Combining Bag-of-Words and Sentiment Features of Annual Reports to Predict Abnormal Stock Returns

Petr Hájek

Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Studentská 84, Pardubice, Czech Republic

e-mail: petr.hajek@upce.cz, tel.: +420 466 036 147, fax: +420 466 036 010

Abstract. Automated textual analysis of firm-related documents has become an important decision support tool for stock market investors. Previous studies tended to adopt either dictionary-based or machine learning approach. Nevertheless, little is known about their concurrent use. Here we use the combination of financial indicators, readability, sentiment categories and bag-of-words (BoW) to increase prediction accuracy. This paper aims to extract both sentiment and BoW information from the annual reports of U.S. firms. The sentiment analysis is based on two commonly used dictionaries, namely a general dictionary Diction 7.0 and a finance-specific dictionary proposed by Loughran and McDonald [1]. The BoW are selected according to their *tf-idf*. We combine these features with financial indicators to predict abnormal stock returns using a multi-layer perceptron neural network with dropout regularization and rectified linear units. We show that this method performs similarly as Naïve Bayes and outperforms other machine learning algorithms (Support Vector Machine, C4.5 decision tree, and k-nearest neighbour classifier) in predicting positive/negative abnormal stock returns in terms of ROC. We also show that the quality of the prediction significantly increased when using the correlation-based feature selection of BoW. This prediction performance is robust to industry categorization and event window.

Keywords: stock return, prediction, text mining, sentiment, neural network.

1 Introduction

The past decade has seen the rapid development of textual analysis of many financial problems, such as the modelling of abnormal stock returns [2–4], volatility modelling [1, 5, 6], liquidity [7], market-to-book ratio [8], fraud detection [1, 9] and financial-distress prediction [10–12]. The findings of these studies have shown that the automated textual analysis of firm-related documents lead to more accurate financial predictions. It is therefore becoming increasingly difficult to ignore the contribution textual analysis may have in finance. In fact, textual sources such as news stories and corporate annual reports carries complementary qualitative information about the firm's current and future prospects. This information is reflected in the expectations of market participants [13].

In recent literature reviews [14–16], two general approaches have been used to analyse firm-related text: (1) dictionary-based and (2) machine learning. The former approach calculates overall word category (positive, negative, certainty, etc.) based on the frequency of words chosen by financial experts, thus addressing the context-specific nature of financial vocabulary better than using general dictionaries like Harvard IV-4 [1]. Machine-learning approaches such as Naïve Bayes [17] or Support Vector Machines (SVM) [13], on the other hand, automatically construct word lists and their weights based on a classification of texts (for example, positive vs. negative texts). This approach may provide more accurate predictions, but it is problem-specific and difficult to interpret. Both approaches have shown promising results in predicting the reactions of financial markets.

For the example of annual reports as textual sources, Li [18] demonstrated that changes in sentiment about risk (uncertainty) in annual reports significantly affects future earnings and stock returns. Li [19] found some evidence that managers may hide adverse information from investors by using harder-to-read language in annual reports. Feldman et al. [20] also reported that market reactions (two days after the U.S. Securities and Exchange Commission filing date)

are significantly associated with the tone (net positive) of the Management Discussion and Analysis (MD&A) section of the annual report. Davis and Tama-Sweet [21] examined language used in two types of managerial disclosures, namely earnings press releases and MD&A. Their results suggest that MD&A provides information incremental to that in the corresponding earnings press release. Specifically, negative association was found between the level of pessimistic language in the MD&A and future firm performance.

Machine-learning approaches to the textual classification of annual reports have also been reported in the literature. For example, Balakrishnan and Srinivasan [22] found that significantly positive, size-adjusted returns can be achieved by using the predictions of a machine-learning model. More specifically, textual information was reported to affect investors' use of price momentum, which then became a key determinant of these excess returns. Butler and Keselj [23] converted annual reports to character n-gram profiles and combined this approach with readability scores in a SVM classification model.

In this study, we use a hybrid textual analysis combining dictionary-based and machine learning approaches. Here, the dictionary-based approach is based on two commonly used complementary dictionaries, a general Diction 7.0 [24] and a finance-specific dictionary developed by [1]. Diction 7.0 uses a series of 35 dictionaries to calculate five general semantic features, namely activity, optimism, certainty, realism and commonality. Diction 7.0 has become a useful tool in strategic management research to examine both language usage in organizations and possible linkages between management's narratives and organizational performance [25]. However, the use of finance-specific dictionaries has shown significantly higher prediction accuracy compared to the use of general dictionaries [1, 4, 26]. Moreover, Loughran and McDonald [16] reported that general dictionaries were especially inappropriate for sentiment analysis of financial disclosures, causing a high percentage of sentiment misclassification. The dictionary by [1] has become particularly dominant in the literature for

finance-related analysis. Loughran and McDonald [1] reported that event period excess returns are positively affected by a frequent use of litigious terms (but only in cases of proportional weights of terms), whereas other financial dictionaries (negative, positive, uncertainty, weak and strong modal) have negative effects for both proportional and *tf-idf* (term frequency-inverse document frequency) weights of terms. Negative, uncertainty, weak and strong modal word lists displayed statistically significant effects for both weighting schemes.

The aim of this paper is to predict abnormal stock returns using the analysis of text in the annual reports of U.S. firms. Most studies in the field tended to focus on either dictionary-based or machine learning approach, paying little attention to their synergistic effects. Here we use the combination of financial indicators, word categories and bag-of-words (BoW) to increase prediction accuracy. First, adopting the approach of prior studies, we employ predefined dictionaries to show the effect of sentiment (tone) on abnormal stock returns. We show that the chosen word categories displayed in the annual reports negatively affect abnormal stock returns, with the exception of sentiment tone. Second, we use a BoW representation to detect the most relevant terms in the annual reports. We examine the effect of *N*-grams (word combinations) to capture the more complex underlying semantics of annual reports [13]. To perform the prediction of abnormal stock returns, we employ a Neural Network (NN) with dropout regularization and rectified linear units [27] and compare it with four machine learning approaches commonly used in text classification [28], namely Naïve Bayes, SVM, C4.5 decision tree, and k-nearest neighbour (k-NN) classifier. We demonstrate that the NN performs best using the combination of financial indicators and BoW approach.

This paper is a significantly extended version of [29]. The previous version was limited to banking industry, whereas here we use a more recent dataset for a wide range of industries. The extension further includes an in-depth literature review and an experimental analysis of the combined effects of (1) general and finance-specific dictionaries, (2) BoW and dictionary-based

features, and (3) financial indicators and linguistic features. We also examine the effect of three dimensionality reduction methods on the accuracy of abnormal stock returns' predictions. Finally, this paper compares the results with two state-of-the-art text classification methods and studies the robustness of the proposed approach to industry categorization and event window.

The remainder of this paper is organised in the following way. Section 2 outlines finance-specific aspects of textual analysis and provides a review of the relevant literature investigating the relationship between textual analysis and stock return prediction. Section 3 presents the corpus of documents and the results of its pre-processing. The prediction of abnormal stock returns is performed in Section 4. In addition to textual information in annual reports, the financial indicators of firms are used for analysis, in line with previous literature. Section 5 discusses the obtained results and concludes the paper.

2 Textual Analysis in Stock Return Prediction – Literature Review

Kearney and Liu [14] classified the sources of textual information in the financial domain into three categories: corporation-expressed, media-expressed, and Internet-expressed. Corporation-expressed information is usually extracted from annual reports [1, 18] or from earnings press releases and conference calls [2]. MD&A sections of annual reports are widely considered to be the most important source of insider information, because they provide management's perspective on past performance, current financial positions and future prospects [20]. These sections may therefore be particularly important for the prediction of firm performance and stock prices. Researchers have shown increasing interest in the analysis of firm-related narratives partly due to the requirements of the U.S. Securities and Exchange Commission (SEC) for electronic filings. 10-K filings (forms) provide both audited financial statements and a comprehensive overview of the firm's business and financial condition. Therefore, they are the most widely used source of data. However, the information provided by

management may be rather subjective and not entirely true, making analysis difficult. Moreover, simultaneously released informative signals may affect the impact of managerial textual content [30].

Li [31] examined the MD&A sections of 10-K (and 10-Q) filings using a Naïve Bayes method, demonstrating that a positive tone in the documents indicates positive future earnings. General dictionaries, on the other hand, failed to predict future financial performance. Demers and Vega [32] examined the impact on future earnings of net optimism and uncertainty of managerial communications regarding a firm's quarterly earnings results, suggesting that net optimism is positively associated with future earnings, whereas uncertainty indicates a decrease in future earnings. Similarly, Doran et al. [4] found that the tone of quarterly conference call dialogue has significantly explanatory power for abnormal stock returns. This tone may result in immediate stock price reaction and two-quarter delayed reaction, respectively [33]. Davis et al. [34] calculated net optimism in earnings press releases, finding that this measure (1) is positively associated with future return on assets and (2) generates a significant market response in a short window of time around the date of the earnings announcement. Moreover, sentiment obtained from MD&A is reported to provide information incremental to that extracted from the corresponding earnings press releases [21].

In contrast to the abovementioned studies, which used a general dictionary, Loughran and McDonald [1] developed a finance-specific dictionary to measure the sentiment in company-related textual documents. They reported that general dictionaries misclassified many negative words, such as "taxes" or "liabilities", thus adding noise to prediction models. Moreover, other industry-specific words ("oil", "cancer") do not carry the generally negative connotation they do in general language. In addition to negative words, Loughran and McDonald [1] considered other effects by using five other word classifications (positive, uncertain, litigious, strong modal, and weak modal). Taken together, higher sentiment (across all word categories) in

annual reports significantly and negatively affected future abnormal returns, whereas it significantly and positively impacted both abnormal volume and return volatility. In general, context specific dictionaries seem to be more powerful than general dictionaries [26, 35].

Meanwhile, media-expressed information is the information of outsiders contained in news stories and analyst reports [6]. Tetlock et al. [3] studied the effect of news stories on future earnings and stock returns, demonstrating that the fraction of negative words in firm-related news stories predicts both low earnings and low stock returns. Li et al. [36] demonstrated that news stories can be utilized to improve the accuracy of prediction on stock returns in intra-day trading. Schumaker and Chen [37] examined a SVM approach for financial news articles analysis using several textual representations: BoW, Noun Phrases, and Named Entities. Hagenau et al. [13] and Geva and Zahavi [38] use similar approaches and improve the performance of prediction models using feature selection procedures. Engelberg et al. [39] showed that there is a significant increase in short selling after news events, providing an information advantage to informed traders. Garcia [40] found that news content helps predict stock returns only during recessions. Moreover, the impact of media may also vary according to firm characteristics and article content [41]. The majority of the sources used are major news websites such as The Wall Street Journal [5] and Yahoo! Finance [42].

Internet-expressed sentiment is used to extract the information from small investors [5]. For example, in their stock price prediction model, Li et al. [43] combined news information with the information obtained from online financial discussion boards. Similarly, Yu et al. [44] have investigated content from the social media, including blogs, forums and Twitter. Their findings suggest that social media has a stronger impact on firm stock performance than conventional media.

Finally, several researchers have investigated a variety of firm-related textual documents. For example, Kothari et al. [45] examined corporate reports, analyst disclosures and briefings, and

disclosures made in the general business press. Their results showed that favourable disclosures have a significantly negative effect on firm's perceived risk (as proxied by the cost of capital, stock return volatility, and analyst forecast dispersion).

Table 1

3 Data and Research Methodology

Our study encompasses 1402 U.S. firms listed on the New York Stock Exchange (NYSE) or Nasdaq and with a reported stock price of at least 3 USD before the 10-K filing date (usually within 90 days after the end of the firm's fiscal year). This limit was chosen to reduce the contribution of bid/ask bounce in reaction to 10-K filing [1]. We also required market capitalisation of at least 100 million USD to reduce the effect of risk factors for stocks [51]. We downloaded all 10-Ks for such firms from the EDGAR system for the period 2013. To control for variables that have shown significant impacts on abnormal stock returns in prior literature [2, 26], we collected corresponding data from the Marketwatch database for the following variables: (1) liquidity ratio (daily trading volume/shares outstanding), (2) Beta (dependence of the behaviour of the share price on the stock indices), (3) log of the market capitalisation (lnMC), (4) price-earnings ratio (P/E), (5) price to book value (P/B), (6) return on equity (ROE), (7) total debt to total assets (TD/TA), and (8) a dummy variable for NYSE versus Nasdaq listing. ε-SVR (Support Vector Regression) was used for the imputation of missing values (with average RMSE=5.21). All attributes except the missing one were used to estimate the missing value. The completed data on financial indicators were used afterwards to predict abnormal stock returns.

Following previous studies [2], abnormal returns were calculated as accumulated returns in excess of the return on the CRSP (Center for Research in Security Prices) equal-weighted market portfolio. Consistent with related studies, we also adopted a three-day event window, from day t-1 to t+1, where t represents the 10-K filing day. The U.S. firms were categorized

into two classes, with positive (762 firms) and negative abnormal returns (618 firms), indicating an imbalanced dataset. Table 2 shows basic descriptive statistics of the sample. NYSE listings predominated in the data at 59.84 % of considered firms.

Table 2

In accordance with prior studies [31], we extracted only the most important textual section from the downloaded 10-Ks, namely Item 7: Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A). This section provides managements' perspective on their firms' past, current and future financial performance [14].

To obtain their tone, we compared the extracted documents with two complementary word categorisations: (1) a general Diction 7.0 [24], and (2) a finance-specific developed by [1]. A series of 35 Diction 7.0 word categories were used to calculate five general semantic features as follows:

certainty = (tenacity + leveling + collectives + insistence) – (numerical + ambivalence + self-reference + variety), (1)

optimism = (praise + satisfaction + inspiration) – (blame + hardship + denial), (2)

activity = (aggression + accomplishment + communication + motion) – (cognitive + passivity + embellishment), (3)

realism = (familiarity + spatial awareness + temporal awareness + present concern + human interest + concreteness) – (past concern + complexity), (4)

commonality = (centrality + cooperation + rapport) – (diversity + exclusion + liberation). (5)

Loughran and McDonald [1] have addressed two major drawbacks of previous finance-specific word lists [2], namely (1) the limited number of words contained in each category, and (2) ignoring other important word categories besides positive and negative. As a result, extensive word lists of 354 positive and 2,329 negative words were included by [1]. In addition, word

categories for uncertainty (291 words), litigious (871 words), and modal (19 modal strong + 27 modal weak words) were created as part of their work.

The use of negative words seems unambiguous, whereas the use of positive words in a negative statement has been one of the main challenges addressed in the literature on sentiment analysis [16]. To handle the problem of negations, we followed the approach proposed by [1], performing a collocation analysis with positive words to detect one of six negation words (no, not, none, neither, never, nobody) occurring within three words preceding a positive word. The frequency of net positive words was then calculated as the positive term count minus the count for negation (positive terms are easily qualified or compromised). Although this procedure should provide a more accurate measurement of positive tone, previous studies have shown that positive word lists can generally locate only a little incremental information [1, 6].

Following previous studies [4, 30, 39, 40], we used the raw term frequency of word categories. This is words in each category were regarded as synonyms. To consider the length of documents, we normalized the word category counts by the length of the MD&A. In addition to the abovementioned word categories, we also calculated the overall tone, defined as the count of positive words minus the count of negative words, divided by the sum of both positive and negative word counts [2]. Table 3 shows that firms with a stronger negative sentiment (and overall negative tone) performed worse. In other words, the overall tone was higher for the firms with positive abnormal return. In addition, realism and certainty in managerial sentiment were also taken positively by investors. This study also controls for the readability of the documents. We used the Gunning fog index as the most commonly applied readability measure [52]. For example, De Franco et al. [53] showed that the readability of analysts' reports increase trading volume reactions. The Gunning fog index can be calculated as follows:

$$\text{Gunning fog index} = 0.4 \times (\text{words per sentence} + \text{percent of complex words}), \quad (6)$$

where complex words are words with three syllables or more.

Table 3

To match the data from the EDGAR system and Marketwatch database, we used the ticker symbols of the firms (see Table 4 for a data sample).

Table 4

To identify a set of useful *N*-grams, we first removed stop-words, performed stemming using the Snowball stemmer, and converted all word tokens to lower case letters. Finally, all unigrams, bigrams and trigrams were identified in the training data and ranked according to their weights. Therefore, one central issue to be addressed is the choice of an appropriate term-weighting scheme to evaluate how important a word is within a document in a corpus [15, 54]. Using raw term frequency, all terms are considered equally important. However, this scheme assigns higher weights to terms that occur frequently in the text and it does not consider, moreover, the length of the document. Therefore, we used the smooth version of the most common term-weighting scheme, *tf-idf*, in which weights w_{ij} are defined as follows:

$$w_{ij} = \begin{cases} (1+\log(tf_{ij}))(1+\log\frac{N}{df_i}) & \text{if } tf_{ij} \geq 1 \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

where N represents the total number of documents in the corpus, df_i denotes the number of documents with at least one occurrence of the i-th term, and tf_{ij} is the frequency of the i-th term in the j-th document.

For our experiments, we used the top 200, 500, 1000, and 2000 *N*-grams in a BoW fashion. The most relevant 2000 words was reported to be enough to classify document classes [55]. Moreover, previous studies indicated overlaps and potential value provided by bigrams [13] and trigrams (e.g., flow, cash flow, future cash flow) [29].

Using top *N*-grams may also be considered as a feature selection method in text categorization problems. In fact, the high dimensionality of the feature space is another major issue to be

addressed. It is therefore desirable to reduce the original feature space in order to enable more effective operation of classification algorithms and thus improve their accuracy [56–58]. It was reported that feature selection can effectively remove more than 90% of the terms without sacrificing the classification accuracy [59, 60]. Therefore, to reduce the original feature space of 2000 *N*-grams we further used two dimensionality reduction methods, namely a correlation-based feature selection (CBFS) [61], latent semantic analysis (LSA) [62] and LSA combined with cosine similarity (LSA+cosine) [63]. In the correlation-based feature selection, the optimality of a feature subset is based on Pearson's correlation coefficient. Thus, irrelevant features are removed due to low correlations with the class, and redundant features are removed owing to high correlations with one or more of the other features. In this study, particle swarm optimization (PSO) (with 20 particles in the swarm, mutation probability of 0.01, individual weight of 0.34, inertia weight of 0.33, and social weight of 0.33) was used as a search procedure in the correlation-based feature selection. The feature selection was performed separately on 10 training datasets (10-fold cross-validation was used to avoid overfitting) to alleviate feature selection bias. On average, 30.3 features were selected for unigrams, 18.7 for bigrams, and 24.6 for trigrams.

The LSA was performed using singular value decomposition in order to transform the original feature space to a low-dimensional semantic space, in which documents with the same semantic concepts can be detected [62]. The number of concepts (26 for unigrams, 27 for bigrams, and 28 for trigrams) to retain was based on a proportion of total singular values to account for (set to 0.99). The LSA+cosine method combined the results obtained from the LSA with cosine similarity to documents separately for positive and negative class. Our research methodology is presented in Figure 1.

Figure 1. Research methodology

4 Experimental Results

The survey on text mining for stock market prediction [15] concludes that SVM and Naïve Bayes are heavily favoured by researchers, whereas NNs are significantly under-researched in the field of stock market predictive text-mining at this stage, despite that NNs have shown promising potentials for textual classification and sentiment analysis. NNs equipped with advanced techniques such as rectified linear units, AdaGrad and dropout regularization have been reported to be particularly effective compared with state-of-the-art approaches to text classification [64].

In our experiments, we examined multilayer perceptron NN with dropout regularization and rectified linear units (Figure 2). Dropout regularization [27, 65] was utilized because fully connected NNs are prone to overfitting. This regularization randomly sets a given proportion of the activations to the fully connected layers to zero during training. Thus, hidden units that activate the same output are decoupled. This largely improves generalization ability and prevents overfitting [66]. However, note that this effect is still not clear in deep NNs, for example in convolutional and pooling layers [67].

Rectified linear units have attracted increased attention because traditional sigmoidal units suffer from the vanishing gradient problem, which may cause slow optimization convergence to a poor local minimum [68]. The synergistic effects of combining rectified linear units with dropout regularization have been demonstrated by [69].

Figure 2. Multilayer perceptron with dropout regularization (crossed neurons dropped)

We trained this multilayer perceptron NN using stochastic gradient descent algorithm with the following parameters: input layer dropout rate = 0.2, hidden layer dropout rate = 0.5, number of hidden layers = {1, 2}, number of units in the hidden layer = {10, 20, 50, 100, 200}, learning rate = {0.05, 0.10}, size of each mini-batch used in computing gradients = 100, and the number of iterations = 1000. The structure and parameters of the NN learning were found using grid search procedure. The large number of neurons in the hidden layer was examined due to the

high number of input features (more than 2000). However, adding too many neurons was not necessary because it would lead to modelling the noise in the training data, eventually causing poor generalization performance.

To demonstrate the effectiveness of this NN, we compared the results with four methods commonly used in text classification tasks, namely Naïve Bayes, SVM, C4.5 decision tree, and k-NN classifier.

Naïve Bayes is the most commonly used generative classifier in text classification. The posterior probabilities of classes are calculated based on the distribution of the words in the document. The main assumption of Naïve Bayes is that the words in the documents are conditionally independent given the class value.

Further, we used the SVMs trained using stochastic gradient descent algorithm (SGD). This algorithm was reported to outperform traditional sequential minimal optimization (SMO) in related document categorization tasks [10]. Since SVMs are robust to high dimensionality, they are well suited for text classification because of the sparse high-dimensional nature of the text. We examined the SGD for the number of epochs = 500 and learning rate = {0.01,0.05}.

Error based pruning algorithm was used to train the C4.5 decision tree. This algorithm uses single-attribute splits at each node. The feature with the highest information gain is used for the purpose of the split. For this algorithm, confidence factor is used when pruning the tree. The following parameters of C4.5 were examined to obtain the best classification performance: confidence factor = {0.1,0.25,0.4}, minimum number of instances per leaf = {1,2, … ,5}, and number of folds = 3.

Linear nearest neighbour search algorithm with Euclidean distance function was used for the k-NN classifier. The number of neighbours was set to 3. The main idea is that documents belonging to the same class are likely to be close to one another based on a similarity measure.

It was reported that the use of common classification performance criteria such as accuracy may yield misleading conclusion in the case of class imbalance [70]. More accurate measures such as ROC (receiver operating characteristic) curve have been predominantly used for imbalanced datasets. Therefore, we measured the quality of abnormal return prediction using the area under the ROC curve. To demonstrate the classification accuracy on each class, we also report true positive (TP) and true negative (TN) rates. To avoid overfitting, all experiments were performed using 10-fold cross-validation.

In the first set of experiments, we used the financial, sentiment and BoW features separately. Table 5 shows the classification performance on the abnormal bank stock returns dataset. We report the Average±Std.Dev. values of ROC from the 10-fold cross-validation. The best performance of the algorithm is marked in bold. In addition, we report average TP and TN rates in Table 6.

SVM and k-NN algorithms performed generally better on the lower dimensional datasets (BoW with 200 and 500 features), whereas NB, C4.5 and NN performed best for the BoW with 2000 features. In case of the NB, this suggests a high variance in the data. Moreover, the quality of the prediction increased when using bigrams and trigrams. However, the best classification performance for the NB and k-NN methods was achieved using only unigrams and bigrams, respectively.

We employed Student's paired *t*-test at p=0.05 to test the differences in ROC. The results show that the NB and NN models performed particularly well on the BoW datasets.

Table 5

Table 6

In the second set of experiments, we examined the effect of the three dimensionality reduction methods, CBFS, LSA and LSA+cosine, on the accuracy of abnormal stock returns' predictions. As can be seen from Table 7, ROC classification performance increased for all methods when

using the CBFS method for dimensionality reduction (compared with the original feature space used in Table 5). Again, the performance was superior for BoW features selected from bigrams and trigrams, respectively. In contrast, dimensionality reduction using the LSA method did not improve the classification performance in terms of ROC. The LSA+cosine method performed better for NB and C4.5, while worse for the other classifiers. These results suggest that the classification performance of the algorithms was mainly deteriorated due to the presence of irrelevant and redundant features. This was particularly true for the majority class (positive abnormal return), see Table 8. Therefore, we used only the BoW features selected by the CBFS in subsequent analyses. Again, NB and NN performed best across all datasets according to the Student's paired *t*-test.

Table 7

Table 8

In the third set of experiments, we combined the categories of features to demonstrate the synergistic effect of financial, sentiment and BoW information. Specifically, we examined the following combinations: (1) BoW features (unigrams + bigrams + trigrams), (2) financial, sentiment and readability, (3) financial and BoW, (4) sentiment, readability and BoW, and (5) financial, sentiment, readability and BoW features. Table 9 shows that the classification performance of all algorithms (except C4.5) increased compared with both single approaches (Table 5) and dimensionality reduction methods (Table 7). For SVM, C4.5 and NN, the performance was best when the financial, sentiment and readability indicators were combined with the BoW approach. Specifically, Table 10 shows that the TN rates (i.e. for the minor class) particularly increased compared with dimensionality reduction methods (Table 8). In contrast, the financial, sentiment and readability indicators were not important predictors for the NB classifier. In terms of ROC, the NN and NB methods significantly outperformed the remaining methods in all four sets of experiments. Using the predictions of the best performing models

for each method, we were able to calculate the average stock return on testing data. When comparing the results with the average stock market return (0.30%) and a trivial majority classifier [13] (0.70%) as benchmarks, the average stock return for the portfolio selected by the used classifiers (i.e. the portfolio of firms classified as positive abnormal return) was as follows: NB (0.85%), SVM (0.97%), C4.5 (0.82%), k-NN (0.72%), and NN (1.01%).

Table 9

Table 10

To compare the performance of the NN (model with Fin.+Sentim.+Readab.+CBFS_uni+bi+tri) with other state-of-the-art methods, we selected multinomial inverse regression (MNIR) [71, 72] and sparse matrix factorization (SMF) [73, 74]. The MNIR uses multinomial regression to map from BoW to the class space via relevant variables. Here, we applied the MNIR to 2000 *N*-grams identified using phrase counts weighting scheme. Specifically, we regressed the weights of the *N*-grams onto three-day stock returns and market returns. Thus, two-dimensional sufficient reduction statistics were obtained for each document. Logistic regression was then used to predict abnormal stock return. In the SMF, abnormal stock price returns are correlated with features extracted from corporate annual reports. In agreement with related studies [73, 74], we used 2000 unigrams to build a latent factor model. A sparse group lasso regularization term was included to eliminate irrelevant unigrams. Table 11 shows that the MNIR performed similar to the NN in terms of ROC. In contrast, the SMF was significantly outperformed. These differences can be partly explained by the use of bigrams and trigrams in the NN and MNIR models, respectively. Another possible explanation for this is that sentiment and readability features were not included in the SMF and MNIR models.

Table 11

To test for the robustness of the NN results, we first examined the NN performance across industries. Manufacturing and finance firms predominated in the data set (Table 12). The results

presented in Table 12 demonstrate that the NN performed well for all SIC (standard industrial classification) categories of industries.

Table 12

It is also important to note that several related studies have used different event windows, ranging from three [1, 3] to ninety trading days [75]. We therefore examined different event windows. Specifically, the abnormal stock returns was predicted by using a seven-day (and an eleven-day) event window, this is from day t-3 to t+3 (and from t-5 to t+5), again centred on the date of the 10-K filing day. Table 13 shows that the NN performed best for the three-day event window. However, the performance did not deteriorate with increasing event window. When examining the results in terms of TP and TN rates, it is obvious that the overall performance is largely dependent on the class ratio rather than event window. Indeed, positive class predominated in the three-day data set in contrast to the eleven-day event window (Table 13). The overall performance of the NN was good for all periods. The average stock return for the portfolio selected by the NN was 0.99% and 2.20% for the seven-day and eleven-day event window, respectively. Again, average stock market return (0.51% and 0.96%) and the trivial majority classifier (0.50% and 0.00%) were significantly outperformed by the proposed model.

Table 13

5 Conclusion

A strong relationship between textual information extracted from annual reports and abnormal stock return has been reported in the literature. This study set out with the aim of assessing the synergistic effects of sentiment analysis and machine learning approach in predicting abnormal stock returns. The results of this study indicate that machine learning approaches using BoW provide more accurate predictions than the aggregate indicators of sentiment categories. However, the elimination of irrelevant and redundant features seems to be critical in the BoW approach. Moreover, the combination of sentiment analysis and machine learning approach

showed increase in ROC accuracy compared with the pure machine learning approach. On one hand, this increase was not significant, suggesting that BoW sufficiently incorporate sentiment-related terms. On the other hand, we demonstrated that this combination outperformed the SMF model that ignores the evaluation of sentiment.

It is plausible that a number of limitations may have influenced the results obtained. Similarly to most previous studies, the sample used in this study was limited to major U.S. stock exchanges. Although the sample can be considered representative of the North America region and similar to Europe region, recent empirical evidence suggests that other regions' stock markets exhibit specific behaviour, such as Japan [76]. Moreover, this study has only investigated a limited time period. In the year 2013, major U.S. stock exchanges returned to growth, including a more optimistic investor sentiment compared with previous years. This trend has remained to the present day, suggesting that the proposed model may perform well for a longer period of time. However, caution must be applied, as managers begin to be aware of the importance of their comments on investors' behaviour. Substantially more experiments should therefore be conducted to generalize our findings.

Another important finding was that NN with dropout regularization and rectified linear units performed particularly well on this prediction task, suggesting that this method may be well suited for text classification tasks working with sparse high-dimensional data. Therefore, further research should be done to investigate the use of this NN model in related text classification tasks. Future research should also concentrate on different feature selection procedures, especially for high-dimensional imbalanced data [77]. Furthermore, a future study investigating the syntactic structure and additional semantic features of firm-related text documents would be interesting. In fact, alternative NN structures such as convolutional NNs have recently been proposed to model sentiment-specific word embedding [78, 79]. Moreover,

further research might explore the role of tone dispersion within managerial comments as this has been identified as an important indicator of current and future corporate financial performance [80]. Finally, the current study was limited by the use of financial fundamental indicators. Future research should therefore concentrate on the investigation of technical analysis [81] in conjunction with the linguistic features of firm-related documents.

The experiments in this study were carried out in R 2.12.0, Statistica 12 and Weka 3.7.13 using the MS Windows 7 operation system.

Funding: This study was funded by the scientific research project of the Czech Sciences Foundation (grant number GA16-19590S).

Conflict of Interest: The authors declare that they have no conflict of interest.

References

1. Loughran T, Mcdonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Finance 66:35–65. doi: 10.1111/j.1540-6261.2010.01625.x
2. Henry E (2008) Are investors influenced by how earnings press releases are written? J Bus Commun 45:363–407. doi: 10.1177/0021943608319388
3. Tetlock PC, Saar-Tsechansky M, MacSkassy S (2008) More than words: Quantifying language to measure firms' fundamentals. J Finance 63:1437–1467. doi: 10.1111/j.1540-6261.2008.01362.x
4. Doran JS, Peterson DR, Price SM (2012) Earnings conference call content and stock price: The case of REITs. J Real Estate Financ Econ 45:402–434. doi: 10.1007/s11146-010-9266-z
5. Antweiler W, Frank MZ (2004) Is all that talk just noise? The information content of Internet stock message boards. J Finance 59:1259–1294. doi: 10.1111/j.1540-

6261.2004.00662.x

6. Tetlock PC (2007) Giving content to investor sentiment: The role of media in the stock market. J Finance 62:1139–1168. doi: 10.1111/j.1540-6261.2007.01232.x
7. Bodnaruk A, Loughran T, McDonald B (2015) Using 10-K text to gauge financial constraints. J Financ Quant Anal 50:623–646. doi: http://dx.doi.org/10.2139/ssrn.2331544
8. Myskova R, Hajek P (2016) The effect of managerial sentiment on market-to-book ratio. Transform Bus Econ 15:80–96.
9. Hajek P, Henriques R (2017) Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. Knowledge-Based Syst 1–14. doi: 10.1016/j.knosys.2017.05.001
10. Hajek P, Olej V (2013) Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In: Iliadis L, Papadopoulos H, Jayne C (eds) Commun. Comput. Inf. Sci. Springer, Berlin Heidelberg, pp 1–10
11. Hajek P, Olej V, Myskova R (2014) Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. Technol Econ Dev Econ 20:721–738. doi: 10.3846/20294913.2014.979456
12. Hajek P, Olej V (2016) Intuitionistic neuro-fuzzy network with evolutionary adaptation. Evol Syst 1–13. doi: 10.1007/s12530-016-9157-5
13. Hagenau M, Liebmann M, Neumann D (2013) Automated news reading: Stock price prediction based on financial news using context-capturing features. Decis Support Syst 55:685–697. doi: 10.1016/j.dss.2013.02.006
14. Kearney C, Liu S (2014) Textual sentiment in finance: A survey of methods and models. Int Rev Financ Anal 33:171–185. doi: 10.1016/j.irfa.2014.02.006

15. Khadjeh Nassirtoussi A, Aghabozorgi S, Ying Wah T, Ngo DCL (2014) Text mining for market prediction: A systematic review. Expert Syst Appl 41:7653–7670. doi: 10.1016/j.eswa.2014.06.009

16. Loughran T, Mcdonald B (2016) Textual analysis in accounting and finance: A survey. J Account Res 54:1187–1230. doi: 10.1111/1475-679X.12123

17. Huang AH, Zang AZ, Zheng R (2014) Evidence on the information content of text in analyst reports. Account Rev 89:2151–2180. doi: 10.2308/accr-50833

18. Li F (2006) Do stock market investors understand the risk sentiment of corporate annual reports? Gene 1–53. doi: 10.2139/ssrn.898181

19. Li F (2008) Annual report readability, current earnings, and earnings persistence. J Account Econ 45:221–247. doi: 10.1016/j.jacceco.2008.02.003

20. Feldman R, Govindaraj S, Livnat J, Segal B (2010) Management's tone change, post earnings announcement drift and accruals. Rev Account Stud 15:915–953. doi: 10.1007/s11142-009-9111-x

21. Davis AK, Tama-Sweet I (2012) Managers' use of language scross alternative disclosure outlets: Earnings press releases versus MD&A. Contemp Account Res 29:804–837. doi: 10.1111/j.1911-3846.2011.01125.x

22. Balakrishnan R, Qiu XY, Srinivasan P (2010) On the predictive ability of narrative disclosures in annual reports. Eur J Oper Res 202:789–801. doi: 10.1016/j.ejor.2009.06.023

23. Butler M, Kešelj V (2009) Financial forecasting using character n-gram analysis and readability scores of annual reports. In: Gao Y, Japkowicz N (eds) Lect. Notes Comput. Sci. Springer, Berlin Heidelberg, pp 39–51

24. Hart RP (2001) Redeveloping DICTION: Theoretical considerations (new). In: West MD (ed) West, M. D. (Ed.). (2001). Theory, Method, Pract. Comput. Content Anal.

Westport, CT Ablex. pp 43–60

25. Short JC, Palmer TB (2008) The application of DICTION to content analysis research in strategic management. Organ Res Methods 11:727–752. doi: 10.1177/1094428107304534

26. Price SM, Doran JS, Peterson DR, Bliss BA (2012) Earnings conference calls and stock returns: The incremental informativeness of textual tone. J Bank Financ 36:992–1011. doi: 10.1016/j.jbankfin.2011.10.013

27. Hinton GE, Srivastava N, Krizhevsky A, et al (2012) Improving neural networks by preventing co-adaptation of feature detectors. ArXiv e-prints 1–18. doi: arXiv:1207.0580

28. Baharudin B, Lee LH, Khan K (2010) A review of machine learning algorithms for text-documents classification. J Adv Inf Technol 1:4–20. doi: 10.4304/jait.1.1.4-20

29. Hajek P, Bohacova J (2016) Predicting abnormal bank stock returns using textual analysis of annual reports – A neural network approach. In: Jayne C, Iliadis L (eds) Commun. Comput. Inf. Sci. Springer, Aberdeen, pp 67–78

30. Demers E, Vega C (2014) Understanding the role of managerial optimism and uncertainty in the price formation process: evidence from the textual content of earnings announcements. Available SSRN 1152326. doi: http://dx.doi.org/10.2139/ssrn.1152326

31. Li F (2010) The information content of forward-looking dtatements in corporate filings - A Naïve Bayesian machine learning approach. J Account Res 48:1049–1102. doi: 10.1111/j.1475-679X.2010.00382.x

32. Demers E, Vega C (2010) Soft information in earnings announcements: News or noise? INSEAD Bus Sch World 1–70. doi: 10.2139/ssrn.1153450

33. Huang X, Teoh SH, Zhang Y (2014) Tone management. Account Rev 89:1083–1113.

doi: 10.2308/accr-50684

34. Davis AK, Piger JM, Sedor LM (2012) Beyond the numbers: Measuring the information content of earnings press release language. Contemp Account Res 29:845–868. doi: 10.1111/j.1911-3846.2011.01130.x

35. Henry E, Leone AJ (2016) Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. Account Rev 91:153–178. doi: 10.2308/accr-51161

36. Li X, Huang X, Deng X, Zhu S (2014) Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. Neurocomputing 142:228–238. doi: 10.1016/j.neucom.2014.04.043

37. Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news. ACM Trans Inf Syst 27:1–19. doi: 10.1145/1462198.1462204

38. Geva T, Zahavi J (2014) Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. Decis Support Syst 57:212–223. doi: 10.1016/j.dss.2013.09.013

39. Engelberg JE, Reed A V., Ringgenberg MC (2012) How are shorts informed?. Short sellers, news, and information processing. J financ econ 105:260–278. doi: 10.1016/j.jfineco.2012.03.001

40. García D (2013) Sentiment during Recessions. J Finance 68:1267–1300. doi: 10.1111/jofi.12027

41. Li Q, Wang T, Li P, et al (2014) The effect of news and public mood on stock movements. Inf Sci (Ny) 278:826–840. doi: 10.1016/j.ins.2014.03.096

42. Schumaker RP, Zhang Y, Huang CN, Chen H (2012) Evaluating sentiment in financial news articles. Decis Support Syst 53:458–464. doi: 10.1016/j.dss.2012.03.001

43. Li Q, Wang T, Gong Q, et al (2014) Media-aware quantitative trading based on public

Web information. Decis Support Syst 61:93–105. doi: 10.1016/j.dss.2014.01.013

44. Yu Y, Duan W, Cao Q (2013) The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decis Support Syst 55:919–926. doi: 10.1016/j.dss.2012.12.028

45. Kothari SP, Li X, Short JE (2009) The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. Account Rev 84:1639–1670. doi: 10.2308/accr.2009.84.5.1639

46. Hanley KW, Hoberg G (2010) The information content of IPO prospectuses. Rev Financ Stud 23:2821–2864. doi: 10.1093/rfs/hhq024

47. Mayew WJ, Venkatachalam M (2012) The power of voice: Managerial affective states and future firm performance. J Finance 67:1–44. doi: 10.1111/j.1540-6261.2011.01705.x

48. Li X, Xie H, Chen L, et al (2014) News impact on stock price return via sentiment analysis. Knowledge-Based Syst 69:14–23. doi: 10.1016/j.knosys.2014.04.022

49. Wisniewski TP, Yekini LS (2015) Stock market returns and the content of annual report narratives. Account Forum 39:281–294. doi: 10.1016/j.accfor.2015.09.001

50. Feuerriegel S, Ratku A (2016) Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In: Bui TX, Sprague RH (eds) 49th Hawaii Int. Conf. Syst. Sci. IEEE, Kauai, pp 1072–1081

51. Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. J financ econ 33:3–56. doi: 10.1016/0304-405X(93)90023-5

52. Loughran T, Mcdonald B (2014) Measuring readability in financial disclosures. J Finance 69:1643–1671. doi: 10.1111/jofi.12162

53. De Franco G, Hope OK, Vyas D, Zhou Y (2015) Analyst report readability. Contemp Account Res 32:76–104. doi: 10.1111/1911-3846.12062

54. Escalante H, Ponce-López V, Escalera S (2016) Evolving weighting schemes for the Bag of Visual Words. Neural Comput Appl 1–15. doi: 10.1007/s00521-016-2223-x

55. Dhillon IS, Mallela S, Kumar R (2003) A divisive information-theoretic feature Clustering algorithm for text classification. J Mach Learn Res 3:1265–1287. doi: 10.1162/153244303322753661

56. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17:491–502. doi: 10.1109/TKDE.2005.66

57. Hajek P, Michalak K (2013) Feature selection in corporate credit rating prediction. Knowledge-Based Syst 51:72–84. doi: 10.1016/j.knosys.2013.07.008

58. Glezakos TJ, Tsiligiridis TA, Iliadis LS, et al (2009) Feature extraction for time-series data: An artificial neural network evolutionary training model for the management of mountainous watersheds. Neurocomputing 73:49–59. doi: 10.1016/j.neucom.2008.08.024

59. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Mach. Learn. Work. Then Conf. pp 412–420

60. Li Z, Lu W, Sun Z, Xing W (2016) A parallel feature selection method study for text classification. Neural Comput Appl 1–12. doi: 10.1007/s00521-016-2351-3

61. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224. doi: 10.1145/1014052.1014149

62. Crain SP, Zhou K, Yang S-H, Zha H (2012) Dimensionality reduction and topic modeling: From latent semantic Indexing to latent dirichlet allocation and beyond. In: Aggarwal CC, Zhai C (eds) Min. Text Data. Springer, New York, pp 129–161

63. Egozi O, Markovitch S, Gabrilovich E (2011) Concept-based information retrieval using explicit semantic analysis. ACM Trans Inf Syst 29:1–34. doi: 10.1145/1961209.1961211

64. Nam J, Kim J, Loza Mencía E, et al (2014) Large-scale multi-label text classification - Revisiting neural networks. In: Calders T, Esposito F, Hullermeier E, Meo R (eds) Lect. Notes Comput. Sci. Springer, Berlin Heidelberg, pp 437–452

65. Barrow E, Eastwood M, Jayne C (2016) Selective dropout for deep neural networks. In: Akira, H., Seiichi O, Doya K, et al (eds) Int. Conf. Neural Inf. Process. Springer, Kyoto, pp 519–528

66. Srivastava N, Hinton G, Krizhevsky A, et al (2014) Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958. doi: 10.1214/12-AOS1000

67. Wu H, Gu X (2015) Towards dropout training for convolutional neural networks. Neural Networks 71:1–10. doi: 10.1016/j.neunet.2015.07.007

68. Maas AL, Hannun AY, Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. In: Dasgupta S, McAllester D (eds) Proc. 30 th Int. Conf. Mach. Learn. JMLR, Atlanta, pp 1–6

69. Jaitly N, Hinton G (2011) Learning a better representation of speech soundwaves using restricted boltzmann machines. In: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. IEEE, Prague, pp 5884–5887

70. Chawla N V, Japkowicz N, Drive P (2004) Editorial : Special issue on learning from imbalanced data sets. ACM SIGKDD Explor Newsl 6:1–6. doi: http://doi.acm.org/10.1145/1007730.1007733

71. Taddy M (2013) Multinomial inverse regression for text analysis. J Am Stat Assoc 108:755–770. doi: 10.1080/01621459.2012.734168

72. Taddy M (2015) Document classification by inversion of distributed language representations. In: Proc. 53rd Meet. Assoc. Comput. Linquistics. pp 45–49

73. Wong FMF, Liu Z, Chiang M (2014) Stock market prediction from WSJ: Text mining

via sparse matrix factorization. In: 2014 IEEE Int. Conf. Data Min. IEEE, pp 430–439

74. Sun A, Lachanski M, Fabozzi FJ (2016) Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. Int Rev Financ Anal 48:272–281. doi: 10.1016/j.irfa.2016.10.009

75. Guay W, Samuels D, Taylor D (2016) Guiding through the Fog: Financial statement complexity and voluntary disclosure. J Account Econ 62:234–269. doi: 10.1016/j.jacceco.2016.09.001

76. Fama EF, French KR (2012) Size, value, and momentum in international stock returns. J financ econ 105:457–472. doi: 10.1016/j.jfineco.2012.05.011

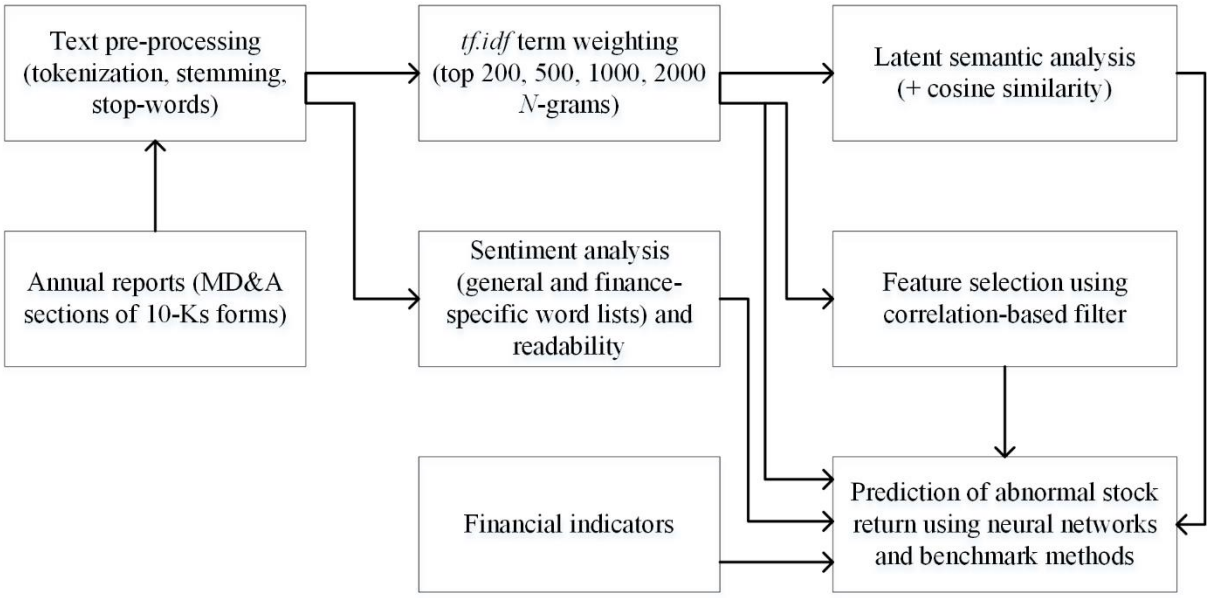
77. Yin L, Ge Y, Xiao K, et al (2013) Feature selection for high-dimensional imbalanced data. Neurocomputing 105:3–11. doi: 10.1016/j.neucom.2012.04.039

78. Tang D, Wei F, Yang N, et al (2014) Learning sentiment-specific word embedding for twitter sentiment classification. In: Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Association for Computational Linguistics, Baltimore, pp 1555–1565

79. Wang P, Xu B, Xu J, et al (2016) Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing 174:806–814. doi: 10.1016/j.neucom.2015.09.096

80. Allee KD, DeAngelis MD (2015) The structure of voluntary disclosure narratives: Evidence from tone dispersion. J Account Res 53:241–274. doi: 10.1111/1475-679X.12072

81. Thenmozhi M, Sarath Chand G (2016) Forecasting stock returns based on information transmission across global markets using support vector machines. Neural Comput Appl 1–20. doi: 10.1007/s00521-015-1897-9



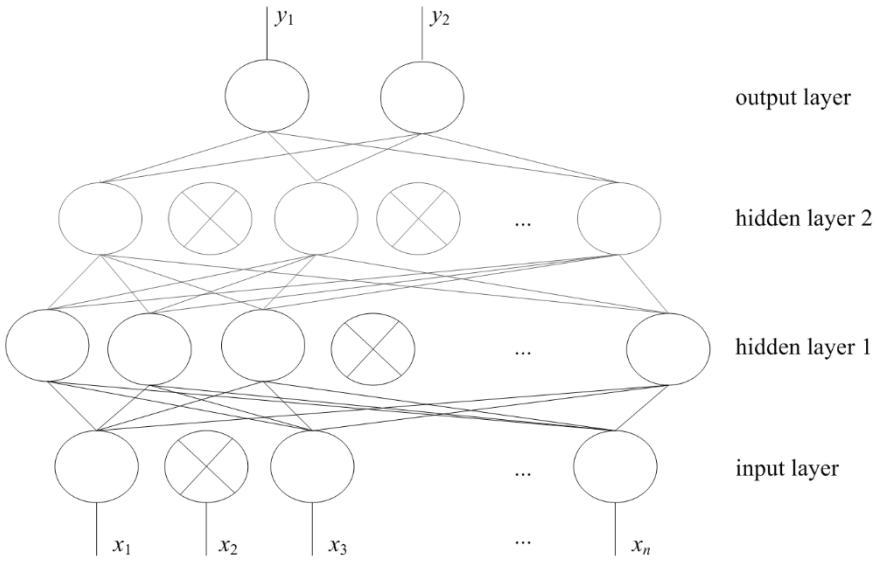


Table 1: List of prior studies on stock return prediction using textual analysis

Study	Textual source	Features	Method	Key findings
[5]	internet stock messages	Bullishness and agreement indexes	NB+SVM	The effect of stock messages on stock returns is statistically significant but economically small
[6]	news stories	77 word categories from Harvard General Inquirer	FA+LR	Negative effect of media pessimism on market prices
[3]	news stories	Negative word category from Harvard General Inquirer	LR	Stock prices briefly underreact to negative sentiment
[2]	earnings press releases	Positive, negative, tone, numerical terms and verbal complexity	LR	Tone influences investors' reactions
[37]	news stories	BoW, noun phrases and named entities	SVM	The model containing both article terms and stock price performed best

[20]	10-Ks	Positive and negative word categories from Harvard General Inquirer	LR	Market reactions around the SEC filing are significantly associated with the tone of MD&A
[46]	IPO prospectuses	Product market, accounting, legal, corporate strategy, patent, marketing, valuation, corporate governance, positive and negative from Harvard General Inquirer	LR	Informative content of IPO prospectuses improves pricing accuracy
[1]	10-Ks	Positive, uncertain, litigious, strong modal and weak modal	LR	Wordlists developed for other disciplines misclassify common words in financial text
[21]	10-Ks, earnings press releases	Positive and negative word categories from Diction [24]	LR	Pessimistic language in MD&A provides information incremental to that in the corresponding earnings press release
[34]	earnings press releases	Positive, negative and tone word categories from Diction [24]	LR	Net optimism generates a significant market response in a short window around the earnings announcement date
[39]	news stories	Positive and negative word categories from Harvard General Inquirer and Loughran and McDonald [1]	LR	Public news provides valuable trading opportunities for short sellers who are skilled information processors
[26]	earnings conference calls	Positive, negative and tone from Harvard General Inquirer and Henry [2]	PCA+LR	Conference call linguistic tone is a significant predictor of abnormal returns and trading volume
[42]	news stories	Polarity (positive, negative and neutral), tone (subjective, objective and neutral)	SVM	Subjective news articles are easier to predict in price direction
[4]	earnings conference calls	Positive, negative and tone from Harvard General Inquirer and Henry [2]	PCA+LR	The tone of the conference call dialogue has significant explanatory power for the abnormal returns
[47]	earnings conference calls	Positive and negative word categories from Loughran and McDonald [1]	LR	Positive (negative) managerial affect is positively (negatively) related to contemporaneous stock returns and future unexpected earnings
[40]	news stories	Positive, negative and tone from Loughran and McDonald [1]	LR	News content helps predict stock returns, but only during recessions
[44]	social media and news stories	BoW	NB+LR	Social media has a stronger relationship with firm stock performance than conventional media
[30]	earnings conference calls	Uncertainty and tone from Harvard General Inquirer, Loughran and McDonald [1], and Diction [24]	PCA+LR	The effect of textual features is greater when text is accompanied by a pro forma earnings figure, managerial earnings forecast, and more numerical data
[33]	earnings press releases	Positive, negative and tone from Loughran and McDonald [1]	LR	Abnormal positive tone positively affects the immediate stock price reaction to earnings announcements
[13]	news stories	BoW	SVM	Market feedback is used to select text features
[17]	analyst reports	BoW	NB+LR	Investors react more strongly to negative than to positive text
[43]	news stories	BoW	SVM	A simulation trading return was up to 166.11 %

[41]	news stories	BoW	SVM	Firm-specific news articles can enrich the knowledge of investors and affect their trading activities
[48]	news stories	BoW, positive and negative sentiment from Harvard General Inquirer and Loughran and McDonald [1]	SVM	Models with sentiment word categories outperform the BoW model
[36]	news stories	BoW	SVM	Integrating market news and stock price information improves the prediction accuracy of stock returns
[38]	news stories	BoW, sentiment score, business event categorization	MLP, DT	MLP performed best using market data, simple news item counts, categorization into business events and calibrated sentiment scores as predictors
[49]	10-Ks	Word categories from Diction [24]	LR	Activity and realism predict subsequent price increases
[35]	earnings press releases	Tone from Harvard General Inquirer, Henry [2], Loughran and McDonald [1], and Diction [24]	LR	Context-specific financial disclosure wordlist is a better predictor than general wordlists
[50]	news stories	40 news topics	LDA	Some topics have no resulting effect on abnormal stock returns, whereas other topics, such as drug testing, exhibit a large effect

Legend: DT – decision trees, FA – factor analysis, LDA – latent dirichlet allocation, LR – linear regression models, MLP - multi-layer perceptron neural network, NB – Naïve Bayes, PCA – principal component analysis, SVM – support vector machine.

Table 2. Descriptive statistics on financial indicators

Class	Positive		Negative	
Var.	Mean	Std.Dev.	Mean	Std.Dev.
Liquidity	0.0078	0.0068	0.0085	0.0077
Beta	3.06	13.80	3.29	14.63
lnMC	7.83	1.66	7.82	1.74
P/E	56.09	273.52	51.91	118.73
P/B	12.11	73.06	8.30	22.61
ROE [%]	5.12	21.99	6.42	24.46
TD/TA [%]	31.37	20.13	32.60	21.36

Table 3. Descriptive statistics on linguistic variables

	Class	Positive		Negative	
Dictionary	Var.	Mean	Std.Dev.	Mean	Std.Dev.
Diction 7.0 [24]	Certainty	0.0179	0.0070	0.0177	0.0067
	Optimism	0.0200	0.0040	0.0199	0.0042
	Realism	0.3005	0.0289	0.2984	0.0288
	Activity	0.0235	0.0060	0.0234	0.0057
	Commonality	0.0277	0.0043	0.0275	0.0041
Loughran and McDonald [1]	Positive	0.0132	0.0023	0.0131	0.0024
	Negative	0.0270	0.0047	0.0272	0.0051
	Uncertainty	0.0136	0.0026	0.0135	0.0027
	Litigious	0.0110	0.0031	0.0108	0.0029
	Modal	0.0005	0.0004	0.0005	0.0004
	Overall tone	-0.3400	0.0941	-0.3445	0.1036
Readability	Gunning fog index	10.55	0.83	10.52	0.82

Table 4. Data sample of financial and sentiment indicators

Ticker	Liquidity	Beta	…	Certainty	Optimism	…	Class
A	0.0092	1.04	…	0.0107	0.0210	…	Neg
AAN	0.0074	0.92	…	0.0234	0.0151	…	Pos
AAP	0.0107	0.96	…	0.0169	0.0200	…	Pos
…	…	…	…	…	…	…	…
ZUMZ	0.0116	1.01	…	0.0073	0.0177	…	Pos

Table 5. Comparison of ROC classification performance – single approaches

	NB	SVM	C4.5	k-NN	NN
Financial	0.5200±0.0514	0.5034±0.0053	0.4983±0.0055	0.4964±0.0439	0.5138±0.0644
Sentim.+Readab.	0.4926±0.0603	0.5000±0.0000	0.5000±0.0000	0.5007±0.0503	0.5030±0.0364
BoW_200uni	0.5195±0.0416	0.5151±0.0147	0.5033±0.0426*	0.5089±0.0501	0.5285±0.0599
BoW_200bi	0.5127±0.0468	0.4946±0.0230*	0.5241±0.0486	0.5041±0.0304	0.5142±0.0332
BoW_200tri	0.5280±0.0423	**0.5392±0.0461**	0.5265±0.0574	0.5200±0.0396	0.5420±0.0511
BoW_500uni	0.5185±0.0399	0.4975±0.0259	0.4853±0.0363*	0.4783±0.0462*	0.5154±0.0436
BoW_500bi	0.5212±0.0455	0.5002±0.0519*	0.4969±0.0512*	**0.5250±0.0348**	0.5286±0.0427
BoW_500tri	0.5284±0.0462	0.5148±0.0251	0.5145±0.0455	0.5226±0.0365	0.5353±0.0343
BoW_1000uni	0.5220±0.0455	0.4722±0.0276*	0.5051±0.0285	0.4683±0.0602*	0.4735±0.0496*
BoW_1000bi	0.5209±0.0478	0.5049±0.0374	0.5091±0.0518	0.5159±0.0514	0.5171±0.0480
BoW_1000tri	0.5333±0.0488	0.5149±0.0280	0.5121±0.0317	0.4937±0.0322*	0.5348±0.0354
BoW_2000uni	**0.5373±0.0505**	0.4791±0.0367*	0.4848±0.0564*	0.4833±0.0401*	0.4904±0.0578*
BoW_2000bi	0.5218±0.0402	0.4941±0.0375*	0.5096±0.0321	0.5054±0.0595	0.5015±0.0495
BoW_2000tri	0.5308±0.0376*	0.5351±0.0512	**0.5387±0.0452**	0.5169±0.0581*	**0.5554±0.0493**

* ROC significantly lower at p=0.05.

Table 6. Comparison of TP and TN rates [%] – single approaches

	NB		SVM		C4.5		k-NN		NN	
	TP	TN	TP	TN	TP	TN	TP	TN	TP	TN
Financial	14.8	86.4	99.9	0.8	98.7	1.0	63.3	36.4	99.9	0.2
Sentim.+Readab.	74.8	23.1	100.0	0.0	100.0	0.0	57.0	43.7	99.6	0.5
BoW_200uni	52.8	49.0	56.9	46.2	61.7	38.5	60.5	40.4	95.9	4.0
BoW_200bi	55.0	46.6	57.6	41.3	58.4	45.8	61.5	39.0	94.5	6.6
BoW_200tri	64.3	38.5	66.3	41.6	53.4	48.8	60.2	41.7	69.2	38.2
BoW_500uni	52.9	49.8	47.2	52.3	55.6	42.7	59.2	38.2	91.6	7.4
BoW_500bi	58.5	44.3	54.2	45.8	57.6	41.3	62.9	38.7	63.5	39.5
BoW_500tri	63.1	38.7	62.2	40.8	55.9	47.9	61.0	42.7	62.6	41.2
BoW_1000uni	54.7	47.7	48.3	46.1	57.2	43.4	61.0	34.9	62.2	34.1
BoW_1000bi	57.7	45.4	52.9	48.1	56.3	44.2	64.4	36.6	61.2	40.1
BoW_1000tri	62.5	41.9	56.4	46.6	56.4	46.4	55.9	42.7	62.7	41.1
BoW_2000uni	58.4	46.1	53.9	41.9	53.3	43.7	59.4	38.2	58.8	40.9
BoW_2000bi	59.8	44.3	56.1	42.7	56.2	45.6	61.9	38.3	61.5	39.3
BoW_2000tri	61.4	42.4	58.0	49.0	60.6	46.8	60.2	41.1	60.4	44.0

Table 7. Comparison of ROC classification performance – dimensionality reduction methods

	NB	SVM	C4.5	k-NN	NN
CBFS_uni	0.5549±0.0419	0.5122±0.0253*	0.5434±0.0534	0.5063±0.0552*	0.5587±0.0381
CBFS_bi	**0.5682±0.0264**	0.5543±0.0247	0.5427±0.0513*	0.5133±0.0432*	0.5717±0.0354
CBFS_tri	0.5655±0.0436	**0.5552±0.0332***	0.5458±0.0488*	**0.5299±0.0635***	**0.5845±0.0546**
LSA_uni	0.4977±0.0683	0.4966±0.0162	0.5000±0.0000	0.4836±0.0529	0.4978±0.0451
LSA_bi	0.5176±0.0413	0.5030±0.0264	0.4977±0.0072	0.4819±0.0281*	0.5194±0.0491
LSA_tri	0.5331±0.0496	0.5154±0.0325	0.5022±0.0230*	0.4636±0.0554*	0.5275±0.0359
LSA+cosine_uni	0.5425±0.0750	0.5030±0.0088*	**0.5523±0.0756**	0.5135±0.0478*	0.5537±0.0319
LSA+cosine_bi	0.5206±0.0478	0.5034±0.0084	0.5165±0.0423	0.5000±0.0428	0.4960±0.0562
LSA+cosine_tri	0.5356±0.0434	0.5092±0.0195	0.5350±0.0524	0.5220±0.0188	0.5140±0.0725

* ROC significantly lower at p=0.05.

Table 8. Comparison of TP and TN rates [%] – dimensionality reduction methods

	NB		SVM		C4.5		k-NN		NN	
	TP	TN	TP	TN	TP	TN	TP	TN	TP	TN
CBFS_uni	70.9	37.4	85.8	16.6	68.1	39.0	67.3	34.3	80.1	26.1
CBFS_bi	70.3	39.6	77.5	33.3	61.5	47.5	62.6	39.5	97.1	4.4
CBFS_tri	70.5	38.5	79.4	31.7	59.7	51.0	65.2	42.2	82.8	27.2
LSA_uni	77.7	21.8	81.9	17.4	100.0	0.0	58.7	37.1	100.0	0.3
LSA_bi	72.0	32.0	78.3	22.3	94.9	4.7	55.9	40.8	81.0	21.3
LSA_tri	71.1	34.8	76.2	26.8	90.7	9.8	55.6	39.3	99.9	0.2
LSA+cosine_uni	88.9	13.9	90.3	10.3	89.3	12.9	57.2	45.2	49.9	50.2
LSA+cosine_bi	91.7	8.9	90.5	10.2	98.3	1.8	59.2	43.7	38.0	65.0
LSA+cosine_tri	91.2	12.0	89.7	12.1	91.5	10.4	59.2	42.5	60.0	40.0

Table 9. Comparison of ROC performance – combinations of approaches

	NB	SVM	C4.5	k-NN	NN
Fin.+Sentim.+Readab.+ CBFS_uni+bi+tri	0.6010±0.0361	**0.5944±0.0232***	**0.5497±0.0351***	0.5284±0.0343*	**0.6184±0.0580**
CBFS_uni+bi+tri	**0.6018±0.0365**	0.5804±0.0397*	0.5368±0.0240*	0.5259±0.0338*	0.6159±0.0532
Fin.+CBFS_uni	0.5455±0.0434	0.5198±0.0384*	0.5233±0.0555	**0.5298±0.0575**	0.5477±0.0429
Fin.+CBFS_bi	0.5812±0.0342	0.5621±0.0328	0.5329±0.0530*	0.5135±0.0359*	0.5764±0.0254
Fin.+CBFS_tri	0.5937±0.0452	0.5468±0.0358*	0.5156±0.0607*	0.5108±0.0598*	0.5913±0.0547
Fin.+Sentim.+Readab.	0.5152±0.0546	0.4967±0.0124	0.4983±0.0055	0.4755±0.0391*	0.5163±0.0511
Fin.+Sentim.+Readab.+ CBFS_uni	0.5893±0.0460	0.5535±0.0435*	0.5209±0.0553*	0.4993±0.0543*	0.5997±0.0398
Fin.+Sentim. +Readab.+ CBFS_bi	0.5342±0.0439*	0.5528±0.0299	0.5268±0.0490*	0.5118±0.0576*	0.5736±0.0305
Fin.+Sentim.+Readab.+ CBFS_tri	0.5746±0.0334	0.5511±0.0355*	0.5489±0.0426*	0.4974±0.0404*	0.6027±0.0595
Sentim.+Readab.+CBFS_ uni	0.5938±0.0438	0.5600±0.0310*	0.5313±0.0655*	0.5110±0.0517*	0.5944±0.0311
Sentim.+Readab.+CBFS_ bi	0.5398±0.0311*	0.5553±0.0271*	0.5181±0.0527*	0.5181±0.0455*	0.5751±0.0302
Sentim.+Readab.+CBFS_ tri	0.5857±0.0219	0.5545±0.0336*	0.5278±0.0548*	0.5112±0.0422*	0.5894±0.0469

* ROC significantly lower at p=0.05.

Table 10. Comparison of TP and TN rates [%] – combinations of approaches

	NB		SVM		C4.5		k-NN		NN	
	TP	TN	TP	TN	TP	TN	TP	TN	TP	TN
Fin.+Sentim.+Readab.+ CBFS_uni+bi+tri	42.0	70.7	72.3	46.6	59.7	49.7	69.8	34.3	76.4	40.8
CBFS_uni+bi+tri	70.6	46.6	72.4	43.7	59.2	49.0	70.2	33.0	75.7	40.8
Fin.+CBFS_uni	13.9	88.4	82.8	21.2	67.8	38.5	67.3	40.0	76.0	27.0
Fin.+CBFS_bi	19.3	87.2	76.4	36.0	59.6	46.7	63.4	40.0	79.1	31.9
Fin.+CBFS_tri	27.7	79.9	78.3	31.0	57.1	47.9	65.9	36.9	77.6	34.6
Fin.+Sentim.+Readab.	15.5	87.6	96.6	2.8	98.7	1.0	56.6	40.3	95.5	5.3
Fin.+Sentim.+ Readab.+CBFS_uni	29.1	80.1	73.4	37.4	58.0	46.3	63.6	35.8	77.0	35.9
Fin.+Sentim.+Readab.+CBFS_bi	14.7	87.6	73.9	36.7	65.9	41.2	62.6	41.3	73.5	33.8
Fin.+Sentim.+ Readab.+CBFS_tri	20.4	86.1	74.0	36.2	58.5	48.5	59.7	39.0	76.1	37.4
Sentim.+ Readab.+CBFS_uni	71.6	41.8	76.1	35.9	59.4	47.1	65.5	35.5	73.8	40.0
Sentim.+ Readab.+CBFS_bi	66.9	39.1	76.0	35.1	67.6	39.6	63.3	44.0	72.6	32.9
Sentim.+ Readab.+CBFS_tri	69.3	43.7	76.0	34.9	57.7	45.9	62.5	39.6	73.8	37.4

Table 11. Comparison of NN performance with SMF and MNIR

	NN	SMF	MNIR
ROC	**0.6184±0.0580**	0.5451±0.0601*	0.5958±0.0438
TP rate	76.4	75.1	62.5
TN rate	40.8	28.3	50.5

* ROC significantly lower at p=0.05.

Table 12. Comparison of NN accuracy across industries

SIC category	Relative frequency	Accuracy of prediction [%]
B - mining	1.2	70.59
C - construction	3.4	65.96
D - manufacturing	30.2	61.15
E - transportation	18.0	62.90
F+G - wholesale and retail trade	10.7	58.11
H - finance	20.1	65.83
I - services	16.3	57.78

Table 13. Comparison of NN performance over different event windows

	3-day [t-1, t+1]	7-day [t-3, t+3]	11-day [t-5, t+5]
ROC	**0.6184±0.0580**	0.5865±0.0409*	0.6085±0.0504
TP rate	76.4	59.6	62.5
TN rate	40.8	52.7	53.4
Positive class ratio [%]	55.22	51.09	47.32

* ROC significantly lower at p=0.05.