The current reference for this work is as follows:

Treiblmaier, H. and Mair, P. (2016) "Applying Text Mining in Supply Chain Forecasting: New Insights through Innovative Approaches", *23rd EurOMA Conference*: Interactions, 17-22 June, Trondheim, Norway.

If you have any questions, would like a copy of the final version of the article, or would like copies of other articles we've published, please contact me directly, as follows:

• **Professor Horst Treiblmaier**
- Email: horst.treiblmaier@modul.ac.at
- Website: https://www.modul.ac.at/user/treiblmaier/

Applying Text Mining in Supply Chain Forecasting: New Insights through Innovative Approaches

Horst Treiblmaier (horst.treiblmaier@fh-steyr.at)
University of Applied Sciences Upper Austria,
Logistikum
Wehrgrabengasse 1-3
Steyr, Upper Austria, Austria

Patrick Mair
Harvard University
33 Kirkland Street
Cambridge, MA 02138

Abstract

As is evidenced by the variety of papers being published in leading OM/SCM journals, these communities not only cover a broad range of topics, but also apply a multitude of methods in order to best tackle the research questions at hand. It therefore comes as a surprise, that so far the vast potential of automatically analyzing huge amounts of textual data (text mining, text analysis) has hardly been utilized. We therefore show how such techniques can be applied in order to mine qualitative data which was gathered in 19 interviews with SCM professionals on the topic of Supply Chain Forecasting.

Keywords: Supply Chain Forecasting, Text Mining, Text Analysis

Introduction

To date, the application of automated analysis for big chunks of text data (text analysis, text mining) has been largely ignored by the OM/SCM communities. Notable exceptions include the application of such techniques for literature reviews (Ghadge et al., 2012) or the summarization of conference abstracts (Rozemeijer et al., 2012). Furthermore, text mining has also been acknowledged as being an important part of demand planning (Cecere, 2013). More recent research has tapped into the huge potential of text data for other scopes of application. Abrahams et al. (2015), for example, mine user-generated social media content in order to discover product defects. Alfaro et al. (2016) illustrate how to detect opinion trends via the analysis of weblogs. Building on the latter ideas we illustrate how to gain knowledge from text data in Supply Chain (SC) Management. The main purpose of this paper is to show how various techniques of text mining can be successfully integrated into the overall portfolio of OM/SCM research methods. In the following sections we briefly discuss the research design of our study, followed by a short

description of the respective methods and their application in the context of a SC research project. We conclude with a brief outlook and suggestions for further research.

Methodology

Our text data are based on 19 interviews in three companies with Supply Chain and Logistics professionals from various ranks, working in the areas of manufacturing and wholesaling. Eight interviews were carried out in company A, six in company B and five in company C. The interviews were focused on current and pending problems in SC forecasting (Treiblmaier, 2015). The semi-structured and partly narrative interviews were taped and fully transcribed by the research team. The subsequent analysis was performed with the statistical open source software R (R Core Team, 2015) and included various steps of data preparation.

Initially, the transcripts were imported into R (R Core Team, 2016), to separate questions from answers, and some basic data cleanup techniques (e.g., removing special characters) were performed. This was done using the *qdap* package (Rinker, 2015). Subsequently, the texts were stored as corpora and the data was prepared using the *tm* package (Feinerer et al., 2008). All characters were converted to lower case and stopwords (i.e., words without any significant meaning), punctuations and numbers were removed. For each company we created a separate text corpus. An element within a corpus consists of the answers given by a particular manager. The interviews were originally conducted in German. We used the *translatoR* package (Lucas and Tingley, 2014) for performing a word-by-word translation using Microsoft's Translation API. The resulting English corpora were used for all further text analyses.

Results

In the following sections we discuss the results from five different methods: word clouds, sentiment analysis, topic models, correspondence analysis and multidimensional scaling. All these techniques were applied on the same corpora and R was used throughout. The methods themselves are well-established within various academic communities and our focus was exclusively on illustrating their applicability for OM/SCM research. More detailed information about the respective procedures and the source code is available upon request from the authors.

Word Clouds

To get an overview of the interview contents, word clouds (sometimes also referred to as tag clouds) were produced for each company separately, using the *wordcloud* package (Fellows, 2014). We set the minimum word frequency to 10 and the maximum number of words in each plot to 100 (see Figure 1). Generating a useful word cloud is a stepwise procedure, since the researchers have to decide on the stopwords in an iterative process. Existing lists of such words exist, but it turned out that further refinement was necessary. This was done in an iterative process until a final agreement among the researchers was reached.

Word clouds are a very popular tool to visually represent text data and they became quite popular on the Internet. Their interpretation is straightforward: the bigger a word, the more often it was mentioned throughout the interviews. In our case it gave the companies a rough idea about the most pending issues which emerged during the interviews. The word clouds provided an ideal starting point for further discussions. For instance, the relative size of the words (topics) can be used to critically assess the importance of the respective topic within the organization. Additionally, it is possible to discover 'missing' topics.

Figure 1 - Word clouds for companies A, B, and C.

Sentiment Analysis

In order to identify and extract subjective information about managers' responses, we used a sentiment analysis. This technique can be applied to determine the general attitude of an interviewee. In this study we were especially interested in the polarity of each interview based on words which are classified as positive, neutral or negative. We used the polarity dictionary by Hu and Liu (2004) as implemented in the *qdap* package. Positive words were tagged with a value of 1, negative words with a value of -1, and neutral words with 0. We computed the polarity score on a document (i.e., interview) level but this would have also been possible on a sentence level.

Table 1 - Interview Sentiment Scores

	Company A	**Company B**	**Company C**
1	0.49	0.33	0.20
2	0.65	0.06	-0.04
3	-0.17	0.65	1.25
4	-0.30	-0.58	0.08
5	0.82	-0.06	0.68
6	0.46	0.45	
7	0.99		
8	0.22		

The individual interview polarity scores are given in Table 1. The average sentiment scores at a company level are 0.39 for Company A, 0.14 for Company B, and 0.44 for Company C. The sentiment score heavily depends on the topics being discussed (e.g., exciting new growth strategy vs. problematic project failure) but it also allows for a general assessment of the interviewee's attitude which in turn is heavily influenced by the daily work routine.

Topic Models

The remaining analyses are based on a document-term matrix (DTM). A DTM is a frequency table with the documents (in our case the interviews) in the rows and the terms (words used in the interviews) in the columns. We computed a DTM for each company separately. DTMs are typically large and very sparse (i.e., they have lots of 0 entries):

- Company A: 2929 terms (sparsity of 77%)
- Company B: 2158 terms (sparsity of 71%)
- Company C: 2592 terms (sparsity of 68%)

We then reduced the complexity of the DTM by considering the most frequent words only. In order to do so, one can set a raw frequency cutpoint or, a bit more sophisticated, use the 'term frequency-inverse document frequency' (tf-idf) for keeping frequent words which reflects how important a word is to a document (see Blei and Lafferty, 2009).

For the topic models in this section we decided to use a tf-idf median cut. That is, we only kept the top 50% of the words which reduces Company A's DTM to 1537 columns, Company B's DTM to 1114 columns, and Company C's DTM to 1453 columns.

Next, we extracted clusters of words, which are also called topics. We used a latent Dirichlet allocation (LDA) (Blei et al., 2003) to perform the computation. LDA is a generative statistical model and is the most popular approach for computing topic models. It is implemented in the *topicmodels* package (Grün and Hornik, 2011). For each company we computed 5 topics and showed the top 10 words belonging to each topic. The results are given in Tables 2, 3, and 4.

Table 2 - Topics Company A

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	logistics	data	time	frequency	logistics
2	forecast	system	orders	branch	deliver
3	positions	forecast	business	pickup	stock
4	term	time	procurement	business	truck
5	tours	product	stock	channels	extreme
6	stock	business	items	forecast	logistic
7	leave	laughs	ordered	data	ordered
8	talks	volume	account	positions	distribution
9	thing	development	internal	quality	major
10	truck	evaluate	professional	extrapolation	sale

Table 3 - Topics Company B

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	budget	overdue	bottle	mio	bottle
2	stand	forecast precision	count	effect	march
3	market	days	mio	affected	mio
4	board	form	deviation	danger	rolling
5	budgeting	moment	takes	quarter	giant

6	contract	past	pcs	advantage	classic
7	directors	campaign	budget	carton	increases
8	rolling	claims	case	controlling	middle
9	theoretical	figure	hand	goods	rebuilt
10	contracts	any	larger	input	drip

Table 4 - Topics Company C

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
1	logistics	data	time	frequency	logistics
2	forecast	system	orders	branch	deliver
3	positions	forecast	business	pickup	stock
4	term	time	procurement	business	truck
5	tours	product	stock	channels	extreme
6	stock	business	items	forecast	logistic
7	leave	laughs	ordered	data	ordered
8	talks	volume	account	positions	distribution
9	thing	development	internal	quality	major
10	truck	evaluate	professional	extrapolation	sale

The results clearly show how different words make up various topics and how they differ between companies. In our research project, which focused mainly on the topic of SC forecasting, it was interesting to see which topics were related to the forecasting process within the companies. Similar to word clouds, these models can be used a starting point for further discussion, but they also allow for additional insight into what the pending issues within organizations actually are.

Correspondence Analysis

Next, we fit a correspondence analysis (CA) (Greenacre, 2007) which simultaneously shows associations among words and managers. We did this for each company separately. At the core of a simple CA computation is a singular value decomposition (SVD) on the frequency table which results (after some normalization) in scores for the row and column categories.

This time we used a simple frequency cutoff for the words. In order to avoid the resulting CA maps getting cluttered, we used the 20 most frequent words from the respective interviews. For each company we fit a two-dimensional solution using the *anacor* package (De Leeuw and Mair, 2009b) and produced the symmetric CA maps given in Figure 2. The within-rows distances and the within-column distances correspond to $\chi 2$-distances. M_x stands for the respective employees and the distance to the topics shows to what extent they "correspond" with those themes. Similar to a principal component analysis or (exploratory) factor analysis, a correspondence analysis is a useful multivariate graphical technique for exploratory research. It helps to detect relationships among categorical variables. In our study, the positioning of the employees allows for an easy identification of different work priorities.

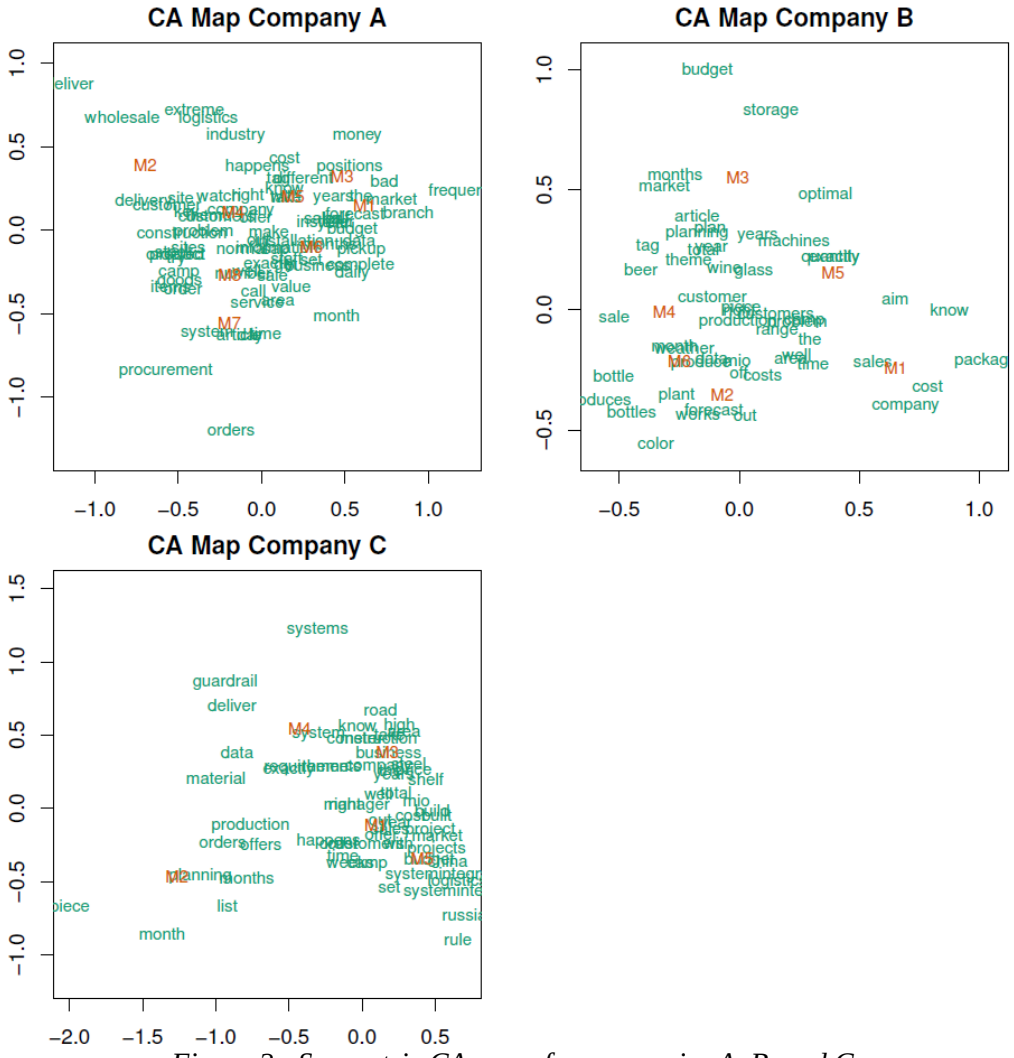

Figure 2 - Symmetric CA maps for companies A, B, and C.

Multidimensional Scaling

The last component of our analysis involved multidimensional scaling (MDS) (Borg and Groenen, 2005) and a subsequent hierarchical cluster analysis. MDS aims to represent input proximities (typically dissimilarities) between objects by means of fitted distances in a low-dimensional space. It therefore visualizes the level of (dis)similarity of cases in a dataset.

For this analysis we merged all the interviews into a single text corpus and strived to represent similarities between managers. The starting point is again a DTM with 19 rows (i.e., the total number of interviews) and 5059 columns reflecting all the words used in the interviews (minus the ones eliminated through data preparation). The first step was to choose a proper dissimilarity measure, applied across columns since we are interested in scaling the managers. We picked the cosine distance, which is popular in text mining applications (see e.g. Chen et al., 2009) since it normalizes for document length. The resulting 19 × 19 matrix acts as input for the MDS computation.

The MDS approach we used is called SMACOF and implemented in the package of the same name (De Leeuw and Mair, 2009a). We fit a two-dimensional, ordinal solution which leads to a stress-1 value of 0.175 (i.e., a standardized stress value which is independent of the scales being used). Figure 3 shows the resulting configuration plot and

we see that MDS nicely recovers three clusters of managers, according to the company they work for.

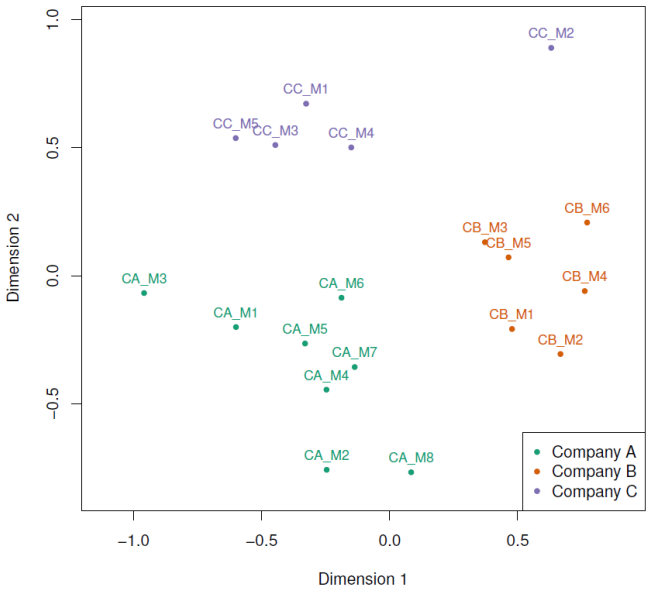

Figure 3 - MDS configurations on full dataset.

This clear cluster separation based on the words used in the interviews is confirmed by a simple hierarchical cluster analysis based on the fitted MDS distance matrix. By cutting the dendrogram (see Figure 4), based on Ward clustering at a value of 3, we see that the resulting clusters perfectly separate the managers from each other. In this example, the clusters represent the companies they are working for.

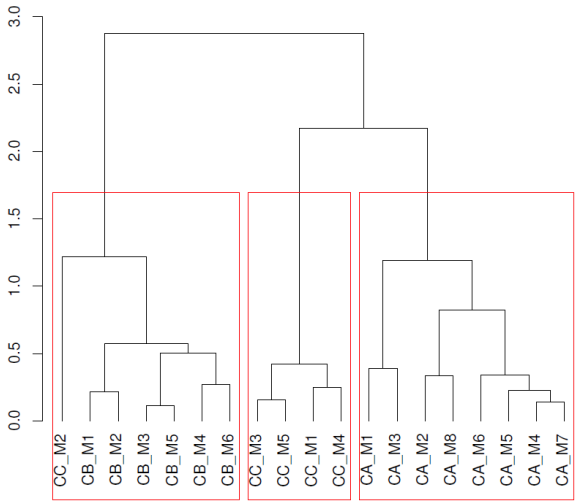

Figure 4 - Dendrogram Ward clustering on MDS distances (3-cluster solution).

Conclusions and Further Research

In this paper we introduce several techniques for automated text analysis which are rarely used in OM/SCM research but have a huge potential for identifying patterns being buried in vast amounts of text data. We present the results of an empirical research project about SC forecasting in order to briefly illustrate how different text analysis techniques may be used and interpreted appropriately.

Apart from applying existing, albeit state-of-the-art procedures of text analysis, this study is highly explorative, since we also test for the suitability of such techniques for OM/SCM research. By using automated text analysis, data can be analyzed in a fast and impartial way which does not depend on human judgement. Specifically, we applied the following techniques.

- Word clouds: Our analysis shows which topics (words) were most frequently mentioned.
- Sentiment analysis: We compute polarity scores based on the sentiments which the respondents used in their interviews. We were especially interested whether the average polarity scores differ across companies.
- Topic models: Using the document-term matrix as a starting point, we cluster the texts based on dominating topics by means of Latent Dirichlet Allocations (LDA). Each resulting cluster contains topics ordered in relation to their importance.
- Correspondence analysis: We computed associations between the words being used and the managers. A visual representation allows for a comparatively easy interpretation how managers and topics are related.
- Multidimensional scaling (MDS): We were interested in representing similarities between the interviews. Clusters of managers emerged which clearly showed differences between companies.

The application of text analysis is by no means limited to the techniques shown in this extended abstract. Amongst others, further options include a dynamic sentiment analysis which can be used to identify changes in the sentiments over time and a readability analysis, which indicates the ease with which readers can actually understand written texts.

In light of the many interesting qualitative studies in OM/SCM research, be it literature reviews (e.g., Cao and Lumineau, 2015), case studies (e.g., Barratt et al., 2011) or, as shown in this paper, interview data (see also e.g., Smith et al., 2009), we suggest that OM/SCM researchers add automated text mining techniques as a valuable tool set to their portfolio of research methods in order to be able to make the most out of the data they have at their disposal.

References

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z., and Jiao, J. (2015), "An integrated text analytic framework for product defect discovery", *Production & Operations Management*, Vol. 24, No. 6, pp. 975–990.

Alfaro, C., Cano-Montero, J, Gómez, J., Moguerza, J. M., and Ortega, F. (2016), "A multi-stage method for content classification and opinion mining on weblog comments", *Annals of Operations Research*, Vol. 236, No. 1, pp. 197-213.

Barratt, M., Choi, T. Y., and Li, M. (2011), "Qualitative case studies in operations management: Trends, research outcomes, and future research directions", *Journal of Operations Management*, Vol. 29, No. 4, pp. 329-342.

Blei, D. M. and Lafferty, J. D. (2009), "Topic models," in Srivastava, A. and Sahami, M. (Eds.), *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Press, pp. 71-93.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Borg, I. and Groenen, P. J. F. (2005), *Modern multidimensional scaling: Theory and applications*, 2nd edition, Springer, New York,

Cao, Z. and Lumineau, F. (2015), "Revisiting the interplay between contractual and relational governance: A qualitative and meta-analytic investigation", *Journal of Operations Management*, Vol. 33, pp. 15-42.

Cecere, Lora (2013), "A practitioner's guide to demand planning", *Supply Chain Management Review*, Vol. 17, No. 2, pp. 40–46.

Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A., and Cazzanti, L. (2009), "Similarity-based classification: Concepts and algorithms", *The Journal of Machine Learning Research*, Vol. 10, pp. 747–776.

De Leeuw, J. and Mair, P. (2009a), "Multidimensional scaling using majorization: SMACOF", *R. Journal of Statistical Software*, Vol. 31, No. 3, pp. 1-30.
De Leeuw, J. and Mair, P. (2009b), "Simple and canonical correspondence analysis using the R package anacor". *Journal of Statistical Software*, Vol. 31, No. 5, 1–18.
Feinerer, I., Hornik, K., and Meyer, D. (2008), "Text mining infrastructure in R", *Journal of Statistical Softwar*e, Vol. 25, No. 5, pp. 1-54.
Fellows, I. (2014), *wordcloud: Word clouds*. R package version 2.5.
Greenacre, M. (2007), *Correspondence Analysis in Practice*. 2nd edition, Chapman & Hall/CRC, Boca Raton, FL.
Ghadge, A., Dani, S., and Kalawsky, R. (2012), "Supply chain risk management: Present and future scope", *International Journal of Logistics Management*, Vol. 23, No. 3, pp. 313-339.
Grün, B. and Hornik, K. (2011), "topicmodels: An R package for fitting topic models", *Journal of Statistical Software*, Vol. 40, No. 13, pp. 1-30.
Hu, M. and Liu, B. (2004), "Mining opinion features in customer reviews", in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Vol. 4, pp. 755–760.
Lucas, C. and Tingley, D. (2014), *translateR: Bindings for the Google and Microsoft translation APIs*. R package version 1.0.
R Core Team (2015), "R: A language and environment for statistical computing", *R Foundation for Statistical Computing*, Vienna, Austria, https://www.R-project.org/, accessed December 6, 2015
Rinker, T. W. (2015), *qdap: Quantitative discourse analysis package.* R package version 2.2.4.
Rozemeijer, F., Quintens, L., Wetzels, M., and Gelderman, C. (2012), "Vision 20/20: Preparing today for tomorrow's challenges", *Journal of Purchasing & Supply Management*, Vol. 18, No. 2, pp. 63–67.
Smith, A. D., Plowman, D. A., Duchon, D., and Quinn, A. M. (2009), "A qualitative study of high-reputation plant managers: Political skills and successful outcomes", *Journal of Operations Management*, Vol. 27, No. 6, pp. 428-443.
Treiblmaier, H. (2015), "A Framework for Supply Chain Forecasting Literature", *Acta Technica Corviniensis – Bulletin of Engineering*, Vol. 7, No. 1, pp. 49-52.