6-15-2020

TOWARDS AUTOMATED ANALYSIS OF FINANCIAL ANALYST COMMUNICATION: THE INDUCTION OF A DOMAIN-SPECIFIC SENTIMENT DICTIONARY

Matthias Palmer
University of Goettingen, matthias.palmer@uni-goettingen.de

Jan Roeder
Chair of Electronic Finance and Digital Markets, jan.roeder@uni-goettingen.de

Jan Muntermann
University of Goettingen, muntermann@wiwi.uni-goettingen.de

TOWARDS AUTOMATED ANALYSIS OF FINANCIAL ANALYST COMMUNICATION: THE INDUCTION OF A DOMAIN-SPECIFIC SENTIMENT DICTIONARY

Research paper

Palmer, Matthias, University of Goettingen, Goettingen, Germany, matthias.palmer@uni-goettingen.de

Roeder, Jan, University of Goettingen, Goettingen, Germany, jan.roeder@uni-goettingen.de

Muntermann, Jan, University of Goettingen, Goettingen, Germany, muntermann@wiwi.uni-goettingen.de

Abstract

The flexibility of general-purpose sentiment dictionaries has led to their extensive application in many different research fields. While these sentiment dictionaries are easy to apply, they are typically inferior to text classification approaches based on machine learning or compared to domain-specific dictionaries. Nevertheless, both approaches generally come along with additional manual data analysis. We address this problem by extending a domain-specific sentiment dictionary utilizing regularized linear models. We induce a dictionary extension that is trained on an extensive dataset and therefore particularly fitted for its specific purpose but can also straightforwardly be applied due to easy-to-use polarity word lists in both research and industry use cases. We develop this dictionary extension based on nearly 15,000 reports from financial analysts and demonstrate that the dictionary measures the sentiment of financial analysts more accurately than other finance-specific (but more general-purpose) dictionaries. We thus contribute to an improved analysis of the sentiments of financial analysts, who are the subject of many research projects as well as highly respected financial experts. Further, we show that our approach realizes context specificity while avoiding extensive manual data analysis.

Keywords: Sentiment Analysis, Dictionary Induction, Financial Analysts, Analyst Reports.

1 Introduction

In the past, research has already demonstrated that qualitative data can provide additional information value over pure financial metrics (Henry, 2006; Tetlock, 2007; Tetlock et al., 2008). In their literature review on natural language-based financial forecasting, Xing et al. (2018) argue that more domain-specific resources should be made available and specifically mention word lists. Cambria et al. (2017) identify sentiment analysis as a multi-layered field of application that requires differentiated consideration. To improve the determination of text sentiment values in the context of sentiment analysis, it is typically superior to use specific word lists (i.e., domain-specific sentiment dictionaries) that are tailored to a particular topic or profession (Loughran and McDonald, 2015). Nonetheless, general-purpose dictionaries are still used in many studies. Oftentimes there is no domain-specific dictionary readily available in the specific research context and its construction typically requires extensive efforts with regards to manual data analysis (Loughran and McDonald, 2016). These sentiment dictionaries contain polarity words, i.e., words indicating a positive or negative opinion. While general-purpose dictionaries such as the General Inquirer (Stone et al., 1966) can be applied to many different contexts, single words from the dictionaries may not be adequately assigned to the positive or negative category in specific cases. For example, when analyzing a politician's speech in terms of the underlying sentiment, words can have a different meaning compared to analyzing a press release of a company. Furthermore, to analyze texts relating to an individual company, it may be relevant to consider the author(s) of a document, e.g., the press department of a company, managers, journalists, private individuals or financial analysts. Due to the large number of documents that are automatically analyzed today, especially in the financial sector, it is important to also consider author-related information in the context of sentiment analysis. Moreover, the preparation of texts for sentiment analysis legitimately plays an important role, as it is difficult to carry out sentiment analyses without well-pre-processed texts. But the best preparation of the text is useless if the word lists of the sentiment dictionaries are not appropriate for a specific domain. This is particularly troublesome when unsuitable word lists are used on a large scale and in an automated fashion. In our view, all the above-mentioned points lead to the conclusion that it is sensible to further the development of domain-specific dictionaries.

The comprehensive sentiment-related literature reviews of Kearney and Liu (2014) and Loughran and McDonald (2016) argue for an intensified domain-specific development of sentiment dictionaries. Twedt and Rees (2012) measure sentiment in analyst reports and point out that analysts play an important role in the interpretation of company-related data. Accordingly, Huang et al. (2017) show that analyst reports offer additional information to earnings-related conference calls by applying topic mining. Financial analysts discuss topics from conference calls in analyst reports in 61 % of the cases, but in 31 % new topics are discussed that receive little or no attention in the previously conducted conference calls (Huang et al., 2017). This implies that analysts provide new information that should be used and analyzed in terms of the analyst's sentiment towards this information. Moreover, the value of conference call interpretations by financial analysts is higher when analysts use their analyst-specific language (Huang et al., 2017). In general, Huang et al. (2014) can show that texts of analyst reports help to interpret quantitative capital market data and qualitative company disclosures. For positive quantitative indicators, the market reacts strongly, especially when the sentiment values of the analyst reports are particularly positive (Huang et al., 2017).

To automatically determine the sentiment of texts, a fundamental distinction can be made between dictionary-based and machine learning-based approaches (Liu, 2012). The first approach is based on statistical models that have been trained using large amounts of data. The latter approach uses fixed word lists to determine sentiment. From a technical perspective, Huang et al. (2014) achieve particularly high accuracies for sentiment classification by utilizing machine learning methods and recommend further use of these methods. In contrast, Loughran and McDonald (2016) emphasize the drawbacks of black-box algorithms, as the arising inaccuracies and opaqueness may overshadow their added value. By comparing machine learning and dictionary-based sentiment measures, Henry and Leone (2015) observe only slight differences between the results and therefore favor the dictionary-based approaches, which they find easier to interpret and apply.

We see a disadvantage in models that cannot be used repeatedly without a sufficiently large dataset being available for model training. A reasonable approach in getting in between might be the development of a sentiment dictionary with the help of machine learning. Pröllochs et al. (2015) are among the first to present an approach for automatic development of domain-specific sentiment dictionaries based on Bayesian learning. They suggest an approach that constructs a new sentiment dictionary by identifying an appropriate list of words. Here, previously generated and more general-purpose dictionaries are disregarded, even though they were found to be useful in different application domains. Against this background, we aim to explore if this knowledge can be exploited by extending general-purpose sentiment dictionaries by adding domain-specific words identified in a semi-automated fashion. Therefore, we ask the following research question *RQ 1: How to effectively develop finance-specific sentiment dictionaries by extending more general-purpose dictionaries?*

Regarding our research context, we feel encouraged in our idea, since Henry and Leone (2015) explicitly advocate the induction of a sentiment dictionary for analyst reports. Consequently, we raise the following research question *RQ 2: Does an automatically generated domain-specific sentiment dictionary provide a superior assessment of document sentiments compared to more generic dictionaries in the context of analyst communication?*

The paper is structured as follows. In section 2, we set forth the basics of domain-specific sentiment analysis and present a brief introduction to financial analysts. In section 3, we describe the process of dictionary development and induce the domain-specific dictionary for financial analysts. We close with a discussion of the results in section 4 and a conclusion and future research opportunities in section 5.

2 Related Literature

2.1 Sentiment Analysis

Sentiment analysis measures the emotional tendencies of a text. It can be determined whether a sentence is objective or subjective and whether the text expresses a positive, neutral or negative sentiment. Aspect-based sentiment analysis can be used to show, for example, on which products or topics opinions are expressed. As a subcategory of Natural Language Processing (NLP), sentiment analysis allows large quantities of unstructured texts to be analyzed automatically. Sentiment analysis is an intuitive task, but the more the methods of sentiment analysis are adapted to the topic of the text or the background of the author, the more complex they become. For example, the word meanings in different domains can be quite different. Literature reviews on textual analysis by mentioning different fields of application and methods provide Li (2010b), Kearney and Liu (2014), and Das (2014). Sentiment analysis can be carried out using two different analytical approaches: machine learning-based and dictionary-based sentiment analysis (Kearney and Liu, 2014).

Machine learning-based approaches use supervised learning techniques that require labeled training and test datasets. Different models, e.g., linear, rule-based or probabilistic models, are trained and subsequently tested based on the dataset. The labeling of the data, e.g., sentences, paragraphs or documents, can be carried out manually. Antweiler and Frank (2004) use this procedure to evaluate Internet bulletin board messages, Das and Chen (2007) analyze stock message board postings, and Li (2010a) measures the sentiment in management discussions and analysis disclosures. Huang et al. (2014) label sentences in analyst reports and train a classifier with Naïve Bayes. However, there is a risk of a poor data basis due to incorrect labeling. Also, the sentiment classification particularly depends on the training dataset. Jegadeesh and Wu (2013) train a sentiment classifier based on stock market reactions to 10-K reports and thus avoid the subjective labeling of sentences. In some cases, regression analysis or neural networks might extract features that are just proxies for other measures, e.g., they just approximate dummies for industries (Loughran and McDonald, 2016), if the model does not control for such factors.

Dictionary-based sentiment analysis is based on word lists. Utilizing dictionaries compared to more advanced model-based learning approaches is appealing since word lists are easy to handle. The words of the different categories (mostly positive and negative) are counted in the text and a sentiment score

is calculated from the ratio of the polarity words. Word lists can also be used for aspects such as uncertainty or objectivity. However, this is rarely practiced in finance research (Loughran and McDonald, 2016). Sentiment dictionaries can easily be applied to different datasets and do not need to be re-trained as they are fitted classifiers. Applying the same word list and the same data preparation steps will have the same result. There are relatively general dictionaries, which were developed for social sciences, such as the General Inquirer (Stone et al., 1966) and DICTION (Hart, 2000). The creation process of those often contains deductive reasoning, while also inductive components like statistical word occurrences are used. Furthermore, specific dictionaries for product reviews (Hu and Liu, 2004) or the analysis of microblogging data (Oliveira et al., 2017) exist. Henry (2008) developed a dictionary for earnings announcements with 105 positive and 85 negative words using a Thesaurus-based approach. A disadvantage of this dictionary is the non-exhaustive listing of negative words as outlined by Loughran and McDonald (2011). Hence, Loughran and McDonald (2011) conceived a dictionary using word counts of 10-K filings. It contains 354 positive and 2,329 negative words. Both dictionaries are commonly used in financial research and practice. In the following, we refer to the two dictionaries as the Henry and the LM dictionary. It has been shown that domain-specific dictionaries work better in their designated domain. For instance, almost three-fourths of the negative words from the non-domain-specific General Inquirer cannot be assigned to negative sentiment in a financial context (Loughran and McDonald, 2011). Other approaches successfully combine existing word lists (Demers and Vega, 2014; Rogers et al., 2011). Pröllochs et al. (2015) use a (largely automated) regularized regression analysis to empirically determine polar terms for financial news disclosure. Furthermore, there are dictionaries for different languages, e.g., Chinese (Peng et al., 2017), Arabic (Mahyoub et al., 2014), and German (Remus et al., 2010), and also bilingual approaches (Lu et al., 2011).

2.2 Financial Analysts and Analyst Reports

Financial analysts are industry experts who analyze companies and assess their prospects. For this purpose, analysts examine financial ratios, business models, management and its decisions, the industry as well as the overall economic situation. In this context, sell-side analysts try to closely communicate with company representatives to obtain new information and pass it on to the clients of their analyst firm (Soltes, 2014). To this end, analysts publish relevant information in analyst reports at regular intervals (Twedt and Rees, 2012) and support both information discovery and interpretation (Chen et al., 2010). In these reports, analysts give both general recommendations as to whether a share of a company should be bought and stock price targets for a specific period. Analysts provide both concise assessments of the latest company developments and in-depth analyses of the business model or market development. Analyst reports are usually written by a team of analysts headed by a senior analyst who has many years of industry expertise. The word choice in these reports is consequently shaped by many finance and management-specific terms. Besides, there is a distinct style of writing in the reports, which differs from corporate publications, public financial reporting or management texts. For this reason, it appears imperative to develop a sentiment-dictionary that is specifically trained to the language of financial analysts (Henry and Leone, 2015).

3 Induction of a Domain-Specific Sentiment Dictionary

3.1 Sentiment Dictionary Induction

In a comprehensive literature review, Mengelkamp (2017) points out the increasing number of newly developed sentiment dictionaries within the last years. This review examines the steps in which the different approaches to creating sentiment dictionaries can be summarized and which correspond to the steps in the approach of Pröllochs et al. (2015), which serves as an orientation in the implementation of our approach. A distinction is made between the *initial construction* phase and the *extension phase* of a sentiment dictionary. For both phases, the procedure is the same: words are selected, then polarized, and subsequently evaluated. In the extension phase, features might be edited or newly added, e.g., synonyms,

antonyms or other mood indicators such as emoticons. In addition, the basic dataset can be extended, or additional data labeling can be carried out.

Appropriate data sources need to be chosen in the *selection phase* (see section 3.2). Existing sentiment dictionaries, adequate text corpora, the knowledge of native speakers, or lexical-semantic databases can be used here. Depending on the data selection, the data must be formatted further. After a *dimensionality reduction* of texts, e.g., removing stop words, irrelevant text parts, duplicates, and whitespace, the *feature extraction* follows, e.g., by utilizing bag-of-words, N-grams (i.e., sequences of N adjacent words), or Part-of-Speech tagging (Nassirtoussi et al., 2014). This is followed by the *feature representation* in which usually a term-document matrix (TDM) is created using a binary, term frequency – inverse document frequency (tf–idf), or chi-squared approach. Depending on the text corpora used, it is possible to develop domain-specific dictionaries (Kearney and Liu, 2014).

In the *polarization phase* (see section 3.3), it is determined whether a word is positive, neutral or negative. The classification depends on the purpose of the dictionary. Previous studies mostly build on existing polarity lists (Loughran and McDonald, 2016). At the same time, word polarity can be derived from other text classifications. Here, for example, product ratings or capital market returns may be used.

The *evaluation phase* (see section 3.4) consists of three approaches (Mengelkamp, 2017). Firstly, native speakers can evaluate dictionaries. Secondly, labeled evaluation datasets can be utilized to assess classification accuracy. Depending on the accuracy, a sensitivity analysis can show how the pre-processing, data representation or the classification needs to be adjusted. Thirdly, a comparison can be made to other dictionaries belonging to a related domain. For this comparison, an evaluation dataset is needed, consisting of data that is more recent than the data on which the dictionary is based.

3.2 Data Selection, Pre-Processing, and Representation

3.2.1 Analyst Reports

We use a dataset that is based on the equity index Dow Jones Industrial Average (DJIA). The DJIA includes the 30 largest U.S. companies, which from our observations of the Institutional Brokers' Estimate System (I/B/E/S) tend to be those companies that are frequently covered by financial analysts. The time frame for our analysis covers the first quarter of 2009 to the first quarter of 2018. We analyze companies that have been represented in the DJIA for the longest time during that period and for which there are relatively many reports available. Table 1 lists these companies. We obtain the data from Thomson Reuters Advanced Analytics, which results in a total amount of 69,056 analyst reports.

Company names		
3M Co	General Electric Co	Microsoft Corp
American Express Co	Goldman Sachs Group Inc	Nike Inc
AT&T Inc	Home Depot Inc	Pfizer Inc
Boeing Co	HP Inc	Procter & Gamble Co
Caterpillar Inc	Intel Corp	Travelers Companies Inc
Chevron Corp	International Business Machines Corp	United Technologies Corp
Cisco Systems Inc	Johnson & Johnson	UnitedHealth Group Inc
Coca-Cola Co	JPMorgan Chase & Co	Verizon Communications Inc
Dupont De Nemours Inc	McDonald's Corp	Visa Inc
Exxon Mobil Corp	Merck & Co Inc	Walmart Inc

Table 1. Companies included in the dataset

We perform the following data pre-processing for the entire dataset. Since the reports are available as PDF files, we must deal with a data format that is not well suited for storing semi-structured data. However, we address this problem by carefully transforming the data into a tabular structure. Each paragraph is now contained in a single cell. In this way, the basic structures and thematic sections of the individual documents are kept, which we see as an advantage for text processing. We filter the generated document

fragments based on heuristics such as a minimum requirement for the number of words and the ratio of words to numbers or special characters. This results in complete sentences being analyzed and headings and tables deleted. In the next step, we delete the remaining special characters and numbers.

We check the text cells resulting from the analyst reports for exact duplicates. To do this, we hash each cell using the MD5 message-digest algorithm and count the number of duplicates for identical hashes. We assume that identical text cells that occur more than five times do not contain new information. Therefore, we only keep the text cell that occurs chronologically first in the dataset. Also, the dataset contains a non-negligible number of reports with varying similarity to prior reports. For example, in cases when an analyst published a minor update, the report often does not offer significant incremental informational value and mostly repeats previously published information. For this reason, we drop all reports that possess a similarity higher than 70 % to any prior report. We determined the value after testing different scenarios combined with a manual review.

Some text cells differ marginally from each other and were not recognized as duplicates. To address this issue, we compute different combinations of N-grams. We decide to use 15-grams based on the manual inspection of the resulting N-gram lists. We see 15 as a sensible choice for the N-grams as it provides a solution against the identification of too many false positives. We delete cells containing 15-grams that occur each more than 20 times in the entire dataset. To this point, the average length of a text cell in the dataset is 45 words, which serves as an indicator that our filtering methods successfully identified body text. A histogram for the word counts of the filtered analyst reports is shown in Figure 1. This figure illustrates that we were able to reduce the analyst reports to their essential content.

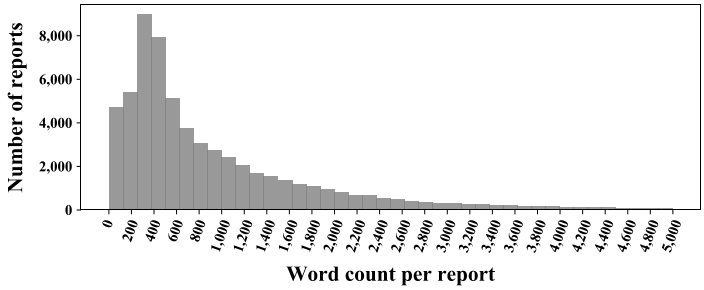

Figure 1. Number of words per analyst report in the dataset

The following dictionary induction requires that only those reports are included in the dataset that are released in the ten days after the day of company quarterly earnings releases. Figure 2 shows the 14,950 analyst reports remaining in the final analysis dataset for the ten days time frame per quarter.

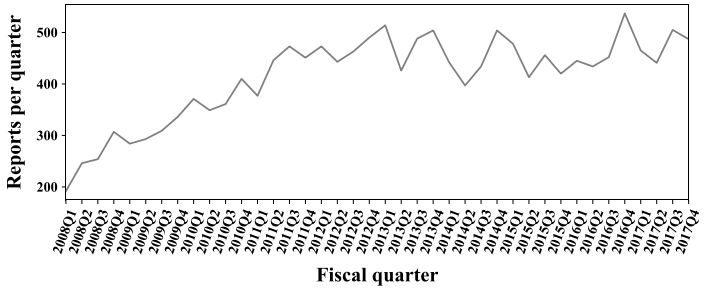

Figure 2. Analyst reports contained in the final dataset. The counts are per fiscal quarter for the ten days after an earnings release of a company.

Since we learned through manual analysis of the analyst reports that essential content can be found on the first pages of the reports, we filter for the first five pages. In the following, we transform the texts to lower case, filter numbers, and delete words with less than three characters. Then, we filter for stop words, date-related words, the names of the companies and the corresponding tickers as well as the names of the analyst firms, mostly broker houses, which publish the reports. For further analysis, we group and merge the text cells at the analyst report level. The transformation of the texts into a TDM

and the resulting reduction and weighting of words often are considered as steps of pre-processing, but in this paper, they take place within the model training. We choose this procedure because the parameter variation for constructing the TDM plays a central role in the dictionary induction. Table 2 contains the pre-processing steps and the corresponding number of words and reports that remain in the dataset.

Pre-processing step	Total word count	Mean word count per text cell	Number of reports
Initial dataset	58,238,105	45.67	69,056
Filter for days 1 to 10	16,138,584	45.27	14,990
Filter for first 5 pages	12,041,145	46.59	14,950
Basic pre-processing steps	7,420,296	28.71	14,950
Stop word removal	6,495,064	25.13	14,950
Custom term removal	6,137,564	23.75	14,950

Table 2. Overview of pre-processing steps with word count and the number of reports calculated after every step

3.2.2 Stock Market Data

The dataset of the stock prices used for the dictionary induction depends on the events, i.e., earnings releases, we have identified to measure abnormal returns and derive word polarities. The earnings release dates were obtained from Thomson Reuters. Earnings releases in our dataset start in January 2008 and last to the fourth quarter of 2017. This amounts to a total of 1,226 event dates. About one year of stock price data before the first earnings release date is required for the following event study. Accordingly, we utilize stock prices from 2007 until the start of the first quarter of 2018 for the companies in our dataset. We use adjusted closing prices downloaded from Thomson Reuters Datastream to account for dividends. We utilize the S&P 500 index as a reference market. More specifically, we chose the total return index, which adjusts for stock splits and dividends. In total, we were able to collect sufficient data points for 1,148 earnings releases (events).

3.3 Word Polarization

3.3.1 Event Study Setup

We relate the texts of the analyst reports to abnormal returns of the companies analyzed in the reports. From our point of view, this is sensible, as both the capital market and financial analysts are said to react to the publication of certain new information. In our research setting, these are the earnings releases of the companies. Huang et al. (2014) find exactly this relationship for analyst reports in their study.

For identifying abnormal capital market returns, the return of a company is compared to a reference market, in our case the S&P500 index. An event study measures whether a previously defined event leads to a change in returns of a company that would not have occurred without the new information of the event. For this purpose, we carry out an event study (MacKinlay, 1997) using the market model and the software EventStudyTools (Schimmer et al., 2014). In line with Henry (2008) and Skinner and Sloan (2002), we choose an event window of three trading days that is centered on the day of the earnings releases. Accounting literature explicitly requires a short event window to exclude confounding events, but also suggests to consider the day before the events, because of tactical reasons, many companies release negative news before earnings releases (Henry, 2008; Skinner and Sloan, 2002). To consider both analyst reports that are published quickly after an earnings release and reports that are published after a few days, we use the analyst reports of the ten days after an earnings release. We set an estimation window of 250 trading days (Thompson, 1995; McWilliams and Siegel, 1997). In the following dictionary induction, we relate the sentiments and abnormal returns by regression. We decided not to divide sentiments and abnormal returns into classes and compare these classifications, because this could result in the loss of important information for the creation of word lists. Even if lists of many unambiguous words might be expected by using only events with significant abnormal returns in the event window,

we refrain from this approach. Firstly, it reduces the list of potentially usable analyst reports and secondly, the dictionary would have been trained for extreme situations. Furthermore, we do not see our contribution in showing that analyst reports can be associated with significant abnormal returns, but the point estimate of the cumulative abnormal returns is relevant for the dictionary development. Figure 3 shows the average abnormal returns across all companies for the days relative to the earnings releases. Moreover, the figure shows the average abnormal returns per day per company and additionally the cumulative abnormal returns per company for the window of ten days around the earnings releases. Especially in the three-day window around earnings releases, abnormal returns are measurable.

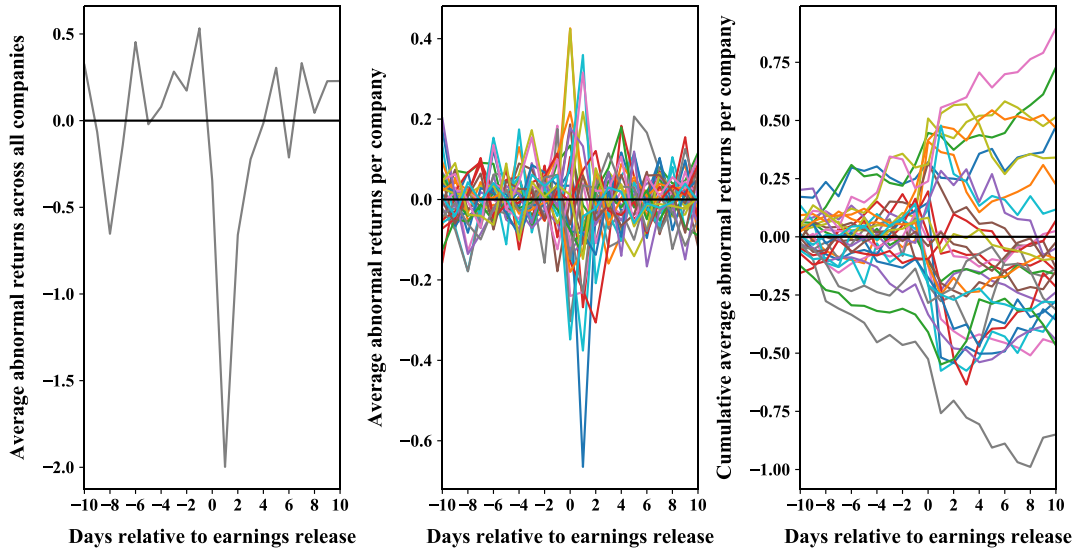

Figure 3. Overview of abnormal returns relative to earnings releases

3.3.2 Model Training

We dedicate 80 % of the earnings releases for word polarization, as the remaining part is needed for the dictionary evaluation. In our setup, the remaining 20 % are the most recent earnings releases. To create the polarity word lists, we split the induction dataset again so that one part can be used to train the dictionary and another part to test it. To increase the generalizability, we make use of the group k-fold cross-validation. This allows the process of word polarization to be repeated among differently composed data samples. The group k-fold procedure ensures that for each fold, a group, in our case a company, is not included in the training and test dataset at the same time. Thereby, we aim to further increase the generalizability of the dictionary. We set the k-fold splitting algorithm to split the polarization dataset into five parts (each contains six companies). By doing so, we try to achieve that companies and quarters are roughly equally represented during data splitting. We calculate the mean performance after the specific dictionary setup has been trained and validated for each of the five folds. Figure 4 shows the splitting of the dataset. The k-fold procedure is repeated for each dictionary setup in the induction process. Different parameters are changed for each setup in the procedure described hereafter. Finally, we select the setup that yields the best results in terms of explaining abnormal stock price returns.

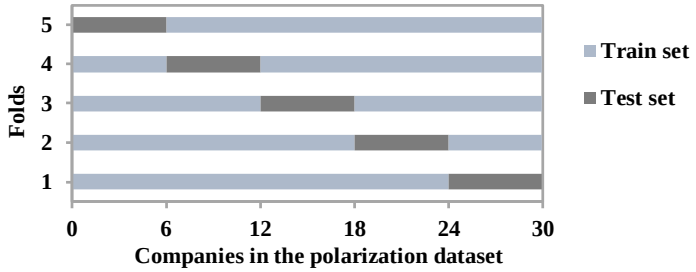

Figure 4. Groupwise k-fold splitting for dictionary model training and testing

Since the data for each fold is composed differently, we must create a TDM from the respective analyst reports. Then, we normalize the data using the tf–idf approach, as recommended for sentiment analysis by Loughran and McDonald (2011). That is, we weight the words according to their frequency in the entire dataset at hand. Words that occur frequently in the dataset are weighted correspondingly weaker per document. Building the TDM also involves setting limits on the percentage of documents in which a word can occur so that it remains in the dataset. On the one hand, we can identify recurring words that do not add any value to the sentiment analysis. On the other hand, we can delete words that can rarely be found in the dataset and therefore might not add relevant information. For dictionary induction, it is conceivable to carry out tokenization for the creation of the TDM not only with 1-grams but also to extend it to 2-grams, i.e., including word combinations of two words. Since we are convinced that considering frequently occurring word-groups might be an improvement to our dictionary, we extend the TDM by 2-grams. For model training, we jointly refer to both 1-grams and 2-grams as features. Utilizing the tf–idf matrix implies that the data is now stored in a bag of words representation. That is, the original order of the words is no longer kept. Subsequently, we label each document in the TDM with abnormal returns of the corresponding earnings releases.

We then perform a regression to determine which features have a particularly strong influence on the explanation of abnormal returns. In our setup, we follow the work of Pröllochs et al. (2015) and use ridge regression. The ridge regression regularizes the coefficients of the features. That is, all features are preserved, but we reduce the magnitude of the coefficients. This regularization approach is especially helpful when the coefficients of a few features are very high, and the model is strongly driven by these features only. The hyperparameter alpha in the ridge regression controls the strength of the rebalancing. If the value for alpha is increased, the magnitude of the coefficients decreases. This type of regression allows us to minimize the mean squared error of the model and increase the R^2 without reducing the number of features and potentially losing valuable information. We make use of dummy variables for the companies to exclude company-specific influences in the regression. We also control for year fixed effects. Furthermore, we control for industries, as these should not be represented by the dictionary (Loughran and McDonald, 2011). Here, we rely on the North American Industry Classification System (NAICS). Even though controlling for companies should theoretically capture industry effects, we want to make sure that we also account for situations where the classification has changed over time.

To get the best possible R^2 for the regression, we vary the following parameters of the previous steps: the alpha of the ridge regression; the maximum percentage of documents in which a feature can occur (max_df); the minimum percentage of documents in which a feature can occur (min_df). Accordingly, the number of features depends on the parameters min_df and max_df. We have tested more than 400 different setups and determined the model performance. In our view, the most informative examples of the various setups are listed in Table 3. They are arranged according to the height of the test mean R^2. For the best model, we find a test mean R^2 of 0.0205. At first glance, this seems to be relatively small but is also within the range measured by Pröllochs et al. (2015) in a similar setting. The final model setup has the following parameters: alpha: 10; min_df: 0.005; max_df: 0.50; features: 2,842. Although we strive to reduce the number of features, we have found that this is not possible for our dataset without losing model performance. Table 3 demonstrates that it is important to find a good balance between model size and the level of alpha in the ridge regression. Overfitting, i.e. the quality of the test model decreases with increasing quality of the training model, seems to be a problem for setups 4 and 5. Even by adjusting the alpha parameter, this cannot be solved sufficiently.

Setup	Test mean R^2	Train mean R^2	Alpha	Features	min_df	max_df
1	0.0205	0.2031	10	2,842	0.005	0.50
2	0.0186	0.1624	15	1,452	0.010	0.50
3	0.0173	0.1280	20	485	0.030	0.50
4	0.0142	0.1867	15	13,989	0.001	0.50
5	0.0137	0.3847	3	13,989	0.001	0.99
6	0.0073	0.1073	20	250	0.050	0.50

Table 3. Mean R^2 scores for ridge regression using different regression and TDM parameters

To evaluate the results of the ridge regression, we have tested further approaches using our best performing parameter setup. By conducting an ordinary least squares regression, we see overfitting since the train dataset results in a mean R^2 of 0.6159 but we get a mean R^2 of 0.0013 for the test dataset. After optimizing the hyperparameters, support vector regression results in an R^2 of 0.0135 for the test dataset. Application of a random forest regression returns in the best model version an R^2 of 0.0172. The comparison of the models confirms our choice of ridge regression.

3.3.3 Creation of Word Lists

The features of the trained model now consist of 1-grams and 2-grams. Each feature has been assigned a coefficient by the ridge regression by which we can sort them. Table 4 shows the top 30 features with the highest positive and negative influence in the model. Accordingly, we interpret these words as strong indicators of positive or negative sentiment. Loughran and McDonald (2011) note that managers can use word lists with negative polarity words to adapt their texts to their advantage. For this reason, they consider it useful to create relatively exhaustive word lists for their dictionary. It is conceivable that financial analysts may also edit their texts accordingly, but we do see a much weaker incentive here. Following Henry (2008), our goal is to create word lists that are as comprehensible as possible and can be easily interpreted. We choose a middle ground for our dictionary extension by using the polarity words from the LM dictionary but limiting it as much as possible.

Positive features	Negative features
raising: 0.0517, beat: 0.0406, better: 0.0338, raise: 0.0309, strong: 0.0295, impressive: 0.0264, raised: 0.0249, build: 0.0232, better expected: 0.0229, software: 0.0229, positive: 0.022, stronger: 0.0216, strong results: 0.0215, cloud: 0.0215, increasing: 0.0214, order growth: 0.0209, raising estimates: 0.0206, upgrade: 0.0199, acquire: 0.0197, inflection: 0.0195, improving: 0.0191, switching: 0.0185, seems: 0.0184, momentum: 0.0183, feared: 0.0181, improved: 0.0181, valuation: 0.0178, benefits: 0.0177, services: 0.0173, cycle: 0.0172	miss: −0.0515, lowering: −0.0473, disappointing: −0.0434, weakness: −0.04, weak: −0.0349, reducing: −0.033, missed: −0.0325, search: −0.0323, cautious: −0.0321, issues: −0.0279, negative: −0.0276, short: −0.0268, lower: −0.0249, guidance: −0.0235, headwinds: −0.0235, challenges: −0.0233, reset: −0.0228, disappointed: −0.0217, reduced: −0.0212, lowered: −0.0211, shortfall: −0.021, outlook: −0.0206, lowering price: −0.0198, downgrading: −0.0195, lowering estimates: −0.0194, weaker: −0.0193, disappointment: −0.0181, edge: −0.0181, weigh: −0.0178, weaker expected: −0.0178

Table 4. Top 30 positive and negative features (1- and 2-grams) resulting from ridge regression for analyst reports. The numbers are the weights within the ridge regression.

Our approach allows us to automatically identify possible polarity words, but a manual review of these is still necessary to ensure that the dictionary consists of words that can generally be regarded as analyst-specific polarity words. To limit the manual effort, we evaluate how the dataset can be narrowed down beforehand. In the following, we analyze which limits must be set for the selection of positive and negative model features. At this point, the features are separated by the positive and negative coefficients of the ridge regression and sorted by the parameter values within the model. The positive list contains 1,436 and the negative list 1,403 features. We divide the feature lists into 10 % quantiles, vary the number of top positive and negative features, and create a dictionary for each combination. Then, we use the respective dictionaries to calculate the sentiment scores per analyst report. Here, the sentiment score is no longer determined by the parameter values. To determine the score, we subtract the number of negative polarity words from the number of positive polarity words and divide the resulting value by the total number of polarity words contained in the text (Henry, 2008):

$$sentiment\ polarity = \frac{\#\ positive\ words - \#\ negative\ words}{\#\ positive\ words + \#\ negative\ words} \quad (1)$$

We relate the sentiment scores to the previously calculated corresponding abnormal returns in a regression analysis to determine the quality of the dictionary. For that purpose, we utilize the entire polarity dataset. The distributions of sentiment scores and abnormal returns (Figure 5) indicate that the datasets are approximately normally distributed and seem to be useful for analysis.

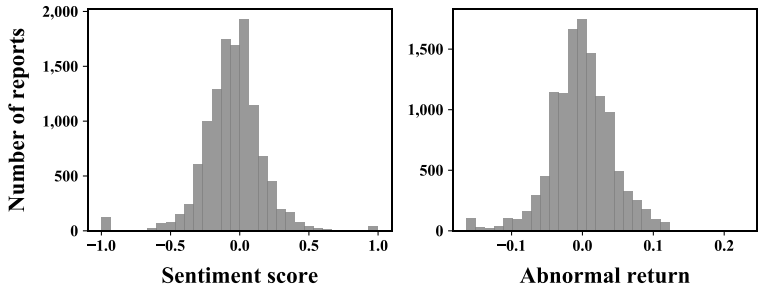

Figure 5. Histograms for sentiment scores (analyst dictionary) and abnormal returns of the polarization dataset

Figure 6 shows the various cut-offs and the associated quality of the dictionaries in terms of explaining abnormal returns. We prune the word lists as far as possible while keeping the highest R^2 of 0.152, i.e., we keep the top 50 % of the positive list (718 features) and the top 30 % of the negative list (421 features). We manually remove all features that can be considered inappropriate for a sentiment dictionary. This reduces the length of the positive list to 169 and the negative list to 103 features.

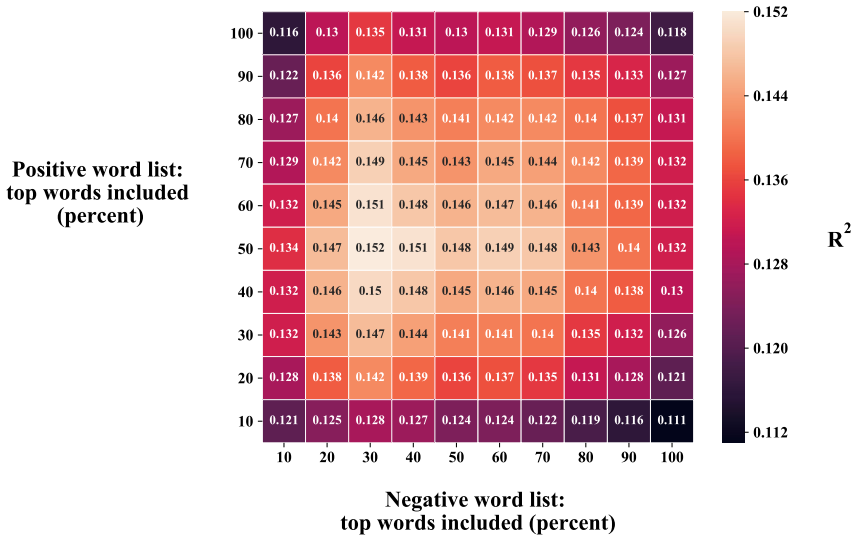

Figure 6. R^2 for explaining abnormal returns with sentiment scores for different combinations of the number of words contained in the polarity word lists

We compare the LM lists with the 2,839 textual features of the ridge regression and prune the LM lists accordingly. We assume that these words are not relevant to analyst communication. Then, we check whether there are words that have been labeled as positive by our approach but are classified as negative in the LM dictionary and remove these from the LM lists. Subsequently, we add the remaining words of the LM lists to our polarity lists and delete duplicates. Duplicates occur if the words are already included in our manually checked lists (38 positive and 36 negative words). With our procedure, we make sure that all LM words that are included in the part of the features we did not check manually are still included in the dictionary (30 positive and 48 negative words). We arrive at relatively short polarity lists of 199 positive and 151 negative words.[1] For the polarity dataset, the analyst dictionary with an R^2 of 0.096 is superior to the reference dictionaries (LM: 0.058; Henry: 0.051; General Inquirer: 0.038).

[1] The polarity word lists are available online: https://doi.org/10.25625/TYUGLF

3.4 Evaluation

We use 20 % of the initial earnings releases that were not part of the polarization. This allows us to easily evaluate how well the dictionary performs when we apply it to previously unknown data. We use the latest data available for the validation dataset. Therefore, we try to evaluate the dictionary as realistically as possible. We compare the induced dictionary with the polarity word lists of the General Inquirer, the Henry, and the LM dictionary. With each dictionary, we calculate sentiment scores based on equation 1 for the analyst reports. In doing so, we calculate relative instead of absolute values for the sentiment scores (Henry and Leone, 2015). Finally, we conduct regression analyses between sentiments and abnormal stock returns. The results show that for our validation dataset, the analyst dictionary has a better ability to explain abnormal returns based on sentiment values (see Table 5). The Henry dictionary and the General Inquirer give values for the R^2 of 0.055 and 0.041. For our dictionary, we get an R^2 of 0.069. This is 15 % better than the LM dictionary, which has an R^2 of 0.060. Through the development steps that led to these results, we demonstrate how an existing dictionary in a finance context can be extended semi-automatically and thereby address *RQ 1*. Moreover, regarding *RQ 2*, the results show that the developed dictionary is superior compared to existing dictionaries currently used in the finance domain.

Dictionary	General Inquirer	Henry	LM	Analyst dictionary
R^2	0.041	0.055	0.060	0.069

Table 5. Comparison between dictionaries applied to the evaluation dataset

4 Discussion

For sentiment analysis, insufficient innovative methods have been implemented and applied (Nassirtoussi et al., 2014). With our presented approach, we meet the demand for more domain-specific sentiment dictionaries (Mengelkamp et al., 2016). With our extension of the already widely used domain-specific (but more general-purpose) LM dictionary, we demonstrate how a domain-specific dictionary can be created based on capital market data (more objective labeling) and with considerably less manual effort. Naturally, the mapping employed in our approach between sentiment scores and abnormal returns must be carried out differently in other contexts (application domains), e.g., regarding sales figures, views, or transactions. From our perspective, such a specific mapping is a strength of the approach we propose. Also, we were able to show that our dictionary extension yields better results than both domain-specific and general-purpose dictionaries. In this context, Huang et al. (2014) compare their Naïve Bayes trained sentiment classifier, which achieves an accuracy of 80.89 %, with other established finance-related sentiment dictionaries for their dataset, which achieved 62.02 % (Loughran and McDonald, 2011) and 65.44 % (Henry, 2008). For the General Inquirer and DICTION, the paper lists 48.40 % and 54.93 % accuracy. Even if we do not solve a classification problem in this paper, but a regression problem, these results point in the same direction as the performance differences that we observe. We do not compare our dictionary with DICTION, because Loughran and McDonald (2015) have already shown that this dictionary is not suitable for applications in finance.

Huang et al. (2014) note that negative words have a stronger influence. Negatively classified analyst reports influence share prices more than positively classified reports. For our dataset and with regard to Figure 6, we do not observe a structurally stronger influence of negative compared to positive polarity words. One way to take different influences of words into account is to use the coefficients determined by our model and thus assign an individual weighting to the individual words of the analyst dictionary. Although this procedure would be conceivable for our dataset, we do not regard this suitable for the further application of the dictionary. Furthermore, Loughran and McDonald (2016) advise against testing word lists with positive words, since negations can have too much of a distorting effect. We leave it to the end-users whether they want to work with these positive word lists and make them available anyway. Theoretically, it might be possible to represent negations via 2-grams in the word lists. However, we refrain from this approach since this would add further complexity to the dictionary induction process and the word lists would probably get considerably longer.

Loughran and McDonald (2011) support a tf–idf weighting to prevent words that occur frequently from carrying too much weight in textual analysis. Henry and Leone (2015), on the other hand, consider this to be a risk and do not advise any weighting, since the word weighting and thus the sentiment depends on the entire dataset and its size and composition. Because the creation of the word lists in our case is based on a large dataset, we have chosen a middle ground and use word weighting in this part of our approach only. For the final dictionary evaluation, we did not use a weighting as we see this critically in the practical application. In potential use cases, there is probably no dataset available that is sufficiently large to carry out a meaningful weighting.

Twedt and Rees (2012) measure report complexity by the Fog Index and De Franco et al. (2015) show that experienced financial analysts write more readable analyst reports. Besides, a positive correlation between the readability and the number of companies covered by an analyst is identified. If readability differs from the readability of other document types in the finance domain, based on which existing sentiment dictionaries were trained, this can be an argument for the induction of the analyst-specific sentiment dictionary. Moreover, the use of finance-specific terminology in analyst reports can be examined. The Hypertextual Finance Glossary is suitable for this purpose (Harvey, 1999), and thus it is possible to draw a comparison with other document types from finance to illustrate differences.

5 Conclusion and Future Research

In this paper, we have developed a domain-specific sentiment dictionary by extending a domain-specific but more general-purpose dictionary in a semi-automated fashion. We show how the developed dictionary, in contrast to more general-purpose dictionaries, provides superior results in the context of analyzing sentiments of analyst reports. Similar to Pröllochs et al. (2015), we deploy regularized linear models to relate textual content to stock prices to extract words that are associated with abnormal stock returns. Words that can be associated with positive (negative) returns get a positive (negative) sentiment label. Regularization allows us to identify particularly influential words and neglect less important words. On this basis, we use this data to extend the LM dictionary and create an analyst-specific dictionary. We evaluate the performance of the sentiment dictionary with a sample of the analyst reports that were not used for model training. Moreover, we compare our developed analyst-specific sentiment dictionary with polarity word lists from established dictionaries, i.e., the General Inquirer (Stone et al., 1966), Henry (2008), and Loughran and McDonald (2011) (our baseline model). In this test, our dictionary proves to be better suited to measure the sentiment in analyst reports related to abnormal stock returns. Our R^2 is 15 % higher than the R^2 of the baseline model. We thus provide a tool that can be used immediately in research as well as in practical applications.

For the induced dictionary, it must be noted that the sentiment word lists highly depend on the analyst reports used. First, the data sample is U.S.-centric. Moreover, if a report is written particularly negatively, but there is a particularly positive abnormal return, the usability of the sentiment dictionary may be limited. In that case, positive words would be associated with negative abnormal returns or vice versa. We assume that abnormal returns are related to sentiment. A disadvantage in our experimental setting is that the dictionaries we used for comparison were not induced by abnormal stock returns. Additionally, the dictionary considers sentiment analysis but not the broader spectrum of emotion detection. In a further development of the dictionary, the evaluation might be extended for this reason. This might show that the dictionary may also be suitable for texts that are not associated with abnormal returns. For this purpose, paragraphs from analyst reports can manually be labeled with the tags positive and negative based on the sentiment they contain. Also, polarity words might be tagged with these labels. On the one hand, it is possible to compare the tagged words with those of the dictionary. On the other hand, an automated classification of the paragraphs can be carried out, which can then be compared with the manual classification. A general extension of the dataset brings the advantage that the model can also be trained and tested for other companies, industries, and time frames. Furthermore, the suitability of the dictionary for the contributions of analysts to conference calls following the publication of quarterly results might be analyzed.

References

Antweiler, W. and M. Z. Frank (2004). "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *The Journal of Finance* 59 (3)**,** 1259–1294.

Cambria, E., S. Poria, A. Gelbukh and M. Thelwall (2017). "Sentiment Analysis is a Big Suitcase." *IEEE Intelligent Systems* 32 (6)**,** 74–80.

Chen, X., Q. Cheng and K. Lo (2010). "On the Relationship Between Analyst Reports and Corporate Disclosures: Exploring the Roles of Information Discovery and Interpretation." *Journal of Accounting and Economics* 49 (3)**,** 206–226.

Das, S. R. (2014). "Text and Context: Language Analytics in Finance." *Foundations and Trends in Finance* 8 (3)**,** 145–261.

Das, S. R. and M. Y. Chen (2007). "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web." *Management Science* 53 (9)**,** 1375–1388.

De Franco, G., O. K. Hope, D. Vyas and Y. Zhou (2015). "Analyst Report Readability." *Contemporary Accounting Research* 32 (1)**,** 76–104.

Demers, E. A. and C. Vega 2014. Understanding the Role of Managerial Optimism and Uncertainty in the Price Formation Process: Evidence from the Textual Content of Earnings Announcements. Available at SSRN 1152326.

Hart, R. P. (2000). *Diction 5.0.* URL: http://rhetorica.net/diction.htm (visited on 03/30/2020).

Harvey, C. R. (1999). *Campbell R. Harvey's Hypertextual Finance Glossary.* URL: http://people.duke.edu/~charvey/Classes/wpg/glossary.htm (visited on 03/30/2020).

Henry, E. (2006). "Market Reaction to Verbal Components of Earnings Press Releases: Event Study Using a Predictive Algorithm." *Journal of Emerging Technologies in Accounting* 3 (1)**,** 1–19.

Henry, E. (2008). "Are Investors Influenced by How Earnings Press Releases are Written?" *The Journal of Business Communication* 45 (4)**,** 363–407.

Henry, E. and A. J. Leone (2015). "Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone." *The Accounting Review* 91 (1)**,** 153–178.

Hu, M. and B. Liu (2004). "Mining and Summarizing Customer Reviews." In*: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177.

Huang, A. H., R. Lehavy, A. Y. Zang and R. Zheng (2017). "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach." *Management Science* 64 (6)**,** 1–23.

Huang, A. H., A. Y. Zang and R. Zheng (2014). "Evidence on the Information Content of Text in Analyst Reports." *The Accounting Review* 89 (6)**,** 2151–2180.

Jegadeesh, N. and D. Wu (2013). "Word Power: A New Approach for Content Analysis." *Journal of Financial Economics* 110 (3)**,** 712–729.

Kearney, C. and S. Liu (2014). "Textual Sentiment in Finance: A Survey of Methods and Models." *International Review of Financial Analysis* 33**,** 171–185.

Li, F. (2010a). "The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (5)**,** 1049–1102.

Li, F. (2010b). "Textual Analysis of Corporate Disclosures: A Survey of the Literature." *Journal of Accounting Literature* 29**,** 143–165.

Liu, B. (2012). "Sentiment Analysis and Opinion Mining." *Synthesis Lectures on Human Language Technologies* 5 (1)**,** 1–167.

Loughran, T. and B. McDonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (1)**,** 35–65.

Loughran, T. and B. McDonald (2015). "The Use of Word Lists in Textual Analysis." *Journal of Behavioral Finance* 16 (1)**,** 1–11.

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

Lu, B., C. Tan, C. Cardie and B. K. Tsou (2011). "Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora." In*: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 320–330.

MacKinlay, A. C. (1997). "Event Studies in Economics and Finance." *Journal of Economic Literature* 35 (1)**,** 13–39.

Mahyoub, F. H., M. A. Siddiqui and M. Y. Dahab (2014). "Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning." *Journal of King Saud University – Computer and Information Sciences* 26 (4)**,** 417–424.

McWilliams, A. and D. Siegel (1997). "Event Studies in Management Research: Theoretical and Empirical Issues." *Academy of Management Journal* 40 (3)**,** 626–657.

Mengelkamp, A., S. Wolf and M. Schumann (2016). "Data Driven Creation of Sentiment Dictionaries for Corporate Credit Risk Analysis." In*: Proceedings of the 22nd Americas Conference on Information Systems,* San Diego.

Mengelkamp, A. J. 2017. *Informationen zur Bonitätsprüfung auf Basis von Daten aus sozialen Medien.* Dissertation, Cullivier: Goettingen.

Nassirtoussi, A. K., S. Aghabozorgi, T. Y. Wah and D. C. L. Ngo (2014). "Text Mining for Market Prediction: A Systematic Review." *Expert Systems with Applications* 41 (16)**,** 7653–7670.

Oliveira, N., P. Cortez and N. Areal (2017). "The Impact of Microblogging Data for Stock Market Prediction: Using Twitter to Predict Returns, Volatility, Trading Volume and Survey Sentiment Indices." *Expert Systems with Applications* 73**,** 125–144.

Peng, H., E. Cambria and A. Hussain (2017). "A Review of Sentiment Analysis Research in Chinese Language." *Cognitive Computation* 9 (4)**,** 423–435.

Pröllochs, N., S. Feuerriegel and D. Neumann (2015). "Generating Domain-Specific Dictionaries Using Bayesian Learning." In*: Proceedings of the 23rd European Conference on Information Systems,* Münster.

Remus, R., U. Quasthoff and G. Heyer (2010). "SentiWS – A Publicly Available German-language Resource for Sentiment Analysis." In*: Proceedings of the 7th International Conference on Language Resources and Evaluation,* Malta.

Rogers, J. L., A. Van Buskirk and S. L. Zechman (2011). "Disclosure Tone and Shareholder Litigation." *The Accounting Review* 86 (6)**,** 2155–2183.

Schimmer, A., A. Levchenko and M. S. (2014). *EventStudyTools.* St. Gallen. URL: http://www.event-studytools.com (visited on 03/30/2020).

Skinner, D. J. and R. G. Sloan (2002). "Earnings Surprises, Growth Expectations, and Stock Returns or Don't Let an Earnings Torpedo Sink Your Portfolio." *Review of Accounting Studies* 7**,** 289–312.

Soltes, E. (2014). "Private Interaction Between Firm Management and Sell-Side Analysts." *Journal of Accounting Research* 52 (1)**,** 245–272.

Stone, P. J., D. C. Dunphy, M. S. Smith and D. M. Ogilvia (1966). "The General Inquirer: A Computer Approach to Content Analysis." *American Sociological Review* 32 (5)**,** 859–860.

Tetlock, P. C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance* 62 (3)**,** 1139–1168.

Tetlock, P. C., M. Saar-Tsechansky and S. Mackskasssy (2008). "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *The Journal of Finance* 63 (3)**,** 1437–1467.

Thompson, R. (1995). "Empirical Methods of Event Studies in Corporate Finance." *Handbooks in Operations Research and Management Science* 9**,** 963–992.

Twedt, B. and L. Rees (2012). "Reading Between the Lines: An Empirical Examination of Qualitative Attributes of Financial Analysts' Reports." *Journal of Accounting and Public Policy* 31 (1)**,** 1–21.

Xing, F. Z., E. Cambria and R. E. Welsch (2018). "Natural Language Based Financial Forecasting: A Survey." *Artificial Intelligence Review* 50 (1)**,** 49–73.