Automated news reading: Stock Price Prediction based on Financial News Using Context-Specific Features

Michael Hagenau  
University of Freiburg  
michael.hagenau@is.uni-freiburg.de

Michael Liebmann  
University of Freiburg  
michael.liebmann@is.uni-freiburg.de

Markus Hedwig  
University of Freiburg  
markus.hedwig@is.uni-freiburg.de

Dirk Neumann  
University of Freiburg  
dirk.neumann@is.uni-freiburg.de

Abstract

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 and the employment of market feedback as part of our word selection process. In our study, 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 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 is based on the firm’s expected future cash flows. 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 substantially during the last decades. Since it is getting increasingly difficult for investors to consider all available information, 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 rarely exceeded 58% – an accuracy level still close to guessing probability for a binary predicted parameter (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 ([5], [8], [18]), the development of suitable text representations is still part of ongoing research [16]. 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 ([17], [12]) 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 [11] or just a minimum occurrence of a word combination [16].

Thus, we expect potential for improvement in two areas: First, more complex and 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 with previous results, we rebuild previous approaches in our evaluation to allow for a direct same-data benchmarking. We use a data set of corporate disclosures from two different sources. These disclosures only contain firm-value relevant facts and therefore are very suitable for developing, improving...
2. Related work & Research questions

In this section, we give an overview on existing literature and pinpoint the differences to our approach. Independent of the application area, we can differentiate related work along following topics:

- **Data set** – the textual message base and corresponding exogenous feedback – a suitable data set that conveys corresponding and verifiable effects with the message set is crucial to validate the quality of the text mining approach

- **Feature processing** – automated process step to generate machine readable information that best represents the content of the text

- **Machine learning** – information content of text is classified based on the output of Feature processing

The first topic to differentiate related work is the data set being used for analysis (Table 1 – DATA SET). Text classification always combines two separate sets of data: First, the news (text message base) and second, the resulting effect on the stock market, i.e. the exogenous feedback.

The Feature processing is a crucial part of text mining and can be characterized by the three common preparatory steps, being Feature Extraction, Feature Selection and Feature Representation. Feature Extraction typically denotes the process step in text mining to define the type of features that best reflect the content of the message and second parse all messages to extract features. Possible features are:

- Single discrete words – not capturing semantics between words, this feature type is also used in bag-of-words approach

- **N-Grams** – a sequence of $N$ words¹ (i.e. distance between words is zero)

- Noun-phrases – a phrase whose head is a noun or a pronoun, optionally accompanied adjectives or other determiners (e.g. “the big black cat”)

- Statistical information on a message (e.g. number of words, position of words)

The subsequent step – Feature Selection – reduces the number of features by cleaning all redundant to obtain the optimal subset, i.e. the subset being which contains the relevant information. An extensive overview of different Feature selection methods is provided by [5]. Three different approaches for feature selection can be observed in literature:

1. **Dictionary-based**: The dictionary-based approach which uses an existing and established dictionary, where domain experts have manually identified the most relevant words ([12], [17]).

2. **Feature selection based on endogenous information in the text message base**: Instead of using a dictionary, features are derived solely from information in the message corpus. Besides very simple measures for relevant words requiring a minimum occurrence as in [16], literature also employs more sophisticated methods. For example, [11] selects the features based on the concept of TF-IDF (i.e. term frequency - inverse document frequency) where occurrences of one term in the processed document are related to the occurrences in all documents of the data set [15]. However, these approaches only base feature selection on endogenous information in the corpus.

3. **Feature selection using exogenous market feedback**: Besides endogenous information in the corpus, feature selection can use market feedback as exogenous effect to select the most relevant features discriminating between positive and negative messages [5]. Market feedback can provide an objective view on the importance of features.

The step of Feature Representation is a technical processing step required to transform the relevant information in a computer-readable format (e.g. document vectors).

For the **machine learning** step, literature shows many different approaches that have been applied in the area of forecasting on the basis of textual messages. Machine learning refers to the discipline where large amounts of data are analyzed by computerized algorithms. One major peculiarity of machine learning approaches (e.g. artificial neural networks, SVMs) is

---

¹ Besides to words, the term ‘N-Gram’ also refers to a sequence of $N$ letters or $N$ syllables
that the algorithms automatically learn to identify patterns using a training set ([8], [14]). Those patterns can be used for classification of data different than the training set (i.e. validation set).

However, comparing different approaches in previous work, it seems that results are not as dramatically dependent on the applied machine learning approach. Instead, the results are crucially dependent on – as previously mentioned – the Feature Selection of the underlying messages. We make use of an SVM approach as it is still the most advanced machine learning technique and most suitable for text classification ([5], [8], [18]).

The main metric to measure the performance of the decision making task is the accuracy – the number of messages classified correctly. With over 65%, we reach considerably better results than any previous work. It is important to note that this accuracy was reached on the validation set. Related work struggled to achieve 58% which is close to guessing probability.

Table 1 provides a summary of our related work and is clustered along the three described topics (data set, feature processing, and decision making).

Our work is most closely related to [16] 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 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 [16] 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 Noun Phrases and Named Entities in [16]. 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 [12] who also focus on German Adhoc announcements to have verifiable stock price effects. However, the

---

**Table 1 - Main differences to related work**

<table>
<thead>
<tr>
<th>Author</th>
<th>Text Base</th>
<th>Text Mining – Feature Processing</th>
<th>Machine Learning</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muntermann et al. 2009 [12]</td>
<td>German Adhoc messages</td>
<td>Stock prices (Daily)</td>
<td>Bag-of-words</td>
<td>Only stopword removal</td>
</tr>
<tr>
<td>Antweiler et al. 2004 [1]</td>
<td>US message postings</td>
<td>Stock prices (Intraday) and Volatility</td>
<td>Bag-of-words</td>
<td>Minimum information criterion</td>
</tr>
<tr>
<td>Tetlock et al. 2008 [17]</td>
<td>US financial news</td>
<td>Stock prices (Daily)</td>
<td>Bag-of-words</td>
<td>Pre-defined Dictionary</td>
</tr>
<tr>
<td>This paper</td>
<td>German Adhoc messages</td>
<td>Stock prices (Daily)</td>
<td>Word combinations</td>
<td>Chi²-based feature selection</td>
</tr>
</tbody>
</table>

2 Results not directly comparable since 3 states (Positive, negative, neutral) are predicted. Precision rates for positive (6%) and negative (5%) events are very low. Accuracy for positive and negative events can be calculated from provided figures and is at only 2.5%.
authors' research can hardly be generalized due to its small sample size of only 423 messages which need to be divided into training and validation set. In addition, their work is influenced by the fact that the authors refrain from performing any feature selection and rather use all words after having removed stopwords (such as e.g. “the” and “of”). Despite relying on the same data source as our work, results are in the range of guessing probability. As the authors also mention, an accuracy of above 50% is only achieved because of an unbalanced data set containing more positive than negative news. Always guessing ‘positive’ would deliver the same accuracy as the proposed SVM-based approach. Unlike [12], [11] uses a feature selection to focus on relevant words: The TF IDF score which relates the occurrences of one term in the 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, but can be estimated to be lower than other previous work².

[1] use internet stock message postings to predict market volatility and stock return. Similar to [11], they refrain from using a dictionary and select the features from the message corpus by applying the minimum information criterion. They find that the effect of messages on stock returns is statistically significant, but economically small. Prediction accuracies are not specified.

[17] 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 R² 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 [7] and [3]. [7] predict intraday market risk based on German Adhoc announcements and uses single words as features. Like [12], the authors do not perform any Feature Selection besides the removal of stopwords. Accuracy values are not comparable due to different classification tasks, i.e. the absolute accuracy values may seem higher, but are achieved on subsets of the data. [3] predict one-year stock price developments relative to a benchmark based on historic annual reports. The authors use N-Grams as features and select features based on a minimum occurrence. For the classification task, the authors use a proprietary statistical measure. Accuracies reach relatively high values for subsets (up to 78%) of the data, but are not comparable to our classification task.

In summary, it can be stated that results crucially depend on the underlying text mining technology – in particular, the Feature Extraction and Selection. Features should be more complex and expressive than being just single words and they should be retrieved from the message corpus and employ exogenous feedback which discriminates between positive and negative price effects.

With improved text mining technology and a relevant data set, we achieve prediction accuracies significantly higher than in literature.

In summary, 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 [16], 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 not 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 [4]). 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
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:

1. 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 [13]. Thus, we extract only word stems. These word stems are either directly used as features in single word representation or combined to more complex features such as N-Grams or word combinations. Noun Phrases are extracted using the Stanford Parser [9].

2. 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 ([5]). Chi-Square compares the observed frequency $O_{ij}$ of the feature $i$ within the set of positive messages with its expected frequency $E_{ij}$, and normalizes the squared deviation. This deviation will be calculated for all four possible outcomes $j$, i.e. feature in positive/negative message and feature not in positive/negative message. The sum of all four normalized deviations constitutes the $X^2$-statistic.

$$X^2 = \sum_{j=1}^{4} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Through the $X^2$-statistic each feature receives a value for higher or lower deviation from the expected. Words that usually influence investor’s decision making carry higher values. Words that investors tend to appraise indifferently will receive lower values. To evaluate whether a word carries higher or lower explanatory power we calculate the p-value based on the Chi-Square Test. We cut-off the feature list at a p-value of 5%, i.e. we obtain a feature list with at least 95% confidence level that the average investor bases his investment decision also on these features.

Comparing our word list to the negative word list of the Harvard-IV-4 dictionary reveals superiority of incorporating market feedback into the feature selection. On the one hand, the dictionary does not reflect specific subject lingo like “bankruptcy”, “insolvency” and “lawsuit”. All of them have very negative meanings in the economic field. On the other hand, the dictionary assumes a negative meaning for words which can be positive in a certain context. The words “cancer” and “disease” are part of the negative word list, but are assigned a positive meaning in our approach. Cancer is a very serious disease, but also a fast-growing market segment for pharmaceuticals companies. Table 2 shows exemplary 2-word combinations of our feature list with high explanatory power (i.e. low p-value). The 2-word combination “however due” indicates that semantics within a sentence also contain value. Subordinate clauses introduced with “however” or “due” might justify a negative development.

<table>
<thead>
<tr>
<th>NO.</th>
<th>WORD</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>result loss</td>
<td>&lt;1.00 E-16</td>
</tr>
<tr>
<td>2</td>
<td>cell cancer</td>
<td>1.09 E-15</td>
</tr>
<tr>
<td>3</td>
<td>situation deteriorated</td>
<td>1.2 E-15</td>
</tr>
<tr>
<td>4</td>
<td>reason delay</td>
<td>3.54 E-14</td>
</tr>
<tr>
<td>5</td>
<td>growth outstanding</td>
<td>5.41 E-14</td>
</tr>
<tr>
<td>6</td>
<td>acquire firm</td>
<td>2.32 E-13</td>
</tr>
<tr>
<td>7</td>
<td>however due</td>
<td>3.45 E-13</td>
</tr>
</tbody>
</table>

3. 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.

4. The Machine Learning step classifies each message based on its content into positive or negative. We split this step up into two parts: First, we train a
Support Vector Machine (SVM) on combinations of messages, represented in feature vectors, and their consequent stock price effects. After training, the SVM is capable of classifying whether a represented message will have either a positive or a negative stock price effect. In training, we use the binary measure of the stock price effect as described above, i.e. ‘0’ for negative price effect and ‘1’ for positive. We use 2/3 of all events to train the SVM. The remainder of 1/3 is then used to validate the success on messages the SVM was not particularly trained on – the accuracy analyses will be discussed in the results section. We use a SVM since previous findings confirm it to be the best available machine learning method for text classification tasks ([5], [8], [18]). 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 exogenous market feedback and once without, i.e. simply by requiring a minimum occurrence in the corpus per feature (as e.g. in [3] and [16]). Thereby, we can demonstrate the improvements feasible by selection features based on market feedback. 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 [17])
- Single words retrieved from the corpus – this representation which is also called bag-of-words is most often used in literature (e.g. [7], [11], [12])
- N-Grams – a sequence of \( N \) words, letters or syllables (as in [3]).

- 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 [16])

<table>
<thead>
<tr>
<th>FEATURE TYPE</th>
<th>FEATURES EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single words I: Based on dictionary</td>
<td>record loss</td>
</tr>
<tr>
<td>Single words II: Retrieved from corpus</td>
<td>increase dividend net loss</td>
</tr>
<tr>
<td>2-word combination</td>
<td>guidance [...] upwards expect [...] lower</td>
</tr>
<tr>
<td>Noun Phrase</td>
<td>ongoing positive result in a difficult market environment</td>
</tr>
</tbody>
</table>

For exogenous-feedback-based feature selection, the Chi²-approach is used to choose the most relevant features occurring in the message set. All features with a p-Value of less than 5% were selected. If no special feature selection is performed, only stopwords are removed and all features with a minimum occurrence of 5 are used for representation of text messages. Table 4 shows the number of respective features used for the classification task.

<table>
<thead>
<tr>
<th>FEATURE TYPE</th>
<th>NO FEATURE SELECTION</th>
<th>CHI²-BASED FEATURE SELECTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single words I: Based on dictionary</td>
<td>3,106</td>
<td>-</td>
</tr>
<tr>
<td>Single words II: Retrieved from corpus</td>
<td>3,032</td>
<td>567</td>
</tr>
<tr>
<td>2-Gram</td>
<td>22,899</td>
<td>3,538</td>
</tr>
<tr>
<td>2-word combination</td>
<td>54,392</td>
<td>15,104</td>
</tr>
<tr>
<td>Noun Phrase</td>
<td>7,489</td>
<td>1,038</td>
</tr>
</tbody>
</table>

The number of features depends on the possible combinations for the feature type and the likelihood to occur in a message. Obviously, for features based on more than one word, more combinations are possible.
Most combinations are possible for 2-Grams and 2-word combinations. However, 2-word combinations are more likely to occur in an article; thus, more combinations exceed the thresholds for minimum occurrence and p-Value. With an increasing number of theoretically possible features, a robust Feature Selection becomes increasingly important.

4.2 Dataset – Textual news base

Our data set comprises corporate announcements from Germany and the UK published between 1998 and 2010. The announcements were obtained from two different sources: DGAP (“Deutsche Gesellschaft für Adhoc-Publizität”) and EuroAdhoc.

Regulatory requirements in many countries (e.g. US, UK, and Germany) oblige listed companies to publish any material facts that are expected to affect the stock price by an authorized intermediate publisher, such as the DGAP and EuroAdhoc. Thus, our news set forms a pre-selection of relevant news from the set of available financial news. These typically include facts on deviations of financial results from earlier expectations, management changes, M&A transactions, dividends, major project wins or losses, litigation outcomes and other types.

From the overall data set, we removed penny stocks and 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 10,860 Corporate announcements from our first source DGAP eligible for our experiment. Thereof, we used 2/3 messages for training of our machine learning method and the remainder of 3,620 news articles for validation. Further, we used the 3,430 obtained news articles from our second source EuroAdhoc as an additional validation set in order to examine to what extend our approach can be generalized and to confirm our results. EuroAdhoc covers different companies and includes news from other countries such as the UK.

4.3 Dataset – Ext. feedback: Stock price effects

Since we want to capture the announcement effect on financial markets, we need to process financial stock price data simultaneously. Accordingly, we first need to define the announcement effect which is inherently firm-specific. Thus, it is required to distinguish firm-specific effects from market-related effects. Accordingly, we investigate daily abnormal returns on the event day, i.e. the day the Adhoc announcement was published. We follow the standard event study methodology [10] and use the market model as benchmark. We estimate the coefficients for the market model based on a reference time window that spans over [-260, -21] trading days prior announcement.

For the stock price analysis, we used publicly available daily open and close prices obtained from Datastream. For events during trading hours, the stock price effect is calculated between open and close auction. For events occurring outside trading hours, effect calculation is based on close prices of previous day and open prices. One could argue that not using intraday stock prices for events occurring during trading hours, introduces potential inaccuracies. However, several reasons favor an approximation by open and close prices. First, opening and close auctions have higher volumes and lead to more valid prices. Second, there is no definition how long pricing of new information takes and assumptions need to be made [6]. Third, the high number of events is expected to balance out noise before and after the event.

The stock price effect is used to create a binary measure of the sign and label all text messages as either positive or negative.

4.4 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 5 shows the classification results on the full training (7,240 messages) and the validation set (3,620 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. 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 first validation set (from DGAP) 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 extends the
findings of [5] who relied on single words as text representation and only found limited benefits of feature selection in combination with an SVM as machine learning approach. The findings are confirmed by the second validation set. Again feature selection increases accuracies for more complex feature types. As training was performed on a different data set containing e.g. different companies, classification accuracies are generally lower on containing e.g. different companies, the second validation set.

**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 [16] 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 on the first validation set from DGAP (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 (57.2%) 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 (57.9%) 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 (57.5%). Noun Phrases may include more than two words and partially captures semantics in a sentence. When combined with Feature Selection, Noun Phrases achieve second highest accuracy (63.5%) in the overall field. Still, 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.

<table>
<thead>
<tr>
<th>FEATURE TYPE</th>
<th>SUBSET</th>
<th>DATA I: DGAP</th>
<th>DATA SET II: EUROADHOC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NO FEATURE SELECTION</td>
<td>CHI²-BASED FEATURE SELECTION</td>
</tr>
<tr>
<td>Single words I: Based on dictionary</td>
<td>Training</td>
<td>62.8%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>58.1%</td>
<td>-</td>
</tr>
<tr>
<td>Single words II: Retrieved from corpus</td>
<td>Training</td>
<td>66.9%</td>
<td>62.8%</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>58.6%</td>
<td>58.6%</td>
</tr>
<tr>
<td>2-Grams³</td>
<td>Training</td>
<td>78.3%</td>
<td>66.0%</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>57.2%</td>
<td>62.3%</td>
</tr>
<tr>
<td>2-word combinations</td>
<td>Training</td>
<td>90.4%</td>
<td>81.7%</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>57.9%</td>
<td>65.1%</td>
</tr>
<tr>
<td>Noun Phrases</td>
<td>Training</td>
<td>75.2%</td>
<td>72.1%</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>57.5%</td>
<td>63.5%</td>
</tr>
</tbody>
</table>

³ Performance of 3-Grams was slightly weaker than 2-Grams and is therefore not listed.
learning than there are messages in the training set, the risk of over-fitting increases [4]. Thus, Feature Selection is needed to choose the features with highest explanatory power and allow for high validation accuracies.

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 (see Table 4).

An important remark relates to computational complexity. While Feature Selection, Feature Representation and the final classification by the SVM are of polynomial complexity [2], 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 [9]. Thus, Noun Phrases come at very high cost despite lower validation accuracies than 2-word combinations.

4.5 Case study: Trading simulation

This section provides not only theoretical research metrics, but also value metrics ensuring practical applicability in this domain. We simulate the average achievable return per trade following a simple trading strategy: For positive trading signals, the underlying stock is bought (i.e. long position), for negative signals, the underlying stock is short-sold (i.e. short position). For simplification, we do not account for volume effects, i.e. we only assume that one stock is bought or short-sold. To ensure that simulated returns can be realized in practice, we only select messages published during trading hours. Further, we focus on the top110 most liquid stocks (as in HDAX published by Deutsche Börse AG) to reduce market frictions and better approximate actual returns. Thus, from all messages in the validation set, only ~840 are used for this experiment. In contrast to abnormal returns described in section 4.3, we use actually observed returns for this evaluation. Abnormal returns reduce the impact of overall market effects and are essential to actually assess if a message was positive or negative. However, for evaluating the potential of a trading strategy, actually observed returns on the stock market are needed. Table 6 describes the average returns per trade achieved with the described trading strategy.

Table 6 – Average returns per trade dependent on Feature Selection (FS)

<table>
<thead>
<tr>
<th>FEATURE TYPE</th>
<th>RETURN – NO SPECIAL FS</th>
<th>RETURN – BASED FS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single words I: Based on dictionary</td>
<td>0.4%</td>
<td>-</td>
</tr>
<tr>
<td>Single words II: Retrieved from corpus</td>
<td>0.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>2-Gram</td>
<td>0.3%</td>
<td>0.8%</td>
</tr>
<tr>
<td>2-word combinations</td>
<td>0.5%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Noun Phrase</td>
<td>0.3%</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

Similar to Table 4, Feature Selection (FS) has a clear benefit. Also return performance of different Feature types ranks in the same order as accuracies in Table 4. However, differences arise in the magnitude of the return: Slight percentage changes in accuracy already lead to strong increase in return profits. The 2-word combinations again show highest performance, even when used without Feature Selection. As each message has a stock return of different magnitude, the computed average of stock returns are not fully monotonically related to classification accuracies. Assuming a normal distribution, the 99%-confidence interval for 2-word combinations with feature selection is [0.7%;1.5%]. Thus, profits can be realized with low variability.

The table shows raw returns before transaction costs and market frictions. However, if transaction costs of 0.1% are assumed [17], returns still remain positive. Additionally, market frictions are reduced by only considering the top110 most liquid stocks.

Overall, trading simulation shows that a profitable trading strategy can be established. More complex feature types and employment of a robust Feature Selection significantly increase returns.

Still, accuracy of the simulation could be further increased by employing intraday stock price effects for calculating returns. Additionally, the simulation should include trading volumes and liquidity to assess actual profits in euro-value and fully consider market frictions.
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.

Results are confirmed by an additional separate data set which is used only for validation. The separate data set contains news from a different provider dealing with different companies and also including news from the UK. Having similar results on two different data sets indicates that findings can be generalized onto other news types and countries.

We simulate our approach in a simple, but rewarding trading strategy to demonstrate achievable returns. Thereby, using 2-word combinations and Chi²-based Feature Selection leads to the highest returns in the field.

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 fulfilling the following requirements: First, the text base consists of a sufficiently large number of single text messages with a minimum number of relevant words. The minimum corpus size depends on the variety of content. The higher the variety, the more text messages are needed to allow for sound training and validation. Second, for each text message verifiable exogenous feedback must be available which is directly corresponding to the text message. Difficulties arise if feedback is only provided for multiple text messages, e.g. if multiple messages form a negotiation log and only one outcome for the whole negotiation is available. Application areas fulfilling these criteria are manifold and include marketing, customer relationship management, security and content handling. Future research will transfer our findings to these areas.

References


