WHEN WORDS MATTER MOST: TAILORING DOMAIN-SPECIFIC DICTIONARIES WITH DECISION ANALYTICS

Abstract

This paper aims to better operationalize managerial decision support for the understanding of financial news disclosures. Previous research commonly measures tone in company disclosures by utilizing manually-selected positive and negative word lists, such as the Harvard IV psychological dictionary. However, such dictionaries may not be suitable for the domain of financial news because positive and negative entries could have different connotations in a financial context. To overcome the problem of words that are selected ex ante, we incorporate several Bayesian variable selection methods to select the relevant positive and negative words from financial news disclosures. According to our findings, the interpretation of words strongly depends on the context and managers thus need to be cautious when relying on manually-selected word lists from the related literature.

Keywords: Decision Support, Financial News, Variable Selection, Dictionary Generation, Bayesian Learning, News Sentiment.

1 Introduction

Organizations are constantly looking for ways to improve their decision-making processes in core areas, such as marketing, firm communication, production and procurement (Turban, 2011). While the classical approach relies on having humans devising simple decision-making rules, modern decision support is predominantly based on statistical evidence that originates from analyzing data (Apte, Liu, Pednault, and Smyth, 2002; Arnott and Pervan, 2005; Asadi Someh and Shanks, 2015; Boylan and Syntetos, 2012; Davenport, 2006; Vizecky, 2011). Crucial aspects of data-driven decision support system entail the prediction of future events, such as consumer behavior or stock market reactions to press releases, based on the analysis of historical data (Apte, Liu, Pednault, and Smyth, 2002; Vizecky, 2011). Decision analytics thus frequently utilizes modeling, machine learning and data mining techniques from the area of *predictive analytics*. In fact, predictive analytics can be instrumented for *"generating new theory, developing new measures, comparing competing theories, improving existing theories, assessing the relevance of theories, and assessing the predictability of empirical phenomena"* (Shmueli and Koppius, 2011).

Predictive analytics frequently contributes to managerial decision support as it is the case when predicting the investor reaction to press releases and financial disclosures (Nassirtoussi, Aghabozorgi, Wah, and Ngo, 2014). In this instance, predictive analytics is typically confronted with massive datasets of heterogeneous and mostly textual content, while simultaneously outcomes are of high impact for any business. Until now, decision support for financial news still relies predominantly on dictionaries which contain positive and negative words in order to measure the tone of a written text. In the domain of financial news, there are several dictionaries available (Henry, 2008; Loughran and McDonald, 2011; Stone, 2002) which, however, show large differences in terms of the entries included. As a consequence, choosing the most suitable dictionary for sentiment analysis is challenging and any choice will not be adequate for news from an arbitrary domain. A further problem is the fact that positive and negative word lists typically contain manually-selected words which also, by definition, assume an equal importance of all included words.

The purpose of this paper is to better operationalize managerial decision support for the understanding of financial news disclosures. This paper contributes to the existing literature by proposing the use of different Bayesian approaches in order to create domain-specific dictionaries for sentiment analysis. In contrast to previous research, which typically selects words manually, we implement statistical methods to select decisive words to generate domain-specific dictionaries. Finally, this paper compares the generated dictionaries with existing dictionaries for financial news and evaluates their predictive performance for sentiment analysis on a validation set. Instead of using a subjective measure, e. g. by manually labelling each announcement, we use the stock market reactions of investors as an exogeneous measure. Subsequently, we compute sentiment values for financial news using the different dictionaries and compare the out-of-sample correlation of these sentiment values with the corresponding stock market returns using a separate test dataset.

Bayesian variable selection approaches are particularly suited to generate domain-specific dictionaries. In contrast to other machine learning approaches, Bayesian learning features a high explanatory power and is especially qualified to draw inferences from data. In fact, Bayesian regularization methods allow for the selection of decisive variables in a regression model. Examples include *ridge regression*, the *least absolute shrinkage and selection operator* (LASSO) and *spike and slab regression*. These regularization methods exclude non-informative noise variables, leading to parsimonious and more interpretable models. In addition, these methods overcome the multicollinearity issues of ordinary least squares (OLS) and, by finding a reasonable trade-off between bias and variance, they solve the problem of overfitting, which occurs if the model complexity is too high. As a result, these methods are appropriate tools for the statistical selection of words in order to generate profound domain-specific dictionaries.

The remainder of this paper is organized as follows. Section 2 provides an overview of related literature which utilizes dictionary-based news sentiment or aims to generate domain-specific dictionaries for financial news. In Section 3, we explain Bayesian regression models as a methodology for dictionary generation. Following this, Section 4 evaluates our generated domain-specific dictionaries in comparison to existing dictionaries using financial news according to two

dimensions: first, we compare both positive and negative word lists and, second, we measure the predictive power for sentiment analysis. Finally, Section 5 discusses managerial implications.

2 Related Work

This section provides an overview of previous publications that also study dictionary-based sentiment and dictionary generation for the sentiment analysis of financial news. Overall, the following section provides evidence that investigating methods to operationalize the decision-making of investors in financial markets, is both a novel and relevant research question to the Information Systems community.

Different approaches have been proposed (e. g. Antweiler and Frank, 2004; Ensuli and Sebastiani, 2010; Hagenau, Liebmann, and Neumann, 2013; Li, Shen, Gao, and Wang, 2010; Schumaker and Chen, 2009) to measure the subjective content of written text, often referred to as *sentiment analysis*. The tone of financial news in previous research is typically measured using positive and negative word lists, i. e. dictionaries. A frequent approach is to measure the tone by calculating the ratio of positive and negative words normalized by the total number of words in a document. By utilizing sentiment analysis, various studies have shown that the linguistic content of a document is useful in explaining stock market returns. For instance, sentiment analysis is used to explain stock returns, stock volatility and firm earnings by the tone of newspapers (e. g. Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008) and regulated ad hoc announcements (Groth and Muntermann, 2011; Liebmann, Hagenau, and Neumann, 2012; Muntermann and Guettler, 2007).

While many publications utilize the Harvard IV psychological dictionary in order to measure the tone of documents, previous research indicates that the Harvard IV list may not be a suitable choice for financial content because the positive or negative entries may have different connotations in a financial context (Loughran and McDonald, 2011). Hence, recent studies attempt to manually select positive and negative words for financial news and propose alternative but static dictionaries (Henry, 2008; Loughran and McDonald, 2011). A major drawback of these dictionaries is that words are not weighted according to their relevance, but implicitly assume that

all words are equally important. To overcome this limitation, a different approach incorporates statistical selection methods to weight words based on the reaction of the stock market to 10-K filings by quantifying the subjective sentiment (Jegadeesh and Wu, 2013). Thereby, the authors use positive and negative word lists from the Loughran-McDonald dictionary as regressors to explain stock market return and determine the weights of words. Table 1 depicts our approach and recapitulates related research in a structured fashion.

In contrast, this paper treats every word from our news corpus as a potential regressor in order to statically generate domain-specific dictionaries for financial news. We attempt to overcome the problem of ex ante selected words, which potentially leads to the erroneous exclusion of relevant regressors. Furthermore, we make use of several Bayesian variable selection methods, which filter out non-informative noise variables and, therefore, show how to statistically select words that are relevant to investors. In addition, such regularization overcomes parts of the multicollinearity problem common to OLS.

Reference	Name/Abbreviation	Positive Words†	Negative Words†	Weighted Words	Statistical Selection	Manual Selection	Source	Benchmark
Harvard IV Psychological Dictionary‡	Harvard IV	1316	1793	✗	✗	✓	Harvard's General Inquirer dictionary	–
Henry (2008)	Henry	53	44	✗	✗	✓	1366 annual press releases from 1998–2002	Stock market return
Jegadeesh and Wu (2013)	OLS&LM	–	–	✓	✓ (OLS)	✗	45,860 10-K reports from the EDGAR database from 1995–2010	Stock market return
Loughran and McDonald (2011)	LM	146	883	✗	✗	✓	70,925 10-K reports from EDGAR database from 1994–2007	Stock market return
This Paper		*125	*107	✓	✓ (OLS, ridge regression, LASSO, spike and slab)	✗	61.241 8-K filings from 2004–2013	Stock market return
		*535	*543	✓	✓ (OLS, ridge regression, LASSO, spike and slab)	✗	14,463 ad hoc announcements from 2004–2011	Stock market return

* Results according to ridge regression.
† Remaining words after stemming entries in the corresponding dictionary.
‡ Available from `http://www.wjh.harvard.edu/inquirer/`.

Table 1: Related literature that generates dictionaries aimed at sentiment analysis in financial news.

3 Methodology

This section introduces our research methodology as depicted in Figure 1. In a first step, each announcement is subject to *preprocessing* steps which transform the running text into a document-term matrix, with each term serving as a potential variable for the different regression approaches. Then, we present the underlying regression methods in the form of ridge regression, the LASSO and spike and slab regression for dictionary generation from a Bayesian point of view and compare these to the classical ordinary least squares.

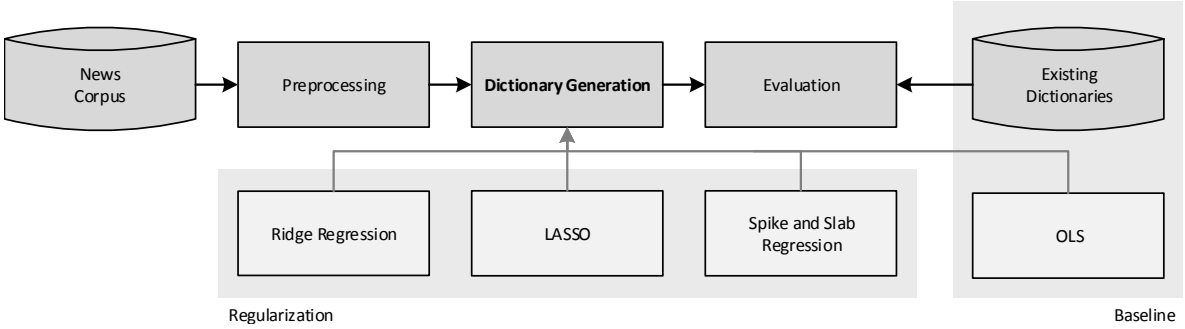

Figure 1: Research model using Bayesian learning to generate domain-specific dictionaries.

3.1 Data Preprocessing

Before performing the actual regressions, several operations are involved in a preprocessing phase. The individual steps are as follows:

1. **Cleaning.** By using a list of cut-off patterns, we omit contact addresses and formatting from ad hoc annoucements in order to extract only the textual components.

2. **Stop word removal.** Words without a deeper meaning, such as *the*, *is*, *of*, ... are named *stop words* (Manning and Schütze, 1999) and can thus be removed. We use a list of 174 stop words (Feinerer, Hornik, and Meyer, 2008).

3. **Stemming.** In computational linguistics, *stemming* refers to the process that reduces inflected words to their stem (Manning and Schütze, 1999). One usually aims to map related words to the same stem, even if this stem is not itself a valid root form, as long as inflected

forms are grouped together. Thus, we annotate words to their stems by using the so-called Porter stemming algorithm (Porter, 1980).

4. **Document-term matrix.** The frequencies of terms that occur in the announcement collection is stored in a document-term matrix. In addition, we remove sparse terms that occur in less than 10 % of all announcements.

5. **Weighting.** The information retrieval approach *term frequency-inverse document frequency* (tf-idf) reflects the importance of a word to a document d in a collection D and allows for the identification of discriminative words (Salton, Fox, and Wu, 1983). Thereby, the raw frequency $f_{t,d}$ of each term t in d is weighted by the ratio of the total number N of documents divided by the number n_t of documents that contain the term t, i. e.

$$tf\text{-}idf(t,d,D) = tf(t,d) \cdot idf(t,D) = f_{t,d} \cdot \log \frac{N}{|\{d \in D \mid t \in d\}|} = f_{t,d} \cdot \log \frac{N}{n_t}. \tag{1}$$

3.2 Dictionary Generation with Elastic Net

Along with our objective of selecting decisive words which have an impact on stock market returns, we utilize the usual linear regression model given by

$$y = \beta_0 + \sum_{j=1}^{P} \beta_j x_j + \varepsilon, \tag{2}$$

with coefficients $\beta = [\beta_0, \beta_1, \ldots, \beta_P]^T$, response $y \in \mathbb{R}^N$, error term $\varepsilon \in \mathbb{R}^N$ and P explanatory variables $x_1, \ldots, x_P \in \mathbb{R}^N$. Then, the ordinary least squares (OLS) estimator calculates the coefficients $\hat{\beta}_{\text{OLS}}$ by minimizing the residual squared error (RSS) via

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} RSS = \arg\min_{\beta} \sum_{i=1}^{N} \left[y_i - \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} \right]^2. \tag{3}$$

Although OLS estimators are the best linear unbiased estimators under the assumptions of the Gauss-Markov theorem, they come along with two major drawbacks (Hastie, Tibshirani, and Friedman, 2009): first, the OLS estimator, generally speaking, features only a moderate prediction accuracy and, second, its interpretability is limited. As a remedy, regularization techniques, such as the elastic net, overcome parts of these problems by sacrificing a reasonable

bias to reduce the variance of the predicted values and therefore improve the overall prediction accuracy. Consequently, prediction accuracy and interpretability can be enhanced by shrinking some coefficients towards zero.

For this purpose, the elastic net extends the classical OLS estimator by introducing l_2- and l_1-norm penalties. Formally, the elastic net calculates its coefficients via

$$\begin{aligned}\hat{\beta}_{\text{ElasticNet}} &= \arg\min_{\beta}\ RSS + \lambda_2 \sum_{j=1}^{P} \beta_j^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j| \\ &= \arg\min_{\beta} \sum_{i=1}^{N} \left[y_i - \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} \right]^2 + \lambda_2 \sum_{j=1}^{P} \beta_j^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j|,\end{aligned} \tag{4}$$

where $\lambda_2 \geq 0$ and $\lambda_1 \geq 0$ are tuning parameter denoting the amount of shrinkage of the coefficients. As a result, the elastic net method includes two special cases, namely, (a) *ridge regression* where $\lambda_2 = \lambda \wedge \lambda_1 = 0$ and (b) the *LASSO* where $\lambda_2 = 0 \wedge \lambda_1 = \lambda$.

The estimated elastic net coefficients $\hat{\beta}_{\text{ElasticNet}}$ are given by the intersection between an ellipse of constant *RSS* and the constraint region given by the shrinkage penalties. In the case of ridge regression, the constraint region is circular in shape with no sharp points. Accordingly, the ridge regression estimates are exclusively non-zero and, thus, ridge regression does not reduce the number of regressors in the model. In contrast, the LASSO yields sparse models which involve only a subset of the variables and, hence, are simpler to interpret. In fact, the LASSO constraint has corners at each of the axes. Consequently, the ellipses are able to intersect the constraint region at an axis which, in turn, sets one of the coefficients exactly to zero.

Regarding the elastic net from a Bayesian viewpoint, the coefficient vector of β follows a prior distribution $p(\beta)$. The likelihood of the data with N observations and P explanatory variables is given by $f(Y \mid X, \beta)$, where $X = (x_1, \ldots, x_P)$, and $Y = (y_1, \ldots, y_N)$. Applying Bayes' theorem yields the posterior distribution

$$p(\beta \mid X, Y) \sim f(Y \mid X, \beta)\, p(\beta \mid X) = f(Y \mid X, \beta)\, p(\beta), \tag{5}$$

where the above equality assumes X to be fixed. In the case of ridge regression, the prior distribution $p(\beta)$ is then given by a normal distribution with zero mean and standard deviation σ

as a function of λ, i. e.

$$p(\beta) = N\left(0, \frac{\sigma^2}{\lambda}\right). \tag{6}$$

In contrast, the LASSO is based on a double-exponential (Laplace) distribution with zero mean and scale parameter as a function of λ, i. e.

$$p(\beta \mid \sigma^2) = \frac{\lambda}{2\sigma} \exp\left\{\frac{-\lambda|\beta_j|}{\sigma}\right\}. \tag{7}$$

We implement ridge regression and the LASSO for dictionary generation as follows: we treat each financial disclosure of the document-term matrix from Section 3.1 as an observation, while we use each column, i. e. each word, as explanatory variables to explain abnormal stock market returns for each announcement. Afterwards, we choose the tuning parameter λ which minimizes the in-sample-error using 10-fold cross-validation. Finally, we re-fit the model using all the observations and the selected value of λ to calculate the ridge regression coefficients. The magnitude of the coefficient estimates $\hat{\beta}_{\text{ridge}}$ and $\hat{\beta}_{\text{LASSO}}$ serve as a measure of variable importance and indicate which words to include in the final dictionary.

3.3 Dictionary Generation with Spike and Slab Regression

The *spike and slab* regression extends a standard linear regression model by performing Bayesian-based variable selection (Varian, 2014). Let $\delta \in \{0,1\}^p$ denote which predictors $1,\ldots,P$ to include in the model. Then, the regression is given by

$$y = \mu + X^{\delta}\beta^{\delta} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{8}$$

where X^{δ} and β^{δ} contain only the non-zero elements of β given by $\delta_i = 1$. Introducing the indicator variable δ_j into the standard linear regression model results into a prior in the form of a mixture of two distributions for each regression coefficient, so-called *spike and slab priors* (Malsiner-Walli and Wagner, 2011). The *spike* component is a distribution with its mass concentrated around zero and the *slab* component is a flat distribution spread over the parameter space. Consequently, variable selection relies on the posterior probability of assigning the corresponding regression effect to the slab component and the selection thus given by the posterior inclusion

probability $p(\delta_j = 1 \mid y)$. We calculate the posterior inclusion probability and the coefficients by Markov Chain Monte Carlo (MCMC) methods (Malsiner-Walli and Wagner, 2011) with $M = 1000$ iterations after a burn-in of 500 draws.

4 Evaluation of Domain-Specific Dictionaries

This section describes our datasets, as well as the experimental setup. Using the methods from the previous section, we proceed by comparing the aforementioned methods to generate domain-specific dictionaries from our news corpora. As presented in Figure 1, we evaluate our methods according to two dimensions. On the one hand, Section 4.2 compares the selected variables (i. e. words) with existing financial dictionaries. On the other hand, Section 4.3 compares the predictive performance between the generated dictionaries and existing dictionaries on a validation set consisting of financial news.

4.1 Dataset

We utilize two separate news corpora from two different markets along with their corresponding disclosure regulations. First, we capture the U. S. market using regulated SEC 8-K filings and, second, we use a corpus of German regulated ad hoc announcements written in English to capture the European market.

For our 8-K filings corpus, we collect all 8-Ks excluding amended documents from the EDGAR website[1] from the years 2004–2013. The complete sample consists of 901,133 filings which then undergoes several filtering steps. First, we select only filings from firms that are listed on NYSE. Second, in order to gain information about the stock market reaction of investors, we remove filings for which we are not able match the SEC CIK numbers to the Thomson Reuters Datastream. Third, we exclude filings that contain less that 200 words (Loughran and McDonald, 2011). These filtering steps result in a final corpus of 76,717 filings.

Our second corpus originates from German regulated ad hoc announcements[2] between January 2004 and June 2011. As a requirement, each announcement must have at least 50 words and be

[1] EDGAR: `www.sec.gov/edgar.html`.

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

written in English. Our final corpus consists of 14,463 ad hoc announcements.

In order to study the stock market reaction, we use the daily *abnormal return* of the corresponding company. Therefore, we use the common event study methodology (Konchitchki and O'Leary, 2011; MacKinlay, 1997) where we determine the normal return, i. e. the return which is expected in the absence of a news disclosure, by a market model. This market model assumes a stable linear relation between market return and normal return. We model the market return using a stock market index, namely, the CDAX, along with an event window of 10 trading days prior to the news disclosure. Finally, we determine the abnormal return as the difference between actual and normal returns. Also here, all financial market data originates from Thomson Reuters Datastream.

4.2 Explanatory Power of Generated Dictionaries

This section extracts words from financial news disclosures that influence the decisions of investors. We start our evaluation with a comparison of the generated dictionaries using the variable selection methods from Section 3.

Word	Ridge	LASSO	SpikeSlab	OLS	LM	Harvard IV	Henry
improv	17.3696	51.4928	55.4023	70.2066	+	+	+
statut	17.0560	8.0725	20.4619	64.8740			
favor	16.9968	33.5073	40.1237	60.2037	+	+	
pleas	12.6671	17.9472	23.1827	52.5895	+	+	+
treat	11.7862	0.9459	10.9117	44.9625		+	
strong	10.1270	20.2586	21.7597	27.6139	+		+
retain	9.2337	7.5925	9.8479	32.8138			
posit	8.9156	6.6232	10.2249	30.0646	+	+	+
electron	7.9246	8.2667	8.5256	28.2501			
own	6.4915	7.3348	10.0435	23.3140			
percentag	5.7296	5.2499	8.7517	19.4002			
increas	5.6114	11.1517	12.8601	22.1028			+
waiver	5.4523	8.4287	9.7955	16.0546			
facil	3.9840	7.8023	9.5347	14.6452			
repurchas	3.7819	7.5322	9.0853	11.7538			

Word	Ridge	LASSO	SpikeSlab	OLS	LM	Harvard IV	Henry
experienc	−17.0564	−30.3721	−37.9384	−49.4450			
although	−16.2789	−27.4814	−33.7679	−44.6959			
declin	−13.4162	−33.8094	−37.1390	−37.5297	−	−	−
howev	−12.6717	−16.4362	−19.9430	−32.5047			
negat	−12.5849	−14.3036	−19.6614	−34.3968	−	−	−
economi	−10.7640	−11.1025	−11.8934	−40.0281			
goodwil	−9.4360	−10.9026	−10.8355	−21.7470			
now	−9.2604	−17.4616	−24.0048	−31.2244			
reduc	−8.5670	−7.2852	−11.9296	−24.9715			
obtain	−8.5508	−8.9842	−14.1417	−37.0206		+	
anticip	−7.9973	−7.6655	−11.5355	−22.3576			
lower	−7.9554	−15.8296	−19.0036	−19.4890	−	−	
decreas	−7.1447	−8.0859	−11.8022	−21.9624	−	−	
review	−6.6823	−7.0251	−9.5494	−15.9370			
impact	−5.7231	−6.9066	−10.3541	−21.6301			

Table 2: Generated dictionaries for the 8-K filings corpus. Left: top 15 positive word stems with the highest coefficient selected according to ridge regression with corresponding coefficients from the LASSO, spike and slab regression and OLS for comparison. A darker background color indicates a higher magnitude. Right: top 15 negative variables respectively. The last three columns denote stems that are included in the positive ("+") and negative ("–") lists of static dictionaries from literature.

Word	Ridge	LASSO	SpikeSlab	OLS	LM	Harvard IV	Henry
posit	8.1100	11.2801	11.2433	20.9920	+	+	+
achiev	6.3779	6.1343	5.8036	16.1572	+	+	+
receiv	6.0038	8.7782	8.8162	14.1496			
reach	5.8657	6.1264	5.9136	12.5493			
includ	5.5461	4.8379	4.6645	14.3105			
strong	5.0331	3.4854	2.8494	12.7406	+		+
base	4.9182	4.4938	4.5190	12.4607			
also	4.2568	1.9484	0.9960	10.5304			
accord	3.8845	2.3572	0.8612	10.9950		+	
key	3.8428	1.0971	0.0000	10.4865			
major	3.7313	3.4113	3.3028	7.5241		+	
rose	3.7306	1.6462	0.4724	9.0222			+
end	3.5033	0.5680	0.0000	15.5818			
increas	3.4439	3.2250	3.1147	12.4237			+
agreement	3.3628	5.7302	5.9535	7.1533		+	

Word	Ridge	LASSO	SpikeSlab	OLS	LM	Harvard IV	Henry
due	−12.1236	−20.3123	−20.3963	−23.9857			
expect	−5.5539	−6.3579	−6.6234	−8.4421			
expens	−5.1379	−3.9766	−3.7220	−9.6116			−
measur	−4.8111	−5.7098	−5.8205	−7.7004			+
adjust	−4.7606	−5.7820	−5.6946	−7.8194			+
loss	−4.1349	−5.4286	−5.3759	−6.0972	−		−
cost	−4.0867	−3.1954	−3.1263	−5.6733	−		−
reduc	−3.8483	−0.9614	0.0000	−4.9157			
oper	−3.7516	−2.4157	−2.4133	−4.7579			
stuttgart	−3.6119	−1.2183	0.0000	−8.9148			
margin	−3.3103	−1.3302	0.0000	−7.2706			−
busi	−3.2037	−2.3751	−2.0357	−3.8965			
level	−3.0985	0.0000	0.0000	−3.9552			
half	−2.9085	−2.4759	−2.2014	−6.5424			
alreadi	−2.8938	−0.1840	0.0000	−6.6026			

Table 3: Generated dictionaries for the ad hoc announcements corpus. Left: top 15 positive word stems with the highest coefficient selected according to ridge regression with corresponding coefficients from the LASSO, spike and slab regression and OLS for comparison. A darker background color indicates a higher magnitude. Right: top 15 negative variables respectively. The last three columns denote stems that are included in the positive ("+") and negative ("–") lists of static dictionaries from literature.

Table 2 lists the top 15 estimated coefficients with the largest positive and negative values according to ridge regression for the 8-K filings corpus. Respectively, Table 3 lists the top 15 positive and negative coefficients for the ad hoc announcements corpus. Since stemming is part of our preprocessing phase, we do not provide the original word itself but its stem. We also state the corresponding coefficients from the LASSO, spike and slab regression and OLS for comparison. In addition, we compare the extracted words with matching entries from static dictionaries. Accordingly, the last three columns denote words which are also included in the static dictionaries from Table 1.

Regarding the selected words from our 8-K filings, we note that the top 15 positive words include many plausible terms, e. g. *improv* or *posit*. These words express a clear positive statement and can be frequently found in sentences such as *"the positive business development was sustainably confirmed"*. In contrast, the top 15 negative words include unexpected outcomes, such as *although* or *however*. Most likely, this result originates from the circumstance that such words commonly occur in connection with a negative statement and often put a positive statement

into a negative perspective, e. g. *"although the loss scenario is slightly brighter than expected, the other reasons are very important"*.

The extracted words from our ad hoc announcements corpus show a similar picture. On the one hand, we find many plausible terms, such as *posit* or *achiev*, which are typically associated with a distinct interpretation independent of the corresponding context. On the other hand, we obtain terms such as *due* or *stuttgart*, that are unexpected at first glance. It is worth noting that these more unexpected terms are mostly attributed to a negative coefficient. Similar to the aforementioned remarks regarding the unexpected terms from the 8-K filings corpus, this is presumably based on the fact that these words frequently express some kind of uncertainty or alleviation of positive statements. In fact, the word *due* can be frequently found in the dataset in sentences like *"turnover falls due to decline in pc sales"* or *"this is particularly due to persistently high oil prices"*. Furthermore, an alleged neutral word like *stuttgart* is contained in sentences such as *"all ordinary shareholders have declared vis-à-vis Porsche Automobil Holding SE, Stuttgart, that they will not participate in the dividend distribution"*.

We observe several additional findings: first, we note that all the regularization methods for each individual corpus do not differ drastically in terms of both the ranking and magnitude of the coefficients. Second, all of the overlapping selected positive or negative variables using ridge regression, the LASSO and spike and slab regression have equally-signed coefficients. Furthermore, this also holds for almost all of the OLS coefficients. Nevertheless, there are significant differences in comparison to the manually-generated dictionaries from related research. For instance, using the ridge regression method, a total of 279 out of 1078 selected words (25.88 %) from the 8-K news corpus are also included in the positive or negative word list of the Harvard IV dictionary. Thereof, 52.69% of the coefficients are equally signed (i. e. are included in the positive word list, if the estimated coefficient is positive, or are included in the negative word list, if the estimated coefficient is negative). In the case of our ad hoc announcements, we are able to classify 53 out of 232 selected words (22.84 %) to the positive or negative word lists from the Harvard IV. Out of these, 66.04 % exhibit the same coefficient sign. The overall correlation of the estimated coefficients with the unweighted word lists from the Harvard IV dictionary is

0.0360 for the selected words from the 8-K filings and 0.1581 for the ad hoc announcements respectively. According to our results, the LM dictionary and the Henry dictionary feature a similar disagreement with the statistically-selected dictionaries. This provides evidence that the positive and negative word lists from the related literature do not fit well into the current financial domain since statistically relevant terms are excluded and words are classified differently to the way in which they are perceived by investors.

Although all the regularization methods do not reveal large differences in terms of sign and magnitude of the coefficients, they differ strongly in terms of the number of included model regressors. All of the regularization methods shrink many coefficients close to zero, whereas the LASSO and spike and slab regression shrink some coefficients exactly to zero (see Table 4). This shrinkage property reduces model complexity but negatively affects the goodness-of-fit. For both news corpora, the OLS approach reveals the highest multiple R^2 value while the regularization methods impede the goodness-of-fit. However, the regularization methods overcome the overfitting and multicollinearity issues of OLS. The number of included positive and negative regressors in the corresponding models, as well as their resulting multiple R^2, is denoted in Table 4. A possible explanation for the generally low R^2-Values is given in a study by Tetlock, Saar-Tsechansky, and Macskassy (2008) which found that *"very few control variables predict next-day returns"* in efficient markets.

Method	8-K Filings			Ad Hoc Announcements		
	#Positive Words	#Negative Words	Multiple R^2	#Positive Words	#Negative Words	Multiple R^2
Ridge	535	543	0.00967	125	107	0.02712
LASSO	33	36	0.00600	36	35	0.01990
Spike and Slab	45	47	0.00691	23	21	0.01973
OLS	584	493	0.01961	145	86	0.03819
Harvard IV	1316	1793	–0.00040	1316	1793	0.01621
LM	146	883	0.00002	146	883	0.00574
Henry	53	44	0.00010	53	44	0.00216

Table 4: Comparison of goodness-of-fit (in-sample set) and the number of positive and negative words per dictionaries.

We note that the ex ante selected word lists from the related literature lead to the erroneous

exclusion of relevant regressors and, therefore, to inferior explanatory power. Moreover, the regularization methods reduce the goodness-of-fit in comparison to OLS. In this regard, the kernel density estimation given in Figure 2 illustrates and compares the shrinkage properties of the regularization methods for the ad hoc announcements corpus. The spike and slab regression leads to the largest amount of shrinkage and, thereby, reduces the R^2 of OLS from 0.0382 to an R^2 of 0.0197. Essentially, the larger the amount of shrinkage, the larger the negative impact is on the goodness-of-fit. However, the results flip for OLS and shrinkage methods once we study their predictive performance in the next section.

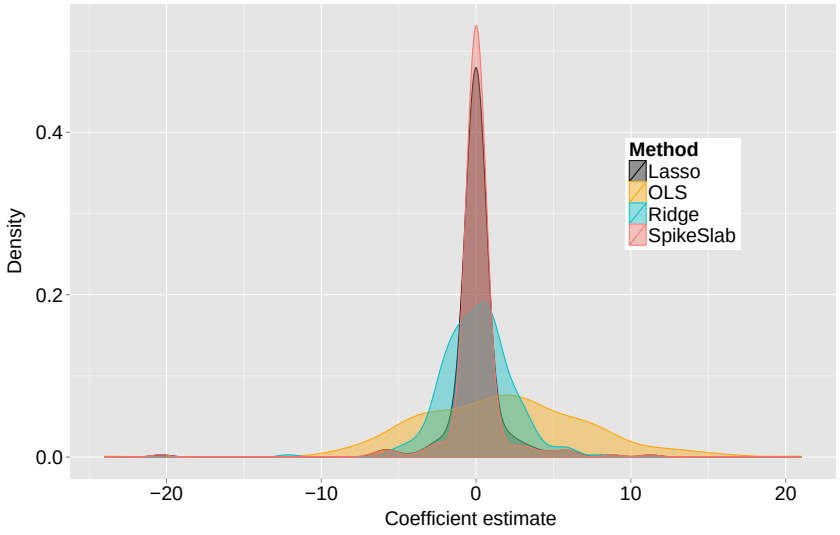

Figure 2: Kernel density estimation of coefficients from generated dictionaries using ad hoc announcements.

4.3 Predictive Performance on Validation Set

This section performs a sentiment analysis to evaluate how the different dictionaries rate in terms of predictive performance. Consequently, we compare the performance of the generated dictionaries (Section 3) with the performance of the existing dictionaries (Table 1). Therefore, we divide our dataset into two subsets: (a) a training set which we use to create our dictionaries, (b) a validation set which we use to validate the dictionaries on out-of-sample data. Regarding the 8-K filings corpus, the training set contains all the filings from the years 2004–2011, i. e. 58,373 observations or 76.09 % respectively. Accordingly, the validation set contains all news disclosures from 2012–2013 resulting in 18,344 observations or 23.91 % respectively. In the case

of the ad hoc announcements corpus, the training set contains all the announcements from the years 2004–2010, giving 12,210 observations (92.96 %) whereas the validation set contains all news disclosures from 2011–2012 giving 925 observations (7.04 %). Afterwards, we generate a dictionary for each method in the aforementioned fashion. Finally, we use the coefficients of the different dictionaries to predict a sentiment value. Then, we calculate the mean squared error and correlation with the corresponding abnormal stock market return. These out-of-sample results for the dictionaries above are presented in Table 5.

Method	8-K Filings		Ad Hoc Announcements	
	MSE	Correlation	MSE	Correlation
Ridge	11.9347	0.0609	81.9695	0.1030
LASSO	11.9286	0.0652	82.1301	0.0948
Spike and Slab	11.9320	0.0626	82.2567	0.0865
OLS	12.2708	0.0480	82.9764	0.0926
Harvard IV	11.9859	−0.0025	82.8238	0.0533
LM	11.9793	0.0245	82.7921	0.0322
Henry	11.9779	0.0318	82.6802	0.0214

Table 5: A validation set is used to evaluate the out-of-sample performance of dictionaries. We compare these in terms of (1) mean squared error and (2) correlation with abnormal stock market returns.

The existing dictionaries from the literature yield generally-speaking inferior results in comparison to the newly generated ones. Out of all the static dictionaries from the related literature, we find the highest correlation with abnormal stock market returns using the Henry Dictionary (Henry, 2008) for the 8-K filings and using the General Inquirer Harvard IV psychological dictionary (Stone, 2002) for the ad hoc announcements. In contrast, regularization methods in the form of ridge regression, the LASSO and spike and slab regression outperform the ordinary least squares estimators in terms of both correlation and mean squared error. Thus, we note that the OLS approach from Jegadeesh and Wu (2013) leads to inferior performance in the current domain.[3] The best performing regularization methods for the individual news corpora are not

[3] We also tested the term importance measure from Jegadeesh and Wu (2013) which utilizes *term frequencies* instead of *term frequency-inverse document frequencies*. Thereby, we find averagely inferior results for ridge regression, the LASSO, spike and slab regression and OLS in terms of mean squared error, and correlation with abnormal stock market returns. The mean squared error of the above methods increases by 0.47 % on average, while correlation with stock market returns reduces by 7.96 % percent on average.

consistent. Regarding the 8-K filings corpus, the LASSO reveals the highest performance on the validation set, yielding an improvement in correlation of 105.03 % in comparison to the best performing dictionary from the related literature. When it comes to the ad hoc announcements, from all the methods, the ridge regression reveals the highest performance on the validation set, yielding an improvement in correlation of 93.25 % in comparison to the existing dictionaries.

5 Discussion and Managerial Implications

We utilize different Bayesian regression methods to generate alternative dictionaries for the sentiment analysis of financial disclosures, with the aim of domain-specificity. The implemented regularization methods reduce the multicollinearity problem of the classical ordinary least squares technique. Moreover, our dynamic approach calculates weights as to how strongly investors are influenced by information in the form of words. This opens a path as to how further research can benefit from statistically-selected instead of manually-selected words in order to study how investors react upon written texts. We note that words classified as positive in dictionaries are not necessarily interpreted positively by investors. Consistent with Loughran and McDonald (2011), we find that linguistically positive words are not directly considered as positive in financial disclosures. In fact, the interpretation of words depends strongly on the context. Consequently, our three main managerial implications of this observation are as follows.

Finding 1. Managers have to be cautious when they frame negative statements using positive terms because of the fact that positive words are not necessarily interpreted as positive.

Finding 2. Generated dictionaries using regularization methods are superior. They outperform static dictionaries from the related literature in terms of goodness-of-fit and predictive performance on a validation set. The most suitable choice regarding the regularization method depends on the characteristics of the data source.

Finding 3. The interpretation of words depends on the context. Therefore, manually-selected dictionaries can lead to misclassification in a financial context. For example, the Harvard IV positive word list contains the term *adjustment* which has negative connotations in financial

news disclosures, e. g. *"the persisting downtrend in demand called for further adjustment of capacity"*, or *"further price adjustments are therefore unavoidable"*.

Altogether, we are able to provide managerial decision support by extracting words in company disclosures that statistically influence the perception of investors in financial markets.

6 Conclusion

The phrasing of news can influence the perception of investors when making decisions on the stock market. To study how investors react to these disclosures, it is common to perform a sentiment analysis based on psychological dictionaries. However, these dictionaries show large differences in terms of included entries and are typically manually-selected. As a consequence, choosing the most suitable dictionary for sentiment analysis is challenging and any choice will not be adequate for news from an arbitrary domain. Thus, this paper generates alternative dictionaries for the sentiment analysis of financial news with the aim of domain-specificity. The statistically selected words open an avenue to better operationalize managerial decision support for the understanding of financial disclosures.

As its main contribution, this paper uses different Bayesian approaches in order to generate domain-specific dictionaries for the sentiment analysis of financial news disclosures. We utilize two separate news corpora and compare the resulting coefficients and the explanatory power of different methods, as well as their predictive out-of-sample performance on a validation set. In terms of predictive performance, statistically-selected dictionaries dominate the existing dictionaries. Moreover, the Bayesian learning implementations outperform the naïve OLS approach from related literature. Regarding the 8-K filings news corpus, the LASSO leads to an improvement in correlation between sentiment value and abnormal stock market return of up to 105.03 % in comparison to the best performing existing dictionary. Altogether, this paper shows that the manually-selected dictionaries from related research can lead to misclassification in a financial context. In fact, by ignoring domain-specific characteristics, these word lists are insufficient in adequately reflecting the perception of investors in financial markets.

References

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

Apte, C., B. Liu, E. P. D. Pednault, and P. Smyth (2002). "Business Applications of Data Mining." *Communications of the ACM* 45 (8), 49–53.

Arnott, D. and G. Pervan (2005). "A Critical Analysis of Decision Support Systems Research." *Journal of Information Technology* 20 (2), 67–87.

Asadi Someh, I. and G. Shanks (2015). "How Business Analytics Systems Provide Benefits and Contribute to Firm Performance?" In: *23rd European Conference on Information Systems (ECIS 2015)*.

Boylan, J. E. and A. A. Syntetos (2012). "Forecasting in Management Science." *Omega* 40 (6), 681.

Davenport, T. H. (2006). "Competing on Analytics." *Harvard Business Review* 134 (1), 98–107.

Ensuli, A. and F. Sebastiani (2010). "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In: *7th International Conference on Language Resources and Evaluation (LREC 2010)*. Ed. by N. Calzolari et al. Valletta, Malta: European Language Resources Assocation, pp. 2200–2204.

Feinerer, I., K. Hornik, and D. Meyer (2008). "Text Mining Infrastructure in R." *Journal of Statistical Software* 25 (5), 1–54.

Groth, S. S. and J. Muntermann (2011). "An Intraday Market Risk Management Approach Based on Textual Analysis." *Decision Support Systems* 50 (4), 680–691.

Hagenau, M., M. Liebmann, and D. Neumann (2013). "Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Capturing Features." *Decision Support Systems* 55 (3), 685–697.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York: Springer.

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

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

Konchitchki, Y. and D. E. O'Leary (2011). "Event Study Methodologies in Information Systems Research." *International Journal of Accounting Information Systems* 12 (2), 99–115.

Li, X., J. Shen, X. Gao, and X. Wang (2010). "Exploiting Rich Features for Detecting Hedges and their Scope." In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 78–83.

Liebmann, M., M. Hagenau, and D. Neumann (2012). "Information Processing in Electronic Markets: Measuring Subjective Interpretation Using Sentiment Analysis." In: *Proceedings of the International Conference on Information Systems (ICIS 2013)*. Association for Information Systems.

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

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

Malsiner-Walli, G. and H. Wagner (2011). "Comparing Spike and Slab Priors for Bayesian Variable Selection." *Austrian Journal of Statistics* 40 (4), 241–264.

Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Muntermann, J. and A. Guettler (2007). "Intraday Stock Price Effects of Ad Hoc Disclosures: The German Case." *Journal of International Financial Markets, Institutions and Money* 17 (1), 1–24.

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

Porter, M. F. (1980). "An Algorithm for Suffix Stripping." *Program: Electronic Library and Information Systems* 14 (3), 130–137.

Salton, G., E. A. Fox, and H. Wu (1983). "Extended Boolean Information Retrieval." *Communications of the ACM* 26 (11), 1022–1036.

Schumaker, R. P. and H. Chen (2009). "A Quantitative Stock Prediction System Based on Financial News." *Information Processing & Management* 45 (5), 571–583.

Shmueli, G. and O. Koppius (2011). "Predictive Analytics in Information Systems Research." *MIS Quarterly* 35 (3), 553–572.

Stone, P. J. (2002). *General Inquirer Harvard-IV Dictionary*. Harvard University, Cambridge, MA.

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

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

Turban, E. (2011). *Business Intelligence: A Managerial Approach*. 2nd Edition. Boston, MA: Prentice Hall.

Varian, H. R. (2014). "Big Data: New Tricks for Econometrics." *The Journal of Economic Perspectives* 28 (2), 3–27.

Vizecky, K. (2011). "Data Mining Meets Decision Making: A Case Study Perspective." In: *Americas Conference on Information Systems (AMCIS 2011)*, Paper 453.