (2014) and Rekabsaz et al. (2017) and use the sentiment of analyst reports to forecast ahead realized volatility of a market index return (and/or firm's stock return - depending on which methodology proves more suitable, cf. explanations below).

Following the idea of Rekabsaz et al. (2017), the text's incremental explanatory power shall then be combined with a "market-based" volatility forecasting model that bases on past realized volatilities (e.g. GARCH and several extensions thereof). Moreover, based on the origin of the

company covered by the analyst report, an additional goal is to uncover potential country differences, i.e. to detect whether market X (or individual stocks in market X) react(s) "heavier" to text content of analyst reports than market Y (or individual stocks in market Y).

My proposed title of the Master-Thesis is therefore:
Using Text Sentiment of Stock Analyst Reports to Forecast Volatility: A Cross-Country Empirical Study

Research Methodology

I initially intend to split the thesis workload into 7 sections:

1. I will download the analyst reports and clean them from HTML code, stop words, tables as well as numbers. Afterwards I will stem the words and provide descriptive statistics about the language used in analyst reports, giving rise for in-depth analysis in subsequent chapters. Moreover, I will download daily levels of several national stock market indices in order to obtain their daily (log or simple) returns, which will later be used for the volatility forecasting task.

2. Following Huang et al. (2014), I aim to split each analyst report into sentences and manually classify a pre-defined number X of sentences into three categories based on their "influence" on future volatility (i.e. either *increasing*, *decreasing*, or *not affecting* future volatility). Using this manual categorization as a training set for a Naive Bayesian classifier and/or a support vector machine (SVM), I seek to label each sentence in each analyst report in one of the three categories via one of these two machine learning algorithms.

3. Based on the frequency/ratio of volatility-increasing and -decreasing sentences in the document, I will construct a "sentiment" index for each analyst report (similar to Huang et al. (2014) who measured the opinion of a report subtracting the percentage of negative sentences in the report from the percentage of positive sentences, i.e. $OPN = PCT_{POS} - PCT_{NEG}$)

4. Using the sentiment indices that belong to a analyst report that covers a company listed in a specific country (or geographical region) i, I will calculate the average opinion score ($AOS_{i,t}$) for all reports issued in a specific week t that cover companies from country i. Moreover, I will calculate the sample standard deviation across the different opinion scores, in order to obtain a measure of "disagreement" between analysts, $DISAG_{i,t}$.

   Consider the short illustrative example: From March 6th 2017 to March 10th 2017 (i.e. calendar week $t = 10$), three analyst reports for three (distinct) German companies were released. The opinion scores of these three analyst reports are, say, 0.10, 0.16, and

0.25, respectively. Then the simple average opinion score $AOS_{Germany,10}$ is equal to 0.17. $DISAG_{Germany,10}$ will be[1]

$$\left(\frac{1}{3-1}\cdot\left[(0.10-0.17)^2+(0.16-0.17)^2+(0.25-0.17)^2\right]\right)^{0.5}=0.0755.$$

5. I will use the average sentiment index across analyst reports as well as their level of disagreement to forecast the volatility of the market index of country/geographical region i one week ahead, i.e. at time $(t+1)$. In other words: I intend to use $AOS_{i,t}$ and $DISAG_{i,t}$ to predict $\sigma_{i,t+1}$, which is the realized volatility of market index of country i five business days (i.e. one week) after the publication of the analyst report. I will follow Rekabsaz et al. (2017) in the definition of realized volatility and use

$$\sigma_{i,t+1}=\sigma_{i,\tau+5}=ln\left(\sqrt{\frac{\sum_{\tau=1}^{5}(r_{i,\tau}-\bar{r_i})^2}{4}}\right),$$

where t indexes weeks and τ indexes trading days while $\bar{r_i}$ describes the average return in the past business week.

6. The forecasting task shall be conducted via either a GARCH(p,q)-model using $AOS_{i,t}$ and $DISAG_{i,t}$ as additional exogeneous regressors or a support vector regression, choosing the more accurate out-of-sample estimator between these two for subsequent analyses. In both specifications, however, other control variables established by past research (e.g. lagged trading volume) shall be included. The research hypothesis is that "analyst optimism" (measured by $AOS_{i,t}$) will have a negative and significant impact on the level of market volatility, while "disagreement across analysts" (measured by $DISAG_{i,t}$) is predicted to have a positive and significant impact on market volatility.

7. Depending on the forecasting accuracy of the volatility for several national stock market indices, I seek to uncover potential differences how "heavy" the market reacts to analyst reports.

However, based on the following potential flaws I expect to incur if using the methodology described above, I propose an alternative research design to meet my forecasting goal (and plan to use the method that performs better). Regardless of the fact that using the textual differences to explain disagreement between analysts seems a promising explanatory variable for

[1]Note that for this example, the number of analyst reports is only $N=3$. Hence, the bias in the sample standard deviation is still significant, even though one uses Bessel's correction for the sample variance, i.e. dividing by $(N-1)$ instead of N. For weeks with a large number of analyst reports issued, i.e. larger N, the bias will reduce.

market volatility, it might be too "approximative". The potential drawbacks of the seven steps mentioned above could be:

- As the number of analyst reports issued in a specific week will differ over time and across countries, the variables $AOS_{i,t}$ and $DISAG_{i,t}$ will not be measured consistently (and sometimes may not be available at all, if no analyst report was issued in this week).
- Even if the number of tracked companies is approximately constant over time and geographical region, the companies analysed in the specific reports will differ from week to week (e.g. in week 10 it will be Apple, Microsoft, and Amazon, while in week 11 it will be Mattel, Harley-Davidson, and Foot Locker). In other words, the composition of $AOS_{i,t}$ and $DISAG_{i,t}$ will differ in both quantitative (number of firms) and qualitative aspects (type of firms). Ignoring the latter heterogeneity is close to the assumption that analyst text about Apple, Microsoft, and Amazon will have the same impact on the S&P 500 volatility as news about Mattel, Harley-Davidson, and Foot Locker. The impact on market volatility will heavily depend on the weights of the analysed companies within the market index (while the first triplet of firms in my simplified example clearly would have much larger influence because of their significantly higher weight in the S&P 500 index).

Considering these issues arising from the "aggregation" of firm-related reports to to market level volatility, I consider taking the opinion score of each firm and include it in a GARCH/SVR model to forecast it's **individual** stock return volatility rather than the overall market volatility. The cross-country comparison shall later be done by observing the forecasting accuracy for each firm and research if geographical clusters are observable. If so, one might consider training the text classification algorithm (Naive Bayes or SVM) on a country-specific training set and test whether the prediction accuracy increases (in a similar fashion as Rekabsaz et al. (2017) have done for different sectors instead of geographical regions/countries).

Note: In order to estimate the incremental effect of text content, I plan to "isolate" the effect of quantitative measures contained in an analyst report in the same fashion as Huang et al. (2014), i.e. constructing variables that indicate whether a change in the earnings forecast, target price or stock recommendation has occurred.

Data Requirements and Sources

I seek to take use of the Investext database via Thomson Eikon (accessible on HSG campus, room 01-U206) for analyst reports, and obtain stock prices / index levels from Thomson Datastream. Moreover, to find earnings forecasts, target prices and stock recommendations I would need to access the Thomson Reuters I/B/E/S database.

References / List of potentially relevant papers

Bholat, D., Hansen, S., Santos, P. & Schonhardt-Bailey, C. (2015). *Text mining for central banks: handbook.* Centre for Central Banking Studies (33), 1-19.

Cao, L. & Tay, F. (2001). *Financial Forecasting Using Support Vector Machines.* Neural Computing & Applications, 10(2), 184–192.

Heidari, M. & Felden, C. (2015). *Impact of Text Mining Application on Financial Footnotes Analysis.*

Huang, A. H., Zang, A. Y. & Zhen, R. (2014). *Evidence on the Information Content of Text in Analyst Reports.* Accounting Review 89(6), 2151-2180.

Gavrishchaka, V.V. & Banerjee, S. (2006). *Support vector machine as an efficient framework for stock market volatility forecasting.* Springer.

Joachims, T. (1998). *Text Categorization with Suport Vector Machines: Learning with Many Relevant Features.* ECML '98 Proceedings of the 10th European Conference on Machine Learning, 137-142.

Kalev, P. S., Liu W., Pham, P.K., and Jarnecic, E. (2004). *Public information arrival and volatility of intraday stock returns.* Journal of Banking & Finance 28, 1441–1467.

Kloptchenko, A., Magnusson, C., Back, B., Visa, A. & Vanharanta, H. (2004). *Mining Textual Contents of Financial Reports.* International Journal of Digital Accounting Research, 4(7), 1-29.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S. & Smith, N. A. (2009). *Predicting Risk from Financial Reports with Regression.* Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09), 272-280.

Li, F. (2009). *The Information Content of Forward-looking Statements in Corporate Filings — a Naive Bayesian Machine Learning Approach.*

Loughran, T. & McDonald, B. (2016). *Textual Analysis in Accounting and Finance: A Survey.* Journal of Accounting Research 54 (4), 1187–1230.

Luss, R. & d'Aspremont, A. (2009). *Predicting Abnormal Returns From News Using Text Classification.*

Robertson, C. S., Geva, S. & Wolff, R. C. (2007). *News aware volatility forecasting: is the content of news important?.* In Christen et al. (2007) Data mining and analytics: proceedings of the sixth Australasian Data Mining Conference (AusDM2007), 157-166.

Zadeh, R. B. & Zollmann, A. (?). *Predicting Market-Volatility from Federal Reserve Board Meeting Minutes.* NLP for Finance.