Research Report

The Verbal Side of Financial Data Analysis – A Study on Machine Learning Capabilities

AS A MATTER OF COURSE, QUANTITATIVE DATA SUCH AS TIME SERIES AND QUARTERLY FIGURES ARE FREQUENTLY USED IN FINANCIAL ANALYSIS. SUCH DATA CAN BE PROCESSED AUTOMATICALLY AND INTERPRETED RATHER EFFICIENTLY. HOWEVER, A SIGNIFICANT PERCENTAGE OF RELEVANT INFORMATION ORIGINATES FROM UNSTRUCTURED SOURCES, PRIMARILY TEXTUAL DATA, WHICH REQUIRE MANUAL (HUMAN) INTERPRETATION. WE EXPLORE EMPIRICALLY HOW MACHINE LEARNING TECHNIQUES CAN PROVIDE SUPPORT FOR ANALYZING AND INTERPRETING SUCH TEXTUAL DATA SOURCES.

Jan Muntermann Sven S. Groth

Introduction

Information management is one of the most challenging tasks for financial institutions. In the last two decades, much progress has been made in the development of quantitative models and approaches. While much effort has been put on the extensive analysis of quantitative data such as historical price series, little intention has been paid to the (automated) analysis of textual data in the past, which undoubtedly represents a large source of information in this context.

Prior empirical research has shown that certain news stories such as corporate disclosures can cause abnormal market behavior subsequent to their publication, which provides further evidence that textual data represents a highly relevant source of new information. On the basis of a dataset that comprises corporate news stories and intraday stock prices, this article explores how such textual data can be analyzed with the help of machine learning techniques.

What is Machine Learning?

Machine learning techniques comprise a family of methods that attempt to allow machines to acquire knowledge for problem solving by showing them historical cases. In a financial context, such historical cases can, for instance, be a sample of news publications that have been mapped to stock price reactions observed on the capital markets. Popular examples of machine learning techniques represent Decision Trees and Artificial Neural Networks, which have already been successfully applied since the 1960s and the 1980s, respectively. The following Figure 1 illustrates how different generations of machine learning techniques have emerged in the last decades. Each of these techniques has specific characteristics, capabilities and shortcomings. These, for example, entail different learning strategies, dataset requirements and computing times. Since textual data is highly unstructured, the automated analysis represents a major challenge. Here, newer machine learning techniques such as Support Vector Machines (SVM) have revealed to be especially promising. In general, these methods aim at forecasting if a certain example (e.g. a newly published news story) belongs to one of two categories such as "relevant" and "not relevant". On the basis of given training examples, the SVM builds a model that for instance aims at predicting if a future news story belongs to the "relevant" or the "not relevant" group. What "relevant" and "not relevant" exactly means can be defined by the analyst, who might wish to evaluate whether a price reaction will be "strong or weak" or "positive, neutral, or negative" (Groth and Muntermann, 2009).

During the following analysis, we will exemplarily show the application of machine learning techniques in the financial domain. Hereby, we aim to forecast whether or not the publication of a new corporate disclosure is followed by extremely high abnormal volatility levels.

Textual Data Does Matter

Pursuing this attempt, our analysis is based on corporate disclosures and corresponding firms' intraday stock prices that were observed prior and subsequent to the disclosures' publication dates. For different 15 minutes intervals, realized volatilities were calculated in order to explore whether or not such publications will trigger significant capital market activities. Furthermore, the volatility observed subse-

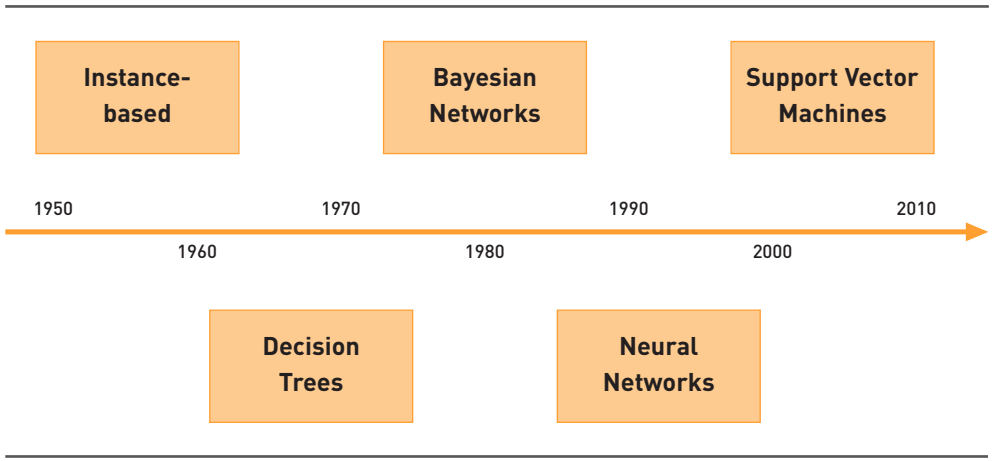

Figure 1: Machine Learning Techniques

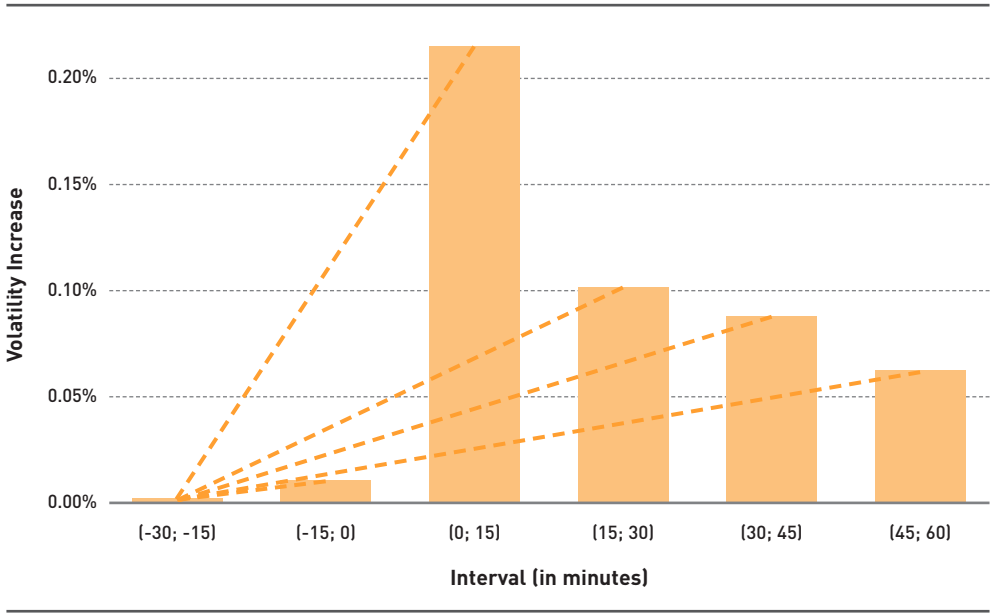

Figure 2: Volatility Shocks following News Publications

quent to the publication of corporate disclosures is adjusted by volatility levels observed during "normal" market phases. Consequently the volatility increases shown in Figure 2 are expected to be above zero during abnormal market phases only. In financial management, such volatility shocks attract much attention, especially in risk management.

As we can also see in Figure 2, negligible volatility increases can be observed for the intervals prior to the publication dates of the corporate disclosures. In contrast, significant volatility increases follow their publication. Over the course of time, the volatility adjusts to a normal level. It may therefore be concluded that the information contained in the corporate disclosures may entail abnormal volatility levels.

"Learning Machines" that Read

Within our samples of corporate disclosures there exist some that resulted in significantly increased volatility, while others seem to attract little attention only. Traditionally, an analyst would manually review the disclosures and assess their relevance. Given the corporate disclosures and the calculated volatility increases, we define a corresponding learning task: Let a computer learn from historical data with the goal to identify those disclosures that resulted in the 25% highest volatility increases during the first 15 minutes following the disclosures' publication. In other words, based on a disclosure's content, automatically assign it to either the class with expected extremely high abnormal volatility levels or the class with expected normal volatility levels. This task is divided into two sub-phases. A learning phase in which the computer develops a model from provided observations and an application phase in which the model is deployed. Here, other observations are evaluated that were not part of the learning dataset. We have used a Support Vector Machine (SVM) algorithm in order to learn a model that can be applied to the textual corporate disclosures. In simple words, the SVM aims to find two distinctive areas in a multidimensional space, where each word represents one of these dimensions and where the two areas divide existing documents into two different classes.

Forecasting Results

Our results provide evidence that the proposed machine learning approach is capable to detect patterns in the disclosure contents. We applied different evaluation metrics to evaluate how well the automated text analysis works. One has to distinguish between how many cases were identified correctly (precision) and how many of all relevant disclosures were identified (recall). Usually, a higher precision can only be achieved by accepting a lower recall and vice versa. In order to influence the inherent trade-off, we additionally included misclassification costs as a steering mechanism. We were for example able to modify settings in such a way that each corporate disclosure that was assigned to the "extremely high volatility class" actually belongs there (i.e. 100% precision). In this scenario, however, merely 33.02% (recall) of all "extremely high volatility"-entailing corporate disclosures were grasped.

Conclusion

Latest machine learning techniques such as Support Vector Machines represent promising approaches to capitalize more efficiently on the massive amounts of available textual data in financial processes. Our empirical analyses have shown that the applied algorithms were able to detect patterns in the disclosure contents and forecasting results were significantly better than random guessing. The field of application in the financial context is manifold (e.g. Fung et al., 2005). One possible application field could be algorithmic trading, where transactions might be triggered based on analyzed textual data. This application can already be observed in latest industry developments. With regard to the observed volatility increases, market monitoring tools may support the management of intraday market risks.

References

Fung, G.P.C.; Yu, J.X.; Lu, H.:
The Predicting Power of Textual Information on Financial Markets, IEEE Intelligent Informatics Bulletin, 5 (1) (2005).

Groth, S.; Muntermann, J.:
Supporting Investment Management Processes with Machine Learning Techniques, Internationale Tagung Wirtschaftsinformatik, Wien (2009).