Using Webcrawling of Publicly-Available Websites to Assess E-Commerce Relationships

Dirk Thorleuchter
Fraunhofer INT - Institute for Technological Trend Analysis
53879 Euskirchen, Germany
dirk.thorleuchter@int.fraunhofer.de

Dirk Van den Poel
Ghent University, Faculty of Economics and Business Administration
9000 Gent, Belgium
dirk.vandenpoel@ugent.be

***Abstract*— We investigate e-commerce success factors concerning their impact on the success of commerce transactions between businesses companies. In scientific literature, many e-commerce success factors are introduced. Most of them are focused on companies' website quality. They are evaluated concerning companies' success in the business-to-consumer (B2C) environment where consumers choose their preferred e-commerce websites based on these success factors e.g. website content quality, website interaction, and website customization. In contrast to previous work, this research focuses on the usage of existing e-commerce success factors for predicting successfulness of business-to-business (B2B) e-commerce. The introduced methodology is based on the identification of semantic textual patterns representing success factors from the websites of B2B companies. The successfulness of the identified success factors in B2B e-commerce is evaluated by regression modeling. As a result, it is shown that some B2C e-commerce success factors also enable the predicting of B2B e-commerce success while others do not. This contributes to the existing literature concerning e-commerce success factors. Further, these findings are valuable for B2B e-commerce websites creation.**

Success Factor; B-B; SVD, Classification; Textmining, Website

I. INTRODUCTION

A well-known topic in literature is to evaluate information systems concerning their impact on the commercial success of a company [1-5]. This is a demanding task because of the cross-linked structure of information systems [6] and because of their intangible benefits and indirect costs [7].

In literature, success factors have been proposed that enable the measurement and the prediction of companies' success [2] and [6]. Some of these factors focus on companies' website quality. This is relevant for e-commerce companies because as shown by many studies, website quality is the main important factor for e-commerce companies' success [1, 8-10]. Thus, specific e-commerce success factors should be based on website quality aspects [11].

Website usability, offered price savings, online payment possibility, or an implemented human computer interaction are examples for these e-commerce success factors [12]. The evaluation of these success factors shows that an e-commerce company is more successful in the B2C environment if these factors are implemented on company's website [1].

The implementation of success factors on a company's website can be done in many different ways however; most success factors are mentioned on the website content in form of textual patterns. An example is the implementation of a money-back guarantee. On the website content it could be described that a refund will be made if the customer is not satisfied about the product. A further way of expressing a money-back guarantee is a description that the company offers the return of customer's money within a specific period of time. As shown by these examples, e-commerce success factors can be represented by textual patterns identified on company's website. These examples also show that different textual patterns are a representative for the same e-commerce success factor even if they do not have a term in common. This is because the textual pattern are not syntactically but semantically related that means they share the aspect of meaning. The collection of all textual patterns sharing one aspect of meaning is defined as a semantic textual pattern. This semantic textual pattern can be used to represent an e-commerce success factor e.g. to represent the aspect that a money-back guarantee is offered by the company. A further example is the high usability of company's website that can be represented by a semantic textual pattern describing the aspect of website usage in detail. Additionally, a well-known brand as success factor can be identified by the high frequently occurrence of a semantic textual pattern where great stress upon the name of a company or a product is laid.

The identification of semantic textual patterns assumes the use of a specific semantic approach. Latent semantic indexing (LSI) is an example for such an approach. It considers the aspects of meaning and identifies semantic textual patterns from textual information automatically. Using the content of B2B e-commerce companies' websites as input lead to the calculation of semantic textual patterns that occur on companies' websites. If a semantic textual pattern represents an e-commerce success factor than the success factor is implemented in the corresponding companies' websites. This enables the automated identification of success factors on companies' websites.

The identified success factors are evaluated concerning their commercial success whereby the commercial success is

defined as the sustainability of commerce transactions between businesses companies. In detail, if the volume of sales exceeds a specific threshold between two business companies over a specific period of time than the commerce transactions are sustainable and the commercial success is assumed.

Analyzing a large number of commerce transactions leads to the classification of companies according to the group of successful B2B commerce transactions (positive examples) or to the group of non-successful B2B commerce transactions (negative examples). Based on this assignment, logistic regression as modeling technique [13] is used to predict the positive examples in contrast to the negative examples based on the calculated semantic textual patterns. As known from above, these patterns represent success factors. Thus, this LSI-based prediction modeling approach enables the evaluation of success factors concerning the successfulness of B2B companies.

For the extraction of the textual content of companies' websites, web mining [14] is used. A website of a company consists of several web pages. However, not all web pages are relevant for crawling. To identify these relevant web pages a web structure mining approach is used e.g. to identify web pages that deal with company's history where we expect to find information about company's founding and awarded certifications. These information are valuable e.g. for identifying company's trustfulness as success factor.

In sum, the provided methodology can be used to evaluate e-commerce success factors concerning the successfulness of B2B commerce transactions. The evaluation of the success factors gives useful insights for decision makers in B2B. Factors that are evaluated successfully can be implemented in the company's website to improve commercial success. This contributes to an improved B2B website creation and to the existing e-commerce success factor literature.

II. BACKGROUND

A. *Success factors for information systems*

Today, information systems are necessary tools for business, however their procurement, implementation, and processing is cost-consuming. To identify the successfulness of these investments made by companies, one has to evaluate information systems concerning their impact on the commercial success of these companies. Measuring this impact is a challenge task because of the different aims to what information systems are referred: productivity increase, competitiveness improvement, and costs reduction. This can only be measured indirectly [15].

Although much researcher focus on modeling information systems within companies to show their impact on companies' success [4, 16-18], a further direction of research identifies success factors and combines several of them to a success measure that can be used to predict companies' success [3] and [6]. A well-known approach for this further direction is the information system success model from DeLone and McLean [3]. This model contains success factors as well as its dependencies from other success factors. DeLone and McLean conclude that quality is an important success factors. This includes the quality of information as well as the quality of information systems. A second important success factor is the usability of a system. The more an information system makes the use by a customer easier, the more satisfied is the customer, and the more is the commercial success. The last important success factor is the cross-linkage of information systems with people and organizational structures within a company.

B. *Website Success factors*

As seen above, decision makers of companies are interested in evaluation the success of information systems. For e-commerce companies, the quality of their websites as the interface between companies and customers is very import for companies' success [1, 8-10]. Thus, many investments have been done by e-commerce decision makers to build and improve their e-commerce website and they are also interesting in evaluating website's impact on companies' success. Literature introduces several e-commerce success factors that enable an increased website quality [1, 19-26].

McKinney et al. [27] takes over results from DeLone and McLean [3] transforming them to specific e-commerce needs. One success factor is the high quality of the used content information on the website and a further success factor focus on the high quality of a website system (e.g. a content management system). These two success factors increase internet customer satisfaction and that again will lead to commercial success.

The usefulness of the content information that is presented to a customer is a success factor as stated by Devaraj et al. [28]. Further, he proposes that the successful website should be easy to use, the website should be able to respond customers' questions in short time, and the website should offer price savings to the customers.

Torkzadeh and Dhillon [29] state that a website increases company's success if it offers a large number of related products to the customers. Then, a customer is offered a wide product choice. A further success factor is the enabling of online payment to increase the internet shopping convenience. The trustworthiness is also identified as success factor because an improved relationship between companies and its customers can be seen if customers trust the vendor.

Torkzadeh and Dhillon also mention two further success factors, the shopping travel on one hand and the internet shipping on the other hand that increases internet shopping convenience.

Based on the website analysis of 260 manufacturing companies, Zhu and Kraemer [30] have identified several success factors. The process of internet purchase should be as easy as possible to increase customer's satisfaction. Website customization is important because it enables to present an individual content to each customer considering his / her specific needs. A further success factor is the cross-linked website with suppliers. This last factor is very important for this research, because it is the only one success factor that is directly related to B2B commercial transactions. This factor is successfully evaluated based on

the 260 manufacturing companies. Thus, it shows that an e-commerce success factor is able to impact commercial success of B2B transactions.

Company's webpage should be consider the different mediums (made-for-the-medium) e.g. internet browsers, smartphones, mobile phones etc as suggested by Argawal and Venkatesh [31]. Additionally, they suggest website promotion initiatives as success factors. A further important aspect is to address the emotions of customers.

Barnes and Vidgen [32] state that the website design is a success factor. Further, customer satisfaction can be improved by considering the empathy with customers.

A success factor as mentioned by Koufaris [33] shows that a website should evoke customer's shopping enjoyment. Beside mentioned that the usability is an important success factor, he also indicates a brief website content that consists of specifically selected information while discarding non-relevant information.

A successful website should offer playfulness to the customer as stated by Liu and Arnett [34]. This is in contrast to the success factor of Koufaris that proposes brief website content without any playfulness.

Loiacono et al. [35] indicates that a website should offer high entertainment to the customer. These effects should lead to an improved relationship between company and customer and thus, to an improved commercial success. Further, the website should offer complementary relationships to further products. This enables a wide product choice for the customer.

A success factor as mentioned by Plamer [36] is the download delay that possibly forces impatient users to change the website. He also introduces an implemented interactivity on the website as a factor for commercial success.

Three success factors are proposed by Webb and Webb [37]. They are the reliability, the security, and the trustworthiness of the company's website and also of the company itself.

Wu et al. [38] also presents three success factors: the cognitive outcome, the technical support, and the visual appearance.

Barki and Hardwick [39] state that companies' websites have to consider user attitude. Robins et al. [40] mention the commercial success of using a money-back guarantee, a well-known brand, and a price reduction on the website. Van den Poel and Leunis [41] focus on the order delivery as an important success factor of a company.

C. Text Classification

Text classification aims at the assignment of pre-defined classes to textual patterns e.g. if an e-commerce success factor is defined as class than textual patterns can be identified in texts that represent this e-commerce success factor. Specifically in this case, it is important to focus on the aspect of meaning rather than on the aspect of words. This is because two textual patterns can represent the same e-commerce success factor without even sharing one word as already mentioned in the introduction.

An assignment that is based on aspects of meaning can be realized by semantic text classification algorithms. They are in contrast to standard approaches that are based on the knowledge structure e.g. Support Vector Machine (SVM), Naive Bayes Classifier, Decision trees, k nearest neighbour (k-NN) classification [42, 43]. The semantic text classification algorithms use eigenvectors as statistical technique to identify textual patterns.

Terms that occur together more frequently than it would be expected by chance in a specific domain or terms that can replace other terms by keeping the aspect of meaning are semantically related. A collection of these semantically related terms can be found in a semantic textual pattern. It enables to identify textual patterns sharing the aspects of meaning. Thus, it also can be used to identify textual information that represents an e-commerce success factor on a company's website.

Latent semantic indexing (LSI) is a well-known example for the semantic text classification algorithms. LSI calculates semantic textual patterns from given textual sources (e.g. the content of companies' websites). Each semantic textual pattern consists of a large number of terms together with the calculated impact of each term on the pattern. Ranked by the highest impact, the reading of the highest ranked terms gives a good impression for human experts to comprehend the aspect of meaning intuitively. Then, the aspects of meaning of the semantic textual patterns can be compared to the aspects of meaning of e-commerce success factor manually. As a result, some semantic textual patterns can be identified that represent e-commerce success factors. LSI enables the assignment of semantic textual patterns to textual patterns on companies' websites. Thus, the assignment of e-commerce success factors to textual patterns is also possible.

III. METHODOLOGY

Here, we propose a methodology that identifies e-commerce success factors from B2B companies' websites and that evaluates their commercial success. The methodology is depicted in Fig. 1 and it consists of different steps. Although related research [44, 45] of the same authors has different aims, the first steps of the provided methodologies in [44] and in [45] are taken over for this methodology.

This methodology starts by collecting customer information from a B2B company. The customers of this B2B company are also companies and the commerce transactions between these companies and the B2B company are investigated. From the customer relationship management system of the B2B company, we extract the website addresses of their business customers and for evaluating purposes, we also extract information about customers' volume of sales. As mentioned in the introduction, the commerce transactions between the B2B company and each of its business customers are classified as successful or as non-successful by the B2B company based on their sustainability. We are taken over this B2B company-intern classification.

Web mining is used to collect the textual information from companies' websites. The information is preprocessed

and the semantic textual patterns are calculated by LSI. For evaluation, a prediction modeling is done that uses logistic regression. It shows that semantic textual patterns can be used to predict successful commerce transactions between the companies. This also shows the general feasibility of evaluating success factors for B2B companies because some semantic textual patterns represent e-commerce success factors. A more detailed analysis focus on the manual identification of e-commerce success factors as represented by semantic textual patterns. Based on the B2B company-intern classification and based on LSI and modeling results, the identified e-commerce success factors are evaluated concerning their success in predicting successful commerce transactions.

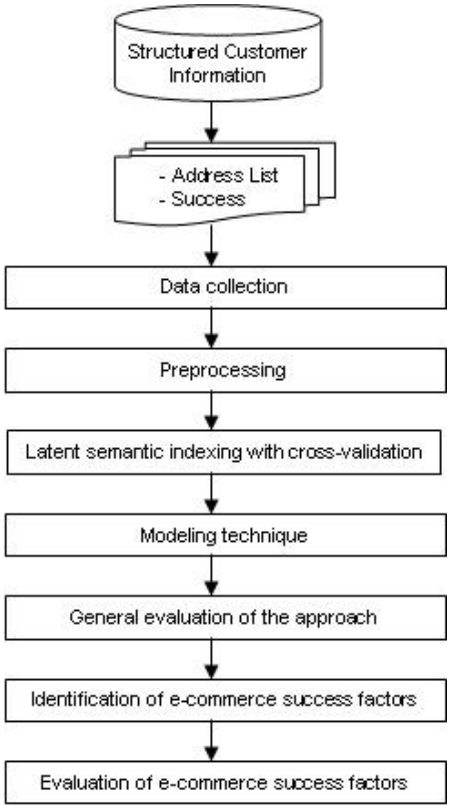

Figure 1. Processing of this approach.

A. *Data collection*

The data collection step extracts textual information from companies' websites. We combine an adapted web structure mining approach and a web content mining approach. The adapted web structure mining approach is necessary to restrict the extraction of information on relevant web pages. Each company's website consist of several web pages and some of them do not contain relevant textual information for this task e.g. 'sitemap', disclaimer', 'data protection policy'. E-commerce success factors normally can be found on pages that are frequently visited by customers. To ensure crawling these web pages, we use the API interface of Google search engine to identify most relevant web pages within a company's website based on the Google ranking algorithm [46, 47]. Thus, the starting page and the most relevant web pages are selected.

We also identify relevant web pages by considering the occurrence of specific terms on the web page. For example, web pages that deal about a company's history are relevant because they contain e.g. information about awarded certifications. Awarded certifications can show e.g. that the company and thus, company's website is trustful. This web page is supposed to be relevant because it possibly contains information about a success factor.

The text that is extracted from relevant web pages of a company consists of highly unstructured information. Thus, a preprocessing step is necessary to enable further processing.

B. *Preprocessing*

The aim of this step is to transform unstructured textual information extracted from customers' websites to structured information [48]. The structure is based on vectors in vector space model [49]. The vectors are used to build a term-by-website matrix in the next step. Preprocessing consists of the steps text preparation, term filtering, term vector weighting, and term vector aggregation.

1) *Text preparation*

Textual information from web pages consists of many elements e.g. images, html- and xml-tags that are removed from the information. Further specific characters are deleted and typographical errors are corrected [50]. With tokenization, the different terms are separated [51]. A case conversion is done by converting all signs in lower case [52].

2) *Term filtering*

The large number of different terms is reduced by using term filtering methods [53]. With part-of-speech tagging, terms of a specific semantic category (e.g. nouns, verbs, and adjective) are selected. Non-informative terms that occur frequently are named stop words and they are discarded. A further reduction of terms is done by use of a stemmer. Based on a dictionary and a set of production rules, all terms are converted to their stem. Additionally, Zipf distribution [49] is applied. Terms that occur just once or twice are non-informative terms and they are discarded.

3) *Term vector weighting*

The selected terms from the term filtering step are used to create a term vector. The standard vector space model is used [54] with one exception: We use weighted frequencies instead of raw frequencies for the term vectors [55]. This is because literature shows a significant improvement in using weighted frequencies [56]. A term has a large weight if it high frequently occurs on a single web page of a company and it also low frequently occurs on the other web pages. [57].

To calculate the weight that is assigned to a term i from the web page j, n is defined as the number of web pages on a customer's website and m is defined as the dimensionality of the term vector. df_i is the number of customers' web pages that contain the term i [58] . Formula 1 shows the calculation of the weight where tf_i as the term frequency is multiplied by

the inverse web page frequency. The denominator represents a length normalization factor [53].

$$w_{i,j} = \frac{tf_{i,j} \cdot \log(n/df_i)}{\sqrt{\sum_{p=1}^{m} tf_{i,j_p}^{\ 2} \cdot (\log(n/df_{i_p}))^2}} \quad (1)$$

4) Term Vector Aggregation

The term vector aggregation step is necessary to aggregate all textual information from the web pages belonging to one customer's website. As a result, each website is assigned to one term vector. The aggregated weight of term i is calculated by the sum of the weights over all web pages on a website [59].

$$Aw_{i,j} = \sum_{k=1}^{r} w_{i,k} \quad (2)$$

C. Semantic textual pattern identification with LSI and singular value decomposition

The aggregated term vectors are used to build a term-by-website (m x n) matrix A with rank r ($r \leqslant \min(m,n)$). The matrix is high dimensional however it can be reduced because it consists of many zero values. LSI is used together with SVD to reduce the number of dimensions. SVD splits A in three matrices:

$$A = U \Sigma V^t. \quad (3)$$

U is the similarity (m x r) matrix of terms and semantic textual patterns. V is the similarity (n x r) matrix of semantic textual patterns and websites. Σ is the diagonal (r x r) matrix with positive singular values. LSI selects the first k singular values. This reduces the rank r to k. However, the selection of parameter k is critical. Estimation about the optimal value of k can be obtained by constructing several rank k-models. Then, the predictive performance of each rank k-model is calculated by prediction modeling and the rank k-models can be evaluated concerning their predictive performance [60]. Modeling is done by using training and test examples that both use the same semantic subspace [61].

D. Prediction modeling

Modeling can be used to show the success of the proposed approach [62-65]. We use logistic regression for modeling to maximize a likelihood function [66] with high computational speed, robustness [67] and simplicity [68]. It is calculated by

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-(w_0 + wx))} \quad (4)$$

Let $T = \{(x_i, y_i)\}$ be the training set with $i = \{1,2,\ldots,N\}$. Let $x \in R^n$ be the n-dimensional input vector, let w the parameter vector, let w_0 the intercept, and let $y_i \in \{0,1\}$ = {commerce transactions between a B2B company and a customer i is not successful, commerce transactions between a B2B company and a customer i is successful} the corresponding binary target labels.

E. Evalutaion Criteria

For the evaluation, we use well-known criteria from the literature: cumulative lift, precision, recall, sensitivity, specificity, receiver operating characteristics curve (ROC) and area under the receiver operating characteristics curve (AUC). With cumulative lift, an increase in density of the positive examples relative to the density of all examples can be measured [44]. Precision measures the exactness [69] and recall measures the completeness [70] of the predicted results. Sensitivity is the proportion of positive cases predicted to be positive and specificity is the proportion of negative cases predicted to be negative. Receiver operating characteristics curve (ROC) is a two-dimensional plot of the sensitivity versus (1-specificity). AUC is the area under the ROC curve [71]. As stated by [72], the AUC is a good performance measure for binary classification.

F. Identification e-commerce success factors and evaluation of their successfulness

As already mentioned, some of the calculated semantic textual patterns represent e-commerce success factors. This is identified by human experts. They compare the aspects of meaning of the semantic textual patterns to the aspects of meaning of e-commerce success factors. As a result, an assignment of some semantic textual patterns to a success factor can be done.

Based on the results of LSI, SVD, and prediction modeling, the predictive accuracy of each semantic textual pattern that represents an e-commerce success factor is used for further evaluation. Three groups of semantic textual patterns can be identified. Members of the first group significantly occur on the websites of companies related to the positive examples. Thus, their occurrence on a company's website can be used to predict that the corresponding commerce transactions between the B2B company and this company is successful. Members of the second group significantly occur on websites related to the negative examples. They also can be used as non-success factors to predict unsuccessful commerce transactions. Members of the third group are related to both, positive and negative examples. Thus, they could not be used as success factors or non-success factors.

IV. Case Study

A. Research data

The empirical verification focuses on two evaluations that are already done in a B2B environment [44, 45]. Thus, we lean on these evaluations (see Sect. IV A, IV B, and IV C) that prove the general feasibility of this approach. For this research, it is very important to identify e-commerce success factors and to evaluate them for predicting successful B2B commerce transactions (see Sect. IV D). This is not part of the evaluation in [44] and [45].

A German mail-order company provides information of about 150,000 business customers. Their customer relationship management (CRM) system includes information about website addresses of their business customers. Business customers that are assigned to the same website address are grouped together. They belong to the same company, however they possibly belong to a different department and thus, the CRM system has stored them separately. Grouping business customers reduces the number of business companies to about 60,000. Based on the volume of sales, the commerce transactions between the B2B company and each of its 60.000 customer companies are classified as successful or as non-successful.

Further, the number of companies is reduced because of language issues. The use of different languages in a text mining analysis causes problems. Therefore, companies with a website in German language are selected and the other are discarded. This results in 35,568 companies. To enable prediction modeling, the commerce transactions between the B2B company and its 35,568 customer companies are split in training examples including validation examples for the parameter estimation and in test examples. The characteristics of the data are shown in Table I [44].

TABLE I. CHARACTERISTICS OF THE DATA

	Number of customer groups	**Relative percentage**
Training and validation set		
Non-successful B2B commerce transactions	11,344	45.56
Successful B2B commerce transactions	13.,553	54.44
Total	24,897	
Test set:		
Non-successful B2B commerce transactions	4,793	44.92
Successful B2B commerce transactions	5,878	55.08
Total	10,671	

B. Optimal dimension selection and interpretation

SVD is used to reduce the dimensions of the term-by-website matrix. After prediction modeling, they are evaluated by a cross-validated AUC [44].

Fig. 2 shows that from of 1–50 dimensions the cross validated AUC increases rapidly and from 50 to 150 dimensions on, it increases less rapidly. 150 dimensions are chosen because the use of more dimensions lead to a more complex model where the cross validated AUC does not increase significantly.

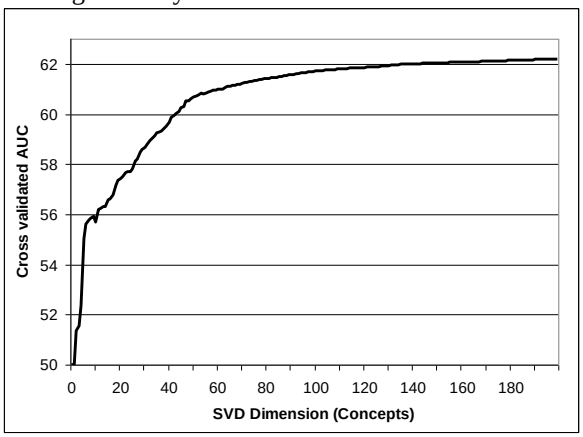

Figure 2. Increase of the cross validated AUC.

C. General Evaluation

The cumulative lift curve is depicted in Fig. 3 and the ROC curve is depicted in Fig. 4. Both figures show the predictive performance of the regression model [45] where the frequent baseline is outperformed. Thus, the evaluation in [44] and in [45] show that in general, this approach is feasible to evaluate semantic textual patterns for the successfulness of B-to-B Commerce transactions.

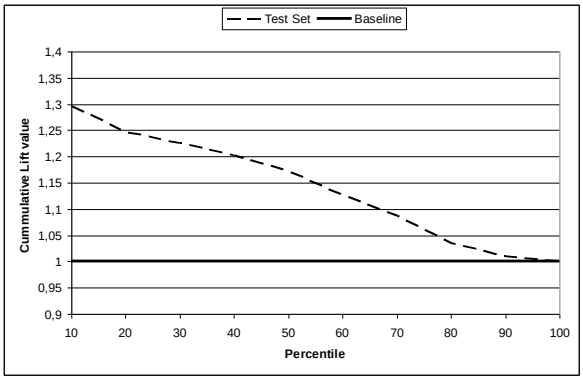

Figure 3. Test set and baseline lift for the logistic regression model

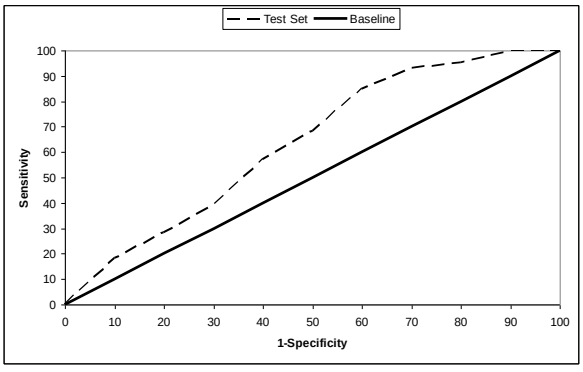

Figure 4. ROC curve

In [44], a further evaluation based on precision and recall is done that leads to the same results as mentioned above.

D. Identification e-commerce success factors and evaluation of their successfulness

Table II shows the results of the identification and evaluation step. All success factors as mentioned in Sect. 2 are listened. If a semantic textual pattern represent a success factor than this success factor is marked as identified in the table. Considering the results of regression modeling, it is also marked that a success factor successfully predicts the negative examples or the positive examples.

TABLE II. RESULTS IN DETAIL

Success factor	**by**	**Identi-fied**	**Neg. Example**	**Pos. Example**
High quality of website content	McKinney et al.			
High quality of website system	McKinney et al.			
Usefulness of website content	Devaraj et al.			
Website Usability	Devaraj et al.	X		
Website responsiveness	Devaraj et al.			
Offering price savings	Devaraj et al.	X		
Wide product choice	Torkzadeh and Dhillon	X		
Online payment	Torkzadeh and Dhillon	X		
Trustworthiness	Torkzadeh and Dhillon	X		X
Shopping travel	Torkzadeh and Dhillon			
Internet shipping	Torkzadeh and Dhillon	X		
Easy purchase processing	Zhu and Kraemer			
Website customization	Zhu and Kraemer	X		
Website based supplier connection	Zhu and Kraemer	X		
Made-for-the-medium	Argawal and Venkatesh			
Website promotion initiatives	Argawal and Venkatesh			
Addressing emotions	Argawal and Venkatesh			
Website design	Barnes and Vidgen			
Shopping enjoyment	Koufaris			
Brief content	Koufaris			
Website playfulness	Liu and Arnett			
Entertainment effects	Loiacono et al.			
Download delay	Plamer			
Website interactivity	Plamer	X		X
Reliability	Webb and Webb	X		X
Security	Webb and Webb	X		X
Website's cognitive outcome	Wu et al.			
Website's visual appearance	Wu et al.			
Website's technical support	Wu et al.	X	X	
Considering user attitude	Barki and Hardwick			
Money-back guarantee	Robins et al.	X	X	
Well-known brand	Robins et al.	X		
Order delivery	Van den Poel and Leunis	X		

Based on the number of all mentioned success factors in Sect. 2, we have identified about half of them in the SVD dimensions. Some of them occur together with others in one SVD dimension. We have expected to obtain this result, because some success factors are not described explicitly in textual information on the websites. Based on the identified success factors, six factors are successful for prediction. Four of them can be used to predict the positive examples: Website interactivity, Trustworthiness, Reliability and Security. Trustworthiness, Reliability and Security are represented by the same semantic textual patterns and this pattern is successful in predicting positive examples. Thus, the three success factors are also successful. As predictor for the negative examples, we identify the money-back guarantee and the technical support of a website. It is hard to understand why money-back guarantee is related to the negative examples. We suppose that today it is usual to offer a money-back guarantee and successful companies do not mentioned that explicitly on their websites.

V. CONCLUSIONS

This work investigates e-commerce success factors concerning their impact on the success of commerce transactions between businesses companies. Success factors are identified based on semantic textual patterns as calculated by LSI, SVD, and predictive modeling. The results show that existing e-commerce success factors can be used to predict successful commerce transactions in the B2B environment. We use a case study that is based on two existing evaluations from literature. These evaluations are adapted to enable an evaluation of this work. The results

contribute to the literature in the field of B2B success factors. Further, they give valuable insights to B2B decision makers e.g. for creating websites of B2B companies. Future work should focus on the success factors that are not identified in this case study.

ACKNOWLEDGMENT

This work uses SAS v9.1.3, SAS Text Miner v5.2, Fraunhofer Idea Web Miner v1.0, and Matlab v7.0.4 for data analysis. Further Google Search API is used for data collection. We would like to thank Joachim Schulze and Joerg Fenner for their constructive technical comments.

REFERENCES

[1] Y. Lee and K. A. Kozar, "Investigating the effect of website quality on e-business success: An analytic hierarchy process (AHP) approach," Decision Support Systems, vol. 42, 2006, pp. 1383–1401.

[2] J. Ballantine, M. Levy, and P. Powell, "Evaluating information systems in small and medium-sized enterprises: issues and evidence," European Journal of Information Systems, vol. 7, 1998, pp. 241– 251.

[3] W.H. DeLone and E.R. McLean, "Information systems success: the quest for the dependent variable," Information Systems Research, vol. 3, 1992, pp. 60– 95.

[4] Z. Irani and P. E. D. Love, "Developing a frame of reference for exante IT/IS investment evaluation," European Journal of Information Systems, vol. 11, 2002, pp. 74–82.

[5] M. Themistocleous and Z. Irani, P.E.D. Love, "Evaluating the integration of supply chain information systems: a case study," European Journal of Operational Research, vol. 159 (2), 2004, pp. 393– 405.

[6] V. Serafeimidis and S. Smithson, "Information systems evaluation as an organizational institution—experience from a case study," Information Systems Journal, vol. 13, 2003, pp. 251– 274.

[7] Z. Irani, "Information systems evaluation: navigating through the problem domain," Information and Management, vol. 40, 2002, pp. 11– 24.

[8] M.C. Carnero, "Selection of diagnostic techniques and instrumentation in a predictive maintenance program: a case study, " Decision Support Systems, vol. 38 (4), 2005, pp. 539– 555.

[9] G.L. Lohse and P. Spiller, "Internet retail store design: how the user interface influences traffic and sales," Journal of Computer Mediated Communication, vol. 5, 1999.

[10] E.W.T. Ngai, "Selection of web sites for online advertising using the AHP," Information and Management, vol. 40, 2003, pp. 233– 242.

[11] M. Zvirana, C. Glezerb, and I. Avnia, "User satisfaction from commercial web sites: The effect of design and use." Information & Management, vol. 43 (2), 2006, pp. 157-178.

[12] P. H. Baecke and D. Van den Poel, "Improving purchasing behavior predictions by data augmentation with situational variables," International Journal of Information Technology and Decision Making, vol.9 (6), 2010, pp. 853-872.

[13] K. Coussement and D. Van den Poel, "Integrating the voice of customers through call center emails into a decision support system for churn prediction" Information & Management, vol. 45, 2008, pp. 164-174.

[14] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Extracting consumers needs for new products - A web mining approach," Proc. World Conference of Knowledge Discovery and Data Mining (WKDD 2010), IEEE Computer Society, 2010, pp. 441.

[15] D. F. Galletta and A. L. Lederer, "Some cautions on the measurement of user information satisfaction," Decision Sciences, vol. 20, 1989, pp. 419-438.

[16] Z. Irani and M. Themistocleous, "P.E.D. Love, The impact of enterprise application integration on information system lifecycles," Information and Management, vol. 41, 2003, pp. 177–187.

[17] L. Mcaulay, N. Doherty, and N. Keval, "The stakeholder dimension in information systems evaluation," Journal of Information Technology, vol. 17, 2002, pp. 241–255.

[18] S. Smithson and R. Hirschheim, "Analysing information systems evaluation: another look at an old problem," European Journal of Information Systems, vol. 7, 1998, pp. 158–174.

[19] P. H. Baecke and D. Van den Poel," Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data," Journal of Intelligent Information Systems, 2011, pp. 367-383.

[20] W. H. Delone and E. R. McLean," Information systems success: the quest for the dependent variable," Information Systems Research, vol. 3 (1), 1992, pp. 60-95.

[21] K. W. DeBock and D. Van den Poel," Predicting website audience demographics for web advertising targeting using multi-website clickstream data," Fundamenta Informaticae, vol. 97, 2009, pp. 1-19.

[22] I. Lopeza and S. Ruiz, "Explaining website effectiveness: The hedonic–utilitarian dual mediation hypothesis," Electronic Commerce Research and Applications, vol. 10 (1), 2010, pp. 49-58.

[23] Y. Lu, L. Zhao, and B. Wang, "From virtual community members to C2C e-commerce buyers: Trust in virtual communities and its effect on consumers' purchase intention," Electronic Commerce Research and Applications, vol. 9 (4), 2010, pp. 346-360.

[24] C. Serrano-Cinca, Y. Fuertes-Callén, and B. Gutiérrez-Nieto, "Internet positioning and performance of e-tailers: An empirical analysis," Electronic Commerce Research and Applications, vol. 9 (3), 2010, pp. 237-248.

[25] D. Van den Poel and W. Buckinx, "Predicting Online-Purchasing Behavior," European Journal of Operational Research, vol. 166 (2), 2005, pp. 557-575.

[26] P. C. Verhoef, R. Venkatesan, L. McAlister, E. C. Malthouse, M. Krafft, and S. Ganesan, "CRM in Data-Rich Multichannel Retailing Environments: A Review and Future Research Directions," Journal of Interactive Marketing, vol. 24 (2), 2010, pp. 121-137.

[27] V. McKinney, K. Yoon, and F.M. Zahedi, "The measurement of webcustomer satisfaction: an expectation and disconfirmation approach," Information Systems Research, vol. 13, 2002, pp. 296–315.

[28] S. Devaraj, M. Fan, and R. Kohli, "Antecedents of B2C channel satisfaction and preference: validating e-commerce metrics," Information Systems Research, vol. 13, 2002, pp. 316– 333.

[29] G. Torkzadeh, and G. Dhillon, "Measuring factors that influence the success of Internet commerce," Information Systems Research, vol. 13, 2002, 87–204.

[30] K. Zhu, and K. Kraemer, "E-commerce metrics for net-enhanced organizations: assessing the value of e-commerce to firm performance in the manufacturing sector," Information Systems Research, vol. 13, 2002, pp. 275–295.

[31] R. Agarwal and V. Venkatesh, "Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability," Information Systems Research, vol. 13, 2002, pp. 168– 186.

[32] S.J. Barnes and R. Vidgen, "An evaluation of cyber-bookshops: the webQual method," International Journal of Electronic Commerce, vol. 6, 2001, pp. 11– 30.

[33] M. Koufaris, "Applying the technology acceptance model and flow theory to online customer behavior," Information Systems Research, vol. 13, 2002, pp. 205– 223.

[34] C. Liu and K.P. Arnett, "Exploring the factors associated with web site success in the context of electronic commerce," Information and Management, vol. 38, 2000, pp. 23– 33.

[35] E.T. Loiacono, D.Q. Chen, and D.L. Goodhue, "WebQualk revisited: predicting the intent to reuse a website," in proceedings of 8th Americas Conference on Information Systems, 2002, pp. 301– 309.

[36] J.W. Palmer, "Web site usability, design, and performance metrics," Information Systems Research, vol. 13, 2002, pp. 151– 167.

[37] H.W. Webb and L.A. Webb, "SiteQual: an integrated measure of web site quality," Journal of Enterprise Information Management, vol. 17, 2004, 430–440.

[38] F. Wu, V. Mahajan, and S. Balasubramanian, "An analysis of e-business adoption and its impact on business performance," Journal of the Academy of Marketing Science, vol. 13, 2003, pp. 425– 447.

[39] H. Barki and J. Hardwick, "Measuring user participation, user involvement, and user attitude," MIS Quarterly, vol. 18 (1), 1994, pp. 59-79.

[40] D. Robins and S. Kelsey, "Analysis of Web-based information architecture in a university library: navigating for known items," Information Technology and Libraries, vol. 21 (4), 2002, pp. 158-169.

[41] D. Van den Poel and J. Leunis, "Consumer Acceptance of the Internet as a Channel of Distribution," Journal of Business Research, vol. 45 (3), 1999, pp. 249-256.

[42] F. Palmieri and U. Fiore, "Network anomaly detection through nonlinear analysis." Comput Secur, vol. 29, 2010, pp. 737-55.

[43] J. Herranza, S. Matwin, J. Nind, and V. Torra, "Classifying data from protected statistical datasets," Comput Secur, vol. 29 (8), 2010, pp. 875-890.

[44] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing, " Expert Systems with Applications, vol. 39 (3), 2012, pp. 2597-2605.

[45] D. Thorleuchter, D. Van den Poel, and A. Prinzie "LSI based Profitability Prediction of new Customers," Proc. SIAM International Workshop on Data Mining for Marketing (DMM 2011), SIAM, 2011, pp. 62-67.

[46] D. Thorleuchter and D. Van den Poel, "Companies Website Optimising concerning Consumer's searching for new Products," Proc. Uncertainty Reasoning and Knowledge Engineering (URKE 2011), IEEE Press, 2011, pp. 40-43.

[47] D. Thorleuchter, J. Schulze, and D. Van den Poel "Improved Emergency Management by a Loosely Coupled Logistic System," Proc. Future Security 2012, Communications in Computer and Information Science, Berlin: Springer, 2012

[48] D. Thorleuchter, "Finding new technological ideas and inventions with text mining and technique philosophy" in Data analysis, machine learning and applications, C. Preisach et al., Eds. Berlin: Springer, 2008, pp. 413-420.

[49] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies," Technol Forecast Soc Change, vol. 77 (7), 2010, pp. 1037-1050.

[50] W. Gericke, D. Thorleuchter, G. Weck, F. Reiländer, and D. Loß, „Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz," Informatik-Spektrum, vol. 32 (2), 2009, pp. 102-109.

[51] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Mining Innovative Ideas to Support new Product Research and Development," in Classification as a Tool for Research, H. Locarek-Junge and C. Weihs, Eds. Berlin : Springer, 2010, pp.587-594.

[52] R. Feldman and J. Sanger. "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data," Cambridge : University Press, 2007, p. 318.

[53] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining," LDV Forum, vol. 20 (1), 2005, pp. 19-26.

[54] D. Thorleuchter and D. Van den Poel, "Semantic Technology Classification - A Defence and Security Case Study," Proc. Uncertainty Reasoning and Knowledge Engineering (URKE 2011), IEEE Press, 2011, pp. 36-39.

[55] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Mining Ideas from Textual Information," Expert Syst Appl, vol. 37 (10), 2010, pp. 7182-7188.

[56] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," J Doc, vol. 28 (1), 1972, pp. 11-21.

[57] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inform Process Manag, vol. 24 (5), 1988, pp. 513-523.

[58] G. Salton, J. Allan, and C. Buckley, "Automatic structuring and retrieval of large text files," Commun ACM, vol. 37 (2), 1994, pp. 97-108.

[59] K. Coussement and D. Van den Poel, "Integrating the voice of customers through call center emails into a decision support system for churn prediction," Inform Manage, vol. 45 (3), 2008, pp. 164-174.

[60] D. Thorleuchter and D. Van den Poel "Using NMF for Analyzing War Logs," Proc. Future Security 2012, Communications in Computer and Information Science, Berlin: Springer, 2012

[61] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," J Am Soc Inform Sci, vol. 41 (2), 1990, pp. 391-407.

[62] D. Thorleuchter and D. Van den Poel, "High Granular Multi-Level-Security Model for Improved Usability," Proc. System Science, Engineering Design and Manufacturing Informatization (ICSEM 2011), IEEE Press, 2011, pp. 191 – 194.

[63] D. Thorleuchter, G. Weck, and D. Van den Poel, "Granular Deleting in Multi Level Security Models Security - an Electronic Engineering approach," Proc. International Conference on Mechanical and Electronic Engineering (ICMEE2012), Lecture Notes in Electrical Engineering, Berlin : Springer, 2012, in press

[64] D. Thorleuchter, G. Weck, and D. Van den Poel, "Usability based Modeling for Advanced IT-Security - an Electronic Engineering approach," Proc. International Conference on Mechanical and Electronic Engineering (ICMEE2012), Lecture Notes in Electrical Engineering, Berlin : Springer, 2012, in press

[65] D Thorleuchter and D. Van den Poel, "Rapid Scenario Generation with Generic Systems," in: International Conference on Management Sciences and Information Technology (MSIT 2012), Lecture Notes in Information Technology, Delaware : IERI, 2012, in press

[66] P.D. Allison, "Logistic Regression using the SAS System: Theory and Application," Cary NC : SAS Institute Inc., 1999.

[67] W.R. Greiff, "A theory of term weighting based on exploratory data analysis," Proc. 21st SIGIR Conference (SIGIR 98), ACM, 1998, pp. 11-19.

[68] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," Biometrics, vol. 44 32), 1988, pp. 837-845.

[69] D. Thorleuchter, S. Herberz, and D. Van den Poel, "Mining Social Behavior Ideas of Przewalski Horses," Lecture Notes in Electrical Engineering, vol. 121, 2011, pp. 649-656.

[70] D. Thorleuchter and D. Van den Poel, "Extraction of Ideas from Microsystems Technology," Proc. Computer Science and Information Engineering (CSIE2012), Advances in Intelligent and Soft Computing. Springer, Springer, 2012, in press.

[71] D. Thorleuchter and D. Van den Poel, "Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website," Expert Syst Appl, 2012, in press.

[72] J.A. Hanley and B.J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 131 (1), 1982, pp. 29-36.