# GENERATING DOMAIN-SPECIFIC DICTIONARIES USING BAYESIAN LEARNING

*Complete Research*

Pröllochs, Nicolas, University of Freiburg, Freiburg, Germany, nicolas.proellochs@is.uni-freiburg.de

Feuerriegel, Stefan, University of Freiburg, Freiburg, Germany, stefan.feuerriegel@is.uni-freiburg.de

Neumann, Dirk, University of Freiburg, Freiburg, Germany, dirk.neumann@is.uni-freiburg.de

## Abstract

*This paper aims to operationalize subjective information processing in financial news disclosures. In order to measure news tone, previous research commonly utilizes 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. These domain-specific dictionaries outperform existing dictionaries in terms of both their explanatory power and predictive performance, resulting in an improvement of up to 93.25 % in the correlation between news sentiment and stock market returns. According to our findings, the interpretation of words strongly depends on the context and managers need to be cautious when framing negative content using positive words.*

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

## 1 Introduction

Financial news disclosures are an important source of information for investors when deciding upon an investment. Thus, listed companies are required by law to publish facts that have the potential to influence their valuation. These announcements are not regulated in terms of form and style. Consequently, companies have the opportunity to frame their publications in order to match their own interests. For instance, previous research has shown that companies try to avoid negative information (Bloomfield, 2002) and make negative news more difficult to read and understand (Li, 2008). Further research (Loughran and McDonald, 2011) indicates that companies make use of inflationary positive words and attempt to frame negative terms in positive terms, such as *"did not benefit"*.

These degrees of freedom in form and style are challenging when analyzing the sentiment of financial news (Pang and Lee, 2008). Typically, sentiment analysis utilizes 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 one 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 to all included words.

As a remedy, Bayesian variable selection approaches are particularly suited to overcome parts of these problems in order 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* and the *least absolute shrinkage and selection operator* (LASSO). These regularization methods shrink 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 and their generalization, the *elastic net*, are appropriate tools for the statistical selection of words in order to generate profound domain-specific dictionaries.

The purpose of this paper is to better understand how companies frame their press releases. This paper contributes to the existing literature by using different Bayesian approaches in order to create domain-specific dictionaries for sentiment analysis. In contrast to previous research, which typically manually selects words, we implement statistical methods to select decisive words to generate domain-specific dictionaries. These methods are based on Bayesian variable selection, which vary in the underlying prior distributions. We implement three approaches, namely, (a) ridge regression, (b) the LASSO and (c) elastic net. 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 objective 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.

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.

### 2.1 Dictionary-Based News Sentiment

Different approaches have been proposed (e. g. Antweiler and Frank, 2004; Ensuli and Sebastiani, 2010; 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 (Demers and Vega, 2010; Feuerriegel and Neumann, 2013).

Various studies have shown that the linguistic content of a document is useful in explaining stock market returns. In this context, dictionary-based methods for sentiment analysis are 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), company press releases (Demers and Vega, 2010; Engelberg, 2008; Henry, 2008), regulated ad hoc announcements (Feuerriegel, Ratku, and Neumann, 2015; Groth and Muntermann, 2011;

Muntermann and Guettler, 2007) and 10-K reports (Feldman, Govindaraj, Livnat, and Segal, 2008; Hanley and Hoberg, 2008; Li, 2008).

## 2.2 Dictionary Generation

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 implement word weights 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, thus, show how to statistically select words 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 issued by firms in the telecommunications and computer industry 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, elastic net) | ✗ | 14,463 ad hoc announcements from 2004–2011 | Stock market return |

† 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, LASSO and elastic net for dictionary generation from a Bayesian point of view and compare these to the classical ordinary least squares.
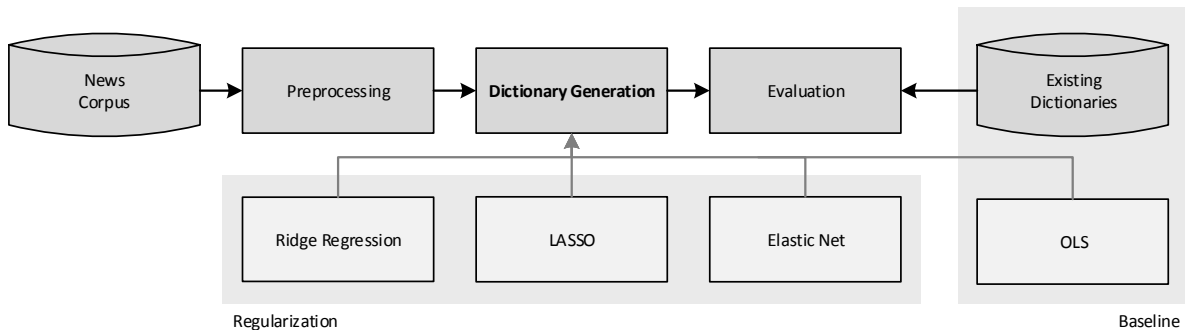


*Figure 1.    Research model using Bayesian learning to generate domain-specific dictionaries.*

### 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.

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

## 3.2  Dictionary Generation with Ridge Regression

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 $\hat{\beta}_{\mathrm{OLS}}$ by minimizing the residual squared error (RSS) via

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

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 ridge regression, 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, ridge regression (Hastie, Tibshirani, and Friedman, 2009) extends the classical OLS estimator by an $l_2$-norm penalty term $\lambda \sum_{j=1}^{P} \beta_j^2$.

Formally, the ridge estimator calculates its coefficients via

$$\hat{\beta}_{\mathrm{ridge}} = \arg\min_{\beta} RSS + \lambda \sum_{j=1}^{P} \beta_j^2 = \arg\min_{\beta} \sum_{i=1}^{N} \left[ y_i - \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} \right]^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{4}$$

where $\lambda \geq 0$ is a tuning parameter denoting the amount of shrinkage of the coefficients. The shrinkage property of the ridge regression approach is visualized in Figure 2 (left). The estimated ridge coefficients $\hat{\beta}_{\mathrm{ridge}}$ are given by the intersection between an ellipse of constant $RSS$ and the constraint region. Since the ridge regression constraint is circular in shape with no sharp points, the ridge regression estimates are exclusively non-zero and, thus, ridge regression does not reduce the number of regressors in the model.

From a Bayesian viewpoint, the coefficient vector $\beta$ 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, the posterior distribution takes the form

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

where the above equality assumes $X$ to be fixed. The prior distribution $p(\beta)$ in the case of ridge regression (Figure 2, right) is then given by a normal distribution with zero mean and standard deviation $\sigma$ as a function of $\lambda$, i. e.

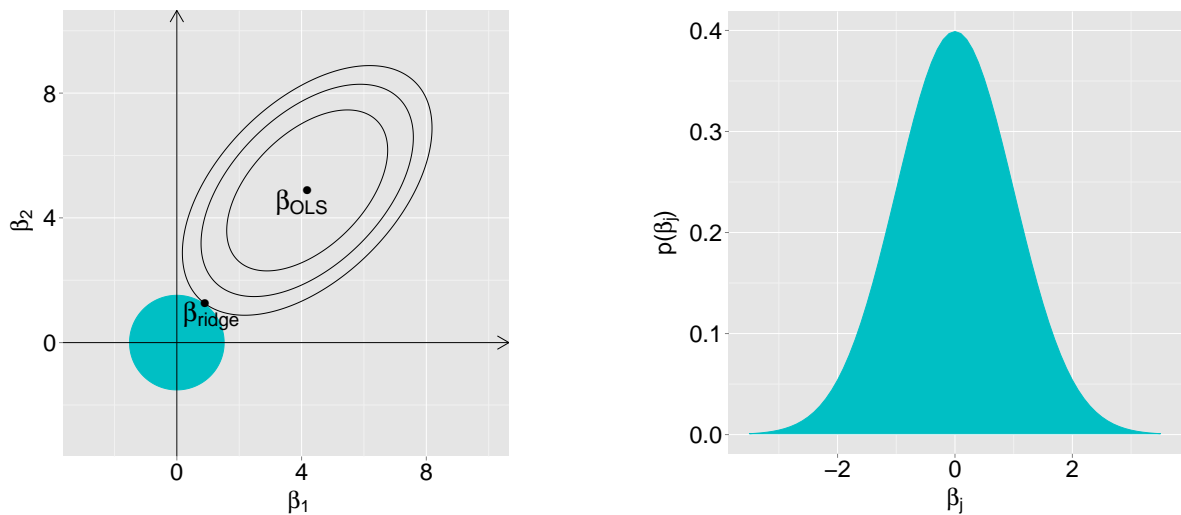$$p(\beta) = N\left(0, \frac{\sigma^2}{\lambda}\right). \tag{6}$$

*Figure 2.* Left: contours of the error and constraint functions for ridge regression. The solid blue area represents the constraint region, while the ellipses show the contours of the residual squared error (RSS). Right: normal distribution represents the prior distribution in ridge regressions.

We implement ridge regression for dictionary generation as follows: we treat each ad hoc announcement 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 $\lambda$ 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 $\lambda$ to calculate the ridge regression coefficients. The magnitude of the coefficient estimates $\hat{\beta}_{\text{ridge}}$ serves as a measure of variable importance and indicates which words to include in the final dictionary.

## 3.3 Dictionary Generation with the LASSO

Ridge regression does not explicitly shrink certain coefficients to zero and so does not allow for the selection of a subset of variables in the final model. To overcome this problem, we present the least absolute shrinkage and selection operator, LASSO for short. Unlike ridge regression, the LASSO method shrinks some coefficients exactly to zero and thus has the property of variable selection (Hastie, Tibshirani, and Friedman, 2009; Tibshirani, 1996). The LASSO replaces the $l_2$-norm penalty term from the ridge regression by an $l_1$-norm penalty term $\lambda \sum_{j=1}^{P} |\beta_j|$. Then, the LASSO coefficient $\hat{\beta}_{\text{LASSO}}$ estimates are given by

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} RSS + \lambda \sum_{j=1}^{P} |\beta_j| = \arg\min_{\beta} \sum_{i=1}^{N} \left[ y_i - \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} \right]^2 + \lambda \sum_{j=1}^{P} |\beta_j|. \qquad (7)$$

The LASSO yields sparse models which involve only a subset of the variables and, hence, are simpler to interpret. The variable selection property of the LASSO approach is visualized in Figure 3 (left). The LASSO coefficient estimates are given by the first point at which an ellipse of constant RSS touches the constraint region. Since the LASSO constraint has corners at each of the axes, the ellipses are able to intersect the constraint region at an axis which, in turn, sets one of the coefficients exactly to zero.

From a Bayesian view, the LASSO is similar to ridge regression (Hastie, Tibshirani, and Friedman, 2009). Instead of using a normal distribution as prior, LASSO uses a double-exponential (Laplace) distribution with zero mean and scale parameter as a function of $\lambda$. Figure 3 visualizes the prior distribution, which is

defined by

$$p(\beta \mid \sigma^2) = \frac{\lambda}{2\sigma} \exp\left\{ \frac{-\lambda |\beta_j|}{\sigma} \right\}. \tag{8}$$
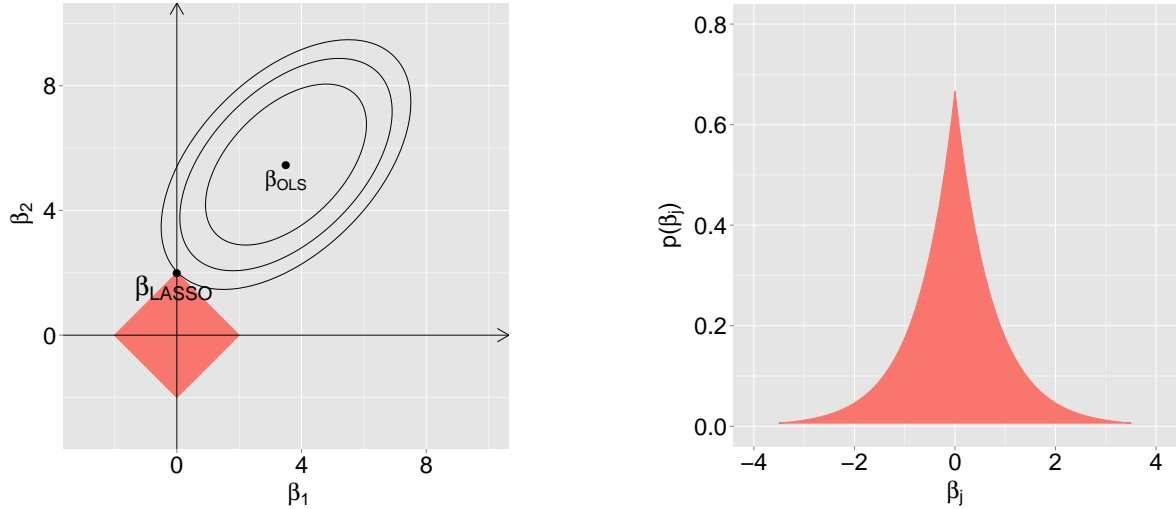


*Figure 3.* *Left: contours of the error and constraint functions for LASSO regression. The solid red area represents the constraint regions, while the ellipses show the contours of the RSS. Right: double-exponential distribution as the prior in the LASSO.*

## 3.4 Dictionary Generation with Elastic Net

If a group of highly correlated variables exists, the LASSO tends to select one variable from a group and ignore the others (Hastie, Tibshirani, and Friedman, 2009). The elastic net overcomes this problem by essentially combining the $l_2$- and $l_1$-norm penalties from both ridge regression and LASSO. Therefore, the elastic net calculates its coefficients via

$$\begin{aligned} \hat{\beta}_{\text{ElasticNet}} &= \underset{\beta}{\arg\min} \; RSS + \lambda_2 \sum_{j=1}^{P} \beta_j^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j| \\ &= \underset{\beta}{\arg\min} \; \sum_{i=1}^{N} \left[ y_i - \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} \right]^2 + \lambda_2 \sum_{j=1}^{P} \beta_j^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j|. \end{aligned} \tag{9}$$

As a result, the elastic net method includes both ridge regression and the LASSO. In fact, they are both special cases, where $\lambda_2 = \lambda \wedge \lambda_1 = 0$ for ridge regression or $\lambda_2 = 0 \wedge \lambda_1 = \lambda$ for LASSO, respectively. In order to combine the two, one introduces $\alpha = \lambda_1 / (\lambda_1 + \lambda_2)$ for reasons of computational tractability. This is equivalent to ridge regression (if $\alpha = 0$) and the LASSO solution (if $\alpha = 1$). From a Bayesian viewpoint, the elastic net prior distribution is given by a compromise between the Gaussian and Laplacian priors from ridge regression and the LASSO, i. e.

$$p(\beta \mid \sigma^2) = \exp\left\{ -\frac{1}{2\sigma^2}(\lambda_2 \beta_j^2 + \lambda_1 |\beta_j|) \right\}. \tag{10}$$

## 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 corpus. 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

Our news corpus originates from German regulated ad hoc announcements[1] between January 2004 and June 2011. The dataset features news that includes financial results, management changes, M&A transactions, dividends, major orders and litigation outcomes, among other disclosures. In related research, ad hoc announcements are a frequent choice (Mittermayer and Knolmayer, 2006) when it comes to evaluating and comparing methods for sentiment analysis. Additionally, this type of news corpus shows several advantages: ad hoc announcements must be authorized by company executives, the content is quality-checked by the Federal Financial Supervisory Authority[2] and several publications analyze their relevance to stock market reactions – finding a direct relationship (e. g. Feuerriegel, Ratku, and Neumann, 2015; Groth and Muntermann, 2011; Muntermann and Guettler, 2007). As a requirement, each announcement must have at least 50 words and be written in the English language. 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. 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.

Table 4 lists the top 15 estimated coefficients with the largest positive and negative values according to ridge regression. Since stemming is part of our preprocessing, we do not provide the original word itself but its stem. We also state the corresponding coefficients from the LASSO 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. Interestingly, the top 15 positive words include many plausible terms, e. g. *posit* or *strong*. The prime positive word *posit* is typically associated with a positive interpretation independent of the corresponding context, e. g. *"the positive business development was sustainably confirmed."*. In contrast, the top 15 negative words include unexpected outcomes, such as *due* or *stuttgart*. Most likely, this result originates from the fact that these negative terms commonly occur in connection with other negative terms. 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"*.

---

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

| Word | Ridge | LASSO | OLS | LM | Harvard-IV | Henry |
|------|-------|-------|------|----|------------|-------|
| posit | 8.2246 | 11.5160 | 21.2947 | + | + | + |
| achiev | 6.4083 | 6.1571 | 16.1932 | + | + | + |
| receiv | 6.0094 | 8.7796 | 14.1281 | | | |
| reach | 5.8315 | 6.0608 | 12.4010 | | | |
| includ | 5.5381 | 4.8170 | 14.2703 | | | |
| strong | 4.9381 | 3.2962 | 12.4780 | + | | + |
| base | 4.9271 | 4.5032 | 12.4943 | | | |
| also | 4.2684 | 1.9581 | 10.5129 | | | |
| accord | 3.8860 | 2.3546 | 11.0324 | | + | |
| key | 3.8675 | 1.1386 | 10.5325 | | | |
| major | 3.7253 | 3.4266 | 7.4819 | + | | |
| rose | 3.7248 | 1.6388 | 8.9633 | | | + |
| end | 3.5282 | 0.6014 | 15.6410 | | | |
| increas | 3.4596 | 3.2608 | 12.4398 | | | + |
| agreement | 3.3520 | 5.7091 | 7.1070 | + | | |

| Word | Ridge | LASSO | OLS | LM | Harvard-IV | Henry |
|------|-------|-------|------|----|------------|-------|
| due | −12.1175 | −20.2709 | −23.9450 | | | |
| expect | −5.5611 | −6.3802 | −8.4687 | | | |
| expens | −5.1539 | −4.0054 | −9.6521 | − | | |
| measur | −4.8057 | −5.6965 | −7.6782 | + | | |
| adjust | −4.7658 | −5.7922 | −7.8425 | + | | |
| loss | −4.1345 | −5.4267 | −6.0693 | − | − | |
| cost | −4.0960 | −3.2162 | −5.7082 | − | − | |
| reduc | −3.8488 | −0.9615 | −4.9228 | | | |
| oper | −3.7953 | −2.4990 | −4.8841 | | | |
| stuttgart | −3.6143 | −1.2239 | −8.8775 | | | |
| margin | −3.3067 | −1.3047 | −7.2766 | | − | |
| busi | −3.2129 | −2.3831 | −3.9363 | | | |
| level | −3.0957 | 0 | −3.8733 | | | |
| alreadi | −2.9373 | −0.2248 | −6.6868 | | | |
| half | −2.9064 | −2.4723 | −6.5031 | | | |

*Table 2. Left: top 15 positive word stems with the highest coefficient selected according to ridge regression with corresponding coefficients from the LASSO 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.*

We observe several interesting findings: first, we note that all the regularization methods do not differ drastically in terms of both the ranking and magnitude of the coefficients. Second, all of the selected positive or negative variables using ridge regression and the LASSO 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. A comparison of the agreement between the above dictionaries is provided in Table 3. Evidently, the generated dictionaries using regularization methods feature a higher similarity to the LM dictionary than the others, while we find the largest disparity in comparison to the Henry dictionary. This large difference is based on the fact that the Henry dictionary contains a relatively small number of words in general and, thus, shares only a few common words with the newly generated dictionaries. Overall, the table shows that the dictionaries from Harvard-IV and Henry in particular do not fit well with the current financial domain by classifying words differently to the way they are perceived by investors.

| | Ridge | LASSO | OLS | LM | Harvard-IV | Henry |
|---|-------|-------|-----|-----|------------|-------|
| **Ridge** | − | (1; 1) | (0.85; 0.79) | (1; 1) | (0.88; 0.85) | (0.73; 0.25) |
| **LASSO** | (1; 1) | − | (1; 1) | (1; 1) | (1; 1) | (0.83; 0) |
| **OLS** | (0.85; 0.79) | (1; 1) | − | (0.92; 0.67) | (0.86; 0.81) | (0.66; 0) |
| **LM** | (1; 1) | (1; 1) | (0.92; 0.67) | − | (1; 1) | (0.83; 0) |
| **Harvard** | (0.88; 0.85) | (1; 1) | (0.86; 0.81) | (1; 1) | − | (0.80; 0) |
| **Henry** | (0.73; 0.25) | (0.83; 0) | (0.66; 0) | (0.83; 0) | (0.80; 0) | − |

*Table 3. Comparison of equal cofficient signs in different dictionaries. The first entry in the bracket denotes the percentage of words which are positive according to both dictionaries. The second entry gives the ratio of words that are regarded as negative in both dictionaries.*

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 only the LASSO method shrinks

some coefficients exactly to zero (see Table 4). This shrinkage property reduces model complexity but negatively affects the goodness-of-fit. In fact, 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. Evidently, both the LM dictionary and the Harvard-IV dictionary include more negative than positive words, while ridge regression results in a dictionary with the opposite traits: more positive than negative words. 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. In this regard, the kernel density estimation given in Figure 4 illustrates and compares the shrinkage properties of the regularization methods. The LASSO reveals the largest amount of shrinkage and, thereby, reduces the $R^2$ of OLS from 0.0382 to an $R^2$ of 0.0199. 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.

| | Ridge | LASSO | OLS | LM | Harvard-IV | Henry |
|---|---|---|---|---|---|---|
| #Positive Words | 125 | 36 | 145 | 146 | 1316 | 53 |
| #Negative Words | 107 | 35 | 86 | 883 | 1793 | 44 |
| Multiple $R^2$ | 0.0271 | 0.0199 | 0.0382 | 0.0057 | 0.0162 | 0.0022 |

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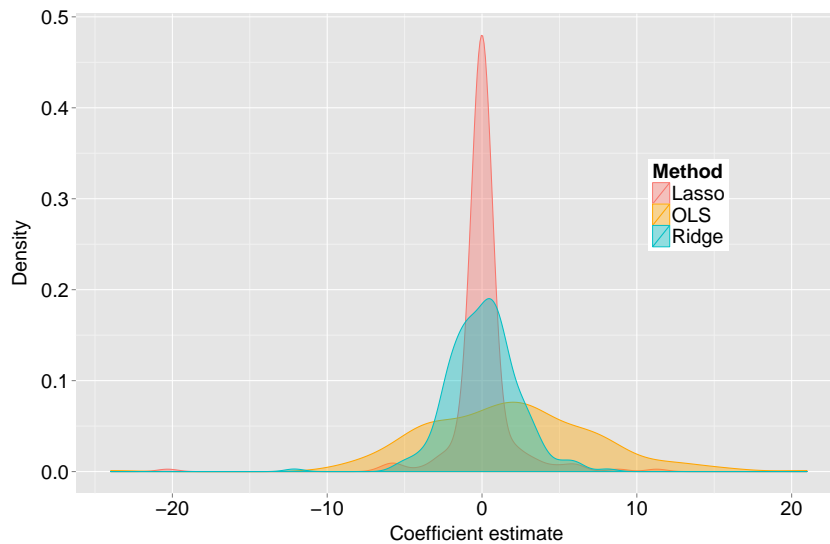


*Figure 4.    Kernel density estimation of coefficients from generated dictionaries.*

## 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. The former training set contains all the announcements from the years 2004–2010, giving 12,210 observations or 92.96 % respectively. The latter validation set contains all news disclosures from 2011–2012 giving 925 observations or 7.04 % respectively. Afterwards, we generate a dictionary for each method in the aforementioned fashion. In addition, we run an OLS regression on the existing dictionaries using all their entries as regressors. 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.

| Criterium | Ridge | LASSO | OLS | OLS&Harvard-IV | OLS&LM | OLS&Henry |
|---|---|---|---|---|---|---|
| Mean Squared Error (MSE) | 81.9695 | 82.1301 | 82.9764 | 82.8238 | 82.7921 | 82.6802 |
| Correlation | 0.1030 | 0.0948 | 0.0926 | 0.0533 | 0.0322 | 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 General Inquirer Harvard-IV psychological dictionary (Stone, 2002). In contrast, regularization methods in the form of ridge regression and the LASSO 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] From all the methods, the ridge regression has the highest performance on the validation set, yielding an improvement in correlation of 93.25 % in comparison to the General Inquirer Harvard-IV psychological dictionary.

Finally, we advance the above regularization methods by using their generalization, the elastic net. The elastic net varies a parameter $\alpha \in [0, 1]$ along a certain interval, which covers a range of potential penalty terms, including those from both ridge regression and LASSO. We compare the correlation with abnormal returns (left) and the out-of-sample mean squared error (right). The results are presented in Figure 5. Both plots show a clear picture: the performance improves when moving from the middle of the interval to its boundaries. We notice clear peaks for both boundaries at $\alpha = 0$ and $\alpha = 1$, representing ridge regression and the LASSO. However, the elastic net achieves the best performance when choosing $\alpha = 0$ and thus results in estimates from the ridge regression. Thus, completely omitting the $l_1$-norm penalty maximizes the correlation with abnormal returns and minimizes the mean squared error. Altogether, we conclude that the best overall performance is obtained by ridge regression.

---

[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 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.
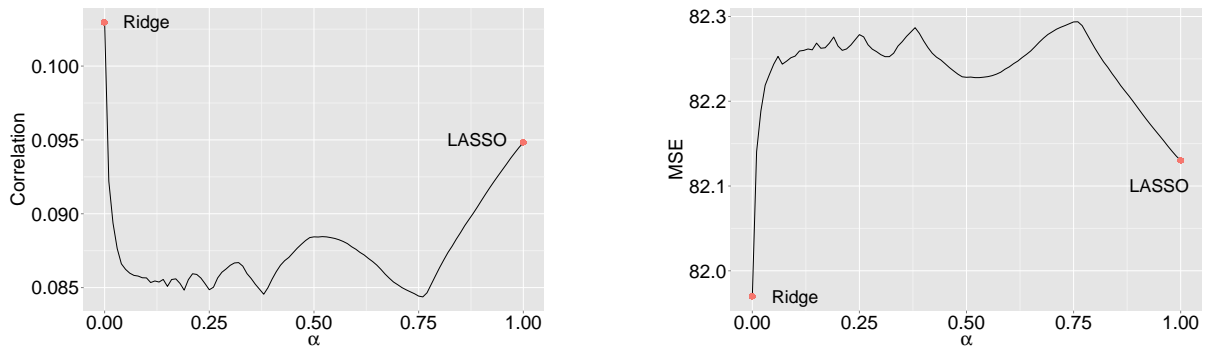
*Figure 5.*    *The elastic net achieves the best performance when choosing the ridge regression estimate. Left: the correlation with abnormal stock market return is maximized at $\alpha = 0$. Right: the mean squared error (MSE) of the elastic net is minimized at $\alpha = 0$. Overall, both metrics improve by omitting the $l_1$-norm penalty.*

## 5   Discussion and Managerial Implications

We utilize different Bayesian regression methods to generate an alternative dictionary for the sentiment analysis of financial disclosures, with the aim of domain-specificity. Furthermore, the implemented regularization methods reduce the multicollinearity problem of the classical ordinary least squares technique. 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. The dynamic approaches calculate weights as to how strongly investors are influenced by information in the form of words. 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. Furthermore, our results indicate that ridge regression outperforms the LASSO. In addition, the elastic net estimate omits the $l_1$-norm penalty term and thus results in the ridge regression solution. This is consistent with prior research (Tibshirani, 1996) which argues that ridge regression dominates the LASSO if multicollinearity among predictors exists.

    **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, our findings indicate that investors expose and recognize the frequent framing of positive into negative words when judging an investment.

## 6   Conclusion

Financial news disclosures are not regulated in terms of form and word choice, providing companies with the opportunity to frame their publications. 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. In order to generate dictionaries dynamically, this paper utilizes different approaches from Bayesian learning in order to statistically select decisive words.

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 then 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, ridge regression dominates the existing dictionaries, leading to an improvement in correlation between sentiment value and abnormal stock market return of 93.25 % in comparison to the best performing existing dictionary. Moreover, the shrinkage property of ridge regression is superior in comparison to the variable selection property of the LASSO. Overall, the Bayesian learning implementations outperform the naïve OLS approach, as well as OLS implementations based on dictionaries from related literature.

In future work, we will advance the above methods in three directions. First, it might be beneficial to include additional news sources from other financial domains to validate the predictive performance of Bayesian variable selection. Second, the inclusion of the spike and slab regression approach, as well as the Γ-LASSO (Taddy, 2013), could open an avenue for further enhancements in variable selection. Third, it is an intriguing notion to integrate this variable selection with negation scope detection (Pröllochs, Feuerriegel, and Neumann, 2015) to reduce the possibility of misclassification due to inverted meanings.

# 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 (3).

Bloomfield, R. J. (2002). "The "Incomplete Revelation Hypothesis" and Financial Reporting." *Accounting Horizons* 16 (3), 233–243 (3).

Demers, E. A. and C. Vega (2010). "Soft Information in Earnings Announcements. News or Noise?" INSEAD Working Paper No. 2010/33/AC. *SSRN Electronic Journal*.

Engelberg, J. (2008). "Costly Information Processing. Evidence from Earnings Announcements." In: *AFA 2009 San Francisco Meetings Paper*.

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)*. (Valletta, Malta, May 17–23, 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 (5).

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal (2008). "The Incremental Information Content of Tone Change in Management Discussion and Analysis." Working Paper. *SSRN Electronic Journal*.

Feuerriegel, S. and D. Neumann (2013). "News or Noise? How News Drives Commodity Prices." In: *Proceedings of the International Conference on Information Systems (ICIS 2013)*. International Conference on Information Systems. (Milan, Italy, Dec. 15–18, 2013). Association for Information Systems.

Feuerriegel, S., A. Ratku, and D. Neumann (2015). "Which News Disclosures Matter? News Reception Compared Across Topics Extracted from the Latent Dirichlet Allocation." *SSRN Electronic Journal*.

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

Hanley, K. W. and G. Hoberg (2008). "Strategic Disclosure and the Pricing of Initial Public Offerings." *US Securities and Exchange Commission Working Paper*.

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). Nic, 363–407 (4).

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

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 (2).

Li, F. (2008). "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2-3), 221–247 (2-3).

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.* Association for Computational Linguistics, pp. 78–83.

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 (1).

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

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

Mittermayer, M.-A. and G. F. Knolmayer (2006). "Text Mining Systems for Market Response to News. A Survey." Working Paper. *SSRN Electronic Journal.*

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 (1).

Pang, B. and L. Lee (2008). "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2 (1–2), 1–135 (1–2).

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

Pröllochs, N., S. Feuerriegel, and D. Neumann (2015). "Enhancing Sentiment Analysis of Financial News by Detecting Negation Scopes." In: *48th Hawaii International Conference on System Sciences (HICSS).*

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

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

Stone, P. J. (2002). *General Inquirer Harvard-IV Dictionary.* William James Hall.

Taddy, M. (2013). "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108 (503), 755–770 (503).

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

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 (3).

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.