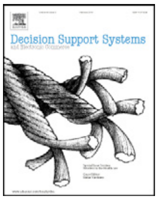
Decision Support Systems 55 (2013) 685–697

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Automated news reading: Stock price prediction based on financial news using context-capturing features

Michael Hagenau *, Michael Liebmann, Dirk Neumann

University of Freiburg, Platz der Alten Synagoge, 79085 Freiburg, Germany

ARTICLE INFO

Article history:
Received 8 May 2012
Received in revised form 25 November 2012
Accepted 4 February 2013
Available online 20 February 2013

Keywords:
Text mining
Financial news
Stock price prediction
Decision support

ABSTRACT

We examine whether stock price prediction based on textual information in financial news can be improved as previous approaches only yield prediction accuracies close to guessing probability. Accordingly, we enhance existing text mining methods by using more expressive features to represent text and by employing market feedback as part of our feature selection process. We show that a robust feature selection allows lifting classification accuracies significantly above previous approaches when combined with complex feature types. This is because our approach allows selecting semantically relevant features and thus, reduces the problem of over-fitting when applying a machine learning approach. We also demonstrate that our approach is highly profitable for trading in practice. The methodology can be transferred to any other application area providing textual information and corresponding effect data.

1. Introduction

When analysts, investors and institutional traders evaluate current stock prices, news plays an important role in the valuation process. In fact, news carries information about the firm's fundamentals and qualitative information influencing expectations of market participants. From a theoretical point of view, an efficient valuation of a firm should reflect the present value of the firm's expected future cash flows. The expectations on the firm's development are crucially dependent on the information set that is available to investors. The information set consists of news that contains qualitative as well as quantitative information from various sources, e.g., corporate disclosures, third party news articles and analyst reports. If financial news conveys novel information leading to adjusted expectations about either firm's cash flows or investor's discount rates, it affects stock returns [4,18]. In the news, not only financial figures have a significant impact on stock price, but also the qualitative textual components impact stock prices [27] when containing new information [14,29].

Due to improved information intermediation, the amount of available information has dramatically increased for the last decades. Since it is getting increasingly difficult for investors to follow and consider all available information, automated classification of the most important information becomes more relevant.

Research in automated classification of textual financial news is, however, in its infancy. Despite numerous attempts and application areas (c.f. [15]), prediction accuracies for the direction of stock prices following the release of corporate financial news rarely exceeded 58% (see Table 1) — an accuracy level hardly above random guessing probability (50%) leaving room for substantial improvements.

Automated classification of textual news comprises text mining which 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 [8,12], the development of suitable text representations is still part of ongoing research [24]. Essentially, text representation techniques refer to the way text is handled. One prominent example is the bag-of-words model, which regards the text as a compilation of unordered single words. In such a case, the feature type ‘single words’ constitutes the text representation. More complex feature types refer to word combinations. Clearly, not all words are needed to reflect a given text; text mining is concerned with the search for the most relevant features to represent the text.

Existing literature on financial text mining typically relies on very simple textual representations, such as the aforementioned bag-of-words model. Further, the list of words used for text representation are created either on the basis of dictionaries [17,28] or retrieved from the message corpus based on actual occurrences of the words. Despite well researched approaches to select the most relevant words or word combinations based on exogenous feedback [8], existing work often relies on frequency-based statistics of the message corpus, such as the information retrieval measure TF-IDF [19] or, even simpler, the minimum occurrence of a word combination [24]. Having in mind that these approaches used in financial text mining are very simple and do not employ state-of-the-art methods, we expect potential for improvement with respect to two areas: First, we need to explore more complex

* Corresponding author. Tel.: +49 761 203 2395; fax: +49 761 203 2416.
E-mail addresses: michael.hagenau@is.uni-freiburg.de (M. Hagenau), michael.liebmann@is.uni-freiburg.de (M. Liebmann), dirk.neumann@is.uni-freiburg.de (D. Neumann).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.02.006

Table 1
Summary of related work (ordered by relevance to our work).

Author	Data set		Text mining — feature processing			Machine learning	
	Text base	Effect	Feature type	Selection method	Market feedback	Method	Accuracy
Schumaker et al. 2009 [24]	US financial news	Stock prices (intraday)	Noun phrases	Minimum occurrence per document	No	SVM	58.2%
Schumaker et al. 2012 [25]	US financial news	Stock prices (intraday)	Noun phrases	Minimum occurrence per document	No	SVR	59.0%
Groth et al. 2009 [10]	German adhoc announcements	Stock prices (daily)	Bag-of-words	Only stopword removal	No	SVM	56.5%
Mittermayr 2004 [19]	US financial news	Stock prices (daily)	Bag-of-words	TF IDF: selecting 1000 terms	No	SVM	—[1]
Wüthrich et al. 1998 [30]	Worldwide general news	Index prices (daily)	Bag-of-words	Pre-defined dictionaries	No	K-nn, ANNs, naïve Bayes	Not comparable
Li 2010 [16]	US corporate filings	Stock prices (daily)	Bag-of-words	Pre-defined dictionaries	No	Naïve Bayes	Not available
Antweiler et al. 2004 [1]	US message postings	Stock prices (intraday) and volatility	Bag-of-words	Minimum information criterion	No	Combination: Bayes, SVM	Not available
Das & Chen 2007 [7]	US message postings	Stock and index prices (daily)	Bag-of-words	Pre-defined dictionaries	No	Combination of different classifiers	Not comparable
Tetlock et al. 2008 [28]	US financial news	Stock prices (daily)	Bag-of-words	Pre-defined dictionary	No	Ratio of negative words	Not available
Groth et al. 2011 [11]	German adhoc announcements	Intraday market risk	Bag-of-words	Chi^2-based feature selection	Yes	SVM	Not comparable
Butler et al. 2009 [3]	US annual reports	1-Year market drift	N-Gram	Minimum occurrence per document	No	Proprietary distance measure	Not comparable

and expressive features (e.g., word combinations) that may be capable of capturing the underlying semantics of the text messages. Second, these features should be combined with a robust selection procedure to pick those features that can best discriminate between news messages entailing positive or negative stock price effects. The assessment of whether or not a message contained positive or negative content requires the reaction of the stock market response to the message as feedback. Thus, an appropriate feature selection method cannot rely on frequency-based statistics of the corpus as the only measure, but has to utilize exogenous market feedback instead.

Most related research in this area suffers from the fact that each researcher uses his proprietary method and evaluates those methods on the ground of a proprietary data set. As a consequence, the results of related methods are vaguely comparable to each other. To make our results comparable, we rebuild previous approaches in our evaluation to allow for benchmarking on the same data set. We use corporate disclosures from two different sources as a data set. These disclosures only contain firm-value relevant facts and thus are very suitable for developing, improving and testing our approach.

Our study shows that features capturing context, i.e., combinations of words, push prediction accuracies significantly above those of related research approaches (up to 76%) when combined with a feature selection that utilizes feedback from the stock market. Practical applicability is demonstrated by a trading simulation based on backtesting results. It turns out that the implemented, albeit simple, trading strategy is highly profitable.

The remainder of the paper is structured as follows: In Section 2, we describe the generic steps in text classification and conduct a comprehensive review of relevant research on the prediction of stock price effects based on qualitative information. The review targets the main differences and exposes the shortcomings of former research. Section 3 presents the design of our own approach for analyzing qualitative information and pinpoints the main innovations compared to existing work. In Section 4, we benchmark our approach by rebuilding existing approaches and discuss the results. Section 5 performs a trading simulation (backtesting) under conditions close to reality. In Section 6, we summarize the paper and outline implications of applying our algorithm outside capital market research.

2. Related work

Existing approaches in financial prediction literature mainly differ in three aspects, being the (i) data set corresponding to a certain application field, the text mining approach, i.e., (ii) the feature processing and (iii) the machine learning algorithm. Accordingly, related work is structured along these three aspects:

i. The *data set* consists of two distinct subsets: the textual message base and corresponding stock market reaction following the announcement of these messages (e.g., stock price reactions).
ii. The *feature processing* task is an automated process step to generate machine readable information that adequately represents the content of the text
iii. The *machine learning* algorithm classifies the text based on the output of feature processing and is used for predicting the stock market reaction.

When comparing the performance of different approaches in literature, it is important to consider the *data set* used for analysis (Table 1 — Data set). As classification tasks vary in difficulty and – depending on the content of the news – some messages are easier to classify than others, performance is only comparable if the same or a very similar data set is used. In the area of financial prediction, the data set always consists of two distinct subsets: on the one hand, the textual message base containing financial news and corporate disclosures and, on the other hand, the corresponding stock market reaction following the announcement of these messages (i.e., the exogenous stock market feedback). Possible stock market reactions range from pure stock price movements for various time-spans to volume and volatility changes.

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 used to define the type of features that best reflect the content of the message and parse all messages to extract features. As previously mentioned the simplest approach – called "bag-of-words" – uses frequencies of single discrete words to represent text. Thus, the bag-of-words model is not capturing any semantics between words. More sophisticated approaches being discussed in literature use combinations of words as features such as Noun-phrases (e.g., "the big black cat") or any sequence of words (i.e., N-Gram).

The subsequent step, commonly denoted as feature selection, reduces the number of features by aiming to remove redundant data to obtain the optimal subset, in a sense that the subset is as small as possible but still retains all the relevant information. An extensive overview of different feature selection methods is provided by [8]. If features are selected thoroughly, it is expected that the feature set will contain only the most relevant information from the input data instead of the full-sized redundant input. Essentially, three different approaches for feature selection can be observed in literature:

1. *Dictionary-based*: The dictionary-based approach employs an established dictionary, where domain experts have manually identified the most relevant words [17,28]
2. *Feature selection without exogenous market feedback*: Instead of using a dictionary, features are derived solely from information in the message corpus. In addition to very simple measures for relevant words requiring a minimum occurrence as in [24], literature also employs more sophisticated methods. For example [19], 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 [23]. However, these approaches only base feature selection on endogenous information in the corpus and does not benefit from exogenous feedback on how messages including certain features were perceived by the stock market
3. *Feature selection employing exogenous market feedback*: Besides endogenous information in the corpus, feature selection can employ market feedback as an exogenous effect to select the most relevant features discriminating between positive and negative messages [8,31].

Having identified the most relevant features, the next step of *feature representation* is a mechanical processing step needed to transform the relevant information in a computer-readable format (e.g., document vectors). Based on the computer-readable format of the feature processing step, the *machine learning* algorithm classifies the information content text which is used for predicting the stock market reaction. One major peculiarity of machine learning approaches (e.g., artificial neural networks, support vector machines (SVMs), and naïve Bayes classification) is that the algorithms automatically learn to identify patterns using a training set [22]. Those patterns can be used for classification to data different than the training set (i.e., validation set). However, comparing different approaches in previous work, it seems that results are not dramatically dependent on the applied machine learning approach. The main metric to measure the performance of the classification task is the accuracy — the number of messages classified correctly.

These three aspects, being data set, feature processing, and machine learning, characterize previous work in financial prediction. Table 1 exhibits related work according to this systematic approach. The main metrics used to express the performance of the approaches is the accuracy, defined as the number of accurately classified messages. This metric reported in Table 1 should be used with caution, as we displayed the accuracy level claimed in the respective papers. A comparison is not possible per-se as different data sets are used. Nonetheless, the numbers give an initial appraisal of the approaches.

For the discussion of related work, we first focus on closely related studies which predict short-term stock price reactions based on financial news by using machine learning algorithms. Our work is most closely related to Schumaker and Chen [24] who are one of the first to explore the impact of different feature types as input for their SVM classification. Beside 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 [24] 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 [24]'s noun phrases and named entities. Our features are based on 2-word combinations which may occur with word distances greater than zero. These word combinations are not limited to nouns, articles, and other determiners, but also may include verbs.

In a later attempt, Schumaker et al. [25] combined their approach with sentiment analysis techniques which slightly improved accuracies to 59.0%.

Another closely related study was performed by [10] who focus on 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. 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 known frequent, but less meaningful words such as stopwords. As the authors admit, their data set is skewed and contains more positive than negative news. Always guessing 'positive' delivered a higher accuracy (~60%) than the proposed SVM-based approach (~56%). Thus, results are even below guessing probability. Unlike the approach of [10], [19] employs a feature selection to focus on relevant words: the TF IDF score relates to the occurrences of one term in a processed document to the occurrence in all documents of the data set. However, the prediction accuracy for positive and negative events is not directly specified in a comparable manner, but can be estimated to be lower than in other previous work.[1]

Other studies are less closely related to our work and differ in classification approach and data set. Thus, results are not comparable to previously discussed studies. Wüthrich et al. [30] were one of the first to use machine learning and text mining techniques supporting financial decision making. They predict stock index movements based on a predefined financial dictionary using different machine learning approaches such as K-nn, artificial neural networks and naïve Bayes. Similarly Li [16], employs several pre-defined dictionaries and a naïve Bayesian approach to predict stock returns based on US corporate filings. Not using financial news, but rather internet stock message postings [1], predict market volatility and stock return. Similar to [19], 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. Like [1], [7] predict stock returns based on US internet stock message postings using a combination of several classifiers. They also make use of pre-defined dictionaries.

Renowned work in the financial literature has been published by [27] who use just 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-

[1] 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%.

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 effects are used by [11] and [3]. Groth and Muntermann [11] predict intraday market risk based on German adhoc announcements and use single words as features. The authors are one of the few in the field to employ Chi^2-based feature selection including exogenous market feedback. However, accuracy values are not comparable to our work due to a different classification task, i.e., the prediction of intraday market risk. Butler and Keselj [3] predict one-year stock price developments relative to a benchmark based on historic annual reports. The authors use N-Gram 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 (up to 70%), but are not comparable to our classification task.

When comparing performance of literature in intraday or daily stock price prediction, accuracy levels below 60% are observed. However, in practice, data sets are often skewed, i.e., contain more positive than negative messages or vice versa [10]. If there are more messages in one class, guessing probability for binary prediction variables is not at 50%, but at the ratio of the class with most messages. Consequently, it is important to account for the mix of positive and negative messages in the data set to assess the added value of the proposed approaches — what previous work often lacked to do. For many real data sets, an accuracy of 50%–60% could be achieved by trivially predicting the majority class. Thus, current accuracy levels below 60% indicate substantial room for improvement in the underlying text mining technology.

3. Research design

In this section, we structure the substantial room for improvement indicated by related work. We formulate three research questions which will guide the design of our text mining approach.

3.1. Research questions

Analysis of literature on intraday or daily stock price prediction indicated potential for improvement of the underlying text mining approach. Despite increasing performance for text classification tasks [8], financial prediction literature has not focused on using robust feature selection with exogenous market feedback to choose the most relevant features. 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 complex features (i.e., more complex than single words)?

Another potential for improvement addresses the fact that prior research has almost exclusively relied on the bag-of-words approach. This approach can hardly capture any semantics of the text and, in particular, cannot capture the context of a word (e.g., if 'increase' is mentioned in conjunction with 'cost' or with 'earnings'). Consistent with [24], we expect better predictive abilities for more complex features also capturing semantics and pieces of context in the text. This leads to our second research question:

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

The large number of possible combinations for complex features (such as 2-Gram, noun phrases or 2-word combinations) drives down actual occurrences per feature 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 accordingly cannot be transferred to the validation set. Over-fitting occurs when a larger number of features are used for learning than messages in the training set (i.e., high number of degrees of freedom [5]). This leads to the third research question which details research question 1 by highlighting one driver behind the improved accuracy:

Question 3: Does feature selection reduce over-fitting?

As previous questions are of theoretical nature, it is desirable that results actually can be exploited in practice. This implies that investment decisions based on generated trading signals by our approach are profitable even if trading commissions and liquidity restrictions are considered. This leads to our fourth research question:

Question 4: Can previous findings be used to establish a profitable trading strategy?

3.2. Text mining with context-capturing features

Based on the stated research questions, this subsection is dedicated to the presentation of our approach and how the research questions can be addressed. 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. In our case, the positive or negative stock price reaction constitutes the exogenous feedback for each news message. For simplification, the corresponding text messages will subsequently be denoted as "positive" and "negative" messages.

We design a four step approach in order to process text messages and combine them with their exogenous feedback (Fig. 1). The four

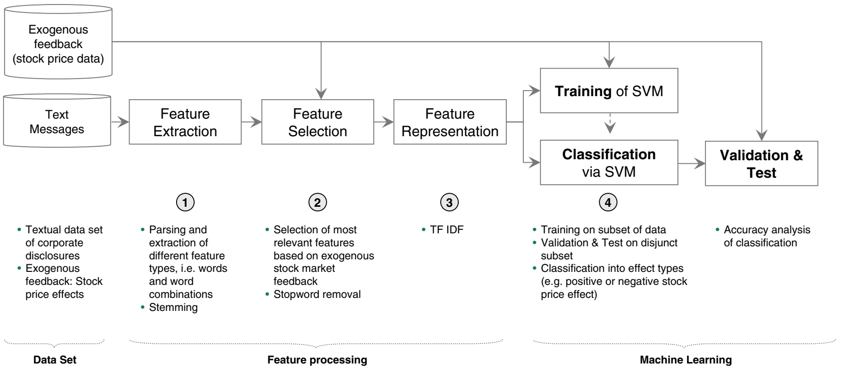

Exogenous feedback (stock price data)
Text Messages
Feature Extraction
Feature Selection
Feature Representation
Training of SVM
Classification via SVM
Validation & Test
1
2
3
4
• Textual data set of corporate disclosures
• Exogenous feedback: Stock price effects
• Parsing and extraction of different feature types, i.e. words and word combinations
• Stemming
• Selection of most relevant features based on exogenous stock market feedback
• Stopword removal
• TF IDF
• Training on subset of data
• Validation & Test on disjunct subset
• Classification into effect types (e.g. positive or negative stock price effect)
• Accuracy analysis of classification
Data Set
Feature processing
Machine Learning

Fig. 1. Overview of methodology for financial news classification.

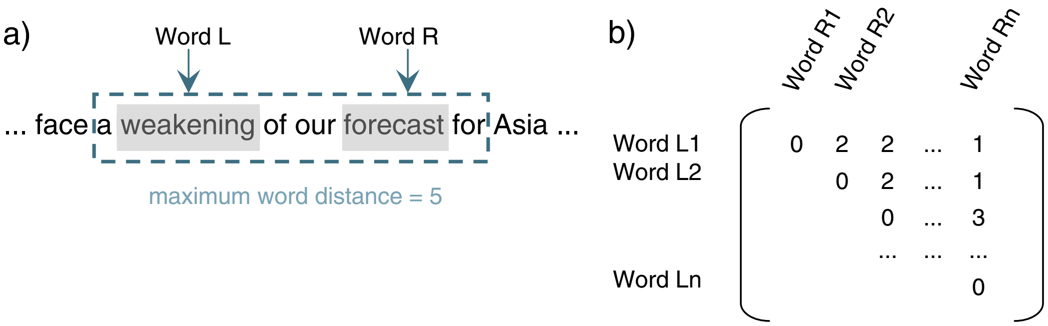


Fig. 2. Overview of methodology for financial news classification.

steps basically represent the three text processing steps, i.e., feature extraction, feature selection, feature representation, and, as the final step, the machine learning: We use a subset of the data 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 effects. Thus, our approach is a state-of-the-art text mining approach with enhanced feature extraction and selection.

The four steps of our algorithm can be described as follows:

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 [21]. Thus, we extract only word stems. For the experiment, each of the following feature types is extracted from the text:
   - Dictionary-approach — for determining the feature list, no features are retrieved from the corpus. Instead, single words from the positive and negative word list in the Harvard-IV-4 psychosocial dictionary are used [27].
   - Single words retrieved from the corpus — this representation which is also called bag-of-words is most often used in literature [10,11,19].
   - N-Gram — a sequence of N words, letters or syllables (as in [3]). Performance of 3-Gram was slightly weaker than 2-Gram since 3-Gram suffer from a high number of combinations causing a rapid decrease in actual frequencies per feature. Thus, 2-Gram were used in our experiment.
   - Noun-phrases — a phrase whose head is a noun or a pronoun, optionally accompanied adjectives or other determiners (as in [24]). Noun phrases are extracted using the Stanford Parser [13].
   - 2-Word combinations — this feature type embodies an extension of the word-based 2-Gram, allowing a word distance greater than zero between two words. In our case, we use a maximum word distance of five to allow for a certain degree of flexibility while limiting combinations across different sub clauses (see Fig. 2a).

     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. As this feature type has not been used in literature yet, it is described in more detail. For creation of the feature list, the occurrences of each combination of two words are counted in a 2×2-matrix with all possible English single words in the corpus for each dimension. As the order of the two words does not matter, only the upper triangle is used (see Fig. 2b). Table 2 gives examples for the described feature types.

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", as these words have no innate differentiation power since they are part of any text document. Second, we calculate the explanatory power by using two known feature selection methods: *Chi-square* and *Bi-normal-separation*. These two methods have been chosen as they both have been found best-performing in [6]. In addition, both methods are structurally different in the way they incorporate exogenous market feedback. *Chi-square* compares the observed frequency O_i of the feature i within the set of positive messages with its expected frequency E_i, 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.

$$\chi^2 = \sum_{j=1}^{4} \frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}.$$

Utilizing this X^2-statistic each feature receives a value for higher or lower deviation from the expected. Words that usually influence investors' decisions carry higher values. Words having less influence on the decision, as they occur uniformly in positive and negative messages, 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.

The second feature selection method is called *Bi-normal separation* (*BNS*) measuring the separation between the prevalence of the features in the class of positive messages and the class of negative messages. Interestingly, it is rarely used in literature, but delivered superior results for higher number of features [8]. BNS is defined as

$$BNS = F^{-1}\left(\frac{O_{i,pos}}{pos}\right) - F^{-1}\left(\frac{O_{i,neg}}{neg}\right)$$

where F^{-1} is the standard normal distribution's inverse cumulative probability function (i.e., z-score) and *pos* being the number of

Table 2
Examples for feature types.

Feature type	Feature example
Single words	Record Loss
2-Gram	Increase dividend Net loss
2-Word combination	Guidance [...] upwards Expect [...] lower
Noun phrase	Ongoing positive result A difficult market environment

Table 3
Excerpt of stemmed feature list (2-word combinations).

No.	Pos. word	P-value	BNS-score	Neg. word	P-value	BNS-score
1	strong grow	0.046%	1.005	uncertainti fiscal	<0.001%	−1.070
2	serv increas	0.160%	0.940	forecast downward	<0.001%	−1.070
3	record ratio	0.160%	0.940	loss because	0.030%	−0.942
4	takeov becom	0.106%	0.964	insolven custom	0.145%	−0.720
5	licens cancer	1.3%	0.794	due difficulti	0.249%	−0.834

positive message and *neg* the number of negative messages. In correspondence to *Chi-square*, $O_{i,pos}$ ($O_{i,neg}$) denotes the observed frequency feature *i* within the set of positive (negative) messages. To avoid the undefined value $F^{-1}(0)$, zero is substituted by a very small number, i.e., 0.0005 [8]. The main structural difference of *BNS* is that it focuses only on the actually occurring words in a document as opposed to *Chi-square* which also includes the count of features not occurring in a message. Accordingly, the absence of a word in a message of a certain context is not attributed a special meaning. Intuitively, this ignorance of absent words could be an advantage as, for example, the absence of the word "share buyback" has no influence in a message signaling that a large contract has been won. 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 may embody a very negative meaning 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 surely part of the negative word list, but are assigned a positive meaning in our approach. The explanation is intuitive: Albeit cancer is a very serious disease, it also represents a fast-growing market segment for pharmaceutical companies. Table 3 shows exemplary 2-word combinations of our stemmed 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.

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 that a feature is best represented using TF IDF [23].
4. 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. For this purpose, the stock price reaction caused by the news event is transformed into a binary measure, i.e., '0' for negative price effect and '1' for positive. We use an SVM since previous findings confirm it to be the best available machine learning method for text classification tasks [8,12,32] and for financial prediction [10,11]. Further, in a pilot study, we compared the performance of artificial neural networks, naïve Bayes and SVMs and found SVMs to be best performing.
While relying on standard approaches for feature representation and machine learning, the main contribution of this paper is the combination of advanced feature types with a feedback-based feature selection. The results of the evaluation in the next chapter show the value-add of feature selection for different feature types.

4. Evaluation

In this section, we apply our text mining approach to a set of corporate disclosures. As we are interested in the impact of the feature selection on performance, we use two advanced feature selection methods, Bi-normal separation (BNS) and Chi-square (Chi^2), with different types of features. For comparison, we benchmark our approach by reproducing approaches in literature and applying them to the same data set.

4.1. Data set

We confront our methodology with real data consisting of two components: the textual news base and stock price effects as exogenous feedback. The textual news base comprises corporate announcements from Germany and the UK published between 1997 and 2011. 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 data set is an excellent choice for evaluating our text mining approach as the text messages and the stock price reactions have a tight logical connection due to regulatory requirements. Additionally, by focusing on material facts, the news set contains a pre-selection of relevant news from the set of all available financial news [20]. The data set features news that include facts on deviations of financial results from earlier expectations, management changes, M&A transactions, major orders and other types (Table 4).

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 10,870 corporate announcements from our first source DGAP eligible for our experiment. Thereof, we randomly (i.e., without temporal distinction) selected 50% of the messages for training of our machine learning method and the remainder for validation. In addition, we used the 3478 obtained news articles from our second source EuroAdhoc as an additional validation and test set in order to examine to what extent our approach can be generalized to confirm our results. The data set can be used for data triangulation as EuroAdhoc covers different companies and includes news from other countries such as the UK.

Based on the publication date and time of the corporate announcement, and the international securities identification number (ISIN) of the company that initiated the disclosure, we obtained the publicly available stock price information from Datastream.

The stock price analysis on the event day is based on daily *open* and *close* prices. For events during trading hours, the stock price effect is calculated between *open* and *close* auctions. For events occurring outside trading hours, the effect calculation is based on close prices of the previous day and open prices. One could argue that not using

Table 4
Content categories of textual news base.

Category	Rel. frequency	Absolute frequency
Financial reports	56.7%	8138
M&A	8.5%	1219
Change in management	4.8%	685
Capital increase	2.9%	417
Share buyback	2.3%	329
Major order	1.7%	244
Cooperation	1.4%	203
Product/manufacturing news	1.1%	162
Dividend	1.0%	138
Joint-venture	0.8%	118
Restructuring	0.8%	111
Other capital measure	0.8%	108
Litigations	0.6%	89
Shareholder structure	0.5%	74
Listing	0.4%	54
Other	15.7%	2259
Grand total	100%	14,348

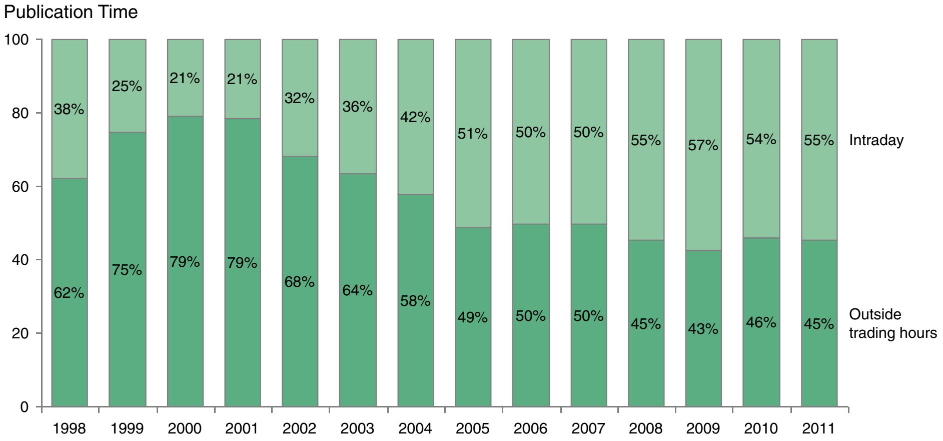


Fig. 3. Share of intraday news over time.

intraday stock prices for events occurring during trading hours, introduces potential inaccuracies. However, several reasons favor an approximation by *open* and *close* prices. First, *open* and *close* auctions have higher volumes and lead to more valid prices. Second, there is no definition of how long pricing of new information takes and assumptions need to be made [9]. Third, the high number of events is expected to balance out noise before and after the event. Based on the stock price effect, a binary measure of the sign is created to label all text messages as either 'positive' or 'negative'.

Companies historically published the larger share of news (~64%) outside trading hours. However, in recent years, this trend has shifted and the majority of news is now published intraday (~55%, see Fig. 3). For assessing classification performance, the full data set is used to allow for an as large as possible data set. Even for news outside trading hours, the stock price effect can be successfully captured by comparing open prices with close prices of the previous day. However, for the trading simulation in Section 5, we focus on intraday news as only those could actually be traded on the stock market.

Our data set is skewed and contains more positive than negative news (Table 5). A trivial majority classifier which would be always guessing 'positive' could reach 58.3% accuracy for the DGAP set and 53.3% for the EuroAdhoc set. These accuracies will form the benchmark for our evaluation.

4.2. Evaluation approach

By reproducing approaches in literature and applying to the same data set, we can reasonably benchmark our approach in a same-data comparison. Every feature extraction approach is conducted once with feature selection based on market feedback and once just based on a frequency-based selection, i.e., simply by requiring a minimum occurrence in the corpus per feature (as e.g., [3,24]). Thereby, we can demonstrate the improvements feasible by selecting features based on market feedback.

For exogenous-feedback-based feature selection, the Chi^2-approach and the bi-normal separation (BNS) are used to choose the most relevant features occurring in the message set. If no special feature selection is performed, only stopwords are removed and all features with a minimum occurrence of 15 are used for representation of text messages. Imposing a cut-off is essential for more complex features due to memory restrictions and to save computational resources. For our corpus, we obtained approximately2 million 2-word combinations making the use of all features impossible. As features are not actually selected, but rather reduced based on the frequencies, we will denote this step "Freq-based feature reduction" in the remainder. Table 6 shows the number of respective features used for the classification task. 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-Gram 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.

Table 5
Distribution of positive and negative messages.

Effect	Source: DGAP		EuroAdhoc
	Training set	Validation set	Validation set
Positive	57.1%	58.3%	53.3%
Negative	42.9%	41.7%	46.7%

4.3. Evaluation results based on prediction accuracies

Results were obtained by running the SVM with a linear kernel which delivered best performance for text classification tasks using a very high number of features [12]. Table 7 shows the classification results on both validation sets. Accuracy is measured as a percentage of correctly classified messages. For all five feature types, we performed training and validation, with each of our two market feedback based feature selection methods and once with frequency-based reduction only. Only for the dictionary approach (single word), did we not perform a feature selection as the dictionary constitutes an alternative feature selection method.

For measuring performance, many recent studies use the F1-measure: the harmonic average of precision (i.e., the percentage of messages classified as positive that actually are positive) and recall (i.e., the percentage of positives that are classified as positive). Depending on the application field, it may be beneficial to focus on precision (if costs of false positives are high, as e.g., filtering out a legitimate e-mail in a spam-filter [8]).

Table 6
Number of features employed.

Feature type	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection
Single words I: based on dictionary	3,106	–	–
Single words II: retrieved from corpus	2,463	1,158	1,259
2-Gram[a]	11,247	6,053	6,652
2-Word combinations	63,902	22,361	29,055
Noun phrases	6,249	2,018	2,817

[a] Performance of 3-Gram was slightly weaker than 2-Gram and is therefore not listed. 3-Gram suffer from a high number of combinations causing a rapid decrease in actual frequencies per feature.

Table 7
Classification accuracies for different feature types.

Feature type	Data I: DGAP			Data set II: EuroAdhoc		
	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection
Single words I: based on dictionary	62.1%	–	–	53.9%	–	–
Single words II: retrieved from corpus	62.0%	63.0%	62.9%	54.7%	54.4%	55.2%
2-Gram	58.0%	65.5%	65.7%	54.1%	56.3%	58.1%
2-Word combinations	62.0%	72.6%	76.3%	54.0%	60.6%	65.4%
Noun phrases	61.3%	63.1%	64.7%	54.7%	54.9%	57.0%
Benchmark: trivial majority classifier	58.2%				53.3%	

Table 8
Classification performance for different feature types measured by R^2.

Feature type	Data I: DGAP			Data set II: EuroAdhoc		
	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection
Single words I: based on dictionary	4.7%	–	–	0.1%	–	–
Single words II: retrieved from corpus	4.7%	4.7%	5.9%	0.4%	1.0%	0.4%
2-Gram[a]	1.3%	8.5%	4.4%	0.3%	2.7%	2.5%
2-Word combinations	4.9%	15.3%	20.2%	0.9%	7.4%	9.4%
Noun phrases	3.8%	6.2%	4.6%	0.4%	0.3%	2.1%

[a] Performance of 3-Gram was slightly weaker than 2-Gram and is therefore not listed. 3-Gram suffer from a high number of combinations causing a rapid decrease in actual frequencies per feature.

However, in our case, a trading engine based on signals of our news classification would face equal misqualification costs, i.e., the cost of getting a positive or a negative message wrong are equal. Thus, we focus on the classification accuracy (number of correctly classified messages divided by the total number of messages) as main performance measure.[2]

In a second performance measure, we also assess the ability to predict the discrete value of the stock return. We use support vector regression (SVR) to predict returns and calculate the R^2 (squared correlation coefficient) between predicted and actually observed return. The optimization behind the SVR is very similar to the SVM, but instead of a binary measure (i.e., positive or negative), it is trained on actually observed returns. While a binary measure can only be 'true' or 'false', this measure gives more weight to greater deviations between actual and predicted returns than to smaller ones. As profits or losses are higher with greater deviations, this measure better captures actual trading returns to be realized.

Table 7 lists classification accuracies for different feature types and feature selection methods. Analysis is performed separately for both validation sets from DGAP and EuroAdhoc. Table 8 lists classification performance based on R^2 between predicted return by the SVM and the actually observed return on the stock market.

In the following, we present our findings along our research questions.

Finding 1: Chi^2-based and BNS-based feature selection improved classification accuracies for all feature types

Results show that all feature types benefited from the Chi^2-based and BNS-based feature selection,[3] through an improved accuracy and R^2 for all validation experiments. The highest performance on the first validation set (from DGAP) with an accuracy of 76.3% and R^2 = 20.2% was achieved for the 2-word combination with BNS-based feature selection. The 2-word combination benefited most from feature selection, single words least. This observation extends the findings of Joachims [12] 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.

BNS-based results are mostly stronger than Chi^2-based results for all feature types. Exceptions are found for 2-Gram and single words where differences in performance are minimal. The findings are confirmed by the second validation set. Again feature selection increases accuracies for more complex feature types. Classification accuracies are generally lower as training was performed on a different data set containing e.g., different companies, different mix of message content categories, and a different ratio of positive and negative messages. Highest classification accuracies were achieved again for 2-word combinations using BNS-based feature selection with an accuracy of 65.4% and R^2 = 9.4%. For this set, BNS-based feature selection outperforms the Chi^2-based version in all cases. The benchmark for all experiments is the trivial majority classifier which always guesses 'positive' as there are more positive than negative messages. Thus, the actual guessing probability is 58.2% for the data set from DGAP and 53.3% for EuroAdhoc. Without feature selection, this benchmark is not substantially exceeded.

Finding 2: Classification accuracy increases with complexity of features when Chi^2-based or BNS-based feature selection is used

Classification performance increases with complexity and expressiveness of features — expressiveness meaning the ability of features to capture and express sentiment and explanatory power. This is consistent with the findings of a previous study [24] 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 the validation set is rather similar for all feature types in both validation sets, i.e., ~62% accuracy for DGAP set and ~54–55% for EuroAdhoc set. Features seem to develop their expressiveness only after selecting the most relevant features and, thus, taking out the noise. 2-Word combinations without feature selection exhibit an even lower performance than single words as they suffer

[2] We list precisions, recalls and F1-measures in detail in Appendix A.

[3] The only exception is single words II with Chi^2-based feature selection on the EuroAdhoc data set.

Table 9
Classification accuracies for different feature types on training set.

Feature type	Data I: DGAP		
	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection
Single words I: based on dictionary	63.7%	–	–
Single words II: retrieved from corpus	67.2%	67.0%	66.0%
2-Gram	80.1%	70.1%	69.3%
2-Word combinations	95.6%	88.0%	92.2%
Noun phrases	79.1%	65.1%	76.6%

from a very high number of random combinations with low expressiveness. The dictionary (single words I) shows a low performance on the first validation set from DGAP (62.3%) and lowest performance on the second validation set from EuroAdhoc (53.9% accuracy, $R^2 = 0.1\%$). Single words retrieved from corpus only perform minimally better than the dictionary. When using feature selection, more complex features, i.e., built of more than one word, performed better than single words. Best performance was achieved for 2-word combinations with BNS, outperforming noun phrases and 2-Gram. 2-Word combinations may carry more expressiveness than 2-Gram as the two are not required to be subsequent and therefore offer a greater flexibility to capture semantics of a sentence. Noun phrases may include more than two words and partially capture semantics in a sentence. However, in contrast to 2-word combinations and 2-Gram, noun phrases lack verbs and adverbs limiting their expressiveness. When combined with feedback-based feature selection, performance of noun phrases is close to, but still below performance of 2-Gram.

After witnessing the performance increase by exogenous feedback based feature selection, it is of interest to know what caused the performance increase. We therefore investigate the accuracies achieved on the training set leading us to our third finding.

Finding 3: Using Chi^2-based and BNS-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-Gram, 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). In particular, when a larger number of features is used for learning than there are messages in the training set, the risk of over-fitting increases [5]. 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 further 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 slightly reduces accuracy values in the training set. However, this could be attributed to the pure reduction in the number of features (see Table 9).

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 between approaches 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, we primarily focus on the other two factors.

Bag-of-words and 2-Gram 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 [13]. A full parsing of our textual data set took one full day on our system. Thus, noun phrases come at a very high cost despite lower validation accuracies than 2-word combinations.

4.4. Discussion of feature types

As results from the previous section demonstrated superior performance for 2-word combinations, we want to shed light on why 2-word combinations are more expressive and why they better capture the context when used for financial text representation. Thus, we will compare each feature type to 2-word combinations and give examples of why 2-word combinations were better able to capture the context and meaning.

For single words, it is obvious that one word alone offers potential for confusion and is very inaccurate in expressing meaning. For example, the word "*good*" occurred 599 times in positive messages and 344 times in negative messages. Chi^2 attested some discriminating power with a p-value at 4.3%. Accordingly, the feature was included in the feature list. However, in 344 cases the word "good" was mentioned in negative messages with the potential for confusion as in "… *results have not been as*

Table 10
Examples for 2-word combinations better capturing the context than other feature types.

Type	Feature	Positive occurrences	Negative occurrences	p-Value	Resp. 2-word combination	Positive occurrences	Negative occurrences	p-Value
Single word	good	599	344	4.30%	highlight good	7	0	0.05%
					need good	2	10	0.52%
Single word	dividend	659	450	0.20%	lower dividend	0	7	0.25%
2-Gram	sharp drop	1	14	0.00%	sharp drop	4	19	0.01%
2-Gram	drop(ped) sharp(ly)	3	5	21.37%				
Single word	short	283	274	0.00%	forecast short	1	14	0.01%
Single word	forecast	1118	761	70.9%				

Table 11
Share of messages containing a minimum number of features.

Minimum # of features	0	1	2	3	4	5	6	7	8
% of messages	100%	86%	78%	70%	63%	56%	49%	43%	38%

good as …". Although the word "*not*" will be a feature as well, the SVM has no information that "*not*" is actually corresponding to "*good*". "*Not*" might just have been part of another sentence in the text.

Similarly, "dividend" has a positive tone (mentioned 679 times in positive and 450 times in negative messages at p-value of 0.2%). However, "lower dividend" for 2-word combinations was never mentioned in positive, but 7 times mentioned in negative messages. The examples illustrate that 2-word combinations were able to capture the meaning and context of a statement while single words were misleading (Table 10).

In contrast to 2-Gram, 2-word combinations may allow words in between and can therefore capture a higher variety of meanings. Thus, statements like "*a slump in demand*" or "… *falls short of its earnings forecast originally communicated* …" can only be captured by 2-word combinations. Further, 2-Gram do not capture order of words: while "sharp drop" vs. "dropped sharply" is the same feature for 2-word combinations after stemming, 2-Gram will capture them as two different features despite equivalent meaning.

The limited flexibility also limits performance of Noun phrases. Many 2-word combinations are not feasible for noun phrases (e.g., "dropped sharply") due to lack of verbs and adverbs and the fact that all words of a noun phrase are subsequent. This is also expressed by the number of discriminating features. Only 133 noun phrases receive a p-value of less than 0.1% while 2-word combinations had 1673 features fulfilling this criterion. With a lower number of discriminating features, it is more difficult to represent a news message in an expressive way.

Finally, dictionaries are not tailored to content and often are based on psychosocial analysis. Thus, the word set is limited and cannot capture all specifics and subject lingo of the underlying domain. Financial dictionaries may add value by adding meaningful and subject relevant terms. However, while some financial terms like "loss" and "insolvency" have a clear and negative meaning, other financial terms like "revenue" and "earnings" need more words to actually describe a good or bad fact.

Despite 3106 words in the dictionary, which is far more than the number of features for single words retrieved from the corpus, not all messages contain a significant number of words in the dictionary. Table 11 lists the share of messages containing at least a certain minimum number of words. For 14% of messages, decision is based on pure guessing as no words from the dictionary occur in these messages at all. For 22% of messages, decision is based on one word or less. As seen from the analysis of single words in this section, a random classification outcome is likely. If we require at least 5 words for a fact-based decision, only 56% of messages are meaningfully classified. Thus, it is not astonishing that the trivial guessing benchmark is hardly beaten.

Table 12
Prediction accuracy by DGAP news content category (ordered by last column).

Category	Count	Accuracy freq-based feature reduction	Accuracy Chi^2-based feature selection	Accuracy BNS-based feature selection
Product/manufacturing news	44	79.5%	88.6%	88.6%
Restructuring	57	61.4%	78.9%	82.5%
Financial reports	3610	63.5%	74.8%	78.7%
Major order	98	71.4%	74.5%	78.6%
Cooperation	91	71.4%	76.9%	76.9%
Joint-venture	49	61.2%	67.3%	75.5%
Share buyback	121	67.8%	71.9%	75.2%
Capital increase	214	56.1%	68.2%	74.8%
Shareholder structure	38	68.4%	78.9%	73.7%
Lawsuit	32	56.3%	68.8%	71.9%
M&A	523	57.9%	68.6%	70.6%
Dividend	52	55.8%	69.2%	69.2%
Other	181	49.7%	63.0%	65.7%
Change in management	292	53.1%	60.6%	65.4%
Other capital measure	32	50.0%	50.0%	59.4%
Total	**5434**	**62.0**%	**72.6**%	**76.3**%

4.5. Content-based evaluation results

When assessing performance, it is also important to look at the underlying message content. On the one hand, with a maximum 76.3% accuracy, a very good performance was achieved, but on the other hand, there is still a substantial gap towards a perfect 100%. This gap can be partially explained by clarity and context sensitivity of the message itself. Financial market participants have expectations about firms. Each incoming news message is compared against previous expectations. Thus, a very positive financial result of a firm might still result in a negative stock market effect when it does not meet expectations of market participants. For some news, even a human expert will have difficulties to classify when not having access to the context of the news. The context sensitivity naturally depends on the content of the news, e.g., financial reports and mergers always have to be seen in context. While it might be easier to assess if a company's quarterly financial reporting is good or bad news just from looking at the text (i.e., financial reports compare figures to previous years and assess good or bad performance in written text), it will be more difficult to assess if the announcement of a Change in Management is positive or negative. It may require extensive knowledge of the company's environment which may not be captured by the text of the message. To illustrate context sensitivity for different news messages, news messages in the validation set have been classified into content-dependent categories based on labels and tags which news emitting companies supplied with the news.

Table 12 illustrates the prediction accuracy broken down by news categories. The 2-word combination experiment with different feature selection methods served as base for the breakdown. It was picked since it provided the highest accuracy values in our evaluation. High accuracies were achieved for news about products, manufacturing, restructuring and financial reports, and a low accuracy for the category "Other capital measure". Intuitively, the categories rank as expected. News about restructuring and financial reports carry more relevant information within the news message while diffuse categories as "Other" and "Other capital measures" or "changes in management" are difficult to assess. Interestingly, different categories rank similarly for all feature selection methods. This underlines that independent of the underlying text mining process different news categories are of a different difficulty to classify.

5. Implications: simulation-based evaluation results

After analyzing classification performance in detail, this section validates if this approach can be applied in practice and how different feature types perform in comparison. 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). The positions are held until the end of the trading day. To ensure that simulated returns can be realized in practice, we only select messages published during trading hours. Further, we focus on the top 110 most liquid stocks (as in HDAX composed by Deutsche Börse AG) to reduce liquidity restrictions and better approximate actual returns. Thus, from all messages in the validation set, only ~1610 from DGAP and ~1340 from EuroAdhoc are used for this experiment. Table 13 describes the average returns achieved with the described trading strategy. For each average return, also the 99.9% confidence interval is achieved assuming that a normal distribution is given.

Similar to Table 7, feature selection has a clear benefit. Also return performance of different feature types ranks in the same order as classification accuracies in Table 7. However, differences arise in the magnitude of the return: Slight percentage changes in accuracy already lead to a

Table 13
Average return and confidence intervals (at α = 0.1%) per trade dependent on feature selection.

Feature type	Data I: DGAP			Data set II: EuroAdhoc		
	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection
Single words I: based on dictionary	**0.6**% (0.3%; 1.1%)	–	–	**0.3**% (−0.1%; 0.7%)	–	–
Single words II: retrieved from corpus	**0.7**% (0.4%; 1.2%)	**0.8**% (0.4%; 1.2%)	**0.8**% (0.4%; 1.2%)	**0.4**% (0.0%; 0.8%)	**0.5**% (0.1%; 0.9%)	**0.5**% (0.1%; 0.9%)
2-Gram	**0.4**% (0.2%; 1.0%)	**0.8**% (0.4%; 1.2%)	**0.7**% (0.3%; 1.1%)	**0.5**% (0.0%; 0.8%)	**0.5**% (0.1%; 0.9%)	**0.7**% (0.3%; 1.1%)
2-Word combinations	**0.8**% (0.4%; 1.2%)	**1.5**% (1.1%; 1.9%)	**1.8**% (1.4%; 2.1%)	**0.4**% (0.0%; 0.8%)	**0.9**% (0.5%; 1.3%)	**1.0**% (0.5%; 1.4%)
Noun phrases	**0.7**% (0.3%; 1.1%)	**0.7**% (0.3%; 1.1%)	**0.7**% (0.3%; 1.1%)	**0.3**% (−0.1%; 0.7%)	**0.4**% (0.0%; 0.8%)	**0.6**% (0.2%; 1.0%)
Benchmark: trivial majority classifier		**0.0**% (−0.4%; 0.4%)			**0.2**% (−0.2%; 0.7%)	

strong increase in return profits. The 2-word combinations again show highest performance. As each message has a stock return of different magnitude, the computed average of stock returns are not fully monotonically related to classification accuracies. For noun phrases, the feature selection only causes an improvement on the second data set. The table shows raw returns before transaction costs and liquidity restrictions such as spread and order impact. However, if transaction costs of 0.1% are assumed [28], returns still remain positive. Additionally, liquidity restrictions are reduced by only considering the top 110 most liquid stocks. Still, it will be difficult to realize profits if returns are only at 0.2% or 0.3%. This becomes even more evident if statistical confidence of results is considered: The confidence intervals demonstrate the large variability in returns, thus, always requiring a higher number of trades to allow for statistical stable results. It is evident that using single words and not employing feature selection put profits at risk. However, 2-word combinations are very likely to produce significant profits, as the 99.9%-confidence intervals are (1.4%; 2.1%) for DGAP and (0.5%; 1.4%) for EuroAdhoc and are therefore still positive after transaction costs and liquidity restrictions are factored in.

The analysis of implied stock returns lacked the consideration of liquidity restrictions which can be substantial for less liquid stocks. To further validate the applicability in practice of our approach, a second backtesting in the form of a trading simulation was run. We use trading returns from our classification run with 2-word combinations and BNS-based feature selection. In the beginning of the simulation, we assume to have 100,000€ in cash. For each trading signal generated, we invest 25,000€ and calculate the profit or loss we would have incurred based on our backtesting stock price information. We again assume a transaction cost of 0.1%, but now also include spread and order impact for each stock as estimated by the German Stock Exchange [26] for a 25,000€ order volume. Resulting realized returns for the full portfolio with this strategy are shown by Table 14. Performance is compared to direct investment into the index HDAX which comprises all stocks which have been in focus for trading.

Table 14
Trading return for full portfolio.

Year	Portfolio value	Simulation	Benchmark: HDAX
2000	100,000	−1.2%	1.1%
2001	86,229	−13.8%	−22.4%
2002	118,644	37.6%	−45.0%
2003	129,955	9.5%	49.6%
2004	157,420	21.1%	5.7%
2005	167,403	6.3%	34.2%
2006	196,168	17.2%	19.9%
2007	213,327	8.7%	−0.7%
2008	251,032	17.7%	−37.9%
2009	276,191	10.0%	31.2%
2010	310,403	12.4%	26.6%
2011	332,982	7.3%	−17.3%
Total (p.a.)		**10.9**%	**1.0**%

Although the burst of the dot-com bubble would have lead to losses in the overall portfolio of more than 13% (with the HDAX losing more than 22%), the approach showed stable performance in the following years. As it trades on an event-by-event basis, it is less dependent on the overall market development. Moreover, it is beating the overall performance of the direct investment into HDAX over the total 11 years despite its conservative setup. Based on the trading simulation we can state our fourth finding:

Finding 4: A profitable trading strategy can be established based on the signals generated by our approach even after considering trading commissions and liquidity restrictions
Thereby, more complex feature types and the 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. Despite the consideration of liquidity (i.e., order impact) it has to be noted, that trading returns of up to 11% p.a. cannot be scaled up to millions of euros. The simulation was carried out with an investment volume of 25,000€ and is precise for that amount. For larger investment volumes, order impact increases reducing profit from each trade.

6. 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 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. When feedback-based feature selection is combined with 2-word combinations, accuracies of up to 76% are achieved. These results were possible as 2-word combinations capture the meaning and context of information pieces in text.

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.

Analysis of the content of the messages indicates that stock price prediction based on news has limitations well below 100% accuracy as stock price effects on capital markets also depend on information not captured by a single financial news message.

To apply our approach in practice, we simulate a simple, but rewarding trading strategy to demonstrate achievable returns. Thereby, using 2-word combinations and BNS-based feature selection leads to the highest returns in the field at a low statistical variance. We do not only simulate average returns, but also a full investment portfolio over 11 years considering actual investment volumes, trading commissions and order impacts. Portfolio simulations show that profitability of our trading strategy is fully competitive compared to absolute return funds and a direct investment into the respective stock index. However, as the trading strategy is based on a large number of orders each subject to liquidity constraints, the approach cannot be scaled up to infinite investment volumes.

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 (like e.g., the stock price reaction in our case) must be available which directly corresponds to the text message. Difficulties arise when 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. Examples for application areas outside capital market research, which fulfill these criteria, can be found in customer relationship management where email communication forms the textual message base and the subsequent consumer action forms the corresponding feedback [6]. Other areas include marketing, security and content handling.

Appendix A

Table 15 detailed listing of precision (P), recall (R), and F1-measure (F1) for results of Table 7.

Feature type	Data I: DGAP			Data set II: EuroAdhoc		
	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection	Freq-based feature reduction	Chi^2-based feature selection	BNS-based feature selection
Single words I: based on dictionary	P: 60.9% R: 89.4% F1: 72.5%	–	–	P: 53.9% R: 94.3% F1: 68.6%	–	–
Single words II: Retrieved from corpus	P: 61.2% R: 87.1% F1: 71.9%	P: 61.6% R: 89.1% F1: 72.9%	P: 61.6% R: 88.8% F1: 72.7%	P: 54.3% R: 91.0% F1: 68.0%	P: 54.4% R: 94.3% F1: 69.0%	P: 54.6% R: 94.4% F1: 69.2%
2-Gram[a]	P: 59.4% R: 80.4% F1: 68.3%	P: 63.0% R: 91.7% F1: 74.7%	P: 63.5% R: 90.1% F1: 74.5%	P: 54.2% R: 89.4% F1: 67.5%	P: 55.2% R: 95.2% F1: 69.9%	P: 56.3% R: 95.5% F1: 70.8%
2-Word combinations	P: 63.0% R: 79.2%% F1: 70.2%%	P: 70.0% R: 89.9% F1: 78.7%	P: 71.8% R: 93.1% F1: 81.1%	P: 54.3% R: 86.4% F1: 66.7%	P: 58.5% R: 89.5% F1: 70.8%	P: 61.2% R: 96.1% F1: 74.8%
Noun phrases	P: 61.7% R: 84.4% F1: 71.3%	P: 61.5% R: 90.4% F1: 73.2%	P: 63.2% R: 87.5% F1: 73.4%	P: 54.1% R: 87.9% F1: 67.0%	P: 54.6% R: 92.1% F1: 68.5%	P: 55.8% R: 92.4% F1: 69.6%

[a]Performance of 3-Gram was slightly weaker than 2-Gram and is therefore not listed. 3-Gram suffer from a high number of combinations causing a rapid decrease in actual frequencies per feature.

References

[1] W. Antweiler, M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, Journal of Finance 59 (3) (2004) 1259–1294.

[2] C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121–167.

[3] M. Butler, V. Keselj, Financial forecasting using character N-Gram analysis and readability scores of annual reports, Advances in AI, 2009.

[4] J.Y. Campbell, R.J. Shiller, Cointegration and tests of present value models, Journal of Political Economy 95 (1987) 1062–1088.

[5] G.C. Cawley, N.L. Talbot, Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters, Journal of Machine Learning Research 8 (2007) 841–861.

[6] K. Coussement, D. Van den Poel, Improving customer complaint management by automatic email classification using linguistic style features as predictors, Decision Support Systems 44 (2008) 870–882.

[7] S.R. Das, M.Y. Chen, Yahoo! for Amazon: sentiment extraction from small talk on the web, Management Science 53 (9) (September 2007) 1375–1388.

[8] G. Forman, An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research 3 (2003) 1289–1305.

[9] G. Gidofalvi, C. Elkan, Using News Articles to Predict Stock Price Movements, Technical Report — Department of Computer Science and Engineering, University of California, San Diego, 2003.

[10] S.S. Groth, J. Muntermann, Supporting investment management processes with machine learning techniques, in: H.R. Hansen, D. Karagiannis, H.-G. Fill (Eds.), Proceedings of the 9. Internationale Tagung Wirtschaftsinformatik, Österreichische Computer Gesellschaft, Wien, Austria, 2009.

[11] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decision Support Systems 50 (2011) 680–691.

[12] T. Joachims, Text categorization with support vector machines: learning with many relevant features, Proceedings of the European Conference on Machine Learning, Springer-Verlag, 1998.

[13] D. Klein, C.D. Manning, Accurate unlexicalized parsing, Proceedings of the 41st Meeting of the Association for Computational Linguistics, 2003, pp. 423–430.

[14] D. Leinweber, J. Sisk, Event driven trading and the "new news", Journal of Portfolio Management 38 (1) (2011) 110–124.

[15] F. Li, Textual analysis of corporate disclosures: a survey of the literature, Journal of Accounting Literature 29 (2010) 143–165.

[16] F. Li, The information content of forward-looking statements in corporate filings — a naïve Bayesian machine learning approach, Journal of Accounting Research 48 (5) (2010) 49–102.

[17] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, Journal of Finance 66 (2011) 35–65.

[18] C. MacKinlay, Event studies in economics and finance, Journal of Economic Literature (1997) 13–39.

[19] M.-A. Mittermayr, Forecasting intraday stock price trends with text mining techniques, Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.

[20] J. Muntermann, A. Guettler, Intraday stock price effects of ad hoc disclosures: the German case, Journal of International Financial Markets, Institutions and Money 17 (1) (2007) 1–24.

[21] M.F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[22] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 3rd Edition Prentice Hall, 2009.

[23] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[24] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACM Transactions on Information Systems 27 (2) (2009).

[25] R.P. Schumaker, Y. Zhang, C. Huang, H. Chen, Evaluating sentiment in financial news articles, Decision Support Systems 53 (3) (2012) 458–464.
[26] S. Stange, C. Kaserer, The impact of order size on stock liquidity — a representative study, CEFS Working Paper No. 2008–9, 2008. (Available at SSRN: http://ssrn.com/abstract=1292304).
[27] P.C. Tetlock, Giving content to investor sentiment: the role of media in the stock market, Journal of Finance 62 (2007) 1139–1168.
[28] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More Than Words: Quantifying Language to Measure Firms' Fundamentals, 63, 2008, pp. 1437–1468.
[29] P.C. Tetlock, All the news that's fit to reprint: do investors react to stale information? The Review of Financial Studies 24 (5) (2011) 1481–1512.
[30] B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, Daily stock market forecast from textual web data, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, 1998.
[31] Y. Yang, J. Pedersen, A comparative study on feature selection in text categorization, International Conference on Machine Learning (ICML), 1997.
[32] Y. Yang, X. Liu, A re-examination of text categorization methods, Proceedings of the 22nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.

Michael Hagenau holds a PhD from University of Freiburg, a diploma in Business Engineering from Karlsruhe Institute of Technology (KIT) and a Master of Science in Computer Science from Georgia Institute of Technology. He is currently a Graduate Researcher at University of Freiburg. Prior to his academic career, he worked for a management consulting firm. His research interests are focused on decision support systems and text mining in financial news. He has (co-)authored research publications at ICIS, ECIS and HICSS.

Michael Liebmann holds a PhD from University of Freiburg, a diploma in Business Engineering from Karlsruhe Institute of Technology (KIT) and a Master of Science in Industrial and Systems Engineering from Georgia Institute of Technology. He is currently a Graduate Researcher at University of Freiburg. Prior to his academic career, he worked for a management consulting firm. His research interests are focused on decision support systems and text mining infinancial news. He has (co-)authored research publications at ICIS and HICSS.

Dirk Neumann is Full Professor with the Chair of Information Systems of the University of Freiburg, Germany. His research topics include Business Analytics, Text Mining and Cloud Computing. He studied information systems in Giessen (Diploma), Economics in Milwaukee, WI, USA (Master) and received a PhD from Karlsruhe Institute of Technology (KIT) in 2004. He has (co-)authored many research publications at European Journal of Operational Research, ACM Transactions on Internet Technology, or Journal of Autonomous Agents and Multi-agent Systems.