INFORMS TutORials in Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Mining Qualitative Attributes to Assess Corporate Performance

Ananda Swarup Das, Aparna Gupta, Gagandeep Singh, L. Venkata Subramaniam

To cite this entry: Ananda Swarup Das, Aparna Gupta, Gagandeep Singh, L. Venkata Subramaniam. Mining Qualitative Attributes to Assess Corporate Performance. *In* INFORMS TutORials in Operations Research. Published online: 04 Nov 2016; 269-281.
https://doi.org/10.1287/educ.2016.0155

TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2016

 | ISBN 978-0-9843378-9-7
http://dx.doi.org/10.1287/educ.2016.0155

Mining Qualitative Attributes to Assess Corporate Performance

Ananda Swarup Das
IBM India Research Labs, New Delhi 110 070, India, anandas6@in.ibm.com

Aparna Gupta
Lally School of Management, Rensselaer Polytechnic Institute, Troy, New York 12180, guptaa@rpi.edu

Gagandeep Singh
IBM India Research Labs, Bangalore 560 045, India, gsingh55@in.ibm.com

L. Venkata Subramaniam
IBM India Research Labs, New Delhi 110 070, India, lvsubram@in.ibm.com

Abstract We present an overview of systems and methods to track ongoing events from sources such as corporate filings, financial articles, expert or analyst reports, press releases, customers' feedback, and news articles that have an effect on corporate performance. In this paper we discuss text analytics and sentiment mining approaches to determine quantitative attributes that can be an indicator of corporate performance. For example, strengths, weaknesses, opportunities, and threats (SWOT) analysis is a well-known structured planning method widely applied to identify the factors determining success or failure of an enterprise. This analysis can be strongly indicative of the business or financial health of the enterprise. It can also provide broader indicators for the firm's business environment, in terms of ease of doing business in the country, as well as for government policies helping (or hurting) the business environment.

Keywords SWOT analysis; text mining; sentiment analysis; opinion mining

1. Introduction

Continual improvement in the quality of life in a country is dependent on the inclusive growth of its economy. It is therefore important to study, monitor, and predict the financial risk factors affecting and determining the health of the business enterprises of an economy. Financial markets provide an integral support to the economic activity of a country. Capital markets, where firms reach out to a broader investor base, provide valuable information regarding the companies that participate in these markets. However, the nature of financial markets makes such predictions regarding the health of a firm extremely difficult solely based on its equity price data. Equity valuation and stock price predictions, as stated in Nassirtoussi et al. [18], are conducted on the basis of firm-specific input data following one of two main themes—namely, (a) technical analysis or (b) fundamental analysis. While historical equity price data are used for technical analysis, fundamental analysis is conducted on the basis of information from news, company reports, social media streams, and any other sources of information that are informative regarding the firm's future prospects.

While technical analysis is almost entirely a quantitative investigation, fundamental analysis must decipher textual content of the sources from which information on the firm, its operations, and future investments is derived. Text mining, on the other hand, is being

widely used in fields such as marketing, political science, sociology, etc. As mentioned in Bholat et al. [5], manually conducted fundamental analysis is bound to have limitations for precise equity valuations because of human constraints in incorporating all the information available for a firm. An obvious suspect for a reason is the complexity involved in transforming volumes of diverse textual data to quantitative indicators. In this tutorial, our primary focus is to utilize extensive textual data in a text analytics framework to perform a detailed analysis for a firm. We show that, using unstructured data primarily extracted from corporate annual reports, expert opinions, press coverage, etc., it is possible to define several qualitative attributes based on the corpus of textual data, which are summaries of firm characteristics that are not directly comprehended as a numeric value. Such an analysis automates the process of identifying and highlighting the risks regarding investment prospects of a particular firm.

Similar to equity analysis, as stated in Edirisinghe et al. [10], bond investors have relied on the accuracy of credit rating agencies as an indicator of credit risk underlying bonds. However, as the financial crisis of 2007 brought to light, the rating agencies run a range of challenges that hinder their ability to generate fully reliable and accurate risk predictions. Rating agencies use both structured and unstructured data, where structured data refers to any data that reside in a fixed field within a record or file (Beal [4]) for the purpose of credit rating assignments. However, unstructured data are not utilized in any formal analytical framework. Text mining techniques, as mentioned earlier, have been widely used in several other disciplines but have not yet been widely adopted in the domain of finance, economics, and regulatory policies (see Bholat et al. [5]). Reasons for this delayed adoption of text mining for these analyses, as mentioned by the authors in Bholat et al. [5], are (i) difficulty of converting textual data into quantitative data, (ii) lack of technical skills needed to handle and manipulate text data, and (iii) ready availability of training for and resources of quantitative data. The recently published One Bank Research Agenda discussion paper (Bank of England [3]) recognized that the analytical potential of textual data has yet to be exploited (Bholat et al. [5]).

Sentiment mining using textual data has been studied for a while. Sentiment mining techniques based on natural language processing (NLP) and general English language dictionaries cannot directly be applied to content from financial reports. This is specifically for the reason that general English language words have specialized interpretation in the financial context. For example, the terms "bullish" and "bearish" are not positive or negative in the English language, while these words have strong positive and negative sentiment, respectively, attached to them in financial lexicon. Research on sentiment analysis based on specialized English dictionary for the financial context, such as Das and Chen [9], Tetlock [24, 25], and Nassirtoussi et al. [18], has focused on mining message dashboards and news headlines to identify the underlying market sentiments with the objective of developing financial trading decisions. Financial reports, such as annual reports, have been examined in the similar dictionary based approach by others (Loughran and McDonald [15], Tetlock et al. [26]) for assessing firm fundamentals. Others have also studied purely statistical approaches for sentiment mining (Feng [11]).

These dictionary-based sentiment analysis techniques are not directly applicable for finding signals for a strengths, weaknesses, opportunities, and threats (SWOT) analysis of a company. Attempting to find forward-looking statements in textual data, as done in Feng [11], is related to what we would need to conduct a SWOT analysis, where a search is conducted for sentences in the text that discuss a firm's future opportunities. Decision support models that utilized text mining to identify SWOT factors from unstructured data sources such as customers' feedback, competitors' press releases, emails, and organisation reports have been studied (Dai et al. [8]).

Opinions and sentiment about a company's current performance, accruals, profit, and liquidity, extracted from textual data, have correlations with several performance factors, such

as stock price and future risk. Research has shown that opinions and sentiments extracted from unstructured data are predictors of corporate performance (Das and Chen [9], Feng [11], Loughran and McDonald [15]).

2. Information Extraction from Financial Documents

The application of text mining techniques in finance has been limited predominantly to predicting financial markets using data from multiple data streams. Several sources of data have been utilized (Nassirtoussi et al. [18]), including firms' financial statements data, currency from the foreign exchange (FOREX) market, financial indexes, as well as textual data from government, regulatory, or monetary policy activities, banking systems, political events, and geographical and meteorological events (e.g., natural or unnatural disasters). Most of these data are in unstructured form, and extracting information from them is a major challenge. The challenge of fundamental analysis is quantifying the textual information. A glance through the literature shows that while dealing with textual information, researchers often mine the content, classifying it into different financial dimensions, and extract the sentiment behind the text. The quantification of the extracted attribute is based on the strength of the sentiment.

As stated in Nassirtoussi et al. [18], the accuracy of sentiment mining in financial text depends predominantly on *feature selection*, *feature reduction*, and *feature representation*. Two approaches to information extraction exist in the literature. Rule-based techniques rely on using financial dictionaries along with several domain-specific extraction rules (Chiticariu et al. [7]). Financial dictionaries such as the Loughran–McDonald financial dictionary (Loughran and McDonald [15]) and AFINN-101 dictionary (Nielsen [19]) aid in identifying the domain-specific words. The dictionary itself may depend on the sector (automobile, infrastructure, finance, etc.). Furthermore, the important terms may depend on the geography and various other contextual factors. Information extraction has been studied in different financial contexts, extracting dividend events from corporate press releases (Malik et al. [16]), where one is looking for specific quantitative information in the unstructured text. Others extract financial entities and the relationships between them (Burdick et al. [6], Chiticariu et al. [7]). Statistical techniques for information extraction classify the content in a financial document along several dimensions including the sector, current performance, accruals, profit, and liquidity.

The sentiment extraction task involves identifying the tone of the discussion within financial documents along the several dimensions listed above. Every discussion at the sentence, paragraph, or document level can be classified into negative, neutral, or positive sentiments. This helps us to understand the outlook along a specific dimension. For example, if we want to determine the outlook for a company along its profit dimension, the first step would be to identify the sentences that refer to profit and then determine the sentiment of the relevant sentences. In Table 1 we tabulate several works in the area in terms of feature selection, dimensionality reduction, feature representation, and the classification algorithm used for sentiment mining. A more extensive literature survey on these aspects can be found in Nassirtoussi et al. [18].

2.1. Statistical Classification Techniques

For content classification tasks, several statistical classification engines have been used, including naïve classifier, linear discriminant analysis, quadratic discriminant analysis, naive Bayesian, Rocchio, etc. A more detailed discussion on each of these classification methods can be found in Alpaydin [1]. We provide brief definitions here:

a. *Naïve classifier*: Each word in AFINN-101 dictionary (Nielsen [19]) carries a positive or negative sign along with a weight depicting the intensity of sentiment it carries. For each sentence extracted from a report, we identify the positive and the negative words present in

TABLE 1. Different approaches utilized in literature for features extraction, reduction, and classification.

Reference	Feature selection	Feature reduction	Feature representation	Classification algorithm
Mittermayer [17]	Bag of words	Selecting a thousand terms	TF-IDF	SVM
Zhai et al. [28]	Bag of words	Top 30 concepts	TF-IDF	SVM with RBF kernel
Peramunetilleke and Wong [20]	Bag of words	Keyword records	TF-IDF	Rule based
Hagenau et al. [12]	Bag of words	Frequency of words	TF-IDF	Association rule
Rachlin et al. [21]	Bag of words	Most influential keywords list	Term frequency	Decision tree
Jin et al. [14]	Latent Dirichlet allocation	Top topics	Topic distribution	Linear regression model
Antweiler and Frank [2]	Bag of words	Top 1,000 words	Boolean TF-IDF	Rule-based naïve Bayes SVM
Feng [11]	Bag of words, tone of content	Predefined dictionary	Binary TF-IDF	naïve Bayes
Das and Chen [9]	Bag of words, triplets	Predefined dictionary	Discrete scores from different classifiers	Majority voting among multiple classifiers

Note. TF-IDF, term frequency–inverse document frequency; RBF, radial basis function.

the sentence, and the final score assigned to a sentence is denoted by $\sum_w x_w$, where x_w is the score for the word w in the dictionary.

b. *Quadratic discriminant analysis*: If the class-conditional densities for a vector $x \in \mathbb{R}^d$ are taken as the normal density, $N_d(\mu_i, \sigma_i)$, the probability of x belonging to C_i can be written as (see Alpaydin [1])

$$p(x \mid C_i) = \frac{1}{(2\pi)^{d/2}(|\Sigma|)^{1/2}} \exp[(-1/2)(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)]. \tag{1}$$

Given a sample $x \in \mathbb{R}^d$, the probability of x belonging to class C_i is given as

$$p(C_i \mid x) = \frac{p(x \mid C_i) \times p(C_i)}{\sum_j p(x \mid C_j) \times p(C_j)}. \tag{2}$$

Since the denominator is same for all the classes, it will be subsequently dropped in the following equations. Define a discriminant function $g_i = \log(p(C_i \mid x))$, which is equivalent to $g_i = \log(p(x \mid C_i)) + \log(p(C_i))$. Therefore,

$$g_i(x) = -(d/2)\log(2\pi) - (1/2)\log(|\Sigma_i|) - (1/2)(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \log(p(C_i)). \tag{3}$$

The first term in the discriminant function is same across all the classes and hence can be dropped.

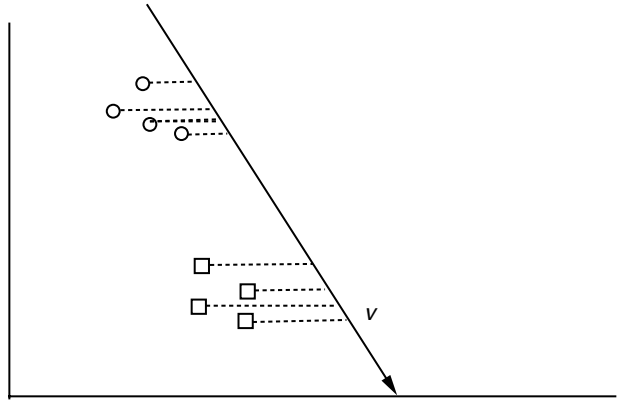
Given the training data for two or more classes, estimates for the means and covariances can be found using the maximum likelihood for each class in the following way:

$$p'(C_i) = \frac{|C_i|}{N}, \tag{4}$$

where N is the total number of samples in the training data and $|C_i|$ is the number of elements from the training sample that belong to class C_i; $m_i = \sum_{x \in C_i} x/|C_i|$. Finally,

$$S_i = \frac{\sum_{x \in C_i}(x-m_i)(x-m_i)^T}{|C_i|}. \tag{5}$$

FIGURE 1. The projection of the elements of the circle and the rectangle classes onto linear discriminant v.



Plugging all these into the discriminant function gives us

$$g_i(x) = -(1/2)\log(|S_i|) - (1/2)(x - m_i)^T S_i^{-1}(x - m_i) + \log(p'(C_i)). \quad (6)$$

On expansion, we get

$$g_i(x) = -(1/2)\log(|S_i|) - (1/2)(x^T S_i^{-1} x - 2x^T S_i^{-1} m_i + m_i^T S_i^{-1} m_i) + \log(p'(C_i)). \quad (7)$$

c. *Linear discriminant analysis*: In case of linear discriminant analysis (LDA), the covariance matrix for all the classes is considered to be the same (see Alpaydin [1]). Thus, Equation (7) can be written as

$$g_i(x) = -(1/2)\log(|S|) - (1/2)(x^T S^{-1} x) + x^T S^{-1} m_i - (1/2) m_i^T S^{-1} m_i + \log(p'(C_i)). \quad (8)$$

The first two terms are the same for all the classes and hence have no discriminatory power. Therefore, they can be dropped. Thus, Equation (8) can be rewritten as

$$g_i(x) = x^T S^{-1} m_i - (1/2) m_i^T S^{-1} m_i + \log(p'(C_i)). \quad (9)$$

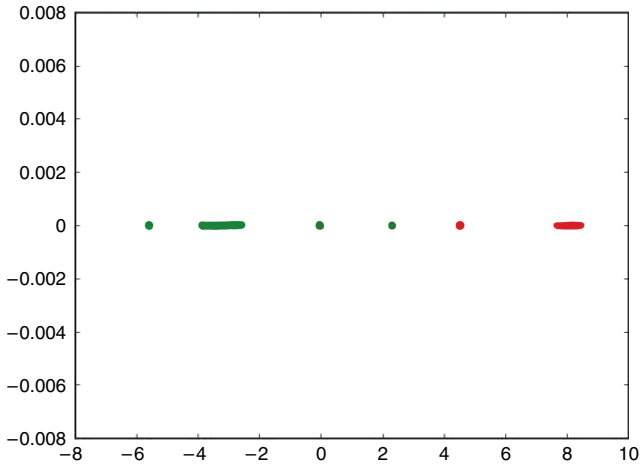
Instead of using a covariance matrix, LDA uses a scatter matrix. The primary objective of LDA is to find a vector, v, that maximizes the separation between the classes after projection on v. To apply LDA, we first map the extracted sentences to a vector space. If there are t words in our lexicon, and each word is assigned a dimension in the vector space, then each extracted sentence can be treated as a t-dimensional unit hypercube.

In the sentiment extraction task, we have primarily three classes: positive, negative, and neutral. A glimpse of the LDA classifier separating positive and negative words is shown in Figure 2. Once the vector v is found, we project the neutral sentences onto v and classify the sentences as positive or negative. Finally, we compute the mean for the neutral sentences falling on the positive side or the negative side. Finally, we compute the means of the positive and negative sentences and project these means onto v.

Given a test vector x, we first find whether it is positive or negative using the vector v by projecting it on to v. If it is positive (respectively, negative), we calculate its distance from the projected means of the positive (respectively, negative) and the mean of the neutrals falling on the positive (respectively, negative) side. On the basis of the distance of x from these means, we assign the final label to x.

d. *Naïve Bayesian classifier*: The use of a naïve Bayesian classifier in the financial domain is not new. Feng [11] used the naïve Bayesian classifier to find forward-looking statements in the management discussion and analysis section of company annual report filings.

FIGURE 2. A glimpse of an LDA classifier separating positive and negative words.



If we assume all the off-diagonal elements of the covariance matrix to be zero, implying that all the features are independent of each other, then the Equation (6) can be written as

$$g_i(x) = -(\frac{1}{2}) \sum_{j=1}^{d} \frac{(x_j - m_{ij})^2}{s_j^2} + \log(p'(C_i)). \tag{10}$$

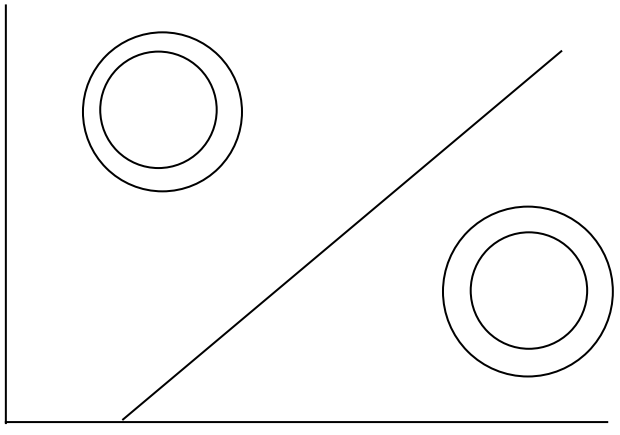
Here, x_j denotes the number of times the word j occurs in the vector x, m_{ij} denotes the mean of the number of times the word j occurs in class C_i, and s_j denotes the standard deviation. This is the naïve Bayesian classifier, where $p(x_j \mid C_i)$ are univariate Gaussian (Alpaydin [1]). Geometrically, the classes are now axis-parallel hyperspheres centered around the mean. See Figure 3 for an illustration.

e. *Rocchio classifier*: If we further assume that the variance for all the features as well as the priors for all the classes to be equal, Equation (6) can be reduced to

$$g_i(x) = -(x - m_i)^T (x - m_i). \tag{11}$$

This classifier is also known as the nearest mean classifier.

FIGURE 3. On assumption that all the features are independent of each other, the classes become axis-parallel hyperspheres centered around the mean.



2.2. Performing SWOT Analysis

Attempting to find forward-looking statements in textual data, as done in Feng [11], is related to what we would need to conduct a SWOT analysis, where a search is conducted for sentences in the text that discuss a firm's future prospects. The information content of forward-looking statements in the management discussion and analysis section of 10-K and 10-Q corporate annual and quarterly report filings assesses the firm's liquidity, capital resources, and operations in a way that many investors can understand. It makes public the information about predictable future events and trends that may affect future operations of the business.

The sentiment (tone) of forward-looking statements from the management discussion and analysis (MDA) section on categories such as current performance, accruals, profit, liquidity, firm age, operations, etc., needs to be determined. A classifier is trained to classify each forward-looking statement into a category, in addition to mining the sentiment (positive, negative, or neutral) for the sentence. To construct the training data for the classifier, randomly selected forward-looking statements need to be manually classified along the content and sentiment dimensions.

Decision support models that utilized text mining to identify SWOT factors from unstructured data sources such as customers' feedback, competitors' press releases, emails, and organizational reports have been studied in the literature (Dai et al. [8]). The category for each sentence in this literature is set to a value from the set: {strength, weakness, opportunity, threat}.

2.3. Financial Opinion Mining

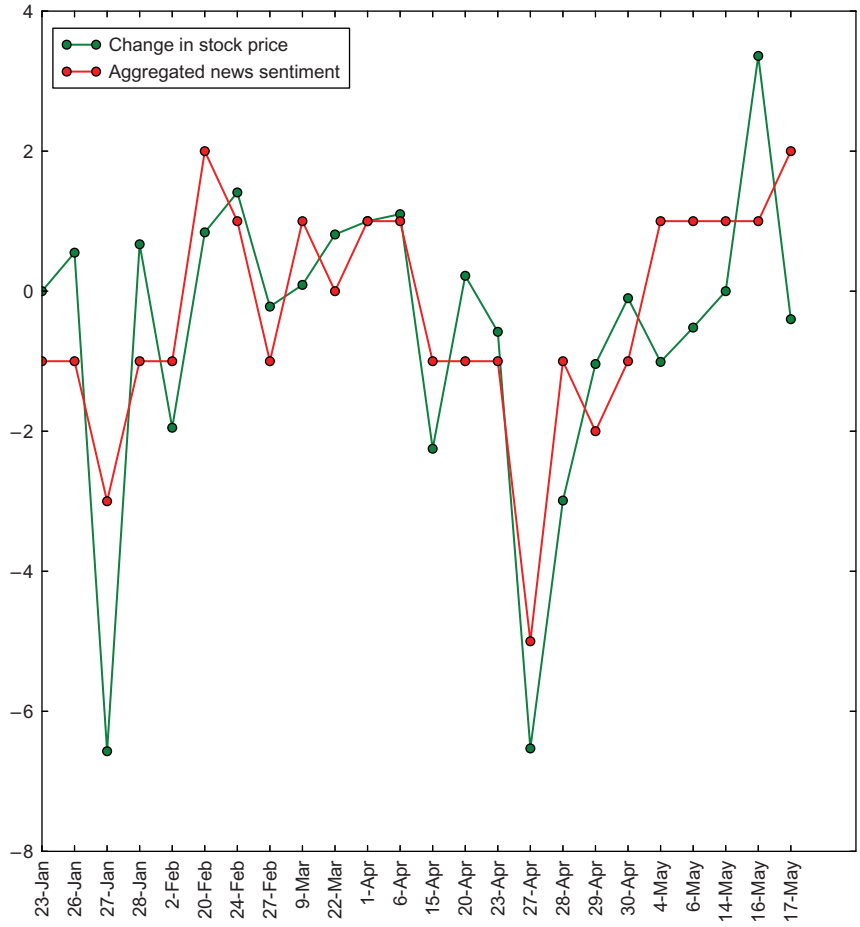
Analysts and industry experts provide detailed analysis of a company, including financial performance, business strategies, competitive position, risk exposures, and the effectiveness of its management. In addition to this textual analysis, they also provide quantitative information in terms of earnings forecasts, stock recommendations, and target prices. The overall opinion in an analyst report is the aggregate of the sentiments (positive, negative, or neutral) across all the sentences in the report (Huang et al. [13]). In this manner, the opinion on a particular stock can be determined from the report. The naïve Bayesian machine learning approach can be used to extract textual opinions from a large sample of analyst reports.

In addition to opinion mining from analyst reports, finding the market sentiment for a company by mining the sentiments of the news headlines can also be beneficial. A detailed study on the impact of breaking news on stock prices has been done in Hagenau et al. [12]. See Figure 4 for an illustration. The red curve in Figure 4 indicates the aggregate sentiment of the market for a company as highlighted by the news headlines. Every positive headline is assigned a score of +1, while every negative headline is assigned −1. A neutral headline is assigned a 0 (zero). The aggregate sentiment is the difference between the number of positive headlines and the number of negative headlines that appeared on a particular day for a particular company. The green curve indicates the daily change of the stock price. It can be seen that on a particular day, when the number of negative headlines is high, the stock price takes a downward plunge.

3. Predictive Analytics Using Text Extracted from Financial Documents

There is some existing work connecting qualitative and quantitative factors affecting corporate performance. The relationship of investor sentiment expressed on stock message boards to stock price movement (Das and Chen [9]) has been shown to exist. The average tone of the forward-looking statements in corporate filings is positively associated with future earnings, even after controlling for other determinants of future performance (Feng [11]).

FIGURE 4. The impact of positive, negative, and neutral headlines on the stock prices of a particular company.



3.1. Linear Regression

Several statistical techniques can be used for prediction. Typically, we need to fit a model to an observed data set of y and X values, where y is a scalar dependent variable and X is one or more explanatory variables (or independent variables). After developing such a model, if an additional value of X is given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.

In linear regression (Seber and Lee [22]), the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Most commonly, the conditional mean of y given the value of X is assumed to be an affine function of X. Less commonly, the median or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Similar to all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

3.2. Risk Prediction

The importance of sentiment words in financial reports toward indicating financial risk can also be quantified (Wang et al. [27]). The goal is to model the relationship between sentiment expressed in financial reports and financial risk. Studies have shown that it is possible to

use a regression model to fit the sentiment information to predict a company's future risk, characterized by its real-value volatility. Wang et al. [27] compared two approaches for risk prediction using volatility as a metric in one approach and model building based on ranking in the other.

In the first approach, for a given collection of financial reports, $D = \mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n$, where each $\mathbf{d}_i \in \mathbb{R}^p$ and is associated with a company c_i, the idea is to predict the company's future risk, which is characterized by its volatility v_i. Such a prediction can be defined by a parameterized function f as follows:

$$\hat{v}_i = f(\mathbf{d}_i; \mathbf{w}). \tag{12}$$

The goal is to learn a p-dimensional vector $\mathbf{w}$ from the training data; $T = \{(\mathbf{d}_i, v_i) \mid \mathbf{d}_i \in \mathbb{R}^p, v_i \in \mathbb{R}\}$.

In the second approach, companies are ranked according to the ranking function modeled as f: $\mathbb{R}^P \to \mathbb{R}$, such that if $f(d_i) > f(d_j)$, then company c_i is ranked higher than company c_j in terms of risk. Ranking support vector machines (SVMs) have been adopted for this task.

3.3. Market Prediction

After the extracted text is transformed into a number of features with a numeric representation, machine learning and regression algorithms can be used Nassirtoussi et al. [18]. The numeric value acts similar to a score or a weight. There are at least six scoring or weighting types that are very popular: Boolean, information gain, chi-square statistics, document frequency, accuracy balanced, and term frequency–inverse document frequency (Taşcı and Güngör [23]). It is possible to extract a sentiment index formed from a time-series accumulation of the sentiment from individual messages (Das and Chen [9]). Das and Chen [9] aggregated sentiment across 24 tech-sector stocks on stock message boards and showed that this sentiment is related to the Morgan Stanley High-Tech Index.

Refer to Figures 5 and 6, where the relationships between some explanatory variables such as (i) the *count of positive, negative, or neutral headlines* and (ii) the *variation in dollar pricing* with a target variable, which is *volume of shares* traded on a day, have been modeled using different regression models. Figure 5 shows the predicted volumes using a linear regression function, while Figure 6 shows the predicted volumes using a polynomial regression function of degree 3.

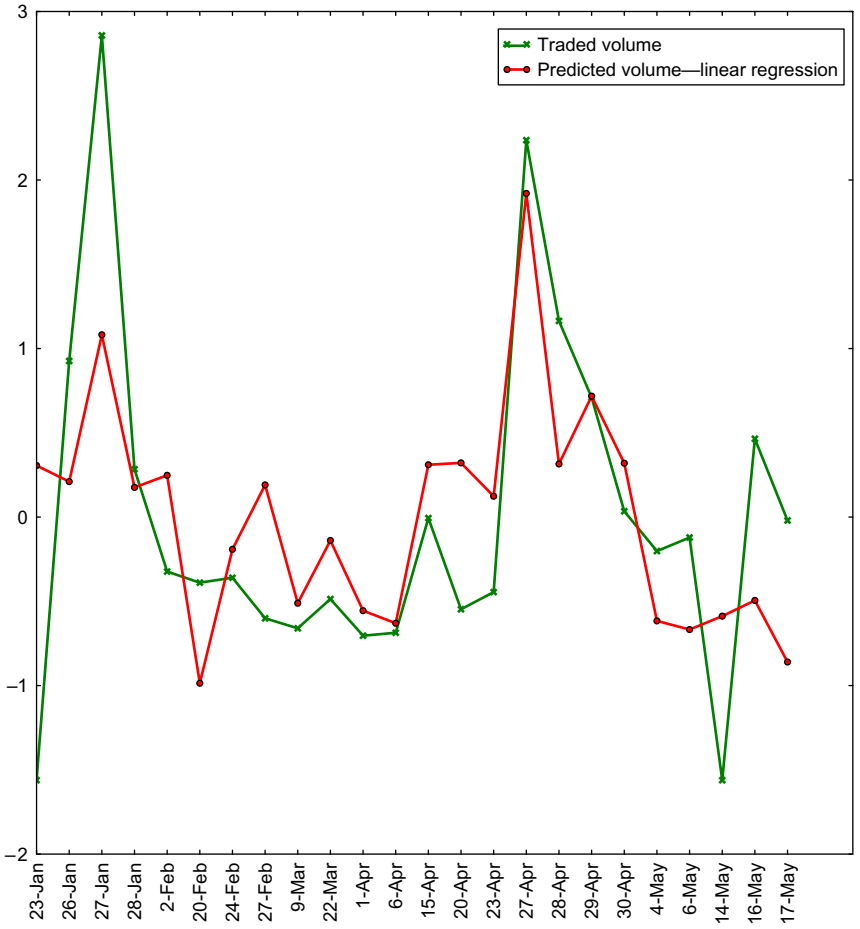
4. General Architecture for Text Mining for Financial Domain

In this section we present the system architecture for a text mining engine. Refer to Figure 7 for an overview. As stated earlier, the most important steps for using text mining in the financial domain are (i) to determine relevant data sources, (ii) extracting the text, (iii) selecting the features, (iv) representing the text, and (v) finding the most relevant features. Next, we elaborate each of these steps.

Data sources: The primary sources of text data, as seen relevant in literature, are annual reports, expert reviews published on trusted web pages, message boards (e.g., Twitter) and discussion boards (e.g., Yahoo! Finance board), and news articles and headlines.

Relevant text: For finding relevant texts, financial dictionaries such as Loughran and McDonald [15] and Nielsen [19] are used. For each sentence extracted from the relevant data source, as a first step, the stop words such as "is," "the," "and," "but," etc., are removed from the sentences. For each remaining word, its presence is searched in the dictionary. If the word is present in the dictionary, the sentence is considered as *relevant*. If none of the words are present in the dictionary, the sentence is dropped without any further consideration. It should be evident from the discussion that the choice of relevant financial dictionary is very important for finding the relevant text.

FIGURE 5. Predicted volumes traded using linear regression.



Selecting features: The choice of the features is sectorspecific, and in general, a set of widely referred keywords is chosen. Each keyword is considered as a dimension.

Representing text: The sentences are scored in a particular dimension by the number of times a specific keyword occurs in the sentence. Thus, the sentences are converted into sparse vectors of fixed length, where the length of the vector is equal to the number of sector-specific keywords chosen.

Finding the most relevant features: To find the most relevant features, often a data set is divided into training and test data sets. The training data are manually labeled. The choice of training versus test data is a crucial step, as too little training data may underfit the model, i.e., make the model less generalizable, while too much training data may lead to overfitting. Once the training data are prepared, one may find the importance of each feature using the *random forest classifier* technique (Alpaydin [1]). It must be noted that the choice of relevant features is specific to the tasks at hand, which in this case are sentiment mining and topic classification.

Once the model for opinion mining engine is trained, the output of the opinion mining engine, which is often a vector maintaining the count or proportions of positive, negative, and neutral sentences that appeared in the text data across all the sources for a company, is passed on to a prediction engine. In addition, other quantitative data such as volumes of traded shares; opening, closing, highest, lowest, and adjusted stock prices; and financial statements quantitative data are also passed to the prediction engine. The scores for

FIGURE 6. Predicted volumes traded using a polynomial regression of degree 3.

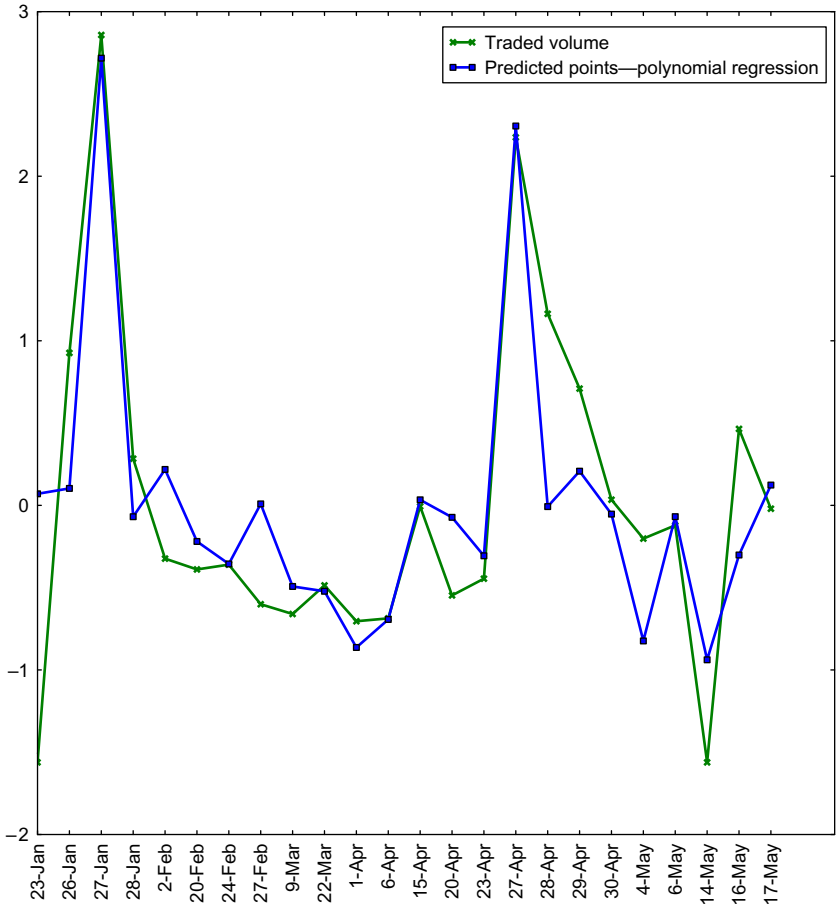
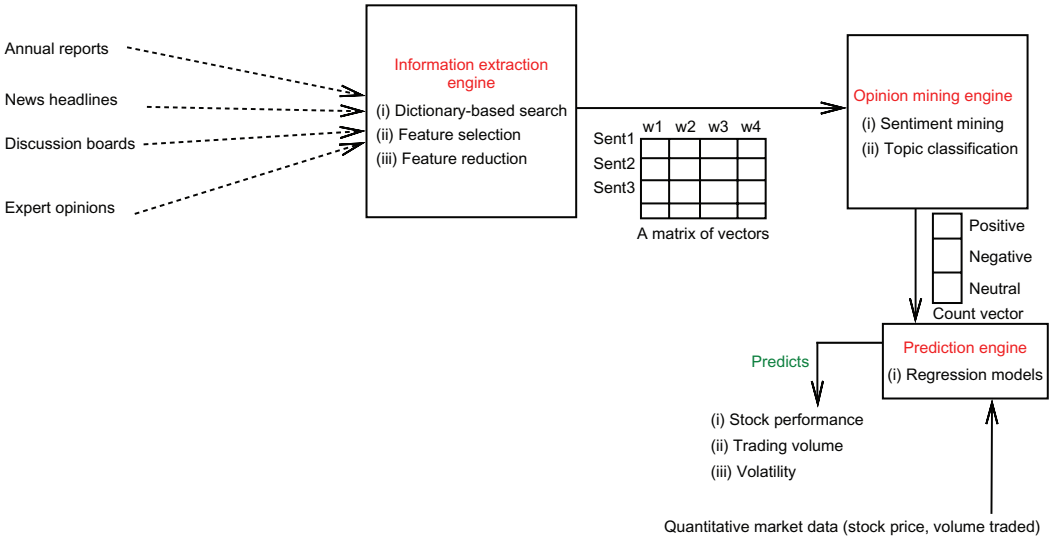


FIGURE 7. A general text mining engine for financial data.



the positive, negative, and neutral data are then considered as explanatory variables, while the stock prices or the trading volumes are treated as the target variable. The relationship (specifically, the weights) between the explanatory variables and the target variable is learned using either a linear regression function or a polynomial regression function of degree d.

Once the right features are selected, opinion mining models are fine-tuned, and the weights for the explanatory variables are learned, the complete system can be used to predict strengths, weaknesses, opportunities, and threats of a company, its stock price, or volumes of shares traded, or its volatility using a new stream of textual data coming for a new company.

5. Conclusions and Outlook

Recent research has shown that it is possible to extract useful financial indicators from textual sources about corporate performance. Text analytics techniques allow the extraction of sentiment about various qualitative corporate performance indicators. It is possible to track this sentiment over time and use it to model various qualitative factors about a company. There are several sources of company information, with several of them containing unstructured information. The information contained in these sources can relate to different aspects such as the stock market sentiment, market reception to the company's products, customer feedback, competitor analysis, etc. While it is easier to utilize quantitative data relating to a company and build mathematical models using these data, the utilization of qualitative information is much more difficult. To be able to utilize qualitative data, first the necessary information needs to be extracted, the sentiment within needs to be accurately measured, and the weight of each of these bits of information needs to be modeled. In this tutorial we have covered some of the ongoing work. We are observing a rising interest in work in this area, where we believe there is lot of scope for automation, by which analysts can be aided beyond resorting to manual analysis of unstructured data sources.

References

[1] E. Alpaydin. *Introduction to Machine Learning*, 2nd ed. MIT Press, Cambridge, MA, 2010.

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

[3] Bank of England. One Bank Research Agenda. Discussion paper, Bank of England, London, 2015.

[4] V. Beal. What is structured data? *Webopedia*, Accessed May 15, 2016, http://www.webopedia.com/TERM/S/structured_data.html.

[5] D. Bholat, S. Hansen, P. Santos, and C. Schonhardt-Bailey. *Text Mining for Central Banks*, Center for Central Banking Studies, Handbook 33. Bank of England, London, 2015.

[6] D. Burdick, M. A. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, S. Vaithyanathan, and S. R. Das. Extracting, linking and integrating data from public sources: A financial case study. *IEEE Data Engineering Bulletin* 34(3):60–67, 2011.

[7] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). Association for Computational Linguistics, Stroudsburg, PA, 1002–1012, 2010.

[8] Y. Dai, T. Kakkonen, and E. Sutinen. MinEDec: A decision support model that combines text mining with competitive intelligence. *International Conference on Computer Information Systems and Industrial Management Applications* (*CISIM*). IEEE, Piscataway, NJ, 2010.

[9] S. R. Das and M. Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53(9):1375–1388, 2007.

[10] C. Edirisinghe, A. Gupta, and W. Roth. Risk assessment based on the analysis of the impact of contagion flow. *Journal of Banking and Finance* 60(November):209–223, 2015.

[11] L. Feng. The information content of forward-looking statements in corporate filings—A naive Bayesian machine learning approach. *Journal of Accounting Research* 48(5):1049–1102, 2010.

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

[13] A. H. Huang, A. Z. Zang, and R. Zheng. Evidence on the information content of text in analyst reports. *Accounting Review* 89(6):2151–2180, 2014.

[14] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, and N. Ramakrishnan. Forex-foreteller: Currency trend modeling using news articles. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD 2013*). ACM, New York, 1470–1473, 2013.

[15] T. Loughran and B. McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66(1):35–65, 2011.

[16] H. H. Malik, V. S. Bhardwaj, and H. Fiorletta. Accurate information extraction for quantitative financial events. *Proceedings of the 20th ACM Conference on Information and Knowledge Management*. ACM, New York, 2497–2500, 2011.

[17] M.-A. Mittermayer. Forecasting intraday stock price trends with text mining techniques. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences* (*HICISS*). IEEE Computer Society, Washington, DC, 2004.

[18] A. K. Nassirtoussi, S. R. Aghabozorgi, Y. W. Teh, and D. C. L. Ngo. Text mining for market prediction: A systematic review. *Expert Systems with Applications* 41(16):7653–7670, 2014.

[19] F. Å. Nielsen. AFINN: A new word list for sentiment analysis on Twitter. *Finn Årup Nielsens Blog* (blog) March 16, https://finnaarupnielsen.wordpress.com/2011/03/16/afinn-a-new-word-list-for-sentiment-analysis/, 2011.

[20] D. Peramunetilleke and R. K. Wong. Currency exchange rate forecasting from news headlines. X. Zhou, ed. *Proceedings of the 13th Australasian Database Conference* (*ADC2002*). Australian Computer Society, Melbourne, Australia, 131–139, 2002.

[21] G. Rachlin, M. Last, D. Alberg, and A. Kandel. ADMIRAL: A data mining based financial trading system. *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining* (*CIDM 2007*). IEEE, Piscataway, NJ, 720–725, 2007.

[22] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis*. John Wiley & Sons, Hoboken, NJ, 2012.

[23] S. Taşcı and T. Güngör. Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications* 40(12):4871–4886, 2013.

[24] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance* 62(3):1139–1168, 2007.

[25] P. C. Tetlock. Does public financial news resolve asymmetric information? *Review of Financial Studies* 23(9):3520–3557, 2010.

[26] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy. More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance* 63(3):1437–1467, 2008.

[27] C.-J. Wang, M.-F. Tsai, T. Liu, and C.-T. Chang. Financial sentiment analysis for risk prediction. *Sixth International Joint Conference on Natural Language Processing,* (*IJCNLP*), 802–808, 2013.

[28] Y. Z. Zhai, A. L. Hsu, and S. K. Halgamuge. Combining news and technical indicators in daily stock price trends prediction. D. Liu, S. Fei, Z.-G. Hou, H. Zhang, and C. Sun, eds. *Advances in Neural Networks—ISNN 2007*, Lecture Notes in Computer Science, Vol. 4493. Springer, Berlin, 1087–1096, 2007.