Paper 123-2008

Employing SAS® Text Miner Methodology to Become a Customer Genius in Customer Churn Prediction and Complaint E-mail Management.

Kristof Coussement
Ghent University, Faculty of Economics and Business Administration, Department of Marketing,
Tweekerkenstraat 2, 9000 Ghent, Belgium

Abstract

Nowadays due to ever increasing internet penetration rate, people are sending more and more emails to companies as a substitute for traditional communication channels like telephone calls or letters. This highly unstructured information contains a lot of valuable information for marketing analysts. However, this textual information is often neglected because no in-house knowledge or ready-to-use framework is available to convert these e-mails into a form which is more suitable for subsequent processing. This project clarifies the text mining methodology as used in the SAS® Text Miner, while two illustrations on real-life data show the beneficial effect of taking into account the textual information sent by customers. It is shown that more profound marketing decisions can be made when one integrates different types of available information.

1. Introduction

Customer Intelligence is a hot topic in today's business environment. Building and growing more profitable customers is the key challenge for a lot of companies. For instance, Customer Intelligence ensures that organizations pick and retain the right customers, makes customers buy more or better products, makes sure the customer is satisfied with the complaint handling system,… However, a lot of these applications use information that is easily extractable from the database – so called structured information. Due to the rapid development of information technology and Internet, customers interact more easily with the company. Nowadays, call centers receive daily huge amounts of e-mails due to the fact that their clients become more used to sending e-mails as a substitute for traditional communication methods like letters, telephone calls,… Consequently, new opportunities arise to improve Customer Intelligence by using these e-mails – so called unstructured, textual information. However, marketing analysts tend to neglect this unstructured, textual information in marketing decision processes because (i) e-mails are not directly applicable in a traditional marketing context due to the additional pre-processing steps, (ii) often there is no in-house knowledge on how to convert this textual information into an analyzable format and (iii) no ready-to-use framework is available which integrates the traditional information with the information from call center e-mails.

This paper shows the beneficial effect of using this unstructured information from client/company e-mails in Customer Intelligence analysis for (i) customer churn prediction and (ii) complaint handling. (i) A first application in Customer Intelligence lies in churn prediction (Coussement and Van den Poel, 2008a). Due to saturated markets and intensive competition, a lot of companies do realize that their existing database is their most valuable asset. Companies start to focus on targeted marketing actions to retain existing customers. Indeed, it is more profitable to keep and satisfy existing customers than to constantly attract new customers who are characterized by a high attrition rate. The idea of identifying those customers most prone to switching – i.e. customer churn prediction - carries a high priority, because it has been shown that a small change in retention rate can result in significant changes in contribution. In order to effectively manage customer churn within a company, it is crucial to build an effective and accurate customer-churn model. By now, marketing managers use traditional marketing information - i.e. information from the marketing database like recency, frequency and monetary value - to predict whether or not a customer will leave the company, while they tend to neglect the client/company interactions from call-center e-mails. Notwithstanding, these e-mails may contain a lot of valuable information for predicting churn. This paper shows that integrating this free-formatted, textual information from call-center e-mails with the traditionally-used information derived from the marketing database increases churn prediction performance. (ii) Another interesting application in the Customer Intelligence domain lies in developing an efficient and better e-mail handling process to improve the client/company relationship (Coussement and Van den Poel, 2008b). More specifically, efficient e-mail handling is becoming a critical key success factor in today's business environment. As a result of the increasing number of e-mails sent, companies start to outsource their customer-e-mail management by relying on customer-call centers to address the voice of customers - i.e. customer complaints and service-information requests. A tool to support efficient processing of customer-complaint e-mails is an automatic e-mail classification system. Automatic text classification labels incoming e-mails into predefined categories – i.e. complaints versus non-complaints in this study. As a consequence, customer-complaint management becomes more successful in mainly two ways. (a) In contrast to manual text classification, automatic text classification is time-saving and thus less expensive in terms of labor costs. It makes the e-mail-handling process more efficient. (b) By classifying incoming e-mails into complaints and non-complaints, one can optimize the complaint-handling process. By making a distinction between complaint e-mails and other e-mail types, the company is able to set up a separate complaint handling department with specially-trained complaint handlers. One can create a separate treatment procedure for complaint e-mails. Consequently, call centers can react more helpful on occurring problems or service failures, which increases customer satisfaction. As such, a complaining customer becomes not necessarily a customer who will churn. In summary, building an automatic e-mail classification system that distinguishes complaints from non-complaints is necessary for optimizing the complaint-handling process within a call center and for increasing customer satisfaction through efficient response on the problem.

2. Methodology

This research project employs the methodology of the SAS® Text Miner. This section throws light on the pre-processing steps needed to convert a raw text e-mail into a form which is more suitable for subsequent processing (see also Coussement and Van den Poel 2008a; Coussement and Van den Poel 2008b). An introduction to the text mining methodology is also given in the book “Getting Started with SAS® 9.1 Text Miner” (SAS Institute Inc., 2004).

2.1. Vector-space approach

This paragraph gives an overview of the conventional text-classification approach using the vector-space approach proposed by Salton (1971). This means that original documents are converted into a vector in a feature space based on the weighted term frequencies. Each vector component reflects the importance of the corresponding term by giving it a weight if the term is present or zero otherwise. The purpose is to select the most informative terms from the number of distinct terms in the corpus dictionary. All documents are traditionally converted from the original format to word vectors following the steps as shown in Figure 1.

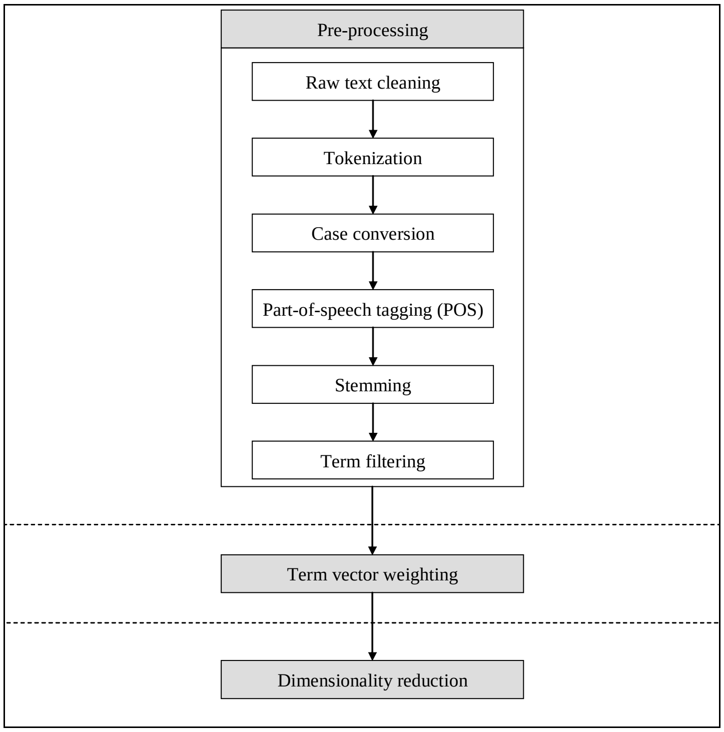

Figure 1: An overview of the conventional vector-space approach for text classification.

2.1.1. Pre-processing

Raw text cleaning converts documents into a form which is more suitable for subsequent processing. In this step, special characters and punctuations are separated from words, while spelling errors are handled by comparing all words in the document with a reference dictionary.

During the *tokenization* step, documents are divided into tokens or words separated by white space characters. Once the text field is divided into words, words are converted to lower case – i.e. *case conversion*.

All words are tagged a *part of speech* based on their syntactic category. All words are summarized into informative and non-informative parts of speech. The non-informative parts of speech contain determiners, conjunctions, auxiliaries, prepositions, pronouns, negative articles or possessive markers, interjections, proper nouns, abbreviations and numbers. On the other hand, words can be part of an informative part of speech like nouns, verbs, adjectives and adverbs.

A next step in the text-preprocessing phase is *stemming* or lemmatization. Word variations are conflated into a single representative form, called the stem. A typical example of a stem is the word ‘connect’ which is the stem for the variants ‘connected’, ‘connecting’, ‘connection’ and ‘connections’. Stemming has two advantages: it reduces the corpus dictionary enormously and it increases the retrieval performance significantly. A dictionary-based stemmer is used throughout this study. The huge advantage is that all morphological variations are treated naturally by comparing them with a reference dictionary. When a corpus term is unrecognizable, the stemmer applies some standard decision rules to give the term the correct stem.

In order to reduce the number of irrelevant terms in the corpus dictionary, a number of *term filtering* tasks are performed. Firstly, rare words are left out from further analysis because they are unable to aid in future classification. Consequently, all words appearing less then three times over the entire document corpus are eliminated for further analysis. Additionally, overly common words like for instance ‘a’ or ‘the’ are also removed from the corpus dictionary. These type of words or stopwords appear so often that they are not discriminative anymore. A stoplist is language and domain specific, as a consequence a standard stoplist is often manually adapted to avoid the risk of removing relevant words. Furthermore, only words that are part of an informative part of speech are included, because these words contain relevant information to aid in future classification. In the end, the temporary dictionary is manually checked and irrelevant words are removed from the dictionary.

The result is a high-dimensional term-by-document matrix where each cell in the matrix represents the raw frequency of appearance of a term in a document. To correct for the importance of a term in a document and its importance in the corpus dictionary, the term vectors in the term-by-document matrix are weighted.

2.1.2. Term-vector weighting

Term-vector weighting is often done by determining the product of the term frequency (*tf*) and the inverse document frequency (*idf*) (e.g. Sparck Jones, 1973). The *tf* reflects the importance of the term within a particular document, while the *idf* is a measure for the general importance of the term in the complete document corpus. The weight of term *i* in document *j* proportionally increases with the number of times term *i* appears in document *j* (tf_{ij}) but is offset by the frequency of the term *i* in the complete corpus (idf_i).

Thus, the weight of term *i* in document *j* (w_{ij}) is given by

$$w_{ij} = tf_{ij}\ idf_i \qquad (1)$$

with tf_{ij} equal to the term frequency of term *i* in document *j*, idf_i equal to the inverse document frequency of term *i* and

$$tf_{ij} = \log_2(n_{ij} + 1) \qquad (2)$$

with n_{ij} equal to the frequency of term *i* in document *j* and

$$idf_i = \log_2\left(\frac{n}{df_i}\right) + 1 \qquad (3)$$

with *n* equal to the total number of documents in the entire document collection and df_i equal to the number of documents where term *i* is present. The result is a weighted term-by-document matrix.

2.1.3. Dimensionality reduction

This weighted term-by-document matrix is a high-dimensional matrix due to the many distinct corpus terms. Moreover, this matrix is very sparse – i.e. it contains a lot of zeros -, since not all documents contain all corpus terms. In order to reduce the dimensionality of the feature space, this study employs Latent Semantic Indexing by using Singular Value Decomposition (SVD) (Deerwester et al., 1990). Latent Semantic Indexing projects documents from the high-dimensional term space to an orthonormal, semantic latent subspace by grouping together similar terms into several distinct concepts *k*. All textual information can be summarized into *k* concepts. Furthermore, these *k* concepts or SVD variables are often used as variables summarizing the textual information. As a consequence, these variables can be used in churn prediction or in traditional text-classification. Literature proposes an operational criterion to find the optimal value for *k*.

2.2 Classification technique

Logistic regression is used as the classification technique within this study, mainly for three reasons: (i) It is a conceptually simple binary classification technique, which maximizes the log-likelihood function in order to become an appropriate fit to the data, (ii) Moreover, a closed-form solution for the posterior probabilities is available. (iii) Logistic regression is a well-known classification technique in marketing applications. Indeed, logistic regression is the most often used classifier by predictive model builders in industry. More information about logistic regression is found in Allison (1999).

3 Examples

This Section describes two real-life examples on how the proposed methodology contributes to the Customer Intelligence domain – i.e. a customer churn prediction and a complaint e-mail management case (Coussement and Van den Poel, 2008a; Coussement and Van den Poel, 2008b).

3.1. Churn prediction

The purpose of customer churn prediction is to identify those customers most prone to switch. Based on a whole bunch of explanatory variables, the marketing analyst tries to identify whether or not a customer will leave the company. This paragraph investigates whether incorporating the information from call center e-mails into a churn prediction model that uses only structured marketing information like recency, frequency, monetary value, age, number of complaints… increases the predictive performance.

Before comparing the predictive performance of the model with only structured marketing information (ModStruc) with the performance of the model which combines the structured marketing information with the textual information from the call center e-mails (ModStruc-Unstruc), the optimal set of structured marketing variables is defined. Employing the forward selection procedure in the logistic regression on the traditional marketing variables, a best subset of 20 marketing variables is obtained including the elapsed time since the last complaint, the monetary value, the length of the current subscription,… .

As such, Modstruc is built using the best subset of the 20 marketing variables, while ModStruc-Unstruc is a combination of those 20 marketing variables with the variables representing the textual information – i.e. in this study, 170 additional variables as extracted using the methodology as described in Section 2.

Table 1, Figure 2 and Figure 3 show that the predictive performance of ModStruc-Unstruc significantly outperforms that of ModStruc in terms of AUC and Top-decile lift.

	AUC	**Top-decile lift**
ModStruc	73.80	2.69
ModStruc-Unstruc	77.75	3.07

Table 1: The performance of ModStruc and ModStruc-Unstruc: AUC and top-decile lift on the test set.

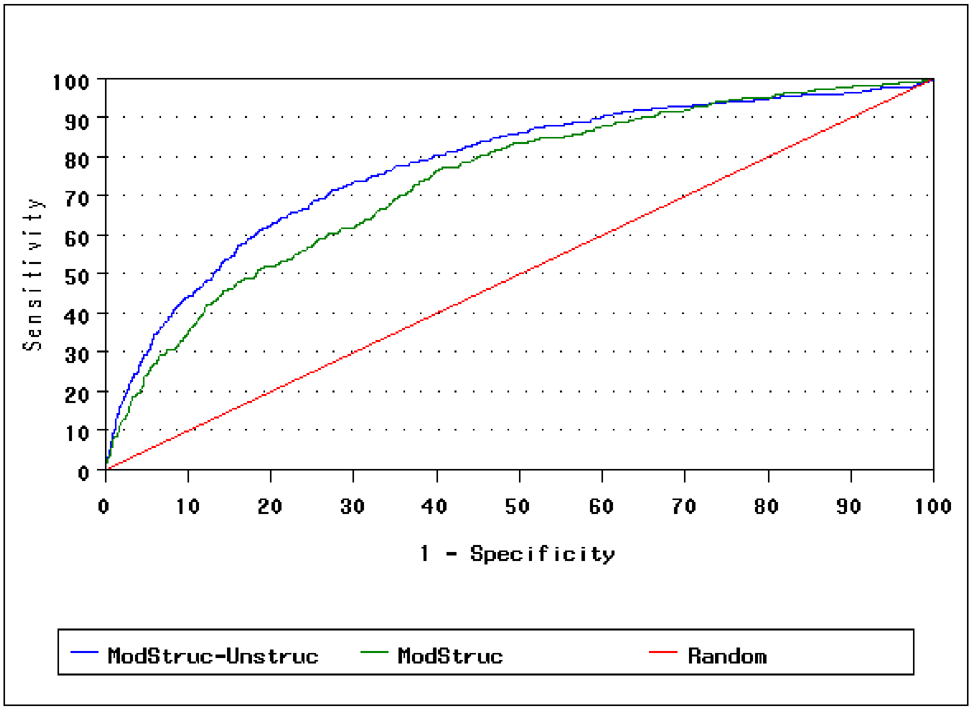

Figure 2: The ROC curves for ModStruc, ModStruc-Unstruc and the Random model (or the zero-information model).

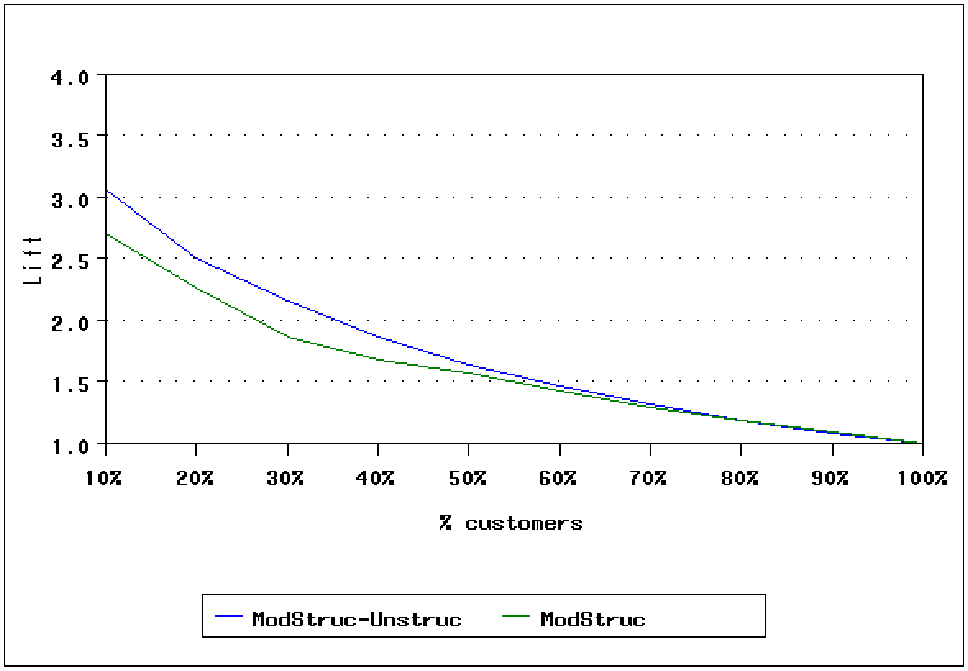

Figure 3: The cumulative lift charts of ModStruc and ModStruc-Unstruc.

As one observes from Table 1, ModStruc-Unstruc clearly outperforms ModStruc. The AUC increases from 73.80 to 77.75 by adding textual information to a traditional churn-prediction model. This improvement of 3.95 AUC-points is significant (χ^2=23.09,df=1,p<0.001). Figure 2 confirms these results graphically. The ROC curve of ModStruc-Unstruc is located further away from Random model than the one of ModStruc, thus the area under the ROC of ModStruc-Unstruc is bigger than that of ModStruc. Consequently, ModStruc-Unstruc is able to better distinguish churners from non-churners. Moreover, the beneficial effect of textual information on predictive performance is also confirmed in terms of top-decile lift (see Figure 3). The cumulative lift curve of ModStruc-Unstuc is situated above the one of ModStruc. In other words, ModStruc-Unstruc is able to identify more customers truly at risk than ModStruc within a specific decile. Lift in the first decile or the 10% top-decile – i.e. the 10% point in Figure 3 – increases from 2.69 to 3.07.

In sum, this study offers a realistic framework to increase the predictive churn performance among customers of whom textual information is available.

3.2. Complaint e-mail management

This paragraph gives an overview on how marketing managers can optimize the complaint handling process by distinguishing complaint from non-complaint e-mails using an automatic e-mail classification system. Using the methodology of the SAS® Text Miner as described in Section 2, e-mails are converted and summarized into *k* variables or SVD variables. Figure 4 shows the predictive performance of the e-mail classification model on the test and validation set in terms of AUC, while a comparison in terms of PCC is shown in Figure 5. The number of SVD-concepts is represented on the X-axis, while on the Y-axis, the performance measure is shown.

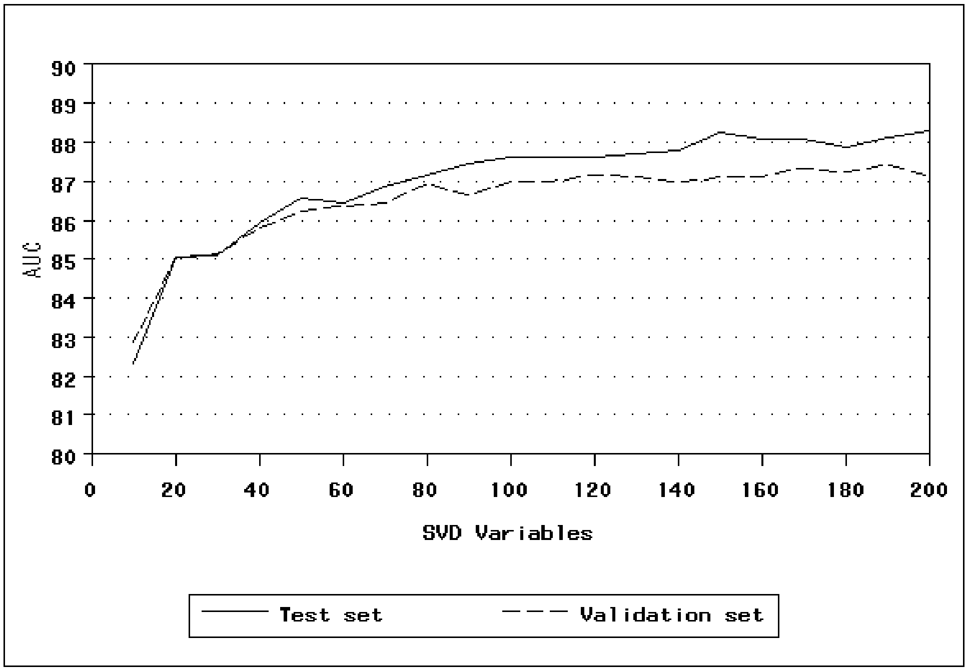

Figure 4: The Area Under the receiver operating Curve (AUC) on the test (solid line) and validation set (dashed line).

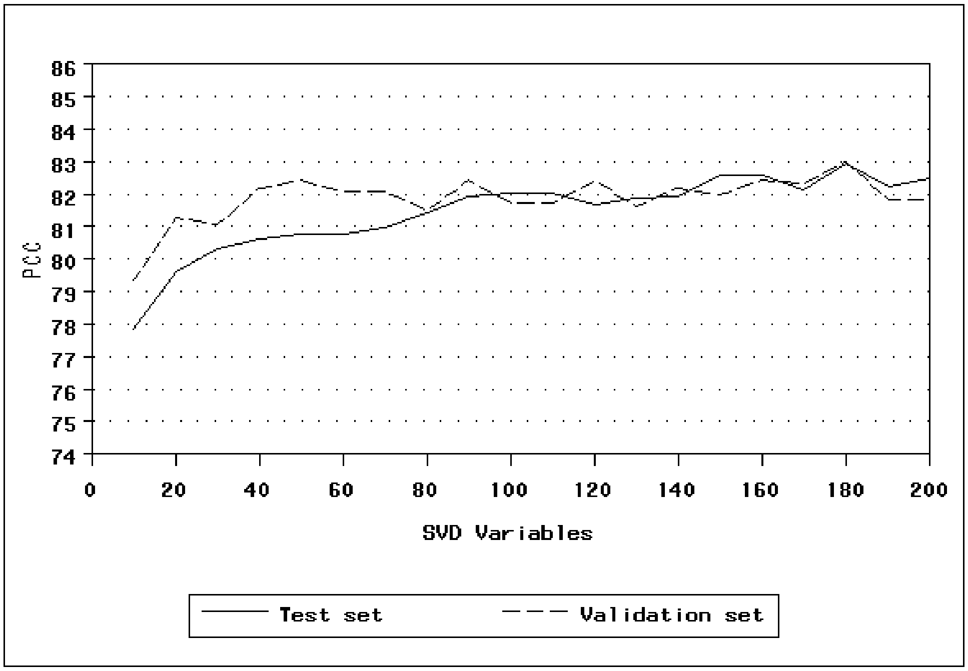

Figure 5: The Percentage Correctly Classified (PCC) on the test (solid line) and validation set (dashed line).

As one observes from Figure 4 and Figure 5, all models perform enormously well in distinguishing complaint e-mails from non-complaint e-mails. Indeed, the AUC performance of all models lies between 82.30 and 88.28, while the PCC lies in the range of 77.83 and 82.99. These results clearly indicate that (i) all models have predictive capabilities in distinguishing complaints from other e-mails, (ii) the more SVD-variables summarize the textual information, the higher the predictive performance and (iii) the models are highly robust because the difference between the test set and validation set performance is small.

In summary, applying these classifiers to a new timeframe does not result in a drastic drop in performance.Implementing the proposed methodology within a call center environment is a valuable strategy to improve customer complaint management.

4. Conclusion

Due to the strong increase in internet penetration, a lot of customers write e-mails as substitutes for traditional communication methods as for instance letters or telephone calls. As a consequence, companies receive daily a huge amount of e-mails. As such, new opportunities arise for Customer Intelligence researchers to use this newly available, but highly unstructured information into their daily analysis using the SAS® Text Miner. This paper contributes the Customer Intelligence domain by offering marketing managers an incentive to use these client/company e-mails as a valuable source to improve marketing decision making.

References

P.D. Allison. 1999. “Logistic Regression using the SAS System: Theory and Application”. Cary, NC: SAS Institute Inc.

Coussement, K. and D. Van den Poel. 2008a. “Integrating the Voice of Customers through Call Center E-mails into a Decision Support System for Churn Prediction”. Information and Management (forthcoming).

Coussement, K. and D. Van den Poel. 2008b. “Improving customer complaint management by automatic e-mail classification using linguistic style features as predictors”. Decision Support Systems (forthcoming).

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and R. Harshman. 1990. “Indexing by Latent Semantic Analysis”. Journal of the American Society for Information Science 41 (6): 391-407.

G. Salton. 1971. “The SMART Retrieval System: Experiments in Automatic Document Processing”. Prentice Hall, Englewood Cliffs, NJ.

SAS Institute Inc. 2004. “Getting Started with SAS® 9.1 Text Miner. Cary, NC: SAS Institute Inc.

K. Sparck Jones. 1973. “Index term weighting”. Information Storage and Retrieval 9 (11): 619-633.

Vitae and Contact Information

Kristof Coussement is CI researcher in the Faculty of Applied Economics and Business Administration at Ghent University (Belgium). He received his Master degree in Applied Economics as well as his Master after Master degree in Marketing Analysis at Ghent University (Belgium). During his research project, he investigates the impact of client/company interactions through verbalized information sources on Customer Relationship Marketing (churn analysis, customer complaint management,…). He published in large international peer reviewed journals like *Decision Support Systems*, *Information and Management*, *Expert Systems with Applications*,… while he has several other papers under review. He attended several international conferences like Benelearn, SIAM-SDM, Marketing Science, ISSPR,… Moreover, he is founder and committee member of BAQMaR (i.e. the Belgian Association for Quantitative and Qualitative Marketing Research) which throws together young marketing professionals by organizing networking activities and a yearly conference.

Your comments and questions are valued and encouraged. Contact the author:

Kristof Coussement
Faculty of Applied Economics and Business Administration (Department Marketing)
Ghent University
Tweekerkenstraat 2
B-9000 Ghent
BELGIUM

E-mail: Kristof.Coussement@UGent.be
Personal Website: http://www.KristofCoussement.com
Department Website: http://www.feb.ugent.be/mareco/ENG/index.asp