“Boom” or “Ruin” – Does it Make a Difference? Using Text Mining and Sentiment Analysis to Support Intraday Investment Decisions

Michael Siering
Goethe-University Frankfurt
siering@wiwi.uni-frankfurt.de

Abstract
Investors have to deal with an increasing amount of information in order to make beneficial investment decisions. Thus, text mining is often applied to support the decision-making process by predicting the stock price impact of financial news. Recent research has shown that there exists a relation between news article sentiment and stock prices. However, this is not considered by previous text mining studies. In this paper, we develop a novel two-stage approach that connects text mining with sentiment analysis to predict the stock price impact of company-specific news. We find that the combination of text mining and sentiment analysis improves forecasting results. Additionally, a higher accuracy can be achieved by using finance-related word lists for sentiment analysis instead of a generic dictionary.

1. Introduction

Financial markets are considered to be complex and to be changing rapidly [8]. For example, financial research shows that stock prices adjust quickly on new information like dividend announcements or other company-related news [9, 24]. In this context, the sentiment expressed in financial news articles is of great importance, too. Different studies provide evidence that it has an impact on the following stock price reactions. For instance, the prevalence of a negative sentiment can lead to a decline in stock prices [18, 31].

Private as well as institutional investors need to react quickly on the publication of company-related news to be able to take profit of possible stock price adjustments. However, they are confronted with a large amount of information that has to be analyzed properly to make favorable investment decisions. To assist this process, decision support systems can be used. Recent studies show that unstructured information in the form of news articles can be a valuable input for these systems to predict future stock market movements [10, 23].

The artifacts which are proposed to analyze unstructured information are usually based on text mining approaches [10, 11, 23]. That is, they use algorithms and methods from the field of machine learning to find patterns in texts that can serve as a basis for predictions [13]. However, these approaches do not adequately take into account the sentiment expressed within news articles. For example, current text mining approaches do not distinguish between opinionated words like “boom” or “ruin” which represent a positive or negative sentiment and words like “turnover” which are not related to a positive or negative sentiment at all.

Nevertheless, the availability of a sentiment measure would allow for a better understanding of financial news articles, especially against the background of the relation between news article sentiment and stock returns [18, 30, 31]. This is also relevant in the field of machine learning. Previous studies have shown that such additional information can be taken into account in machine learning setups to improve forecasting results [17, 25].

As a consequence, we investigate whether sentiment analysis improves the predictability of stock price changes after the publication of financial news articles (research question one). For that purpose, we introduce a novel two-stage approach to combine text mining and sentiment analysis. At first, for each news article, we calculate a sentiment measure based on a dictionary with opinionated words. Second, according to the sentiment measure, we select a local classifier which is trained on documents with the same kind of sentiment to forecast the subsequent price reactions. The results of this two-stage approach are compared with the predictions of a global classifier that is trained on documents containing all kinds of sentiment.

In general, sentiment analysis can be conducted on the basis of different dictionaries containing opinionated words. In the case of corporate disclosures, Loughran and McDonald [18] show that the use of a
domain-specific dictionary can improve results. To our knowledge, this has not been verified for financial news articles yet. Hence, we also investigate whether the use of a domain-specific dictionary instead of a generic dictionary for sentiment analysis improves the predictability of stock price changes after the publication of financial news articles (research question two).

To examine these research questions, the remainder of this paper is organized as follows. In section 2, we briefly describe the previous research on text mining and sentiment analysis in the financial domain. Thereafter, section 3 illustrates our study setup including the text mining approach and the calculation of the sentiment measure. Section 4 presents the results of the classical machine learning evaluation, whereas section 5 provides a domain-specific evaluation in the form of an investment strategy. Finally, Section 6 concludes and gives an outlook on further research purposes.

2. Literature Review

2.1. The use of text mining to predict stock price changes

Several studies apply text mining to predict stock price changes caused by the publication of company-related news. Theoretically, this stream of research relies on the semi-strong form of the efficient market hypotheses (EMH) which states that stock prices adjust quickly to the arrival of new publicly available information [9]. In practice, this process does not proceed immediately [24], which enables investors to take profit from these price reactions. For that purpose, several approaches are proposed that differ concerning the financial instruments of interest, the documents analyzed and the horizon for which a forecast is made.

Once a day, Wuthrich et al. [34] browse a number of web pages and crawl news articles containing financial analyses and stock market information. Taking into account the content of these articles, the corresponding daily price reaction of five stock indices is predicted. For evaluation purposes, it is shown that an investment strategy following the predictions generates higher returns than investing in the subsequent index. In contrast, Mittermayer [23] focuses on single stocks. Within his study, press releases are analyzed in order to predict the stock price reaction within a timeframe of one hour. Additionally, an evaluation in the form of an investment strategy is provided. It is revealed that investing according to the predictions leads to higher returns than buying or selling the stocks randomly.

In comparison to [34] as well as [23], Groth and Muntermann [11] do not analyze news articles that are published voluntarily. Instead, they concentrate on corporate disclosures which have to be published due to legal regulations. After an analysis of related financial studies, it is stated that stock prices react quicker than it was assumed in previous text mining studies. As a result, for predictions of single stocks’ price changes, a timeframe of 15 minutes is used. Concerning the evaluation, the results are similar to [34] and [23]. Schumaker and Chen [29] provide additional evidence that unstructured information can be used to forecast stock price movements which are caused by the publication of financial news articles.

A recent study by Geva and Zahavi [10] combines structured and unstructured information to forecast price changes caused by news items that were published on the web. Therefore, the news articles are also combined with structured information like previous stock returns. Geva and Zahavi [10] predict whether the stock return at the end of the trading day exceeds the S&P500 index by more than 1% or not. They find that the combination of structured and unstructured information as inputs for a decision model can improve the forecasting results.

Although the combination of unstructured and structured information is a first step to improve the performance of classical text mining setups, none of these approaches performs a deeper analysis of the documents. In particular, the sentiment of these documents is not taken into account even though recent studies show that this can be a valuable addition to explain stock price changes [18, 30, 31].

2.2. Sentiment analysis in the financial domain

Sentiment analysis encompasses the investigation of documents like news articles, message board postings or product reviews in order to determine their tone concerning a certain topic [27, 26]. In the financial context, the sentiment which is expressed in documents like news articles covers the opinions, expectations or beliefs of market participants towards certain companies or towards certain financial instruments [4].

In general, there are two broad strategies to perform sentiment analysis: it can be distinguished between supervised and unsupervised approaches [36]. Supervised approaches require a dataset which is manually labeled according to the documents’ sentiment. This dataset is used to train a classifier which thereafter can be applied to determine the sentiment of further documents or sentences. Unsupervised approaches rely on external knowledge such as predefined dictionaries which provide lists of
words that are connected with a positive or negative sentiment. These word lists are usually created manually with a couple of precoded terms and are used to determine a sentiment measure [32].

A supervised approach is conducted by Antweiler and Frank [1] who collect messages posted on two finance message boards. At first, they manually determine the sentiment of a sub-sample of 1,000 messages and use these messages to train a classifier. Afterwards, this classifier is used to assess the sentiment of the remaining messages. The authors find that a disagreement in sentiment among the messages leads to an increase in the number of trades. Additionally, they observe that the number of messages posted during a day can help to predict the stock returns during the following day. Das and Chen [6] follow a similar approach and investigate messages which are published on stock message boards, too. Like Antweiler and Frank [1], they use a manually labeled sub-sample of messages to train different classifiers and subsequently to classify messages according to their sentiment. Next, these classified messages are used to calculate an overall sentiment index. Das and Chen [6] find that the level of this sentiment index has explanatory power for the level of the corresponding stock index. In contrary to this result, they only find weak evidence that the sentiment concerning an individual stock can forecast daily stock price movements.

Apart from these studies, there are also works applying unsupervised approaches. For example, Tetlock [30] analyses the sentiment of a daily wall street journal column. For that reason, he uses the General Inquirer’s Harvard-IV-4 classification dictionary to classify each word of the column according to its sentiment. Afterwards, he uses these classified words to calculate a pessimism factor. The study finds that high pessimism leads to a decline of market prices. Additionally, an abnormal high or low level of pessimism is supposed to predict high trading volumes. Tetlock et al. [31] conduct a similar study. On a daily base, they analyze the news stories published in the Wall Street Journal and in the Dow Jones News Service. Using the Harvard-IV-4 dictionary, they determine the fraction of negative words per news story and find that stock prices are related to this measure. Loughran and McDonald [18] evaluate the sentiment of 10-K company reports. For that reason, they develop different word lists containing positive and negative terms. Subsequently, they calculate the number of negative words per report to determine a negativity measure. They find that in the case of 10-K reports, the negativity measure based on their word lists provides a better explanation for the stock returns during the following days than the negativity measure based on the Harvard-IV-4 dictionary.

The studies presented above provide evidence that sentiment expressed in documents like news articles or message board postings is related to stock returns. However, this influence is too low to serve as a sole source for forecasting future stock returns [1]. Additionally, stock price changes after the publication of news articles occur within minutes rather than days. As a consequence, the combination of an intraday text mining approach with sentiment analysis seems to be promising.

3. Combining text mining and sentiment analysis to predict stock price changes

3.1. General setup

Figure 1 shows the general setup of our study considering the application of text mining and sentiment analysis to forecast stock price movements after the publication of financial news articles.

![Figure 1. Study setup](image)

At first, we acquire a dataset which consists of a large number of financial news articles. To be able to conduct supervised learning, each news article is then labeled according to its stock price impact. Thereafter, a dictionary containing opinionated words is used to determine the news article sentiment. Thereafter, two major text mining procedures are conducted [7]: Subsequent to a document pre-processing step, we make use of one out of three local
3.2. Dataset description

In total, our news article dataset is composed of 11,518 news articles which have been acquired from Dow Jones News. Each news article is published in English and is related to one out of 30 stocks that are constituents of the German blue chip index DAX. The news articles were published from 2010-04-06 until 2011-04-08. Additionally, each news article is assigned to a specific stock by the news provider and has a timestamp which is exact to the second.

To be able to conduct supervised learning, every news article has to be labeled according to the price change after its publication. Consequently, we make use of Thomson Reuters Tick History to acquire the respective tick by tick price data of the trades which were carried out on the electronic securities trading system Xetra. As these prices are only available during the trading hours, we exclude all news articles which have not been published from 9 am to 5 pm. Furthermore, if there has been more than one news article published within the forecast period of 15 minutes, we exclude all news articles that have been published within this timeframe. This avoids possible interferences. Finally, if a news article does not have enough predecessors which can serve as a basis for sentiment calculation, it is excluded, too. As a result, there are 2401 news articles remaining which are used within this study.

3.3. Labeling

For supervised learning, each news article is labeled according to the observed stock price reaction following its publication. At first, the return measure $R_{it}$ can be used to determine the price change of stock $i$ within $t$ minutes after publication:

$$ R_{it} = \frac{P_{it} - P_{i0}}{P_{i0}} $$

(1)

Formula 1 takes into account $P_{i0}$, which is the price of stock $i$ at the time of publication and $P_{it}$, representing the price $t$ minutes after publication. Previous studies find that news is often reflected within stock prices in 15 minutes, so we choose $t = 15$ [24].

Apart from the reaction on company specific news, stock prices can also be influenced by overall market trends and by market wide events like the announcement of macroeconomic news [21]. As these influences on the stock price are also included in $R_{it}$, this measure has to be adjusted before it can be used for proper labeling. For that purpose, a market model is used to estimate “the expected return without conditioning on the event taking place” [19]. The market model relates the past returns of stock $i$ $(R_{it})$ during the period $t$ to the past returns of the market portfolio $m$ $(R_{mt})$ and determines a linear relation [19]:

$$ R_{it} = \alpha_i + \beta_i R_{mt} + \varepsilon_{it} $$

(2)

$\alpha_i$ and $\beta_i$ are stock-specific parameters which are estimated with regression analysis, $\varepsilon_{it}$ denotes the error term. The regression considers the returns of the 50 trading days preceding the publication of the news article. Thereby, the return of the market portfolio is represented by the return of the DAX index.

Next, $\alpha_i$ and $\beta_i$ can be used to calculate the “normal” return which would have been realized if no news article was published. If this normal return is subtracted from $R_{it}$, the part of the stock price change which is caused by the publication of the financial news article is obtained. This measure is denoted as abnormal return and expresses the stock market reaction on the publication of firm specific news [20]. The measure is depicted in formula 3:

$$ AR_{it} = R_{it} - (\alpha_i + \beta_i R_{mt}) $$

(3)

Having completed these steps, it is possible to label each news article according to the corresponding abnormal return. In doing so, we make use of two classes, i.e. “negative” and “positive”. The class “negative” is assigned to a news article if the calculated abnormal return is lower than zero. In all other cases, the class “positive” is assigned. Such a two-class approach is common in text mining studies and is also applied by [10–12].
3.4. Determination of news article sentiment

To determine the sentiment expressed in the financial news articles, we decide to follow an unsupervised dictionary-based approach. Thus, several word lists containing positive and negative words are used to analyze the documents and to calculate a sentiment measure. In comparison to assessing the sentiment with a supervised machine learning-based approach, no additional classifier and no manual labeling of news articles according to their sentiment is necessary. Additionally, the results provided by Tetlock [30] as well as Loughran and McDonald [18] provide evidence that a dictionary-based approach is appropriate. In the following, we make use of two dictionaries, each providing positive and negative word lists. On the one hand, we adapt the word lists from the Harvard-IV-4 classification dictionary. These lists are often used by related studies for sentiment determination [18, 30, 31]. On the other hand, we make use of two word lists which are provided by Loughran and McDonald [18]. These lists are tailored to the financial domain and contain positive and negative terms, too. In the following, these dictionaries are denoted as H-IV-4 and FIN.

For news article sentiment determination, we first make use of a dictionary and count the occurrences of positive and negative words. Therefore, the positive and negative word lists are compared with each news article. Consistent with Loughran and McDonald [18], we consider negations: if there is a negation preceding a word which is identified as positive or negative, its interpretation is reversed, which means that originally positive words are counted as negative, et vice versa.

Next, we adapt a document-level sentiment measure which is depicted in formula 4 [31, 35]:

\[
Sent = \frac{pos - neg}{pos + neg}
\]  

(4)

This measure takes into account \(pos\), which represents the number of positive words and \(neg\), which represents the number of negative words, both calculated as described above. If a news article contains neither a positive nor a negative word, the measure is defined as zero. A positive (negative) value of \(Sent\) indicates a prevailing number of positive (negative) words and consequently a positive (negative) sentiment polarity. Additionally, the value of \(Sent\) represents the strength of the sentiment which is expressed in the news article.

Furthermore, we follow Tetlock et al. [31] and normalize \(Sent\) through subtracting its mean \((\mu_{Sent})\) and through dividing by its standard deviation \((\sigma_{Sent})\), taking into account the five preceding news articles concerning the same stock:

\[
sent = \frac{Sent - \mu_{Sent}}{\sigma_{Sent}}
\]

(5)

On the one hand, normalization is advantageous since it allows comparing sentiment measures which are calculated on the basis of different dictionaries [3, 31]. This is important because the H-IV-4 and the FIN dictionary each contain a different amount and an unequal ratio of positive and negative terms. On the other hand, \(sent\) takes into account the sentiment of a news article compared to the mean sentiment of the preceding news articles. If \(sent\) is extremely high or low, it indicates that the actual news article has a sentiment which is very different compared to the previous news articles’ sentiment. As described above, we exclude all news articles for which too few predecessors are available to calculate the sentiment measure. This is the reason why we do not calculate the mean and the standard deviation taking into account a larger number of preceding news articles.

According to \(sent\), we split the whole news article dataset (denoted as COMPL) into three subsets: POS, NEG and NEUT. The POS-subset consists of the news articles containing the most positive sentiment (25% of COMPL). Correspondingly, the NEG-subset consists of the news articles with the most negative sentiment (25% of COMPL). The remaining 50% are considered to express a neutral sentiment and are contained in the NEUT-subset. This segmentation was chosen because it separates the different subsets quite well, as shown in table 1. The mean values and standard deviations (STDEV) of \(sent\) are reported for the H-IV-4-setup (column 1) and for the FIN-setup (column 2).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>H-IV-4-setup</th>
<th>FIN-setup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STDEV</td>
</tr>
<tr>
<td>POS</td>
<td>1.1814</td>
<td>0.2693</td>
</tr>
<tr>
<td>NEUT</td>
<td>0.0221</td>
<td>0.4275</td>
</tr>
<tr>
<td>NEG</td>
<td>-1.1919</td>
<td>0.3018</td>
</tr>
<tr>
<td>COMPL</td>
<td>0.0084</td>
<td>0.9146</td>
</tr>
</tbody>
</table>

3.5. Document pre-processing

As classic machine learning techniques are not able to deal with plain texts, we perform two generally accepted text pre-processing steps. These are feature extraction and selection as well as feature representation [2, 33].

The step feature extraction and selection aims at identifying a set of features which represents the individual documents [33]. The features are created by splitting every document into single words, whereas each word is used as a feature. Because the documents also contain words with little meaning like “the” or
“a”, we use a stop word list to remove these words before generating the features. Additionally, a porter stemmer [28] is applied to transform the features to their grammatical roots.

The number of features which is obtained after these actions is still numerous, especially for a large number of documents. To reduce the computational effort in the further text mining process, we sort the feature set by the corresponding information gain and, in line with Geva and Zahavi [10], select the top 500 features to represent the documents.

The step feature representation creates a document-feature matrix. Within this matrix, each document is represented by its features. These features are weighted by the tf-idf measure which is common for feature representation. Tf-idf takes into account the term frequency, which is the number of appearances of a term within one document. Additionally, it also considers the inverse document frequency, which includes the number of documents a feature is included in [13].

3.6. Classification

In this study, we make use of support vector machines (SVM) since related studies have shown that SVMs are a good choice for document classification [12, 14]. SVMs represent a machine learning algorithm that usually distinguishes between two different classes. For that purpose, each document including its associated class is represented as a data point in the feature space. Afterwards, a maximum margin hyperplane is constructed that maximizes the distance between itself and the representatives of both classes [16]. As a result, a new document can be classified by investigating on which side of the hyperplane its data point falls on [27].

If it is not possible to linearly separate the data points, they can be transformed to a higher dimensional space where linear separation is possible again. In this context, kernel functions can be used to reduce computational effort. According to previous studies and because of computational efficiency, we make use of a linear kernel [12]. As described above, three local and one global classifier are trained both for the H-IV-4-setup and for the FIN-setup.

4. Domain-independent evaluation

4.1. Domain-independent evaluation setup

In general, the performance of a classifier can be assessed by analyzing whether its classifications of a test dataset are correct or not. In this context, it is important to ensure that the classified dataset has not been used for training before. Otherwise, the results obtained during the evaluation would be too optimistic and could not be reproduced with further datasets [22].

To overcome this problem, k-fold cross validation can be used. Within k-fold cross validation, the whole dataset is split into k subsets, whereas k-1 subsets are used for training and the remaining subset is used for testing. This procedure is repeated k-times. In total, every subset is used once for testing and k-1 times for training. To be able to compare the results of the different iterations, the subsets should be stratified. This means that the proportion of the different classes remains constant across the subsets. Previous research has shown that for real-world datasets (like our news article dataset), 10-fold stratified cross validation performs best [15]. As a consequence, we choose k=10 and evaluate the classifiers by means of 10-fold stratified cross validation.

At the end of each iteration, a contingency table containing the number of correctly and incorrectly classified examples is created. These statistics are often denoted as true positives (TP) and true negatives (TN) as well as false positives (FP) and false negatives (FN). Finally, these values are summed up and are presented in a global contingency table. On this basis, different performance measures can be calculated, which is also known as micro-averaging [5]. In detail, we make use of the following measures:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (7)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (8)
\]

Thereby, accuracy measures the number of correct predictions in comparison to the number of predictions in total [16]. Next to accuracy, we also calculate precision and recall [13]. By definition, these measures are calculated for the class “positive”. However, within our study, the classes “positive” and “negative” are equally important. As a consequence, we also calculate these measures for the class “negative”.

4.2. Domain-independent evaluation results

The results of the domain-independent evaluation are reported in Table 2. Thereby, the datasets POS, NEUT and NEG are classified by the three local classifiers SVM\textsubscript{POS}, SVM\textsubscript{NEUT} and SVM\textsubscript{NEG} as well as the global classifier SVM\textsubscript{COMPL}. For the dataset COMPL, we provide two classification results: On the one hand, we provide the consolidated results (CONS)
which are obtained if each document’s sentiment is calculated and the respective local classifier is used for classification. On the other hand, the results of the global classifier SVM_\text{COMPL} are shown. In general, these results are reported twice: On the left, the results for the H-IV-4-setup are shown, whereas on the right, the results for the FIN-setup are provided. To ensure that no document is part of the training and of the test set at the same time, we make use of the classification results which are obtained during the ten-fold cross validation whenever necessary.

First, it can be noted that CONS has a higher accuracy than the global classifier SVM_\text{COMPL}. Second, the local classifiers SVM_\text{POS} and SVM_\text{NEG} also have a higher accuracy in classifying the POS respectively the NEG datasets in comparison to the global classifier. This is valid both for the H-IV-4-setup and for the FIN-setup. Precision and recall are superior, too. In contrast, the local classifiers perform worse than the global classifier if they classify datasets they are not specialized on. For example, SVM_\text{POS} performs worse than SVM_\text{NEG} and even worse than SVM_\text{COMPL} if the NEG dataset has to be classified.

For the news articles which are contained in the NEUT dataset and which subsequently have a neutral sentiment, the use of a local classifier does not improve the classification accuracy. In the case of the H-IV-4-setup, the local classifier SVM_\text{NEUT} has a slightly lower accuracy in classifying the NEUT dataset than the global classifier SVM_\text{COMPL}.

In summary, considering research question one, it can be noticed that sentiment analysis leads to an improvement in classification accuracy. This applies to the consolidated results as well as to news articles with the most positive and negative sentiment, independent of the dictionary used. As stated above, there is no improvement in the case of news articles containing a neutral sentiment. Considering research question two, it can be noticed that the consolidated results based on the FIN-dictionary are slightly better than the consolidated results based on the H-IV-4-dictionary.

5. Domain-dependent evaluation

5.1. Domain-dependent evaluation setup

Next to the domain-independent evaluation, we also conduct a domain-dependent evaluation in the form of an investment strategy based on the predictions of the classifiers. Subsequently, the performance of a classifier can be measured by the return which is achieved by the corresponding investment strategy. This evaluation approach is necessary because the domain-independent evaluation can reveal misleading results. It is possible that a classifier has high accuracy, but a corresponding investment strategy leads to poor results. This is the case if news articles causing low abnormal returns are predicted correctly but news articles causing high abnormal returns are predicted incorrectly.

Against this background, we propose the following investment strategy: At time of publication, each news article is classified according to the predicted stock price impact. If the class “positive” is assigned, the corresponding stock is bought. Thereafter, the stock is held for 15 minutes before it is sold again. In the case of news articles containing a neutral sentiment, the use of a local classifier does not improve the classification accuracy. In the case of the H-IV-4-setup, the local classifier SVM_\text{NEUT} has a slightly lower accuracy in classifying the NEUT dataset than the global classifier SVM_\text{COMPL}.

In summary, considering research question one, it can be noticed that sentiment analysis leads to an improvement in classification accuracy. This applies to the consolidated results as well as to news articles with the most positive and negative sentiment, independent of the dictionary used. As stated above, there is no improvement in the case of news articles containing a neutral sentiment. Considering research question two, it can be noticed that the consolidated results based on the FIN-dictionary are slightly better than the consolidated results based on the H-IV-4-dictionary.

Table 2. Domain-independent evaluation results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Setup based on H-IV-4 dictionary</th>
<th>Setup based on FIN-dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Class positive</td>
<td>Class negative</td>
</tr>
<tr>
<td>POS</td>
<td>SVM_\text{POS}</td>
<td>71.50</td>
<td>71.29</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{NEUT}</td>
<td>55.05</td>
<td>57.06</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{NEG}</td>
<td>53.80</td>
<td>56.83</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{COMPL}</td>
<td>63.67</td>
<td>62.78</td>
</tr>
<tr>
<td>NEUT</td>
<td>SVM_\text{POS}</td>
<td>53.46</td>
<td>59.11</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{NEUT}</td>
<td>65.11</td>
<td>65.56</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{NEG}</td>
<td>54.69</td>
<td>58.71</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{COMPL}</td>
<td>65.28</td>
<td>65.74</td>
</tr>
<tr>
<td>NEG</td>
<td>SVM_\text{POS}</td>
<td>55.12</td>
<td>60.61</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{NEUT}</td>
<td>54.78</td>
<td>60.61</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{NEG}</td>
<td>70.17</td>
<td>69.50</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{COMPL}</td>
<td>67.67</td>
<td>68.21</td>
</tr>
<tr>
<td>COMPL</td>
<td>CONS</td>
<td>67.97</td>
<td>67.98</td>
</tr>
<tr>
<td></td>
<td>SVM_\text{COMPL}</td>
<td>65.47</td>
<td>65.59</td>
</tr>
</tbody>
</table>

Accuracy (Acc.), Precision (Prec.) and Recall (Rec.) are expressed as a percentage.
change between these two points in time. Consistent with other studies, we assume zero transaction costs [17, 29]. To provide comparability with the results of the domain-independent evaluation, we make use of the predictions which are made during the 10-fold cross validation. This is also consistent with [12].

5.2. Domain-dependent evaluation results

Table 3 presents the mean returns realized by an investment strategy based on the predictions of the different classifiers. Additionally, the corresponding standard deviations are provided (STDEV). Similar to the domain-independent evaluation, the results are reported for the datasets POS, NEUT, NEG and COMPL whereas the predictions are made by the three local classifiers as well as the global classifier. Furthermore, CONS represents the consolidated results of the three local classifiers. Again, the results are provided both for the H-IV-4 as well as the FIN-setup.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>H-IV-4-setup</th>
<th>FIN-setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>SVM_POS</td>
<td>0.0624</td>
<td>0.0894</td>
</tr>
<tr>
<td></td>
<td>SVM_NEUT</td>
<td>0.0020</td>
<td>0.0055</td>
</tr>
<tr>
<td></td>
<td>SVM_NEG</td>
<td>0.0000</td>
<td>-0.0027</td>
</tr>
<tr>
<td></td>
<td>SVM_COMPL</td>
<td>0.0327</td>
<td>0.0559</td>
</tr>
<tr>
<td>NEUT</td>
<td>SVM_POS</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>SVM_NEUT</td>
<td>0.0389</td>
<td>0.0492</td>
</tr>
<tr>
<td></td>
<td>SVM_NEG</td>
<td>0.0029</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>SVM_COMPL</td>
<td>0.0392</td>
<td>0.0416</td>
</tr>
<tr>
<td>NEG</td>
<td>SVM_POS</td>
<td>0.0032</td>
<td>-0.0138</td>
</tr>
<tr>
<td></td>
<td>SVM_NEUT</td>
<td>0.0080</td>
<td>-0.0020</td>
</tr>
<tr>
<td></td>
<td>SVM_NEG</td>
<td>0.0547</td>
<td>0.0461</td>
</tr>
<tr>
<td></td>
<td>SVM_COMPL</td>
<td>0.0566</td>
<td>0.0287</td>
</tr>
<tr>
<td>COMPL</td>
<td>CONS</td>
<td>0.0487</td>
<td>0.0585</td>
</tr>
<tr>
<td></td>
<td>SVM_COMPL</td>
<td>0.0420</td>
<td>0.0420</td>
</tr>
</tbody>
</table>

Return is expressed as a percentage.

At first, these results reveal that the consolidated investment strategy (CONS) realizes higher returns than an investment strategy based on the global classifier’s predictions. This is valid for both sentiment measures. Second, in comparison to the global classifier SVM_COMPL, the local classifier SVM_POS performs better in classifying the POS dataset. Concerning the FIN-setup, this also applies to the NEUT dataset and the respective local classifier. In contrast, the NEUT and NEG datasets are classified slightly worse in the case of the H-IV-4-setup. Third, it cannot be recommended to use a local classifier which has not been trained on the corresponding sentiment category. Fourth, the return which can be achieved with the consolidated investment strategy based on the FIN-setup is higher than the corresponding return based on the H-IV-4-setup.

To confirm these results and to answer research question one statistically, we formulate the following hypotheses:

\[ H_0: \mu(R(data, classifier)) \leq \mu(R(data, SVM_{COMPL})) \]  
\[ H_A: \mu(R(data, classifier)) > \mu(R(data, SVM_{COMPL})) \]  

\[ \mu(R(data, classifier)) \] represents the mean return which is achieved when an investment strategy is performed according to the recommendations of a local classifier which classifies a dataset denoted as data. This mean return is compared with the mean return which is realized when the same documents are classified by the global classifier SVM_COMPL. To test the hypotheses, we perform a two-sample t-test assuming unequal variances with a hypothesized mean of zero, the results are shown in table 4. For brevity, we do not report the results of local classifiers classifying documents they are not specialized in. In all of these cases, \(H_0\) cannot be rejected.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>H-IV-4-setup</th>
<th>FIN-setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>2.0824**</td>
<td>1.8161**</td>
</tr>
<tr>
<td>NEUT</td>
<td>-0.0311</td>
<td>0.6862</td>
</tr>
<tr>
<td>NEG</td>
<td>-1.2086</td>
<td>1.2102</td>
</tr>
<tr>
<td>COMPL</td>
<td>0.8332</td>
<td>2.0476**</td>
</tr>
</tbody>
</table>

** indicates significance at the 5%-level

At first, we consider the classification results of the whole dataset. Concerning the H-IV-4-setup, the null hypothesis that the consolidated investment strategy does not perform better than the investment strategy based on the COMPL-classifier cannot be rejected. In contrast, this null hypothesis can be rejected in the case of the FIN-setup at a 5% level of significance. This provides evidence that the combination of text mining and sentiment analysis can improve stock return predictions.

Additionally, we analyze the performance of the local classifiers in more detail. For both dictionaries, the local classifiers trained on the POS subsample perform better in classifying documents with a positive sentiment than the global classifier. Both results are statistically significant on a 5% level. In contrast, the SVM_NEG classifiers do not perform significantly better...
than the global classifiers. Concerning the FIN-setup, the corresponding p-value is relatively low but fails to be below 0.10. The H-IV-4-dictionary-based results do not allow rejecting the null hypothesis, too. This also applies to SVMSELT. As a result, it can be noticed that sentiment analysis mainly improves the classification of news articles containing a positive sentiment.

To explore if the results based on the FIN-dictionary are superior to the results based on the H-IV-4-dictionary (research question two), we investigate the returns achieved by the consolidated classifiers. As a consequence, we formulate the following hypotheses and perform a two-sample t-test assuming unequal variances with a hypothesized mean of zero, too. Concerning the notation, we add a subscript including the name of the dictionary the consolidated strategy is based on:

\[ H_0: \mu(\text{R(COMPL, CONS}_\text{FIN})) \leq \mu(\text{R(COMPL, CONS}_\text{H-IV-4})) \]  
\[ H_1: \mu(\text{R(COMPL, CONS}_\text{FIN})) > \mu(\text{R(COMPL, CONS}_\text{H-IV-4})) \]  

Table 5. Comparison of the H-IV-4-setup and the FIN-setup

<table>
<thead>
<tr>
<th>H_0</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu(\text{R(COMPL, CONS}<em>\text{FIN})) \leq \mu(\text{R(COMPL, CONS}</em>\text{H-IV-4})) )</td>
<td>1.2144</td>
<td>0.1123</td>
</tr>
</tbody>
</table>

As table 5 shows, the null hypothesis that the results of CONSFIN are worse than the results of CONS_{H-IV-4} cannot be rejected at a 10% level of significance. However, a p-value of 0.1123 shows that the probability of CONSFIN being worse than CONS_{H-IV-4} is 11.23% which is comparably low. As a result, concerning the domain-dependent evaluation, there is only partial support that the FIN-dictionary provides results which are superior to the results provided by the H-IV-4 dictionary.

6. Summary and Conclusion

Financial research shows that stock prices react quickly on novel company-related information. Consequently, recent studies applied text mining to support investment decisions by forecasting the stock price impact of financial news articles. However, news article sentiment has not been taken adequately into account before, although it provides additional information which can be used to improve machine learning setups.

Against this background, we propose a novel two-stage approach to combine text mining and sentiment analysis of financial news articles. First, every news article is analyzed to calculate a sentiment measure. Thereby, a dictionary-based approach is followed. Second, the news articles are categorized according to positive, neutral and negative sentiment. In accordance with this categorization, a local classifier which is trained on this sentiment category is selected to predict the subsequent stock price movement.

The results based on the domain-independent as well as the domain-dependent evaluation reveal that sentiment analysis improves the predictability of stock price changes after the publication of financial news articles. Surprisingly, this improvement is mainly driven by those news articles expressing positive sentiment. In contrast, there is only a small improvement in forecasting price reactions caused by news articles expressing negative sentiment. In the case of neutral sentiment, the results are ambiguous. Moreover, the domain-independent evaluation provides support that a sentiment measure based on a domain-specific dictionary can improve forecasting results. Nevertheless, in comparison to a generic dictionary, the returns realized by a respective investment strategy are not significantly higher.

This paper has several implications for further research. First, sentiment which is expressed in financial news articles can also be measured by a supervised approach. In this context, a comparison concerning the performance of both approaches in the financial domain needs to be conducted. Second, structured information like economic indicators can be added as additional input variables to investigate whether superior results can be achieved. Third, the differences between news articles expressing positive, neutral and negative sentiment need to be investigated in more detail. However, concerning the results of this paper, it can be stated by now that it definitely makes a difference whether a news article expresses positive (“boom”) or negative (“ruin”) sentiment.

7. Acknowledgements

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (grant agreement n. 257928).

8. References

