Application of data mining techniques to stakeholder sentiment analysis towards corporate social responsibility in the social media: a case study on S&P 500 firms

Markus Stiglbauer* and Christian Häußinger

School of Business and Economics,
University of Erlangen-Nürnberg,
Lange Gasse 20, 90403 Nuremberg, Germany
E-mail: markus.stiglbauer@wiso.uni-erlangen.de
E-mail: chr.haeussinger@me.com
*Corresponding author

Abstract: The issue of corporate social responsibility (CSR) and the use of social media have increased in importance over the last few years. Companies are not judged solely on their economic performance because they also have to succeed ecologically and socially. Social media (e.g., Facebook, Twitter, and Google Alerts) exert pressure on companies to meet the expectations of different stakeholders in terms of their CSR. Thus, social media are a useful resource for companies who want to ensure that their CSR self-image matches their CSR public image. Data mining techniques and related software (e.g., RapidMiner) can help companies evaluate their CSR public image and adjust their CSR if there is a mismatch between their self-image and public image. Thus, there may be a continuous CSR (control) cycle that helps companies to gain a long-term competitive advantage over companies who do not manage their CSR strategically.

Keywords: data mining; social media; corporate social responsibility; CSR; stakeholder management; Facebook; Google Alerts; Yahoo! Pipes; RapidMiner; web science.

Reference to this paper should be made as follows: Stiglbauer, M. and Häußinger, C. (2013) 'Application of data mining techniques to stakeholder sentiment analysis towards corporate social responsibility in the social media: a case study on S&P 500 firms', *Int. J. Web Science*, Vol. 2, Nos. 1/2, pp.27–43.

Biographical notes: Markus Stiglbauer is a Professor for Corporate Governance at the School of Business and Economics at the University of Erlangen-Nürnberg (FAU), Germany. He holds a PhD in Economics and Social Sciences and two Master's in Business Administration and Economics from the School of Business, Economics and Management Information Systems at the University of Regensburg, Germany.

Christian Häußinger is an Alumni of the School of Business and Economics at the University of Erlangen-Nürnberg (FAU), Germany. He holds a Master's in Finance, Auditing, Controlling, Taxation (FACT) and Bachelor's in Business Studies both from FAU.

1 Introduction

The ‘sustainability’ of corporate strategy has become an important issue in recent years. Companies are expected to succeed not only economically but also ecologically and socially. Companies can exploit the issue of corporate social responsibility (CSR) to capture new markets and opportunities. However, this can give rise to new requirements because the increased activities of social media stakeholders may exert greater pressure on companies to assume CSR, whereas executives should lead by example. However, ‘responsibility’ and ‘sustainability’ are not just ‘hot topics’ (Porter and Kramer, 2006; Smith, 2007) because managers need to be accountable and they should lead by example. Thus, managers should provide transparent and comprehensive insights into various CSR measures in order to make themselves accountable for their actions in the eyes of stakeholders (Bhattacharya and Sen, 2004; Bhattacharya et al., 2008). Nevertheless, such communication needs to be free of non-binding advertising messages and green- or (white-) washing (Robbins, 2001) if they are to be convincing and if they are to engender trust in a company’s operations. These indicators help companies to manage their internal operations and ensure that their processes are ‘sustainable’. From the outside, these indicators enable stakeholders to make managers responsible (Sen et al., 2006) and increase the pressure on sustainable management and corporate governance (Weaver et al., 1999). Communication about environmental, social, and governmental issues is assumed to reduce the information asymmetry between a company and its stakeholders by helping companies manage their external and internal perceptions, and to meet the expectations of stakeholders in terms of CSR (Eldomiaty and Choi, 2006).

Social media are becoming increasingly important in business life. They function as a communication tool that connects businesses and stakeholders. They allow corporations to obtain rapid feedback from their stakeholders so they can better understand the expectations of their target groups. Social media also allow corporations to promote their products and services directly. However, social media also face risks because resentment and disaffection about specific businesses can spread rapidly throughout the world. Frequently, companies experience difficulties disproving allegations, while their reactions to criticism and the public image of a company can be influenced by social media. For example, social media such as Facebook or Twitter can be used by non-governmental organisations to highlight specific business activities and to communicate their scepticism about the behaviour of companies. Thus, corporations could provide public and moderated channels on social media that control the flow of information based on a transparent information policy. Data mining provides a technological solution that can facilitate the understanding of trends. It is a simple approach for detecting patterns and trends. The interface between statistical methods and computer-based learning provides us with tools for semantic analysis. Information technology can be used to organise data efficiently, extract information from data, and automatically recognise patterns to evaluate information. A wide range of models can be used to apply machine learning. This synthesis of mathematical principles and new technological models can be trained using available data to generate predictions about unseen data. Current societies must acknowledge that there is interdependency between the generation and consumption of information. Given the issues surrounding CSR, data mining can facilitate a better understanding of this interdependency by detecting patterns. Combinations of web services such as Yahoo! Pipes and Google Alerts with data mining

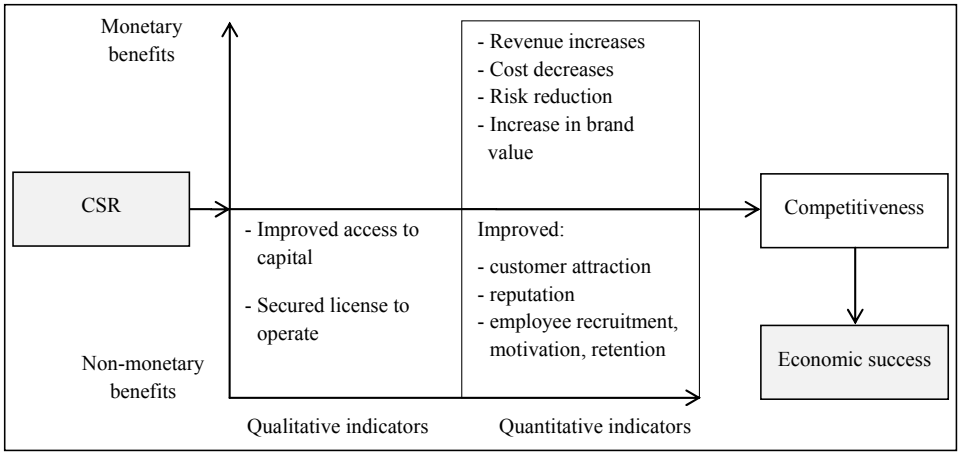
software such as RapidMiner can detect patterns in raw data to make predictions about future data.

2 Theoretical perspectives

2.1 Benefits of CSR

Enterprises should have a process in place to integrate social, environmental, ethical, and human rights, and consumer concerns into their business operations and core strategy while working in close collaboration with their stakeholders with the aim of maximising the shared value created for their owners/shareholders, other stakeholders, and society at large (European Commission, 2011). It is generally acknowledged that CSR gives a competitive advantage (Porter and Kramer, 2006). A closer analysis of the assumed benefits of CSR contributes to a better understanding of its strategic importance for companies. Companies that engage in CSR activities are assumed to accrue non-monetary and monetary benefits, as illustrated in Figure 1.

Figure 1 CSR impact model



Source: Based on Weber (2008, p.250)

The main drivers of CSR are economic and ethical considerations, and other important non-economic aspects such as trust (Mayer et al., 1995) and reputation. These factors can affect a company's sustainability-based brand image (Simmons and Becker-Olsen, 2006) while they may also reduce risks (Derwall et al., 2011) and have ideal benefits (Hermann, 2005). Therefore, perceived CSR is assumed to have an effect on economic benefits, e.g., by increasing the shareholder value or reducing capital costs (through risk reduction) (Münstermann, 2007). CSR also affects customer satisfaction. Customer satisfaction can lead to a company having greater customer loyalty and positive word of mouth (Gruca and Rego, 2005) so customers are willing to pay premium prices (Homburg et al., 2005). This may lead to higher cash flow levels and less volatility in future cash flows, which positively influences the long-term financial performance (Trudel and Cotte, 2009) and the firm's value (Gruca and Rego, 2005). The returns of CSR can be positive or negative

depending on a company's corporate abilities, such as its innovation capability and product quality. Thus, it is important to establish an appropriate combination of CSR activities and product-related capacities (Luo and Bhattacharya, 2006). CSR also has an impact on the cost of equity (El Ghoul et al., 2011) and cost of debt, which can lower the credit risk and allow companies to receive better credit ratings (Bassen et al., 2006). Chang and Shen (2011) found that companies with higher degrees of CSR are perceived as having greater credit-worthiness. Furthermore, El Ghoul et al. (2011) showed that its importance has also increased in recent years [for additional empirical studies see Weber (2008)].

2.2 Assessment of CSR impact and expectations

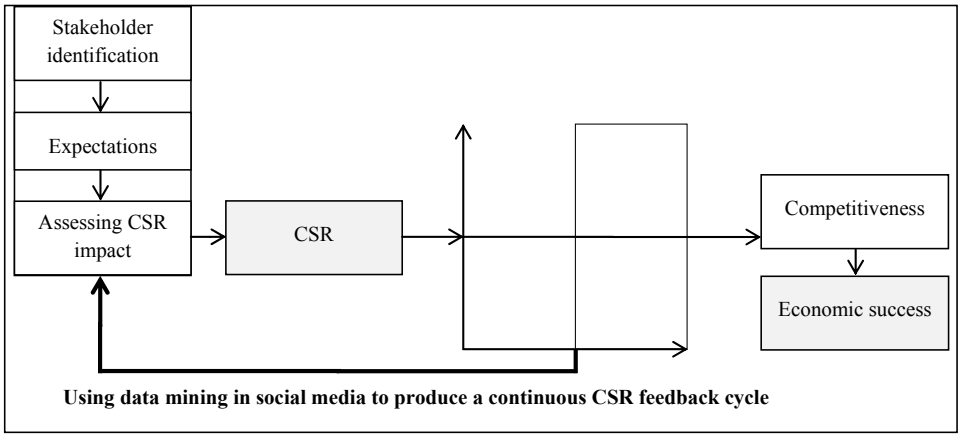
It is useful to explore the different types of stakeholder groups and their expectations about a company in terms of CSR to determine the appropriate content of CSR reports (Finch, 2005). Therefore, CSR strategies can facilitate a frequent and careful dialogue between a corporation and its stakeholders. Thus, people responsible for managing CSR need to consider its strategic impact (Podnar, 2008) and evaluate the success of their CSR projects.

After identifying the key stakeholders and their expectations, companies can assess the qualitative and quantitative impact of their CSR. The qualitative impact can often be measured before the qualitative impact and it provides an indicator of future monetary performance (Kaplan and Norton, 2008). As shown in Figure 2, the 'real' impact of CSR on the competitiveness and performance (via monetary and non-monetary benefits) of a company may initiate a continuous CSR cycle that changes over time. An appreciation of the current state of competitiveness and performance helps companies to select CSR projects and to ensure the ongoing evaluation and monitoring of their long-term CSR activities, which contribute to "a final assessment of impact after the end of each (single CSR) project" [Weber, (2008), p.252]. This approach allows them to identify relevant stakeholders in different contexts, so they can make a fresh assessment of the impact of CSR based on stakeholder expectations, before they modify their CSR activities (if necessary). Anderson and Frankle (1980) found that companies committed to CSR were considered more credible and they were better borrowers that generated higher returns. According to Habisch (2003), this is because companies that demonstrate their commitment to CSR and its reporting are not perceived as focused only on short- and medium-term increases in profits. These interrelationships are of particular significance in regional financial markets. Thus, CSR can contribute to an increase in socially responsible investments (SRI), which have previously been a niche area. In particular, investors who take a long-term view may benefit from CSR because it is focussed on the future potential of the company. Therefore, investing in a company that reports 'good' CSR may pay economic and social dividends. Bhattacharya et al. (2008) found that an organisation with good CSR may acquire business benefits, such as improvements in their image and reputation, if they recruit qualified talent. In this context, Thomas (2005), and Hemingway and MacLagan (2004) argued that several multinational organisations tend to publish their CSR efforts on corporate websites, in recruitment brochures, and in CSR reports so they are perceived as very socially responsible employers.

Moreover, we must strongly advise against companies using CSR as a short-term method for gaining monetary and non-monetary benefits (or simply for public relations). Companies can use CSR to send positive CSR signals to stakeholders, but stakeholders

will expect the maintenance of the same CSR project performance in the long run. If companies are not able to meet these expectations (which are reinforced by their own signals) this can eliminate the positive effects of a single CSR cycle. Furthermore, a loss of trust in one period might taint further CSR cycles/business periods and lower stakeholder expectations about the CSR projects of companies, while it may take a long time to recapture trust in the CSR activities of companies. Thus, we present data mining techniques that may help companies to match their CSR self-image with their CSR public image based on an analysis of social media. This comparison allows companies to take the right CSR action and to meet the expectations of stakeholders.

Figure 2 Strategic stakeholder management cycle by CSR



3 Data mining of social media to produce a continuous CSR feedback mechanism

3.1 Social media: Google Alerts, Yahoo! Pipes, and RapidMiner

The selection of different forms of social media has a strong effect on the web-based analyses used for corporate reporting. In the following empirical analysis, we distinguished two different types of company affinity and three different types of internet resources. Some companies link their web pages with Facebook and Twitter, which are classified as official, and they are moderated by companies as a means of keeping in touch with their customers (Kaplan and Haenlein, 2009). Some companies also publish information about their social responsibility and their fans or followers can comment on these contributions. These official communication channels provide a positive corporate image and useful commentaries. According to a standard code of practice, critical commentaries should be indexed, which means that a specific corporate image could be distorted (Deephouse, 2010).

Other social media resources are viewed as unofficial and they are attributed to non-governmental organisations, news agencies, private persons, and associations. Facebook search results list organisations, companies, and individuals. In the following semantic search, only organisations and other companies related to companies were used

in the evaluation. Some tweets and social media fan pages relate to individual companies, whereas some of those operated by non-governmental organisations provide reports on various companies. These media are often aimed at propounding a contrasting view that provides a counterpoint to corporate publicity. A third major social media resource (in addition to Facebook and Twitter) is Google Alerts. This service provides daily links to further information about companies. This information is published mostly on forums, blogs, or web pages.

RapidMiner is a programme that can be used for data mining and text analytics (Figure 3). RapidMiner was used in the following example to automatically categorise the contributions of social media such as Facebook and Twitter. The content is classified based on its positive and negative characteristics, while the text analytics are based on sentiment analysis. Words with a high emotional meaning and importance are used to classify the textual content of social media, which are known as discriminants. After the algorithm has learnt to classify the text, unknown contributions can be rated based on their positive or negative intentions (Kosorus et al., 2011). Yahoo! offers a cloud service known as 'Pipes' that can be used to process web content. This service is used to join different feeds connected with Google Alerts to generate a single feed. Therefore, RapidMiner needs only one operator to import the feeds, so the computational performance can be optimised during processing. Yahoo! Pipes also contains filters that can be used to separate relevant content via Google Alerts. These filters include terms such as 'CSR' or 'corporate citizenship', which can be used to identify content about the sustainability of selected companies. After running semantic web analyses, RapidMiner performs checks every half an hour to determine whether an updated version of a feed is available, and it selects 100 random contributions from the combined feed. Some companies are more popular than other companies, so Google Alerts will identify more internet resources related to them. The most popular companies often operate in the business-to-consumer sector or they may be in the public eye, such as banks during and after the financial crisis of 2007. Thus, the combined feed weights of articles about popular companies are higher than those of other companies (Lin et al., 2006). To prevent bias in the average weighting of the feed, the contributions of some companies are weighted based on the average number of internet resources found by Google Alerts.

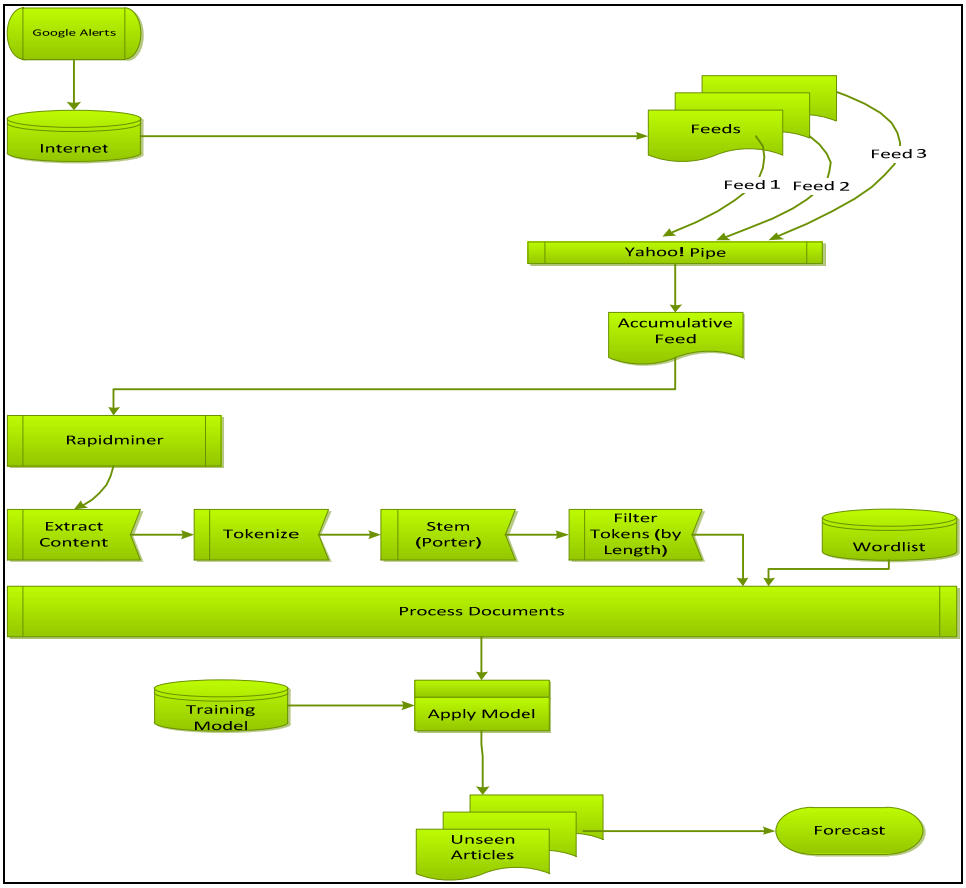
3.2 Theoretical application of the software

3.2.1 Generation of word list

Word lists are the basis of semantic analysis. These lists contain English words and phrases, which are generated in two different ways. First, a word list can be created by using a thesaurus to find synonyms automatically. Second, the CSR reports found in annual reports and on company websites can be used to generate a word list. The thesaurus method produces words with positive and negative connotation, which are manually listed in a column in Microsoft Excel. Excel has a thesaurus function, but this function is normally applied only to a single word. Therefore, we coded a macro programme in visual basic for applications, which applies the thesaurus function to an entire column of words as a pre-processing step (Nora et al., 2010). The macro generates word synonyms for each cell. These synonyms may have positive or negative connotations compared with the original word. Thus, synonyms are only accepted in the word list if they are present in a list of synonyms with at least two primary labelled

words. The original words are selected and labelled manually. If at least two of these words are the basis of two identical synonyms, the probability of finding synonyms with the same positive or negative label is similar to that of the original words. The newly generated synonyms are then used as the basis for the second round to identify further synonyms. The frequency of all words is counted after the second round. The output of the second round of synonym generation is usually identical to the original labelled words and the first round of synonym generation. If the same synonyms are generated repeatedly, their frequency is an indication of an accurate synonym label.

Figure 3 Flowchart of the RapidMiner process (see online version for colours)



3.2.2 Processing the word list

The programme has two components. The function of the first component is to initialise the algorithm before it classifies the text. The imported word list contains English words and phrases, which are classified based on their positive and negative meaning. The programme operator transforms the word list into a collection of documents by outputting a document for each word or phrase in the word list. The newly created text objects serve as inputs for the next operator, which generates a term vector. The new dataset includes

only a single data file; therefore the resulting data is only specified using a single text file. The tokens in the text are used to generate a word vector with the TF-IDF schema. Other sub-operators located within the word vector operators are used to process the word list.

3.2.3 Attribute vectors

The goal of this semantic analysis is to categorise all of the unseen comments found in social media. The algorithm cannot understand connected text, so we create word vectors as an interim step. A word vector also includes characteristic features such as a string attribute. These features may include parts of the entire text with nominal attributes. These attributes form the basis of models that make it easier to compare pieces of text, while the values of string vectors can be equal or unequal. However, the string attributes contain no information about their textual relationships. The word vector also contains the word frequency as a set of additional attributes, which is used to evaluate the meaning of a single word relative to the entire text (Suzuki, 2003). Initially, all of the characters in the word list are transformed to the lower or upper case. Selected regular expressions in tokens can then be removed based on specified replacements. Next, an interior operator splits the word list into an array of tokens. The splitting points are specified by the tokenisation mode 'non-letter character' so a token consists only of a single word. Small English words in English phrases are removed by the next operator. This operator uses an in-built stop word list containing small English words such as 'is' that generally have no effect on the text analysis.

The next inner operator filters tokens that have a length of less than two characters. All of the last five inner operators are part of the word vector operator and they process the word list before creating the word vector. According to the word list, these operators are less important when processing social media content, but they are very helpful for analysing phrases in the word list. The processed word list is one of two outputs from the document processing operator and it is used as an input for a similar operator in the second component of the programme. The processed wordlist is used as a guideline for classifying unknown text in social media content. The second output is a dataset that selects the attributes to be removed or those that should be part of the results. In this programme, all attributes that do not constitute a missing value in any dataset are selected. The next operator uses the dataset as an input and changes the role of an attribute. The initial attribute is located in the word list column with the labels 'positive' and 'negative', where English words and phrases are classified (using a naive Bayesian classifier algorithm). This role is changed into the special attribute 'label', which is used by the learning operator.

3.2.4 Training the algorithm

The learning operator includes a cross-validation process that estimates the performance of the algorithm. This cross-validation operator includes a testing sub-process and a training sub-process. The inner testing operator splits the input dataset into a number of validation subsets. The training sub-process returns a model based on the input dataset. The testing sub-process generates a performance vector using the returned model and it also quantifies its performance (Modha and Masry, 1998). The performance measurement is an appraisal and it is based on the final word list. Therefore, the performance may be

lower when analysing social media. The cross-validation operator has two outputs. It logs the performance and returns the average performance vector. Three other sub-operators function within the cross-validation operator. On the training side, a naive base learner outputs the classification algorithm based on the estimated normal distribution of the training dataset, using Laplacian correction to reduce the effects of zero probabilities (Zhang et al., 2006). On the testing side, an operator applies the model using the trained data, which is based on the imported word list. The trained data are used to classify the future contributions of social media. The other operator used on the testing side is a performance operator, which considers the weight of the input data to calculate the performance of the applied model. It automatically detects the labelled input dataset, which contains two attributes, i.e., one with a role label and another with a role prediction. The second output of the cross-validation operator is the trained algorithm, which is used by the second function of the programme for classifying text (Modha and Masry, 1998). The performance of the cross-validation learning operator can be improved by increasing the number of labelled words and phrases in the word list.

3.2.5 Cross-validation process

The cross-validation model evaluates the predictions of machine learning models to verify the categorisations attributed to the unseen data. Techniques such as cross-validation can also be used to compare the predictive performance of different machine learning methods. The requirements of this process are that all other factors with a possible influence remain unchanged. The cross-validation method divides the data into training and testing data; typically, two-thirds are generally used for training the learning algorithm whereas one-third is used for testing the predictions of the known data (Modha and Masry, 1998). The presence of inappropriate proportions in the classes could lead to a one-sided learning bias, if many or all of the items in a specific class are not present in the two-thirds used as training data. To improve the quality of predictions, we need to select data mining methods with the lowest estimated error levels based on the results of cross-validation estimations.

3.2.6 Extracting relevant contributions from social media content

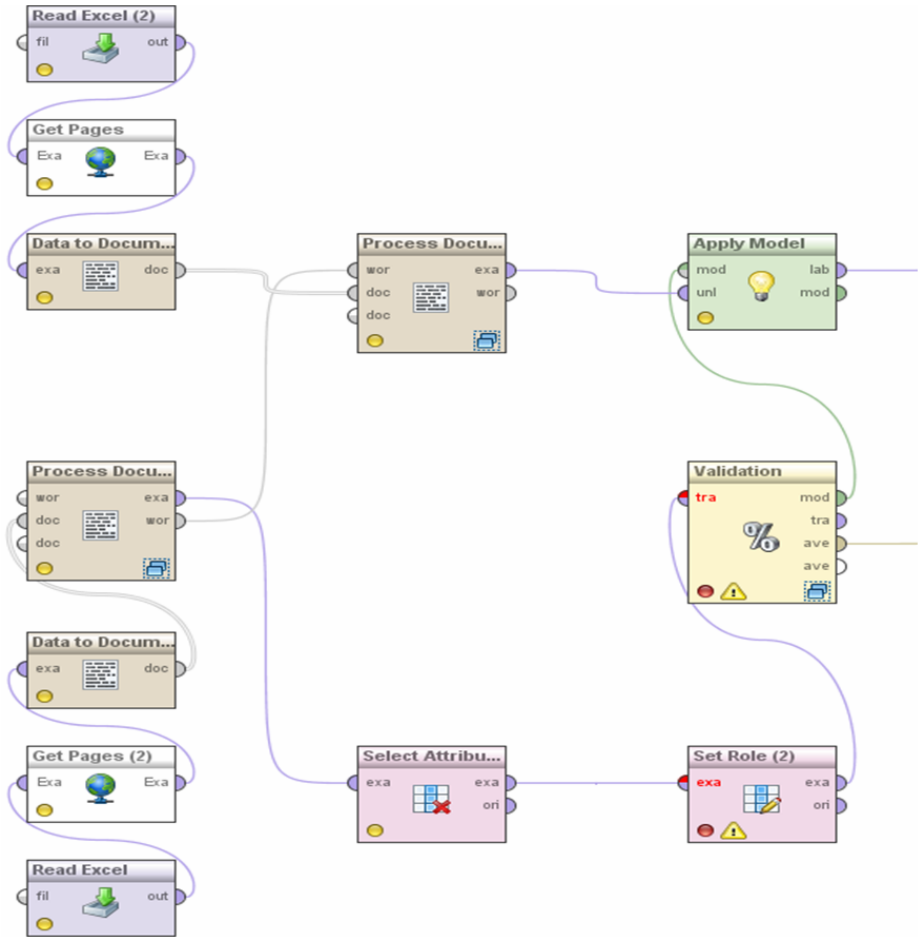
The second component of the programme performs actual data mining of social media content. Initially, an Excel list that contains links to Twitter feeds and Facebook fan pages and forums about the stock market is imported. Each row in this input dataset contains a URL to an RSS feed. The next operator sends a GET request to each of the URLs and saves the pages temporarily using a page-specific attribute. The next two steps are similar to the first part of the programme. One of the operators generates a collection of documents for each URL. The next operator uses the page tokens from the URL to produce a word vector. This operator has two inputs. The first is the list of labelled English words and phrases produced by the first component of the programme. This word list is combined with the pages of the URL to generate a model based on a term vector in the following steps (Nora et al., 2010). The same sub-operators within this operator have to specify lower or upper case characters in a document to split documents into sequences of tokens and filter the tokenised text based on English stop words and the length of words with less than two characters. There is only one other sub-operator that extracts the

textual substance of the social media content from the HTML coding language (Hippner and Rentzmann, 2006).

3.2.7 Application of the algorithm

The final operator uses the trained algorithm to determine the 'positive' or 'negative' intention of social media content. To achieve this, the operator uses the model and the labelled wordlist produced by the first component of the programme. The unlabelled parts of the social media content are used as the second input dataset, which needs to be categorised. This operator joins both parts of the programme, i.e., the labelled word list from the first component of the programme and the unlabelled contributions from the second component of the programme. The word list is one of the most important factors (Nora et al., 2010). The production of a higher number of labelled English words and phrases improves the ability of the learning operator when predicting unknown contributions in social media content. Both the word list and the algorithm are used to analyse social media.

Figure 4 Application of the software (see online version for colours)



3.3 Practical application of the software: a case study on S&P 500 firms

Based on the previous theoretical section, we now report how RapidMiner and Google Alerts were used to determine the CSR reputation of companies. This analysis was based on a sample of 26 firms from the S&P 500.

Table 1 Sample

Companies		
Alcoa	ExxonMobil	Pfizer
Altria Group	General Dynamics	Starbucks
Bank of America	Goldman Sachs	The Coca-Cola Company
Best Buy	IBM	The Dow Chemical Company
CarMax	Johnson & Johnson	Walmart
Chevron	JPMorgan Chase & Co.	WellPoint
Clorox	McDonald's	Wells Fargo
Dean Foods	Nike	Yum!
Disney	Procter&Gamble	

Specific Google Alerts feeds were merged with Yahoo! Finance. The newly joined feed was imported into RapidMiner and we determined the combined CSR reputation values of companies. We also determined the reputation values of specific companies. This analysis did not require Yahoo! Pipes and we imported each separate Google Alerts feed into RapidMiner sequentially.

Figure 5 Yahoo! Pipes (see online version for colours)

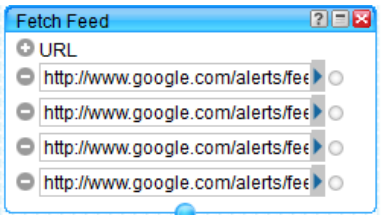


Table 2 Sentiment analysis

Word list		*Process documents from files*		
Classification	*Word*	*Attribute name*	*Total occurrences*	*Document occurrences*
Positive	Honour	Honour	2	2
Positive	Honourable	Honourable	2	1
Positive	Hope	Hope	1	1
Negative	Critical	Critical	4	3
Negative	Demoralised	Demoralised	3	3
Negative	Demotivated	Demotivated	5	4

Initially, we analyse the articles mentioned in Google Alerts using RapidMiner. The CSR sentiment analysis extracted the frequency of classified words from a word list related to The Coca Cola Company. The frequency of words was segmented into the frequency of Google Alerts comments related to the company being analysed and the current comments analysed in Google Alerts.

Table 3 Frequency of new generated synonyms

New generated synonyms	*Frequency*
Frightened	7
Angry	6
Anxious	6
Sad	6
Disheartened	5
Annoyed	5

Table 4 Google Alerts update cycles

Companies	*Number of Google Alerts*	*Update interval*	*No. of updates*
Alcoa	43	2	2
Altria Group	10	6	2
Bank of America	67	1	4
Best Buy	41	2	2
CarMax	3	15	1
Chevron	26	3	3
Clorox	6	9	1
Dean Foods	8	5	1
Disney	27	2	2
ExxonMobil	52	2	2
General Dynamics	20	3	2
Goldman Sachs	17	3	2
IBM	36	1	1
Johnson&Johnson	35	1	1
JPMorgan Chase & Co.	57	1	2
McDonald's	14	2	1
Nike	22	2	2
P&G	86	1	3
Pfizer	9	4	1
Starbucks	45	1	2
The Coca-Cola Company	23	2	1
The Dow Chemical Company	74	1	3
Walmart	84	2	4
WellPoint	5	15	1
Wells Fargo	35	2	2
Yum!	11	1	2

To maintain the quality of the word list, the meanings of the newly generated synonyms were controlled for their frequency. If newly generated words appeared repeatedly, their frequency signalled that the original and new synonyms had the same positive or negative connotation.

The daily CSR reputation values also depended on the frequency of new comments published on Google Alerts. Two criteria were associated with the number of Google Alerts comments available. The update interval indicated how many days were required before new Google Alerts article updates appeared. The number of updates indicated how many published articles were connected with a new update in Google Alerts.

The following sentiment analysis shows the daily reputation values for four of the 26 companies. Larger percentages indicated the higher CSR reputation of a company.

Table 5 Companies' reputation determined by Google Alerts

Date	*Alcoa*	*Altria Group*	*Bank of America*	*Best Buy*
01.02.12	85%	75%	70%	80%
02.02.12	85%	75%	75%	85%
03.02.12	83%	75%	80%	90%
06.02.12	85%	90%	80%	90%
07.02.12	85%	75%	75%	85%
08.02.12	85%	75%	75%	85%
09.02.12	80%	75%	85%	85%
10.02.12	70%	73%	90%	80%
13.02.12	55%	75%	90%	80%
14.02.12	60%	75%	90%	80%
15.02.12	60%	75%	75%	85%
16.02.12	55%	75%	65%	80%
17.02.12	55%	90%	65%	75%
20.02.12	50%	75%	45%	75%
21.02.12	50%	75%	45%	75%
22.02.12	55%	90%	70%	80%
23.02.12	60%	75%	65%	85%
24.02.12	60%	75%	70%	80%
27.02.12	55%	75%	60%	85%
28.02.12	55%	90%	60%	85%

4 Conclusions

Web semantic analysis may allow companies to evaluate their reputation and to improve specific operations related to CSR. The analysis of social media content can extract active and passive feedback about the perceived public image of a corporation. Active feedback can be acquired from available corporate communication channels such as online contact forms and it requires that an individual makes direct contact with corporate representatives (Baldus et al., 2011). This can be facilitated by the semantic analysis of

the comments made in passive internet feedback. Social media such as Facebook, Twitter, or blogs contain numerous public entries about companies. Many global internet users utilise social media to make comments about corporations, particularly their working conditions, environmental pollution record, or their consideration of human rights (Hartmann, 2011). Data mining software such as RapidMiner can identify information that is relevant to companies for forecasting positive or negative perceptions. This information is often more personal and emotional than comments that are sent directly to corporations and they capture public perceptions with greater authenticity. A comparison of a company's self-assessment and the perceptions of others can deliver valuable insights into CSR reputation and the necessary changes that may be required to ensure the maintenance of a good CSR reputation (Hartmann, 2011).

CSR is often regarded as a 'PR exercise', 'green-washing', or 'window-dressing' (Frankental, 2001). This may be because PR and CSR address relationship management with important stakeholder groups such as customers, investors, suppliers, and employees. Thus, it is possible that negative public attitudes towards PR could easily be transferred to CSR. In this context, Clark (2000) stressed that public relations and CSR have similar objectives and that both disciplines aim to enhance the quality of the relationships between an organisation and its key stakeholders. Social media and its analysis using data mining techniques may provide fresh insights into the CSR reputation of specific companies and it may be able to detect poor perception of company behaviour because it integrates different stakeholder perspectives and their experiences with a specific company. Therefore, the analysis of social media related to a company's CSR practices using data mining may allow companies to match their CSR self-image with their CSR public image. Thus, this approach may make it easier to identify the necessary adjustments required to enhance a company's CSR reputation. Factors such as corporate governance and CSR are increasingly important as investment criteria (Derwall et al., 2011) and they are often considered to be important key financial data (Coombes and Watson, 2000). Thus, investors and analysts could use data mining to continuously analyse the CSR reputation of companies via social media and optimise their portfolios.

This is a critical context for evaluating CSR, but the continuous identification of key stakeholders, identifying their expectations about company behaviour, and assessing the impact of CSR are important because this process makes CSR transparent and credible. Previous researchers have stated that it is questionable when companies discuss their responsibilities but they only perform isolated and unconvincing social measures and donation campaigns where the main aim is pure PR (Staples, 2004). Researchers argue that CSR activities should not contradict the corporate philosophy and policy of organisations. However, critics stress that firms often ignore these criteria and prefer to adopt a scattergun approach in their social activities. This may explain why critics from all academic fields argue that CSR pays little more than lip service to social problems and that it distracts society from ethical concerns via PR activities (Robertson and Nicholson, 1996). At present, the media and stakeholders tend to monitor each CSR activity initiated by an organisation critically and sceptically. Thus, companies need to proactively communicate with their stakeholders continuously (e.g., via social networks) to avoid attacks and scepticism from the media and stakeholders. Anticipatory CSR strategies and enhanced CSR communication strategies are indispensible for successful CSR management. Most companies are aware that senior management need to be sensitive to market trends and expectations. However, it is important to remember that the implementation of CSR activities and its communication remain delicate matters,

although sophisticated and long-term CSR approaches can be a source of competitive advantage. The use of data mining techniques to assess the CSR reputation of companies based on social media can provide useful insights that may facilitate the achievement of this goal.

References

Anderson, J.C. and Frankle, A.W. (1980) 'Voluntary social reporting: An iso-beta portfolio analysis', *The Accounting Review*, Vol. 55, No. 3, pp.467–479.

Baldus, A., Müller, S. and Schumann, F. (2011) 'Kunden als Impulsgeber', *Personalwirtschaft*, Vol. 2011, No. 10, pp.56–58.

Bassen, A., Hölz, H-M., Schlange, J., Meyer, K. and Zamostny, A. (2006) *The Influence of Corporate Social Responsibility on the Cost of Capital*, Working Paper; University of Hamburg.

Bhattacharya, C.B. and Sen, S. (2004) 'Doing better at doing good: when, why, and how consumers respond to corporate social initiatives', *California Management Review*, Vol. 47, No. 1, pp.9–24.

Bhattacharya, C.B., Sen, S. and Korschun, D. (2008) 'Using corporate social responsibility to win the war for talent', *MIT Sloan Management Review*, Vol. 49, No. 2, pp.37–40.

Chang, Y. and Shen, C-H. (2011) 'Is corporate social responsibility rewarded by the cost of debt? – Credit ratings view', Working paper, Tamkang University.

Clark, C.E. (2000) 'Differences between public relations and corporate social responsibility: an analysis', *Public Relations Review*, Vol. 26, No. 3, pp.363–380.

Coombes, P. and Watson, M. (2000) 'Three surveys on corporate governance', *McKinsey Quarterly*, Vol. 2000, No. 4, pp.74–77.

Deephouse, D.L. (2010) 'Media reputation as a strategic resource: an integration of mass communication and resource-based theories', *Journal of Management*, Vol. 26, No. 6, pp.1091–1112.

Derwall, J., Koedijk, K. and Ter Horst, J. (2011) 'A tale of values-driven and profit-seeking social investors', *Journal of Banking & Finance*, Vol. 35, No. 8, pp.2137–2147.

El Ghoul, S., Guedhami, O., Kwok, C.C.Y. and Mishra, D.R. (2011) 'Does corporate social responsibility affect the cost of capital?', *Journal of Banking & Finance*, Vol. 35, No. 9, pp.2388–2406.

Eldomiaty, T.I. and Choi, C.J. (2006) 'Corporate governance and strategic transparency: East Asia in the international business systems', *Corporate Governance: International Journal of Business in Society*, Vol. 6, No. 3, pp.281–296.

European Commission (2011) *More Responsible Businesses can Foster more Growth in Europe*, European Commission, Brussels.

Finch, N. (2005) 'The motivations for adopting sustainability disclosure', MGSM Working Paper No. 2005-17.

Frankental, P. (2001) 'Corporate social responsibility: a PR invention?', *Corporate Communications: An International Journal*, Vol. 6, No. 1, pp.18–23.

Gruca, T.S. and Rego, L.L. (2005) 'Customer satisfaction, cash flow, and shareholder value', *Journal of Marketing*, Vol. 69, No. 3, pp.115–130.

Habisch, A. (2003) *Corporate Citizenship – Gesellschaftliches Engagement von Unternehmen in Deutschland*, Springer Verlag, Berlin.

Hartmann, D. (2011) 'Live communication und social media: Die perfekte Symbiose', *Marketing Review St. Gallen*, Vol. 2011, No. 2, pp.34–39.

Hemingway, C.A. and MacLagan, P.W. (2004) 'Managers' personal values as drivers of corporate social responsibility', *Journal of Business Ethics*, Vol. 50, No. 1, pp.33–44.

Hermann, S. (2005) *Corporate Sustainability Branding: Nachhaltigkeits- und stakeholderorientierte Profilierung von Unternehmensmarken*, Gabler Verlag, Wiesbaden.

Hippner, H. and Rentzmann, R. (2006) 'Text mining', *Informatik-Spektrum*, Vol. 29, No. 4, pp.287–290.

Homburg, C., Koschate, N. and Hoyer, W.D. (2005) 'Do satisfied customers really pay more?', *Journal of Marketing*, Vol. 69, No. 2, pp.84–97.

Kaplan, A.M. and Haenlein, M. (2009) 'Users of the world, unite! The challenges and opportunities of social media', *Business Horizons*, Vol. 53, No. 1, pp.59–68.

Kaplan, R.S. and Norton, D.P. (2008) 'Mastering the management system', *Harvard Business Review*, Vol. 2008, No. 18, pp.1–17.

Kosorus, H., Hönigl, J. and Küng, J. (2011) 'Using R, WEKA and RapidMiner in time series analysis of sensor data for structural health monitoring', *DEXA '11: Proceedings of the 2011 22nd International Workshop on Database and Expert Systems Applications*, IEEE Computer Society, Washington, DC, USA, pp.306–310.

Lin, Z., Li, D., Janamanchi, B. and Janamanchi, W. (2006) 'Reputation distribution and consumer-to-consumer online auction market structure: an exploratory study', *Decision Support Systems*, Vol. 41, No. 2, pp.435–448.

Luo, X. and Bhattacharya, C.B. (2006) 'Corporate social responsibility, customer satisfaction, and market value', *Journal of Marketing*, Vol. 70, No. 4, pp.1–18.

Mayer, R.C., Davis, J.H. and Schoorman, F.D. (1995) 'An integrative model of organizational trust', *Academy of Management Review*, Vol. 20, No. 3, pp.709–734.

Modha, D.S. and Masry, E. (1998) 'Prequential and cross-validated regression estimation', *Machine Learning*, Vol. 33, No. 1, pp.5–39.

Münstermann, M. (2007) *Corporate Social Responsibility: Ausgestaltung und Steuerung von CSR-Aktivitäten*, Gabler Verlag, Wiesbaden.

Nora, B., Lemnaru, C. and Potolea, R. (2010) 'Semi-supervised learning with lexical knowledge for opinion mining', *ICCP '10: Proceedings of the Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing*, IEEE Computer Society, Washington, DC, USA, pp.19–25.

Podnar, K. (2008) 'Communicating corporate social responsibility', *Journal of Marketing Communications*, Vol. 14, No. 2, pp.75–81.

Porter, M.E. and Kramer, M. (2006) 'Strategy & society: the link between competitive advantage and corporate social responsibility', *Harvard Business Review*, Vol. 84, No. 12, pp.78–92.

Robbins, P.T. (2001) *Greening the Corporation. Management Strategy and the Environmental Challenge*, Earthscan Publications, London.

Robertson, D.C. and Nicholson, N. (1996) 'Expressions of corporate social responsibility in U.K. firms', *Journal of Business Ethics*, Vol. 15, No. 10, pp.1095–1106.

Sen, S., Bhattacharya, C.B. and Korschun, D. (2006) 'The role of corporate social responsibility in strengthening multiple stakeholder relationships: a field experiment', *Journal of the Academy of Marketing Science*, Vol. 34, No. 2, pp.158–166.

Simmons, C.J. and Becker-Olsen, K.L. (2006) 'Achieving marketing objectives through social sponsorships', *Journal of Marketing*, Vol. 70, No. 4, pp.154–169.

Smith, A.D. (2007) 'Making the case for the competitive advantage of corporate social responsibility', *Business Strategy Series*, Vol. 8, No. 3, pp.186–195.

Staples, C. (2004) 'What does corporate social responsibility mean for charitable fundraising', *International Journal of Nonprofit and Voluntary Sector Marketing*, Vol. 9, No. 2, pp.154–158.

Suzuki, S. (2003) 'Probabilistic word vector and similarity based on dictionaries', in *CICLing'03: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, Springer-Verlag, Berlin/Heidelberg, Germany, pp.562–572.

Thomas, B. (2005) 'Companies report on CSR performance', *Institutional Investor: International Edition*, Vol. 32, No. 7, pp.1–6.

Trudel, R. and Cotte, J. (2009) 'Does it pay to be good?', *MIT Sloan Management Review*, Vol. 50, No. 2, pp.61–68.

Weaver, G.R., Trevino, L.K. and Chochran, P.L. (1999) 'Integrated and decoupled corporate social performance: management commitments, external pressure and corporate ethics practices', *Academy of Management Journal*, Vol. 42, No. 5, pp.539–552.

Weber, M. (2008) 'The business-case for corporate social responsibility: a company-level measurement approach for CSR', *European Management Journal*, Vol. 26, No. 4, pp.247–261.

Zhang, J., Kang, D.K., Silvescu, A. and Honavar, V. (2006) 'Learning accurate and concise naive Bayes classifiers from attribute value taxonomies and data', *Knowledge and Information Systems*, Vol. 9, No. 2, pp.157–179.