Expert Systems With Applications 148 (2020) 113223

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Predicting sentence-level polarity labels of financial news using abnormal stock returns

Bernhard Lutz[a], Nicolas Pröllochs[b,*], Dirk Neumann[a]

[a] University of Freiburg Platz der Alten Synagoge Freiburg 79098, Germany
[b] University of Giessen Licher Str. 62 Giessen 35394, Germany

ARTICLE INFO

Article history:
Received 23 August 2019
Revised 17 November 2019
Accepted 18 January 2020
Available online 25 January 2020

Keywords:
Financial news
Expert systems
Natural language processing
Multi-instance learning
Decision-making

ABSTRACT

Expert systems for automatic processing of financial news commonly operate at the document-level by counting positive and negative term-frequencies. This, however, limits their usefulness for investors and financial practitioners seeking specific positive and negative information on a more fine-grained level. For this purpose, this paper develops a novel machine learning approach for the prediction of sentence-level polarity labels in financial news. The method uses distributed text representations in combination with multi-instance learning to transfer information from the document-level to the sentence-level. This has two key advantages: (1) it captures semantic information of the textual data and thereby prevents the loss of information caused by bag-of-words approaches; (2) it is solely trained based on historic stock market reactions following the publication of news items without the need for any kind of manual labeling. Our experiments on a manually-labeled dataset of sentences from financial news yield a predictive accuracy of up to 71.20%, exceeding the performance of alternative approaches significantly by at least 5.10 percentage points. Hence, the proposed approach provides accurate decision support for investors and may assist investor relations departments in communicating their messages as intended. Furthermore, it presents promising avenues for future research aiming at studying communication patterns in financial news.

1. Introduction

Companies around the world are required by law to publish information that has the potential to influence their valuation (Benston, 1973). These *financial news* serves as an important source of information for investors considering exercising ownership in stock, as they trigger subsequent movements in stock prices (Kearney & Liu, 2014; Loughran & McDonald, 2016). Besides quantitative figures, such as sales volumes or earnings forecasts, financial news also contains a substantial amount of qualitative content (Lupiani-Ruiz et al., 2011; Wang, Zhe, Kang, Wang, & Chen, 2008). Hence, investors are required to perform a quick and accurate evaluation of language and word choice before deciding on whether or not to purchase the stock in question (Cavalcante, Brasileiro, Souza, Nobrega, & Oliveira, 2016).

Due to the sheer amount of available financial information, it is of great importance for investors and financial professionals to possess computerized tools to operationalize the textual content of financial news (Chan & Chong, 2017). Over the last several years, researchers have created a vast number of expert systems to facilitate the automated processing of textual content. The main goal of these systems is to identify relations between the textual content and the reception on the investors' side (e.g., Chan & Chong, 2017; Gunduz & Cataltepe, 2015; Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014). From a methodological point of view, the overwhelming majority of these systems use bag-of-words methods that treat every financial news item as a single document with a label, e.g., the stock market reaction following the publication of a news item (e.g., Cavalcante, Brasileiro, Souza, Nobrega, & Oliveira, 2016; Pröllochs, Feuerriegel, & Neumann, 2016). The predominant approach is then to predict a sentiment polarity label for the document by counting positive and negative term-frequencies based on predefined sentiment polarity dictionaries (Loughran & McDonald, 2016). Alternatively, one employs historic stock market returns following the publication of news items for training a supervised machine learning classifier (e.g., Groth & Muntermann, 2011; Kraus & Feuerriegel, 2017; Schumaker, Zhang, Huang, & Chen, 2012).

A major downside of existing expert systems for the automated processing of financial news is that they only operate at the document-level without drawing inferences on a fine-grained ba-

* Corresponding author.
E-mail addresses: bernhard.lutz@is.uni-freiburg.de (B. Lutz), nicolas.proellochs@wi.jlug.de (N. Pröllochs), dirk.neumann@is.uni-freiburg.de (D. Neumann).

https://doi.org/10.1016/j.eswa.2020.113223
0957-4174/

sis. This ignores the fact that financial news typically entails positive as well as negative information, and that different sentences in a single text are likely to express different polarities (Lerman & Livnat, 2010). Hence, readers of financial news are required to apply the utmost attention and detailed, domain-specific knowledge in order to assess the information on a fine-grained basis. For example, investors might be interested in the specific positive and negative aspects (e.g., "revenue in the US declined"; "production costs could be reduced") communicated by the company before making an informed decision (van de Kauter, Breesch, & Hoste, 2015). In the same vein, companies and investor relations departments are lacking tools to assist them in communicating their message as intended (Loughran & McDonald, 2014).

Despite this, automatically identifying individual positive and negative aspects in financial news can prove challenging. Aspect-based sentiment analysis methods as frequently used outside of the finance domain rely on manual annotations of aspects of individual sentences (e.g., Liu & Zhang, 2012), or other predefined resources such as sentiment lexicons (e.g., Kelly & Ahmad, 2018). These supervised learning approaches require manually labeled data which is usually very costly and time consuming (García-Pablos, Cuadros, & Rigau, 2018; Li, 2010b). In the finance domain, such human labels are also highly subjective and do not necessarily correspond to actual stock market reaction (Pröllochs, Feuerriegel, & Neumann, 2018). As a result, aspect-based sentiment analysis as used in previous works is not readily applicable for sentence-level polarity classification of financial news.

As a solution, we propose a novel machine learning approach that allows one to learn polarity labels for individual sentences in financial news. Our method does not require any kind of manual labeling or predefined sentiment dictionaries, as it is solely trained on the stock market reaction following the publication of a news item. We use a two-step approach: First, *distributed text representations* allow for the preservation of the context-dependent nature of language, thereby overcoming some of the shortcomings of the bag-of-words approach. Second, *multi-instance learning* allows one to train a classifier that can be used to transfer information from the document-level to the sentence-level. In our scenario, a document is represented by a financial news item, whereas the document label is given by the reaction of investors on the stock market. Based on this document-level information, our approach is capable of learning polarity labels for the individual sentences within financial news. Our later analysis shows that this approach yields significantly superior predictive performance, exceeding the performance of alternative approaches by at least 5.10 percentage points. In summary, we contribute to the existing research body as follows:

1. We develop a novel machine learning approach to learn sentence-level polarity labels in financial news solely based on historic stock market returns without the need for manual labeling
2. We evaluate our method on a manually-labeled dataset of sentences from financial news; showing that it significantly outperforms existing approaches in terms of predictive accuracy
3. We demonstrate how our approach can be used to analyze fine-grained communication patterns in financial news and how this may help to enhance existing financial expert systems

Our study has important implications for research and practice. Researchers can use our approach as a tailored means to analyze the structure of financial news on the level of individual sentences. These insights can generate new knowledge regarding financial decision-making and help to improve the performance of financial expert systems. Investors can use our approach to gain a faster overview of positive and negative aspects of financial news which mitigates the risk of being outperformed by competitors (Hirshleifer & Teoh, 2003). This is particularly relevant in the case of negative content, which is known to have a particularly pronounced impact on stock markets (e.g. due to negativity bias (Brown & van Harlow, 1988)). Ultimately, our approach can be used as writing assistance by company executives and investor relations departments that may consider choosing their language strategically to ensure that their message is interpreted as intended.

The remainder of this work is structured as follows. Section 2 highlights the drawbacks of current approaches for studying the sentiment polarity on a fine-grained level and provides an overview of literature that performs sentiment analysis of financial news. Subsequently, Section 3 introduces our data sources and the way in which we integrate distributed text representations and multi-instance learning to predict positive and negative sentence labels in financial news. Section 4 presents our primary results, while Section 5 illustrates how our approach can be used to analyze communication patterns in financial news. Section 6 discusses the implications of our study for researchers and practitioners, and details the limitations of our approach. Section 7 concludes and provides directions for future research.

2. Background

A tremendous amount of literature has examined the extent to which stock market prices correlate with the textual information provided in financial news (e.g., Feuerriegel & Pröllochs, 2018; Loughran & McDonald, 2011). Existing studies in this area typically focus on the analysis of the overall sentiment polarity of documents (Loughran & McDonald, 2016). In this context, sentiment polarity is predominately considered as a measure of the qualitative information, referring to the degrees of positivity or negativity in opinions shared by the authors regarding individual stocks or the overall market (Kearney & Liu, 2014; Loughran & McDonald, 2016). A common approach for predicting sentiment polarity of financial news is to count positive and negative term frequencies in documents based on predefined polarity dictionaries. Examples include the General Inquirer (Stone, 2002) and the Loughran-McDonald dictionary (Loughran & McDonald, 2011). The sentiment polarity of a document is then determined based on the ratio between the number of positive and negative words (Loughran & McDonald, 2016). Corresponding methods in different variations have been used to predict stock market returns based on ad hoc announcements (e.g., Muntermann & Guettler, 2007), newspapers (e.g., Tetlock, 2007) or company press releases (e.g., Henry, 2008). Comprehensive literature overviews regarding the textual analysis of financial news can be found in Nassirtoussi et al. (2014) and Cavalcante, Brasileiro, Souza, Nobrega, and Oliveira (2016), as well as Loughran and McDonald (2016).

Although the aforementioned dictionary-based methods have produced remarkably robust results, they come with several drawbacks; in particular when it comes to predicting the sentiment polarity of individual sentences in financial news. First, dictionary-based approaches rely on lists of words that classify terms as either positive or negative based on human judgments (e.g., Henry, 2008; Loughran & McDonald, 2013). However, the word lists are selected ex ante based on the subjective opinions of their authors. Thus, they can be neither as comprehensive nor as precise as statistical rigor (Pröllochs et al., 2018). Second, the underlying bag-of-words approach severely compromises the results (Li, 2010b). For instance, in the sentence *"the company reduced its costs and increased its profit margin"*, bag-of-words approaches cannot distinguish the meaning from the same sentence with an exchange of *"costs"* and *"profit"*. Third, we will later see that even highly finance-specific dictionaries struggle with the task of sentence po-

larity classification as many sentences in financial news do not contain any of the words from the polarity word lists.

An alternative that circumvents the need for labels from human judges comes in the form of machine learning (e.g., Antweiler & Frank, 2004; Schumaker & Chen, 2009). Supervised learning can be used to train a classifier based on arbitrary text representations using the stock market return as gold standard (Pröllochs et al., 2018). However, training a machine learning classifier at document-level tends to result in low predictive performance at sentence-level (Kotzias, Denil, de Freitas, & Smyth, 2015; Zhou & Zhang, 2003). Alternatively, one could rely on the manual labeling of positive and negative sentences to train a machine learning classifier for sentences in financial news (e.g., van de Kauter et al., 2015). However, this again compromises the results given that human labels are highly subjective and do not necessarily correspond to actual stock market reaction (Pröllochs et al., 2018).

Since the above-mentioned hurdles limit the degree of severity of textual analysis, only a few studies analyzed financial news on a fine-grained level. As one of very few examples, Allee and Deangelis (2015) use the Loughran-McDonald dictionary to study the role of sentiment dispersion in corporate communication. The authors find that the distribution of sentiment is significantly associated with investors' reaction to the textual narratives. Another study by Li (2010a) acknowledges the drawbacks of dictionary-based methods and instead uses a Naïve Bayes approach to train a sentence classifier based on a set of 30,000 manually-labeled sentences drawn from the forward-looking statements in the Management Discussion and Analysis section of 10-K filings. Likewise, van de Kauter et al. (2015) use a support vector machine classifier with human annotations to predict implicit and explicit sentiment in sentences from financial newspapers. Both studies find that the machine learning approach outperforms common dictionary-based approaches when predicting sentence-level sentiment polarity. Nonetheless, apart from the fact that assigned manual labels are highly subjective and do not necessarily correspond to actual stock market reaction, the utilized methods suffer from the bag-of-words drawbacks, such as missing context and information loss (Nassirtoussi et al., 2014).

Apart from the finance domain, researchers frequently utilize aspect-based sentiment analysis methods to assess the semantic orientation of documents on a fine-grained basis (Liu & Zhang, 2012). Instead of classifying the overall sentiment of a text into positive or negative, aspect-based analysis allows one to associate specific sentiments with different aspects mentioned in a text (Liu & Zhang, 2012). A typical application is the assessment of product reviews (e.g., Amplayo, Lee, & Song, 2018; Pham & Le, 2018; Qiu, Liu, Li, & Lin, 2018) where 1 aims at determining the sentiments expressed on different features of the products under evaluation (e.g., battery or performance of a smartphone). For building a prediction model, however, aspect-based sentiment analysis typically requires a significant amount of manually-labeled data (e.g., (Peng, Ma, Li, & Cambria, 2018; Pontiki et al., 2016)) or other language specific resources for training on a particular domain and for a particular language (García-Pablos, Cuadros, & Rigau, 2018; van de Kauter, Breesch, & Hoste, 2015). For example, state-of-the-art models predicting sentiment at sentence-level commonly rely on training examples that contain different sentiment labels for different aspects or targets in the sentence (Xue & Li, 2018). As a result, aspect-based sentiment analysis as used in previous works is not readily applicable for sentence-level polarity classification of financial news. The underlying reasons are two-fold. First, financial news are complex and training examples with fine-grained labels of aspects in financial news are costly and barely available (Li, 2010b). Second, as aforementioned, the market reaction following the publication of a news item does not necessarily correspond to the classifications of words in sentiment lexicons or manual labels of human judges (Pröllochs et al., 2018).

To tackle this problem, this paper compares algorithms to predict the polarity of individual sentences in financial news. As a remedy to the drawbacks of previous approaches, we propose a fine-grained approach based on distributed text representations and multi-instance learning that allows one to transfer information from the document-level to the sentence-level. In contrast to existing approaches, our method does not require any kind of manual labeling or predefined sentiment dictionaries, as it is solely trained on the stock market reaction following the publication of a news item. Although multi-instance learning has been successfully applied for several machine learning tasks – including text classification (Angelidis & Lapata, 2018; Kotzias, Denil, de Freitas, & Smyth, 2015), malware classification (Stiborek, Pevný, & Rehák, 2018), and image classification (Sudharshan et al., 2019) – we are not aware of any publication that utilizes this method to predict sentence polarity labels for financial news.

3. Materials and methods

In this section, we introduce our dataset and present our method of analyzing financial news at sentence-level. Fig. 1 presents a schematic illustration of our proposed expert system. As a first step, the system performs several preprocessing steps using common tools from natural language processing. Second, the textual data is mapped to a vector-based representation using sentence embeddings. Third, we combine the vector representations with the historic stock market returns of companies to train a sentence-level classifier using multi-instance learning. Finally, the predicted sentence labels can provide decision support for investors in placing their orders and assist company executives in writing their corporate disclosures.

3.1. Dataset

Our financial news dataset consists of 9502 German regulated ad hoc announcements[1] from January 2001 and September 2017. As a requirement, each ad hoc announcement must contain at least 50 words and be written in English. In research, ad hoc announcements are a frequent choice (e.g., Groth & Muntermann, 2011; Hagenau, Liebmann, & Neumann, 2013; Pröllochs, Feuerriegel, & Neumann, 2019) when it comes to evaluating and comparing methods for sentiment analysis. Additionally, this type of news corpus presents several advantages: ad hoc announcements must be authorized by company executives, the content is quality-checked by the Federal Financial Supervisory Authority[2] and several publications analyze their relevance to stock market reactions – finding a direct relationship (e.g., Muntermann & Guettler, 2007).

To study the stock market reaction, we use the daily *abnormal return* of the company that has published a financial item. For this purpose, we use the common event study methodology (MacKinlay, 1997) whereby we determine the normal return, i.e., the return which is expected in the absence of a news disclosure, using a market model. This market model assumes a stable linear relation between market return and normal return. Concordant with the related literature (e.g., Pröllochs & Feuerriegel, 2018), we model the market return using a stock market index - namely, the CDAX - along with an event window of 30 trading days prior to the news disclosure. Finally, we determine the abnormal return as the difference between actual and normal returns. Here, all financial market data originates from Bloomberg.

[1] Kindly provided by Deutsche Gesellschaft für Ad-Hoc-Publizität (DGAP).
[2] Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin).

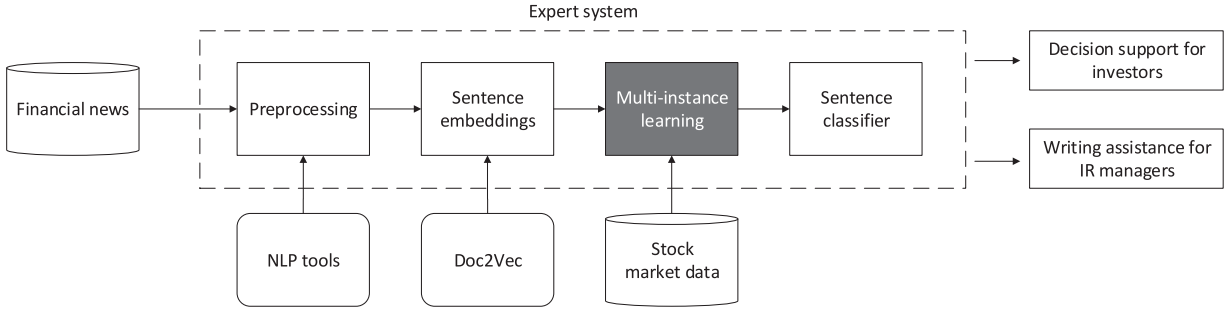

Fig. 1. Schematic illustration of an expert system for sentiment analysis on financial news at sentence-level.

3.2. Preprocessing

We apply several common filtering steps to our dataset which allows us to reduce the effect of confounding influences in our later analysis. Concordant with the related literature, we account for extreme stock price effects by removing penny stocks with a price lower than $1 and by omitting outliers at the 1% level (Zhang, Swanson, & Prombutr, 2012).[3] In addition, we follow previous research by removing all announcements for which the stock market data was not available on Bloomberg, announcements that were published after trading hours (Groth & Muntermann, 2011), and announcements from companies that have issued multiple announcements on the same day. These filtering steps result in a sample of 6360 ad hoc announcements.

Next, we perform several common preprocessing steps on the textual data, in order to remove formatting and noisy content (Manning & Schütze, 1999; Pang & Lee, 2008; Ravi & Ravi, 2015). First, by using a list of cut-off patterns, we omit contact addresses and standard phrases. We then convert each ad hoc announcement to lower case and replace abbreviations. Moreover, we replace URLs, ISIN numbers, dates, times, positive numbers and negative numbers with appropriate tokens. In addition, we replace infrequent terms that appear less than five times in our corpus with a separate token (Kotzias, Denil, de Freitas, & Smyth, 2015). These preprocessing steps reduce the size of our vocabulary from 34,910 words to 10,969 words.

Finally, we use the sentence splitting tool from Stanford CoreNLP (Manning et al., 2014) to split each ad hoc announcement into sentences. It is worth noting that this approach also addresses the frequently encountered challenges in previous works regarding the accurate splitting of financial items into sentences because *"the presence of extensive lists, technical terminology, and other formatting complexities, makes sentence disambiguation especially challenging in accounting disclosures" (Loughran & McDonald, 2016)*. We observe that 93.81% of all ad hoc announcements contain between 5 and 40 sentences, while a few ad hoc announcements are of very short or excessive length. Thus, to ensure comparability, we remove all ad hoc announcements with lengths in the top and bottom percentile from our dataset[4] Our final corpus consists of 6258 ad hoc announcements (i.e., 65.9 % of the original dataset). The total number of sentences across all ad hoc announcements is 90,958.

3.3. Descriptive statistics

Table 1 presents the summary statistics of our dataset. Companies in our dataset have published as few as 1 ad hoc announcement, but also as many as 85, with a median number of 8 per company. The ad hoc announcements were issued by 418 different companies. During our period of study, 177 companies started trading and 97 ended trading. Companies from the financial sector[5] released the most announcements (1214 in total, 11.7 per company), whereas companies from the diversified sector produced the fewest (34 in total, 11.3 per company). The average number of ad hoc announcements published per month is 31.30. The mean length of a single ad hoc announcement is 360.15 words or 14.53 sentences. The average length of a sentence in our dataset is 24.01 words.

Table 1 also contains summary statistics regarding the distribution of end-of-day stock market returns that followed the publication of ad hoc announcements. The nominal and abnormal stock market returns are fairly normally distributed with a mean and median that is shifted slightly to the right. We observe a mean nominal return of 0.72% with a standard deviation of 5.27. The mean abnormal return amounts to 0.65% with a standard deviation of 5.03%. Out of all disclosures, a total number of 3486 ad hoc announcements (55.70%) resulted in a positive abnormal return, whereas 2772 (44.30%) led to a negative abnormal return.

3.4. Distributed text representations

The accuracy of sentiment analysis applications depends heavily on the representation of the textual data and the selection of features (Mirończuk & Protasiewicz, 2018; Pröllochs, Feuerriegel, & Neumann, 2018). To overcome the drawbacks of the frequently employed bag-of-words approach, such as missing context and information loss, we take advantage of recent advances in learning distributed representations for text. For this purpose, we employ the *doc2vec* library developed by Google (Le & Mikolov, 2014). This library is based on a deep learning model which creates numerical representation of texts, regardless of their length. Specifically, the underlying model allows one to create distributed representations of sentences and documents by mapping the textual data into a vector space.

The word vectors being used by this library have several useful properties. First, more similar words are mapped to more similar vectors. For instance, the word *cost* is mapped closer to *debt* than to *company*. Second, the feature vectors also fulfill simple algebra properties such as, for example, *king - man + woman = queen*. Thus, in contrast to the bag-of-words approach, the *doc2vec* library incorporates context-specific information and semantic similarities.

[3] As part of our robustness checks, we performed additional analyses where we included penny stocks and announcements with extreme price movements. The difference in predictive performance is not statistically significant. The results are provided in our supplementary materials.

[4] We evaluate the relationship between announcement length and predictive performance in Section 4.

[5] All sector codes originate from Bloomberg.

Table 1
Descriptive statistics.

	Mean	Median	Min	Max	SD	Skew.	Kurt.
Nominal return (in %)	0.724	0.331	-24.903	27.419	5.273	0.322	2.735
Abnormal return (in %)	0.655	0.323	-19.290	22.443	5.037	0.308	2.747
Announcement length (in sentences)	14.530	12.000	2.000	59.000	8.857	1.674	3.697
Announcement length (in 100 words)	3.602	3.010	0.400	20.430	2.289	2.053	6.139
Announcements per company	14.971	8.000	1.000	85.000	16.864	1.656	2.392

As a further advantage, the feature space of the sentence representations is typically in a relatively small range between 200 and 400 dimensions (as compared to oftentimes several thousands for bag-of-words models). The feature representations created by the *doc2vec* library have been shown to significantly increase the predictive performance of machine learning models for text classification (Le & Mikolov, 2014).

For the training of our *doc2vec* model, we initialize the word vectors with the vectors from the pre-trained Google News dataset[6], which is a common choice in the previous literature (e.g., Lau & Baldwin, 2016; Tang, Fang, & Wang, 2014). Here, we use the hyper parameter settings developed by Lau and Baldwin (2016) during an extensive analysis. Subsequently, we generate vector representations for all sentences in our sample. These vectors are used in the next section as input data to train a sentence-level classifier using multi-instance learning.

3.5. Sentence-level polarity classification using multi-instance learning

The goal of this study is to infer the polarity of individual sentences in financial news using labels that are only provided at document-level. Hence, we are facing a problem in which the observations (documents) contain groups of instances (sentences) instead of a single feature vector, whereby each group is associated with a label (abnormal stock returns). Formally, let $X = \{\boldsymbol{x}_i\}, i = 1 \ldots N$ denote the set of all instances, N the number of instances, D the set of groups and K the number of groups. Each group $D_k = (\mathcal{G}_k, l_k)$ consists of a list of instances $\mathcal{G}_k \subseteq X$ and is assigned a label l_k (0 for negative and 1 for positive). The learning task is to train a classifier y with parameters $\boldsymbol{\theta}$ to infer instance labels $y_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ given only the group labels.

The above problem is a multi-instance learning problem (Dietterich, Lathrop, & Lozano-Pérez, 1997; Kotzias, Denil, de Freitas, & Smyth, 2015), which can be solved by constructing a loss function consisting of two components: (a) a term that punishes different labels for similar instances; (b) a term that punishes misclassifications at the group level. The general loss function $L(\boldsymbol{\theta})$ shown in Eq. (1) is then minimized as a function of the classifier parameters $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{S}(\boldsymbol{x}_i, \boldsymbol{x}_j)(y_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_{\boldsymbol{\theta}}(\boldsymbol{x}_j))^2 + \frac{\lambda}{K} \sum_{k=1}^{K} (A(D_k, y_{\boldsymbol{\theta}}) - l_k)^2, \quad (1)$$

where λ is a free parameter that denotes the contribution of the group-level error to the loss function. The term $\mathcal{S}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ measures the similarity between two instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and $(y_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_{\boldsymbol{\theta}}(\boldsymbol{x}_j))^2$ denotes the squared error in the predictions for instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. In addition, $A(D_k, y_{\boldsymbol{\theta}})$ denotes the predicted label for the group D_k. Hence, the loss function punishes different labels for similar instances while still accounting for a correct classification of the groups.

In order to adapt the loss function to our problem, i.e., classify sentences in financial news into positive and negative categories, we specify concrete functions for the placeholders in Eq. (1) as follows. First, we use a radial basis function kernel to calculate a similarity measure between two sentence representations, i.e., $\mathcal{S}(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}$. This similarity measure is 0 for completely different sentences and 1 for identical sentences.

Second, we need to specify a classifier to predict $y_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$. Here, we choose a neural network with a Long Short Term Memory (LSTM) layer (Hochreiter & Schmidhuber, 1997) followed by a dense layer with sigmoid activation. In a recent article, Kraus and Feuerriegel (2017) demonstrated the advantages of such a model in processing ad hoc announcements.

The LSTM layer of our network consists of 300 neurons which equals the dimension of our sentence embeddings. The weights of the neural network $\boldsymbol{\theta}$ are initialized with random values and optimized using Adam optimization with the default learning rate 0.001. In addition, we perform grid search to optimize the hyper parameter λ. In order to prevent overfitting, we use a validation split of 10% of the training data to determine the optimal stopping point. Further details regarding the parameter settings are provided in our supplementary materials.

Ultimately, we use the above model to predict labels of individual sentences as follows. First, a sentence is transformed in its vector representation $\boldsymbol{x}_i$. Second, we calculate the output $y_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ which is always between 0 and 1. Hence, the model predicts positive if $y_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \geq 0.5$ and negative otherwise.

Evidently, the model is principally also capable of making predictions at document-level. For this purpose, it chooses the most frequent label of all the sentences contained in the document, i.e., positive documents are expected to contain a greater number of positive sentences than negative sentences and vice versa.

4. Results

This section evaluates our method for predicting sentence-level polarity labels in financial news. First, we present our model and illustrate an example of how the resulting inferences can aid investors and financial practitioners. Subsequently, we compare the predictive performance of our method with several baseline approaches and validate the robustness of our results.

4.1. Extraction of sentence labels

We use the methodology as described in the previous sections to infer sentence labels from ad hoc announcements. The result of the learning procedure is a dataset containing documents that consist of groups of sentences in vector representations. In this context, each sentence in a document is assigned a positive or negative polarity that is inferred from the document label, i.e., the abnormal stock market returns.

We proceed by presenting summary statistics of the resulting dataset. We find that a majority (73.38%) of all sentences are assigned a positive polarity, whereas the remaining 26.62% are as-

[6] Available from the Google code archive at https://code.google.com/archive/p/word2vec/.

Table 2
Distribution of positive and negative sentences for different stock market reactions.

		Sentence label	
		positive	negative
Market reaction	positive	38,945 (78.07%)	10,943 (21.93%)
	negative	27,796 (67.68%)	13,274 (32.32%)

[...] As shown in LEONI AG's interim report, consolidated external sales amounted to EUR 350 million on 31 March 2005. The figure is therefore up about 23 percent on the same quarter one year earlier (EUR 284.8 million) despite difficult market conditions. However, given the strong sales in quarters three and four of the previous year, it will not be possible to sustain this high rate of growth over 2005 as a whole. LEONI therefore reaffirms its sales forecast for fiscal 2005 of EUR 1.43 billion (up from EUR 1.25 billion in the previous year). Earnings before interest and taxes (EBIT) were up from EUR 9.5 million in the first quarter of 2004 to EUR 17.2 million in the same period this year, equating to growth of 81 percent. Consolidated net income increased by almost 44 percent, from EUR 5.5 million to EUR 7.9 million. In terms of operating earnings before interest and taxes (EBIT), the Company is a still aiming for a margin of seven percent over the year as a whole. However, the insolvency of LEONI's customer MG Rover must be expected to incur exceptional charges of between five and seven million euros. It is not possible at this time to state the extent to which it might be possible to offset these charges during the current financial year. [...]

Fig. 2. Figure highlights statistically positive and negative sentences in an exemplary ad hoc announcement. Here, positive sentences are colored in light gray, whereas negative sentences are colored in dark gray.

signed a negative polarity. Table 2 shows the number of occurrences of positive and negative sentences in our dataset together with the resulting market reaction. Specifically, we see that positive news items contain 78.07% positive sentences and 21.93% negative sentences. In contrast, news items with a negative market reaction contain only 67.68% positive sentences and 32.32% negative sentences.

Interestingly, we observe that most ad hoc announcements consist of a combination of positive and negative aspects. Specifically, out of all documents, 83.30% contain both positive and negative sentences. In addition, 16.60% of all documents contain only positive sentences, while 0.10% consist solely of negative sentences. We find two possible explanations for this high proportion of positive sentences overall: (1) the document labels feature a positive mean abnormal return, and (2) negative sentences in financial news typically exhibit a greater length compared to positive sentences. We will thoroughly analyze these structural differences in the following sections.

4.2. Illustrative example

We now present an example of how our expert system can aid investors and practitioners in the finance domain. For this purpose, Fig. 2 shows an excerpt of an ad hoc announcement from the cable and harnessing manufacturing firm LEONI AG. This announcement was published on May 12, 2005 and led to an abnormal return of -4.60% at the end of the trading day. The announcement consists of both positive and negative sections. While the positive sections describe increases in net income and margin expectations, the negative sections describe lower expectations for future growth rates and the insolvency of a customer.

According to Fig. 2, our classifier identifies all positive and negative parts correctly, including negated text fragments. Interestingly, applying traditional bag-of-words approaches would be misleading in this case. For instance, because of a failure to account for context, the first negative sentence would be classified positively, as it contains many positive words, such as *"strong"*, *"possible"*, or *"growth"*. Overall, this example illustrates the challenges of accurate sentence classification in financial news. The identification of positive meaning is highly context-dependent and can result in entirely different interpretations when relying solely on word frequencies. As a remedy, our method can process complex sentences while preserving context and order of information. In addition, our model is solely trained on an objective response variable and thus, adapts to domain-specific particularities of the given prose.

4.3. Predictive performance on manually labeled dataset

We now evaluate the predictive performance of our method on a manually labeled dataset. For this purpose, we use a disjunct dataset that is labeled by three external subjects with a background in finance. The actual label is determined by majority vote. The dataset[7] consists of 1000 randomly drawn sentences from ad hoc announcements with an equal number of 500 positive and 500 negative sentences. We use this dataset to compare the predictive power of our approach to several baseline methods. First, we employ common sentiment dictionaries for polarity detection, namely the Harvard IV dictionary (Stone, 2002) and the Loughran-McDonald dictionary (Loughran & McDonald, 2011), the latter of which was developed for finance-specific texts. These dictionaries are a frequent choice when it comes to sentiment analysis of financial news (e.g., Garcia, 2013; Loughran & McDonald, 2013; Price, Doran, Peterson, & Bliss, 2012; Tetlock, Saar-Tsechansky, & Macskassy, 2008). Second, we utilize common machine learning classifiers for text categorization, i.e., logistic regression, a support vector machine and a feed forward artificial neural network[8] Third, we train the traditional machine learning models on bag-of-words feature representations. We train all of these models using the dataset that is used in the previous sections.

Table 3 compares the predictive performance of our approach with the baselines. As a first result, we note that the frequently employed dictionaries are not suitable for sentence-level sentiment analysis of financial news. In fact, the last column of Table 3 reveals that the Harvard IV dictionary classifies 22.67% of all sentences as neutral. We observe a similar pattern for the finance-specific Loughran-McDonald dictionary which assigns 53.00% to a neutral category. There are two reasons for this finding: first, dictionary-based approaches assign a neutral categorization if the number of positive polarity words equals the number of negative polarity words. Second, the polarity dictionary does not contain any of the words in a given sentence.

The traditional machine learning models based on bag-of-words representations show a superior performance compared to the dictionary-based approaches. The best performing variant (artificial neural network) yields a predictive accuracy of 58.30% on the manually labeled dataset. Table 3 also shows that sentence classification results in a drastically higher predictive accuracy when being trained on distributed text representations instead of bag-of-words feature representations (accuracy of 66.10%). All of the aforementioned approaches are, however, surpassed by multi-instance learning which yields an accuracy of 71.20%. This result exceeds the accuracy of the baseline methods by at least 5.10 percentage points.

[7] For reasons of reproducibility, the dataset is publicly available via https://github.com/InformationSystemsFreiburg/SentenceLevelSentiment.

[8] We optimize the hyperparameters of these classifiers with a grid search over a discrete parameter space. Further details regarding the model tuning are provided in the supplementary materials.

Table 3
Predictive performance on manually-labeled dataset.

Method	Accuracy	Recall	Precision	F_1-Score	Neutral
Baseline: Dictionaries					
Harvard IV	48.00%	75.33%	48.71%	59.16%	22.67%
Loughran-McDonald	31.67%	25.33%	29.00%	27.05%	53.00%
Baseline: Bag-of-words					
Logistic regression	55.40%	60.40%	54.91%	57.52%	–
Support vector machine	56.40%	63.00%	55.65%	59.10%	–
Artificial neural network	58.30%	55.80%	58.74%	57.23%	–
Distributed text representations					
Logistic regression	64.90%	76.80%	62.04%	68.63%	–
Support vector machine	65.60%	65.80%	65.54%	65.66%	–
Artificial neural network	66.10%	75.80%	63.48%	69.10%	–
Multi-instance learning	71.20%	68.20%	72.55%	70.31%	–

Table 4
Out-of-sample predictive performance at document-level.

Method	Accuracy	Recall	Precision	F_1-Score	Neutral
Baseline: Dictionaries					
Harvard IV	50.00%	99.64%	50.00%	66.59%	0.27%
Loughran-McDonald	51.19%	39.23%	51.56%	44.56%	9.95%
Baseline: Bag-of-words					
Logistic regression	53.38%	45.26%	54.03%	49.26%	–
Support vector machine	53.28%	63.87%	52.71%	57.76 %	–
Artificial neural network	54.20%	61.50%	53.66%	57.31%	–
Distributed text representations					
Logistic regression	55.93%	57.30%	55.77%	56.53%	–
Support vector machine	56.48%	56.93%	56.42%	56.68%	–
Artificial neural network	57.21%	64.23%	56.32%	60.02%	–
Multi-instance learning	56.57%	75.55%	54.76%	63.50%	–

This difference is statistically significant ($p < 0.001$) according to a Mc-Nemar test (Dietterich, 1998).

4.4. Predictive performance at document-level

Next, we evaluate the performance of our model as a document-level classifier. For this purpose, we compare the document-level predictions of our method with the document labels, i.e., the abnormal returns. As a first step, we split our dataset of ad hoc announcements according to a 80:20 ratio for training and testing, respectively (Kraus & Feuerriegel, 2017). This procedure precludes learning anomalies based on information which would only be available ex-post. To facilitate the interpretation of the results, we account for the imbalance between the positive and negative class in our test set by under-sampling the majority i.e., the positive class (Drummond & Holte, 2003). After this step, the balanced test dataset for document-level prediction consists of 1096 documents. Subsequently, we compare the results of our method with the same baseline classifiers from the previous sections, i.e., dictionary-based approaches and machine-learning methods.

According to Table 4, the best performing bag-of-words method (artificial neural network) yields an accuracy of 54.20% on out-of-sample documents. Distributed text representations increase the document-level performance by 3.01 percentage points to an accuracy of 57.21%. Unsurprisingly, we observe a slightly (0.64%) lower document-level performance for the multi-instance learning model. The difference in prediction accuracy is, however, not statistically significant according to the Mc-Nemar test. Interestingly, we see that this sentence-level classifier still outperforms the best performing bag-of-words classifier for document-level prediction. Hence, our approach presents a viable alternative that competes well with traditional models at document-level but, at the same time, achieves higher predictive performance at sentence-level. Moreover, we see that the method is capable of successfully transferring information from document-level to sentence_level, and back again from sentences to documents.

4.5. Robustness checks

We perform several robustness checks to validate our results. We briefly highlight the key takeaways in the following section. Detailed results are provided in the supplementary materials to this paper.

First, we repeat our analysis from the previous sections using nominal returns instead of abnormal returns. We achieve qualitatively equal results and our multi-instance learning approach outperforms all alternative approaches by at least 3.8 percentage points. Second, we examine the suitability of bag-of-words approaches in combination with term frequency-inverse document frequency (tf-idf) instead of term frequencies (tf). We find that the tf-idf feature representation has a slightly higher predictive performance than the tf representation. However, the predictive performance is still inferior to the approaches that use distributed text representations. We further note that the predictive performance using tf-idf features is at least 9.6 percentage points lower than the performance of our multi-instance learning approach.

Third, we study whether the document-level predictive performance of our approach differs with regards to documents of different length. We use the number of sentences in ad hoc announcements to split the documents into two subsets: all ad hoc announcements that are shorter or equal to the median length are considered *short* documents and all ad hoc announcements that are longer than the median length are considered *long* documents. Again, we balance each subset by under-sampling the majority class. We find that the predictive performance on long documents

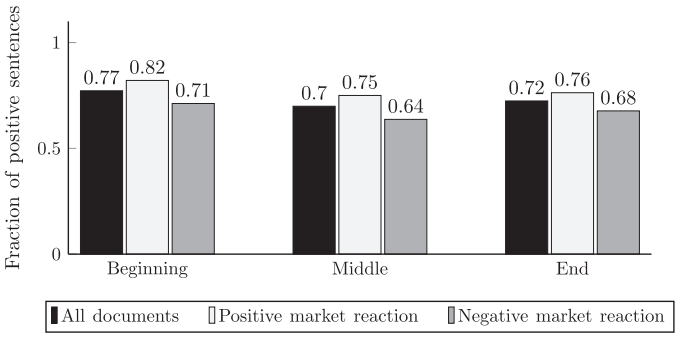

Fig. 3. Fraction of positive sentences among different document parts of ad hoc announcements.

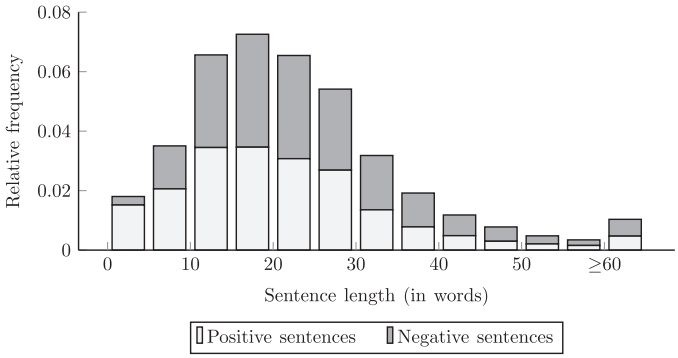

Fig. 4. Relative frequency of sentences for different levels of words included. Positive sentences are displayed using light gray colored bars, whereas negative sentences are displayed using a dark gray color.

exceeds the performance on short documents by 4.63 percentage points. We find a similar pattern for the baseline methods from the previous analyses. This suggests that the stock market reaction to ad hoc announcement is easier to predict if the documents are longer.

Fourth, we analyze whether the predictive performance at document-level depends on the market capitalization of the company that has released the ad hoc announcement. For this purpose, we rank each stock for market capitalization (in USD) and assign it to two different categories. News items from companies with ranks higher than the median market capitalization are categorized as news from large-cap stocks, whereas news from companies with ranks lower or equal to the median market capitalization are categorized as news from small-cap stocks. Here, we find no statistically significant difference in the predictive performance of our method.

5. Supplementary analyses

Inferring sentence-level polarity labels from financial news items is also a valuable tool for analyzing their structure and the effects of managerial impression techniques. This section demonstrates multiple interesting examples of how our method can be used to study investor communication on a fine-grained basis.

We start by investigating the positioning of positive and negative information in financial news. To make the structure of the documents comparable, we normalize the length of each disclosure to a range between 0 and 1. Fig. 3 shows the fraction of positive sentences for the beginning, middle and end of the document with respect to the stock market reaction. Here, a value of 0.5 indicates a neutral part of an ad hoc announcement, i.e., that 50% of all sentences are classified as positive and 50% as negative.

According to the figure, the fraction of positive sentences is at its peak at the beginning of the document and at its lowest in the middle of the document. This pattern also holds if we consider only positive or negative ad hoc announcements. To validate these findings, we perform three χ^2 tests in which we compare the fraction of positive sentences within the individual document parts. We find that all χ^2-values are significant with $p < 0.001$. Thus, we find strong evidence that companies are more likely to place positive information in financial news at the beginning or at the end of an ad hoc announcement.

Next, we analyze how the length of sentences in ad hoc announcements is related to a positive or negative polarity. Fig. 4 visualizes the relative frequency of sentences in terms of the number of words included. Here, positive sentences are displayed using light gray colored bars, whereas negative sentences are displayed using a dark gray color. According to the figure, short sentences are more likely to feature a positive polarity, whereas longer sentences are more likely to feature a negative polarity. The mean length of positive sentences is 23.16 words, whereas the mean length of negative sentences is 26.32 words. In order to validate this finding, we perform a t-test to examine whether the difference in the mean length of sentences is significantly greater than zero. We observe a t-value of 24.09, which is significant with $p < 0.001$. Accordingly, we reject the null hypothesis of no difference between means and find strong evidence that negative sentences exhibit a greater length as compared to positive sentences.

Finally, we analyze the role of linguistic negations in financial news. For this purpose, we compare the number of occurrences of negation words in sentences with a positive or negative polarity. Here, we use a list of explicit negations as proposed by Jia, Yu, and Meng (2009). Similar to Pröllochs, Feuerriegel, and Neumann (2016a), we find that 5.61% of all sentences in ad hoc announcements contain at least one negation word. In addition, we see that 2.89% of all positive sentences contain negations, whereas this number amounts to 13.14% for negative sentences. This difference is statistically significant with $p < 0.001$ when performing a χ^2-test (χ^2-value of 3524.7). Thus, we find strong support that sentences with a negative polarity contain a higher number of negations as compared to positive sentences.

It is worth noting that these results also largely coincide with previous psychological research. For example, the *serial-position effect* suggests that that senders of information are more likely to place negative content in the middle of a text (Legg & Sweeny, 2014). However, our evidence is collected in an automated manner outside of an artificial laboratory setting.

6. Discussion

This section discusses the implications of our study for research and practice. Our study not only provides an expert system to assist investors in their decision-making, but is also highly relevant for investor relations departments when they publish financial information. In addition, this section addresses the limitations of our study.

6.1. Implications for research

This work leads to multiple implications for research on expert systems that allow for an automatic processing of financial news. First, it shows that common bag-of-words approaches that count positive and negative term-frequencies are not adequate for analyzing financial news on a fine-granular basis. Corresponding inferences for individual sentences result in lower explanatory power and predictive performance. This also coincides with Li (2010a), who suggests that the *"dictionary-based approach ignores the context of a sentence"* (p. 4). As a remedy, we propose the use of dis-

tributed word representations and multi-instance learning to infer sentences with a positive or negative polarity. By incorporating context and domain-specific features, this methodology can be used to generate insights for individual text fragments in the presence of a document label, such as stock market returns. Our analysis on a manually-labeled dataset demonstrates that our approach outperforms all baselines for sentence-level polarity prediction of financial news, including the ubiquitous Loughran-McDonald dictionary and traditional machine learning approaches.

Furthermore, our method serves as a powerful tool for future research to study communication patterns in financial news. This may help to more accurately explain how financial news are perceived by investors and to infer behavioral implications. For example, our supplementary analyses have shown that companies place information strategically in different parts of the document. We find that positive content is typically placed at the beginning and at the end of financial news, whereas negative information is more likely to be placed in the middle. Concordant with Loughran and McDonald (2011), we also see that companies try to frame negative information by using sentences of greater length as compared to those sentences used to convey positive information. We find two possible explanations for this interesting finding. First, firms with losses or unstable income might need longer texts to describe their situation to investors (Li & Hitt, 2008). Second, companies might employ non-informative filler or modification words that have the potential to distract readers from negative information (Epelboim, Booth, Ashkenazy, Taleghani, & Steinman, 1997).

6.2. Implications for practitioners

The presented approach can be used to enhance the performance of existing expert systems for news-driven trading. In financial markets, it is crucial to place the orders at a fast pace (Cavalcante, Brasileiro, Souza, Nobrega, & Oliveira, 2016). The proposed approach presents an intriguing tool for improving the automated processing of financial news. For example, our approach can be integrated into graphical tools that are targeting financial professionals or private traders aiming to process large quantities of relevant information sources. Our approach could then be used to highlight positive and negative text fragments of incoming financial news. This would allow traders to gain a rapid overview of the presence of negative information, which is likely to have a stronger impact than positive information (Brown & van Harlow, 1988).

Our approach can also aid companies and investor relations departments by addressing the question of how individual text parts of their corporate communication will be perceived by investors. Among other functions, our method could be integrated into tools that would be capable of assisting authors of financial news by highlighting text fragments that are statistically perceived as positive or negative by the stock market. Hence, managers and investor relations departments can benefit from a self-reflective writing process that avoids noisy signals in their communications, thus helping to prevent unexpected stock market reactions. In a similar vein, they can use our method to analyze communication patterns in their past disclosures and monitor forms and writing styles relative to their competitors.

6.3. Limitations

Our approach faces a number of limitations. Just like any other research in the context of textual analysis of financial news, our method reveals shortcomings when additional background knowledge is needed to comprehend the context or when sections of text refer to information that is known by the market but not contained in the document. For example, consider the statement *"earnings are one percent higher than last year"*. At first glance, it seems reasonable to classify this sentence as positive, however, if the earnings from last year were very low, an increase by one percent might be interpreted as insufficient. The same applies for implicit market expectations regarding earnings results or sales numbers.

Furthermore, our approach struggles to account for subjective characteristics in the perception of investors. For instance, investors might interpret news differently depending on their information processing skills and prior beliefs (Baker & Nofsinger, 2002). Hence, further research is necessary to study the differences in the predictive value of financial news for different groups of investors and firm sectors. Ultimately, our approach cannot account for systematic changes in the writing style and structure of financial news. For example, in 1998, the SEC published a detailed handbook about how to write financial documents in "plain English".[9] Here, a potential remedy is to frequently retrain the model with the most recent data.

7. Conclusion and future research

Expert systems for the automatic processing of financial news commonly operate at the document-level by counting positive and negative term-frequencies. However, this limits their usefulness for investors and financial practitioners as it does not allow for the drawing of inferences on a more fine-grained level, e.g., individual sentences. For this purpose, we proposed a method using distributed text representations and multi-instance learning to predict the polarity of individual sentences in financial news. In contrast to previous approaches that merely predict the stock market reaction in response to news items at document-level, our method allows one to analyze financial news on a fine-grained basis. The model is solely trained based on historic stock market reactions following the publication of news items without the need for manual labeling. According to our results, the proposed approach significantly outperforms common bag-of-words approaches by at least 5.10 percentage points on a manually-labeled dataset of financial news. Hence, the method can help to improve the performance of financial expert systems and generate new knowledge regarding the reception of financial news on the stock market.

Our study provides several promising avenues for future research. First, the proposed method presents opportunities to improve the accuracy and quality of expert systems that aim at the automated processing of large quantities of financial news. For example, an intuitive question is whether the predictions from our approach can be combined with other variables from the stock market in order to enhance predictive performance. Second, we hope that the method presented in this paper will become an important tool in order for yielding novel insights into the reception of financial news on the investors' side. For example, it would be interesting to examine whether investors have a preference for certain argumentation patterns in financial news, like arguments that contain both positive and negative information. Third, future research can analyze characteristics of positive and negative sentences in financial news in more depth; and whether the perception of language differs between different markets. For example, our supplementary analysis suggests that negations are more likely to appear in sentences with a negative polarity. This idea can easily be transferred to studying the usage of past vs. present tense, "we" vs. "they" speech, or passive vs. active voice. Fourth, our approach for predicting sentence-level polarity labels in texts based on a document-level gold standard is not limited to the study of financial news, but can be easily extended to the study of reviews in recommendation systems or on retailer platforms.

[9] See https://www.sec.gov/pdf/handbook.pdf.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eswa.2020.113223.

Credit authorship contribution statement

Bernhard Lutz: Methodology, Software, Formal analysis, Writing - original draft. **Nicolas Pröllochs:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Dirk Neumann:** Resources, Supervision.

References

Allee, K. D., & Deangelis, M. D. (2015). The structure of voluntary disclosure narratives: Evidence from tone dispersion. Journal of Accounting Research, *53*(2), 241–274.

Amplayo, R. K., Lee, S., & Song, M. (2018). Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis. *Information Sciences, 454*, 200–215.

Angelidis, S., & Lapata, M. (2018). Multiple instance learning networks for fine–grained sentiment analysis. *Transactions of the Association for Computational Linguistics, 6*, 17–31.

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

Baker, H. K., & Nofsinger, J. R. (2002). Psychological biases of investors. *Financial Services Review, 11*(2), 97–116.

Benston, G. J. (1973). Required disclosure and the stock market: An evaluation of the Securities Exchange Act of 1934. The American Economic Review, *63*(1), 132–155.

Brown, K. C., & van Harlow, W. (1988). Market overreaction: Magnitude and intensity. *Journal of Portfolio Management, 14*(2), 6–13.

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L. F., Nobrega, J. P., & Oliveira, A. L. I. (2016). Computational intelligence and financial markets: A survey and future directions. Expert Systems with Applications, *55*, 194–211.

Chan, S. W. K., & Chong, M. W. C. (2017). Sentiment analysis in financial texts. *Decision Support Systems, 94*, 53–64.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, *10*(7), 1895–1923.

Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence, 89*(1), 31–71.

Drummond, C., Holte, R. C., et al. (2003). C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In Proceedings of the 10th International Conference on Machine Learning (pp. 1–8).

Epelboim, J., Booth, J. R., Ashkenazy, R., Taleghani, A., & Steinman, R. M. (1997). Fillers and spaces in text: The importance of word recognition during reading. Vision Research, *37*(20), 2899–2914.

Feuerriegel, S., & Pröllochs, N. (2018). Investor reaction to financial disclosures across topics: An application of latent Dirichlet allocation. *Decision Sciences, forthcoming*.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance, 68*(3), 1267–1300.

García-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2vlda: Almost unsupervised system for aspect based sentiment analysis. Expert Systems with Applications, *91*, 127–137.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. Decision Support Systems, *50*(4), 680–691.

Gunduz, H., & Cataltepe, Z. (2015). Borsa istanbul (bist) daily prediction using financial news and balanced feature selection. Expert Systems with Applications, *42*(22), 9001–9011.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685–697.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication, 45*(4), 363–407.

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, financial reporting and disclosure. Journal of Accounting and Economics, *36*(1–3), 337–386.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Jia, L., Yu, C., & Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In Proceedings of the 18th Conference on Information and Knowledge Management (pp. 1827–1830).

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis, 33*(1), 171–185.

Kelly, S., & Ahmad, K. (2018). Estimating the impact of domain-specific news sentiment on financial assets. *Knowledge-Based Systems, 150*, 116–126.

Kotzias, D., Denil, M., de Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. In Proceedings of the 21st International Conference on Knowledge Discovery and Data Mining (pp. 597–606). ACM.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, (104), 38–48.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. In Proceedings of the 1st Workshop on Representation Learning for NLP (pp. 78–86). ACL.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine learning (pp. 1188–1196).

Legg, A. M., & Sweeny, K. (2014). Do you want the good news or the bad news first? The nature and consequences of news order preferences. Personality and Social Psychology Bulletin, *40*(3), 279–288.

Lerman, A., & Livnat, J. (2010). The new form 8-K disclosures. Review of Accounting Studies, *15*(4), 752–778.

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.

Li, X., & Hitt, L. M. (2008). Information Systems Research, *19*(4), 456–474.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. Mining Text Data (pp. 415–463). Springer.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. The Journal of Finance, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. Journal of Financial Economics, *109*(2), 307–326.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance, 69*(4), 1643–1671.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54*(4), 1187–1230.

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. T., & Camón-Herrero, J. B. (2011). Financial news semantic search engine. *Expert Systems with Applications, 38*(12), 15565–15572.

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature, 35*(1), 13–39.

Manning, C. D., Schütze, H., et al. (1999). *Foundations of statistical natural language processing*: Vol. 999. MIT Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (pp. 55–60).

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, *106*, 36–54.

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

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, *2*(1), 1–135.

Peng, H., Ma, Y., Li, Y., & Cambria, E. (2018). Learning multi-grained aspect target sequence for chinese sentiment analysis. *Knowledge-Based Systems, 148*, 167–176.

Pham, D.-H., & Le, A.-C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data & Knowledge Engineering, 114*, 26–39.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., de Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (pp. 19–30).

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance, 36*(4), 992–1011.

Pröllochs, N., & Feuerriegel, S. (2018). Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling. *Information & Management, forthcoming*.

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2016a). Detecting negation scopes for financial news sentiment using reinforcement learning. In *Proceedings of the 49th hawaii international conference on system sciences* (pp. 1164–1173).

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2016b). Is human information processing affected by emotional content? Understanding the role of facts and emotions in the stock market. In Proceedings of the 37th International Conference on Information Systems.

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2018). Statistical inferences for polarity identification in natural language. *PloS ONE, 13*(12), 1–21.

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2019). Learning interpretable negation rules via weak supervision at document level: A reinforcement learning approach. In Proceedings of the 2019 Conference of the North American Chapter

of the Association for Computational Linguistics: Human Language Technologies (pp. 407–413). ACL.
Qiu, J., Liu, C., Li, Y., & Lin, Z. (2018). Leveraging sentiment analysis at the aspects level to predict ratings of reviews. Information Sciences, *451–452*, 295–309.
Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14–46.
Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing & Management, 45*(5), 571–583.
Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458–464.
Zhou, Z.-H., & Zhang, M.-L. (2003). Ensembles of multi-instance learners. In Proceedings of the 14th European Conference on Machine Learning (pp. 492–502).
Stone, P. J. (2002). General inquirer harvard-iv dictionary.
Stiborek, J., Pevný, T., & Rehák, M. (2018). Multiple instance learning for malware classification. Expert Systems with Applications, *93*, 346–357.
Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., & Honeine, P. (2019). Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications, *117*, 103–111.
Tang, T., Fang, E., & Wang, F. (2014). Is neutral really neutral? The effects of neutral user-generated content on product sales. *Journal of Marketing, 78*(4), 41–58.
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., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance, *63*(3), 1437–1467.
van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications, 42*(11), 4999–5010.
Wang, S., Zhe, Z., Kang, Y., Wang, H., & Chen, X. (2008). An ontology for causal relationships between news and financial instruments. Expert Systems with Applications, *35*(3), 569–580.
Xue, W., & Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 2514–2523).
Zhang, Y., Swanson, P. E., & Prombutr, W. (2012). Journal of Financial Research, *35*(1), 79–114.