IMPACT OF FEATURE SELECTION AND FEATURE TYPES ON FINANCIAL STOCK PRICE PREDICTION

Michael Hagenau, Michael Liebmann and Dirk Neumann
Institute of Information Systems Research, University of Freiburg, Freiburg, Germany

Keywords: Text mining, Machine learning, Financial news, Stock price effect.

Abstract: In this paper, we examine whether stock price effects can be automatically predicted analyzing unstructured textual information in financial news. Accordingly, we enhance existing text mining methods to evaluate the information content of financial news as an instrument for investment decisions. The main contribution of this paper is the usage of more expressive features to represent text through the employment of market feedback as part of our word selection process. In a comprehensive benchmarking, we show that a robust Feature Selection allows lifting classification accuracies significantly above previous approaches when combined with complex feature types. That is because our approach allows selecting only semantically relevant features and thus, reduces the problem of over-fitting when applying a machine learning approach. The methodology can be transferred to any other application area providing textual information and corresponding effect data.

1 INTRODUCTION

News plays an important role for investors when judging fair stock prices. In fact, news carries information about the firm's fundamentals and expectations of other market participants. From a theoretical point of view, an efficient valuation of a firm should be equal to the present value of the firm's expected future cash flows. The expectations crucially depend on the information set that is available to investors. The set consists of qualitative and quantitative information of different kind and from various sources, e.g. corporate disclosures, news articles and analyst reports. Due to improved information intermediation, the amount of available information – especially qualitative information – increased during the last decades. Since it is getting increasingly difficult for investors to follow and to take into account all the information available, automated classification of the most important information becomes more relevant.

Research in this area is still in its infancy. Despite numerous attempts, prediction accuracies for the stock price effect (i.e. positive or negative) following the release of corporate financial news never exceeded 58% (see Table 1) – an accuracy level still close to random guessing probability for two predictive states (50%) leaving room for substantial improvements.

Automated text mining translates unstructured information into a machine readable format and mostly uses machine learning techniques for classification. While suitable machine learning techniques for text classification are well established (Forman 2003; Joachims 1998), the development of suitable text representations is still part of ongoing research (Schumaker et al., 2009). In particular, determining the feature type (e.g. single words or word combinations) and choosing the most relevant features to represent text is the crucial part.

Existing literature on financial text mining mostly relies on very simple textual representations, such as bag-of-words (i.e. distinct single words). Further, the list of words or word combinations to actually represent text is selected based on dictionaries (Tetlock et al., 2008) or retrieved from the message corpus based on actual occurrences. Despite well researched approaches to select the most relevant words or word combinations based on exogenous feedback (Forman, 2003), existing work relies on frequency-based statistics of the message corpus, such as TF-IDF (Mittermayr, 2004) or just a minimum occurrence of a word combination (Schumaker, 2009). Thus, we expect potential for improvement in two areas: First, more complex and

Hagenau M., Liebmann M. and Neumann D..
IMPACT OF FEATURE SELECTION AND FEATURE TYPES ON FINANCIAL STOCK PRICE PREDICTION.
DOI: 10.5220/0003665602950300
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 295-300
ISBN: 978-989-8425-79-9

expressive features (e.g. Noun Phrases, word combinations) also capturing semantics should be used for text representation. Second, these features should be combined with a robust Feature Selection procedure to pick those features best discriminating between news messages leading to positive or negative stock price effects. As outside feedback from the stock market is needed to determine if a message was positive or negative, the Feature Selection method cannot rely on frequency-based statistics of the corpus, but has to utilize exogenous market feedback instead.

As every scholar tailors his methodology on his own data set and therefore is only vaguely comparable to previous results, we rebuild previous approaches in our evaluation to allow for a direct same-data benchmarking. We employ a data set of corporate disclosures which only contains firm-value relevant facts and therefore is very suitable for developing, improving and testing our approach.

The remainder of the paper is structured as follows: In section 2, we conduct a review of relevant research on prediction of stock price effects based on qualitative information. Section 3 designs our own approach for analyzing qualitative information and pinpoints the main innovations compared to existing work. In section 4, we present our analyses and findings from the comparison with existing approaches. Section 5 summarizes and outlines directions for further research on media content.

2 RELATED WORK & RESEARCH QUESTIONS

In this section, we give an overview on existing literature and pinpoint the differences to our approach. Our work is most closely related to Schumaker & Chen (2009) who also has the highest accuracy for stock price prediction based on financial news so far. The authors are one of the first to explore the impact of different Feature Extraction methods forming the basis for their Support Vector Machine (SVM) classification. Besides the extraction of single words and named entities, a proprietary tool was used to identify and aggregate noun phrases based on lexical semantic and syntactic tagging. However, feature selection remained rather simple: Only those features were selected that occurred at least three times in a document. Prediction accuracy did not exceed 58.2%. We mainly differ from Schumaker & Chen by applying exogenous-feedback-based Feature Selection to limit our feature set to the most relevant. Additionally, we find value in also including verbs into our features, unlike Schumaker's Noun Phrases and Named Entities. Our features are based on 2-word combinations which may occur with word distances of greater than zero. These word combinations are not limited to nouns, articles, and other determiners, but also may include verbs. Another closely related study was performed by Muntermann et al. (2009) who focus on the same news type (German Adhoc announcements) to have verifiable stock price effects. However, the authors' research can hardly be generalized due to its fairly small sample size of only 423 messages which need to be divided into training and validation set. Despite relying on the same data source as our work, with 56.5% accuracy, results are in the range of random guessing probability. Unlike Muntermann et al. (2009), Mittermayr (2004) employs a feature select to focus on relevant words: The TF IDF score which relates the occurrences of one term in processed document to the occurrence in all documents of the data set. However, prediction accuracy for positive and negative events is not directly specified in a comparable manner. Tetlock et al. (2008) use negative words in Wall Street Journal and Dow Jones News articles to create a content measure and predict stock returns. The content measure classifies messages as positive or negative based on the Harvard-IV-4 psychosocial dictionary – a selection of words widely used in psychological studies. Instead of prediction accuracies, the authors specify an R^2 of 0.24% between their content measure and the observed stock returns. A similar text message base, but different capital market effect predictions are used by Groth et al. (2011). Groth et al. predict intraday market risk based on German Adhoc announcements and use single words as features. Like Muntermann et al. (2009), the authors do not perform any Feature Selection besides the removal of stopwords. Accuracy values are not comparable due to different classification task, i.e. the absolute accuracy values may seem higher, but are achieved on subsets of the data.

With improved text mining technology and a relevant data set, we achieve prediction accuracies significantly higher than in literature. Existing work in prediction of stock prices has rarely used a robust Feature Selection to choose the most relevant features yet. As the number of possible combinations increases for more complex and expressive features, it becomes more relevant to select the features that could discriminate best between positive and negative effects. In our first

research question, we examine the impact of Feature Selection for different feature types:

Question 1: Does Feature Selection improve accuracies for more complex features than single words?

Prior research has almost exclusively relied on bag-of-words approach. Consistent with Schumaker & Chen (2009), we expect better predictive abilities for more complex features also capturing semantics in the text. This leads to our second research question:

Question 2: What is the impact of different feature types on classification accuracy?

The high number of possible combinations for complex features (such as 2-Grams, noun phrases or 2-word combinations) drives down actual occurrences in the overall message corpus increasing the risk of over-fitting. Over-fitting describes the fact that machine learning algorithms learn relations and structural dependencies in the training set which do not exist in reality and therefore can't be transferred onto the validation set. Over-fitting occurs when a larger number of features is used for learning than messages in the training set (i.e. high number of degrees of freedom, Cawley & Talbot, 2007). This leads to the third research question:

Question 3: Does Feature Selection reduce over-fitting?

The following section describes our approach to address these research questions.

3 METHODOLOGY

Analyzing unstructured information in the shape of text requires a complex processing algorithm. In order to classify text, exogenous feedback as base for the classification is required. The feedback has to have a direct cause-and-effect relation to the text. The algorithm can handle any kind of classification with two states as long as there is a direct relation between text and exogenous effect. For simplification of the two predictable states based on the text messages, in the following, just "positive" and "negative" will be mentioned. Consequently, the corresponding text messages will be named positive and negative messages.

We design a four step approach in order to process text messages and combine them with their exogenous feedback. The four steps can be separated into three steps of text processing, Feature Extraction, Feature Selection, Feature Representation, and the final step of the actual machine learning: We use a subset of the data (i.e. text-effect combinations) to train the machine learning algorithm. After training, the Support Vector Machine (SVM) is able to classify the remaining text messages into positive and negative. We measure the accuracy by comparing our classification results to the observed effect. The four steps of our algorithm can be described as following:

In **Feature Extraction**, we first define the feature type (e.g. words or word combinations) that best reflects the content of the message and second parse all messages to extract their features. We base our features on all words transported within the body of each message, i.e. we remove tables and graphs. During the parsing we extract each word separately. In order to remove redundancy between words with the same word stem, but a different commoner or inflexional ending, we employ the Porter Stemmer (Porter, 1980). Thus, we extract only word stems. For the experiment, the following feature extraction methods are used:

- Dictionary-approach – no features are extracted from the corpus. Instead, single words from the positive and negative word list in the Harvard-IV-4 psychosocial dictionary are used (see Tetlock et al., 2008)
- Single words retrieved from the corpus – this representation which is also called bag-of-words is most often used in literature (e.g. Groth et al. 2011; Mittermayr 2004; Muntermann 2009)
- N-Grams – a sequence of *N* words, letters or syllables (as in Butler et al. 2009). Performance of 3-Grams was slightly weaker than 2-Grams. Thus, 2-Grams were used.
- 2-word combinations – this feature type forms an extension of the word-based 2-Gram, allowing a word distance greater than zero between two words. In contrast to Noun Phrases, this feature type is not limited to certain parts of speech, but may also contain verbs and adverbs – as long as the Feature Selection attests high explanatory power. This feature type has not been used in literature yet
- Noun-phrases – a phrase whose head is a noun or a pronoun, optionally accompanied adjectives or other determiners (as in Schumaker & Chen 2009). Noun Phrases are extracted using the Stanford Parser (Klein & Manning 2003).

In **Feature Selection**, we exclude features that are of a lower explanatory power. As explanatory power we define the ability to differentiate between positive and negative messages. First, we take out stopwords, such as *"and"* and *"or"*. Second, we calculate the explanatory power by using a Chi-

Square based method as in (Forman, 2003).

In **Feature Representation,** we design a vector for each message based on all selected features in step 2. There are numerous methods of representing a feature within a vector. We found a feature best represented when using the logarithm of the feature's frequency within one message.

In the **Machine Learning** step, we use a Support Vector Machine (SVM) on combinations of messages, represented in feature vectors, and their consequent stock price effects. We transform the stock price effect into a binary measure, i.e. '0' for negative price effect and '1' for positive. We use a SVM since previous findings confirm it to be the best available machine learning method for text classification tasks (Forman 2003; Joachims 1998). Further, in a pilot study, we compared the performance of Artificial Neural Networks, Naïve Bayes and SVMs and found SVMs to be best performing.

Previous work mostly relies on the bag-of-words scheme, i.e. uses simple single words to represent text. The main contribution of this paper is the combination of advanced Feature Extraction methods with a customized Feature selection. The results of the evaluation in the following chapter show the value-add of Feature Selection for different Feature types.

4 EVALUATION

4.1 Evaluation Approach

In this evaluation, we apply our methodology to a set of corporate disclosures. We apply the Chi²-based Feature Selection to different types of features which have already been described in literature.

By reproducing approaches in literature and applying to the same data set, we are able to benchmark our approach in a same-data comparison. Every feature extraction approach is conducted once with feature selection based on market feedback and once without, i.e. simply by requiring a minimum occurrence in the corpus per feature (as e.g. in Schumaker & Chen, 2009; Butler et al., 2009). Thereby, we can demonstrate the improvements feasible by selection features based on market feedback. For exogenous-feedback-based feature selection, the Chi²-approach is used to choose the most relevant 10% of features occurring in the overall message set. If no special feature selection is performed, only stopwords are removed and all features with a minimum occurrence of 3 are used for representation of text messages.

Our data set comprises ~11,000 German Adhoc news published between 1998 and 2007. We removed penny stocks and extreme values (based on a 99%-interval). We required each message to have a minimum of 50 words in total. We impose these filters to limit the influence of outliers and avoid messages that only contain tables. Finally, we obtained 9,150 Adhoc announcements with consistent stock price information eligible for our experiment.

For capturing the announcement effect on financial markets, it is required to separate firm-specific effects from market-related effects. Therefore, we investigate daily abnormal returns on the day the Adhoc announcement was published (MacKinlay, 1997). The stock price effect is used to create a binary measure of the sign and label all text messages as either positive or negative.

4.2 Results

Results were obtained by running the SVM with a linear kernel which delivered best performance for text classification tasks using a high number of features (Joachims 1998). Table 1 shows the classification results on full training (7,100 messages) and validation set (3,050 messages). Accuracy is measured as percentage of correctly classified messages. For all five Feature types, we performed training and validation, once with our customized Feature Selection and once without (i.e. using all features with a minimum frequency). Results are stated as classification accuracies. Only for the Dictionary approach (single word) we did not perform our approach as the Dictionary itself is already a kind of Feature Selection.

In the following, we present our findings that are directly related to our research questions.

Finding 1: Chi²-based Feature Selection improved classification accuracies for all feature types

Results show that all feature types benefited from the Chi²-based Feature Selection, through an improved accuracy for all validation experiments. The highest performance on the validation set with 65.2% was achieved for the 2-word combination with Chi²-based Feature Selection. The 2-word combination performed slightly better than 2-Grams (62.6%) and Noun Phrases (63.7%) and significantly better than the single word approaches. The 2-word combination benefited most from Feature Selection, single words least. This observation extend the findings of Forman (2003) who relied on single

Table 1: Classification results for different feature types.

FEATURE TYPE	SUBSET	ACCURACY WITHOUT SPECIAL FEATURE SELECTION	ACCURACY WITH CHI²-BASED FEATURE SELECTION
Single words I: Based on dictionary	Training	62.8%	-
	Validation	58.1%	-
Single words II: Retrieved from corpus	Training	71.6%	62.8%
	Validation	58.6%	58.7%
2-Grams	Training	78.3%	69.7%
	Validation	56.8%	62.3%
2-word combinations	Training	87.2%	81.7%
	Validation	58.3%	65.1%
Noun Phrases	Training	75.2%	72.1%
	Validation	57.7%	63.5%

words as text representation and only found limited benefits of feature selection in combination with an SVM as machine learning approach.

Finding 2: Classification accuracy increases with complexity of features when Feature Selection is used

Classification performance increases with complexity and expressiveness of features – expressiveness meaning the ability of features to capture and express content and explanatory power. This is consistent with the findings of a previous study (Schumaker & Chen 2009) showing an increased performance for Noun Phrases compared to single words. However, this performance increase can only be observed when a Feature Selection is employed. Without exogenous-feedback-based Feature Selection performance on validation set is rather similar for all feature types. Features seem to develop their expressiveness only after selecting the most relevant features and, thus, taking out the noise.

The dictionary (single words I) shows slightly lower performance (58.1%) than the single words II retrieved from corpus (58.6%) due to its limited word set which cannot capture all specifics and subject lingo of the underlying domain. An even lower accuracy was achieved by the 2-Grams without Feature Selection (56.8%) which suffer from a high number of random combinations with low expressiveness. Only after selecting those with highest explanatory power, better accuracies were reached (62.3%). Without Feature Selection, the 2-word combinations perform better (58.3%) than 2-Grams, but slightly worse than the single words. 2-word combinations may carry more expressiveness than 2-Grams, but compared to single words, they also suffer from a high number of random combinations when used without Feature Selection. Slightly better performance than 2-Grams was achieved for Noun Phrases. Due to the high number of possible combinations, we mostly found low frequencies for each Noun Phrase in the corpus (i.e. 95% of features with less than five occurrences). A low number of features representing a text message limits the ability of the SVM to correctly classify. Further, in contrast to 2-word combinations, Noun Phrases lack verbs and adverbs limiting their expressiveness.

Finding 3: Using Chi²-based Feature Selection indicates to reduce over-fitting

When using Feature Selection, we observe lower accuracy values in the training set. However, we also observe higher accuracy values on the validation set for complex feature types. This indicates that over-fitting in the training set has been reduced.

The risk of over-fitting increases for more complex features, such as 2-Grams, noun phrases or 2-word combinations. For these features, the higher number of possible combinations leads to a higher number of features (but with low frequency in the corpus). Thus, Feature Selection is needed to choose the features with highest explanatory power and allow for high validation accuracies.

We are cautious in stating a full causal relationship between Feature Selection and the reduction in over-fitting. It is obvious that just a reduction of features (without selection the most relevant) will decrease training accuracy values. However, just reducing the number of features

compromises accuracy on the validation set. Feature Selection reduces the number of features, but increases accuracy, since it only takes out less relevant features. Thus, over-fitting might be actually reduced by Feature Selection.

For single words, Feature Selection is not beneficial. It still reduces accuracy values in the training set. However, this could be attributed to the pure reduction in the number of features.

An important remark relates to computational complexity. While Feature Selection, Feature Representation and the final classification by the SVM are of polynomial complexity (Burges 1998), major differences arise for Feature Extraction. Computational cost is mainly driven by the number of words per text message, number of used features and the corpus size, i.e. the number of total messages. As the corpus size is a linear complexity factor for all Feature Extraction methods, it's not considered in detail.

Bag-of-words and 2-Grams run in $O(M*F)$ with M as the number of words per message and F as the number of considered features. For extraction of 2-word combinations, complexity increases to $O(M*W*F)$ with W as the maximum distance between two words. However, the time consumed by the part of speech tagger task cannot be bounded by a polynomial (Klein & Manning 2003). Thus, Noun Phrases come at very high cost despite lower validation accuracies than 2-word combinations.

5 CONCLUDING REMARKS

In summary, our research shows that the combination of advanced Feature Extraction methods and our feedback-based Feature Selection boosts classification accuracy and allows improved sentiment analytics. Feature Selection significantly improves classification accuracies for different feature types (2-Gram, Noun Phrases and 2-word-combinations) from 55-58% up to 62-65%. These results were possible because our approach allows reducing the number of less-explanatory features, i.e. noise, and thus, may limit negative effects of over-fitting when applying machine learning approaches to classify text messages.

Our text mining approach was demonstrated in the field of capital markets – an area with numerous, direct and verifiable exogenous feedback. Such feedback is essential to develop, improve and test a text mining approach. However, since our approach is multi-applicable, it can be used on different data sets like marketing, customer relationship management, security and content handling. Future research will transfer our findings to these areas.

REFERENCES

Burges, C. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery* 2, pp. 121-167

Butler, M., Keselj, V. 2009. "Financial Forecasting using Character N-Gram Analysis and Readability Scores of Annual Reports", *Advances in AI*

Cawley, G., Talbot, N. 2007. "Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters", *Journal of Machine Learning Research* 8, pp.841-861

Forman, G, 2003. "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research* 3, pp. 1289-1305

Groth, S., Muntermann, J. 2011. "An Intraday Risk Management Approach Based on Textual Analysis", *Decision Support Systems* 50, p. 680

Joachims, T., 1998. "Text categorization with support vector machines: Learning with many relevant features", *Proceedings of the European Conference on Machine Learning*

Klein, D. & Manning, C. D. 2003. "Accurate Unlexicalized Parsing", *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

MacKinlay, C. A. 1997. "Event Studies in Economics and Finance", *Journal of Economic Literature*, S. 13-39.

Mittermayr, M.-A. 2004. "Forecasting Intraday Stock Price trends with Text Mining techniques", *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*

Muntermann, J., Guettler, A., 2009. "Supporting Investment Management Processes with Machine Learning Techniques", *9. Internationale Tagung Wirtschaftsinformatik*

Porter, M. F. 1980. "An Algorithm for Suffix Stripping", *Program*, 14(3): 130–137

Schumaker, R. P., Chen, H. 2009. "Textual analysis of stock market prediction using breaking financial news: the AZFin Text System", *ACM Transactions on Information Systems* 27

Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S, 2008. "More than words: Quantifying Language to Measure Firms' Fundamentals", *The Journal of Finance*, Volume 63, Number 3, June 2008 , pp. 1437-1467