Salespeople message boards

application of practical text mining techniques for understanding message board conversations

Borst, L.

Award date:
2013

Eindhoven, April 2013

Salespeople message boards: Application of practical text mining techniques for understanding message board conversations

by
ing. Lennart (L.) Borst

BEng Business management and technology
Student identity number 0722278

In partial fulfillment of the requirements for the degree of

Master of Science
in Innovation Management

Supervisors:
dr. A. de Jong, TU/e, ITEM
dr.ir. R.M. Dijkman, TU/e, IS

TUE. School of Industrial Engineering.
Series Master Theses Innovation Management

Subject headings: Text mining, text categorization, message boards, salespeople, content analysis

Management summary

The goal of this thesis is to analyze the content of conversations by sales representatives on online message boards, to thereby understand what subjects are discussed, the sentiment which salespeople hold towards these topics and if relationships between online discussions and the stock prices are observable.

To achieve this goal, we propose a roadmap to utilize two text mining techniques: text categorization and association rule learning. This roadmap proposes the standard text mining procedures (Höppner, 2005; Howland & Park, 2008; Manning & Schütze, 1999; Mitkov, 2005) in a way they can be applied in desktop friendly applications, that enable managers to quickly perform these analysis themselves. The text mining techniques are thereby not new; our goal is to acquire useful quantitative data from salespeople interactions on Cafepharma through readily available tools that might enable managers to relatively easy acquire insights into salespeople behavior and opinions. Therefore the novelty is not in the techniques used but in the insights into online salespeople interactions, as this has remained unstudied.

For this study we have manually labeled discussions from a pharmaceutical sales representative message board with category and sentiment tags. This data is used to teach self-learning text categorization models to label message board conversations automatically. For the explorative study of relations between categories, sentiment and stock prices we have applied association rule learning.

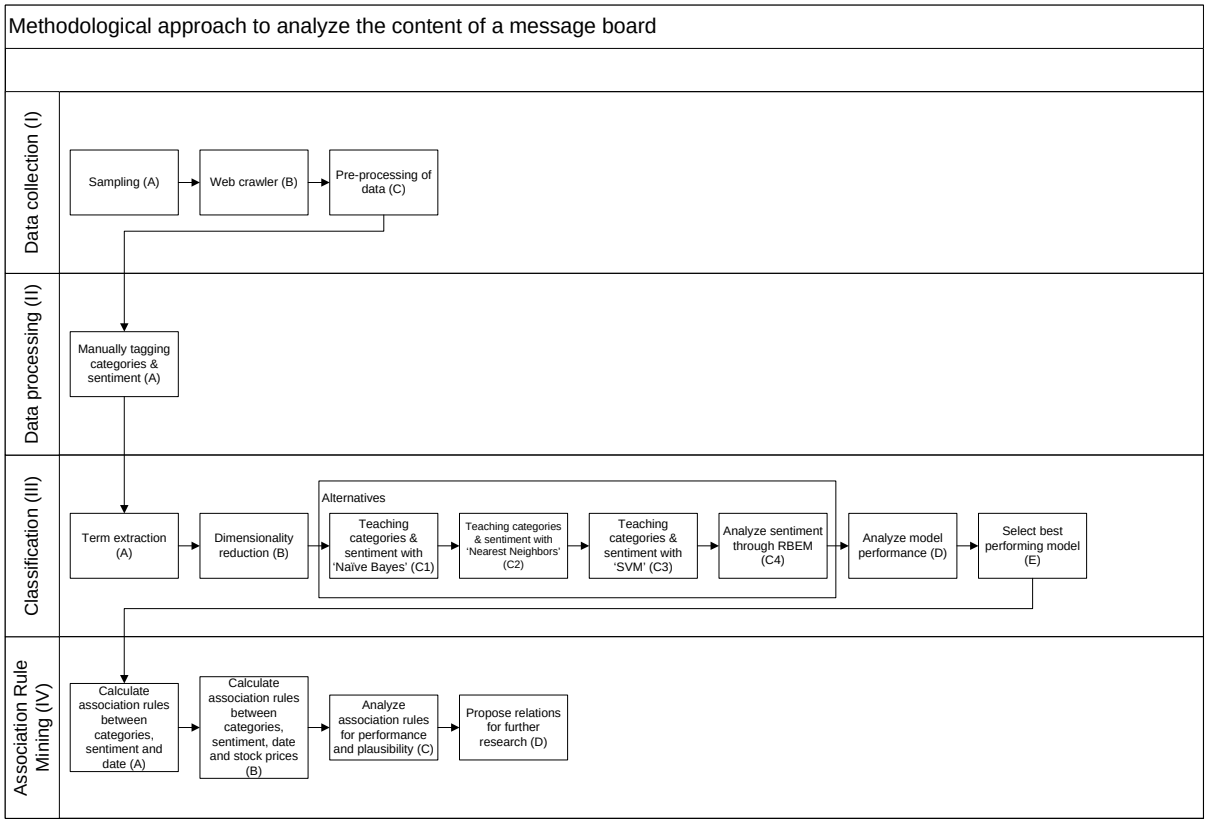

Figure 1 Roadmap for analyzing the content of salespeople message boards

The depth and complexity of roadmap was in some respect limited by the tools we applied (RapidMiner); this set a realistic scenario for the practical application.

The roadmap provide a sequential process in which data is collected (Step I), data is processed by labeling it with categories and sentiment labels (Step II), teaching and evaluating self-learning models (Step III) and finally looking for relations through association rules (Step IV).

Step I: Several boards were selected for sampling from the message due to their size and the size of the companies they were associated with. The conversations were retrieved from the message board through a web crawler and were stripped from any HTML code.

Step II: The sample data was labeled with a restricted set of four categories, a more detailed set of ten categories and three sentiment labels.

Step III: The labeled data was used as input for self-learning models. Three modeling techniques were selected due to their performance with text categorization. These techniques were: naïve Bayes, K nearest neighbor and Support Vector Machine (SVM). Dimensionality reduction based on Singular Value Decomposition (SVD) was applied to increase model performance and efficiency. Additionally we discovered that the model accuracy did not accurately reflect performance due to the overrepresentation of certain categories and sentiment labels, this was overcome by randomly deleting some of these cases to create a more balanced data set.

Step IV: Possible relations within the data set and with stock prices were acquired through an explorative study based on association rule learning. For relations within the data set we used categories, sentiment labels and dates. For the relations with stock prices we acquired historical stock prices on the companies in our sample.

For Step I we required custom tools due to the fact that the features within RapidMiner for acquiring data from message boards were not sufficient.

Based on our own experience with Step II we propose to let this step be performed by multiple people in parallel, this would overcome biases of an individual tagging large volumes of message board data (Fleiss et al., 1969; Fleiss, 1971).

The best performing model in Step III were SVM models for the first level category and the sentiment. For a more detailed level of categories no adequate model was found. In this study we found that the manual sentiment analysis was more accurate (Chmiel et al., 2011) than the RBEM sentiment analysis. The quality of the sentiment analysis could be improved by either applying the improvements proposed in Step II or applying RBEM with respect to the structure of topics on Cafepharma.

In Step IV no relations were found between topics on Cafepharma and the stock prices of the pharmaceutical companies involved. The only insight we gained was that we observed a negative trend for the sentiment over the last couple of years. For better insights into message board conversations in relation to performance we propose the use of a data source that is more closely related to sales, for instance sales KPIs or turnover.

In this study, we made use of a publicly available performance measure of companies: stock prices. Companies themselves have much more accurate and specific performance measurements down to sales groups and products. Through these performance measurements and – for instance – the category *Task: New Product* combined with the sentiment an organization can get insights into the reception of its salespeople of new products. Without categorization it would be very hard to find related topics on a message board and to get quantitative data on the sentiment towards a certain category. With this data it becomes easier for companies to find – for instance – criticism of salespeople on certain products or measure the effects of organizational performance on salespeople sentiment. Thereby managers are able to gain quantitative insights into salespeople behavior and opinions through using this roadmap with readily available PC applications. The current limit of these readily available applications is that they are not yet able to categorize text with word lists.

Key words

Text mining, text categorization, message boards, salespeople, content analysis

Acknowledgement

Firstly I wish to thank my mentor dr. Ad de Jong for offering me the possibility of this study, the support and the sometimes much needed demand for deadlines. When initial thesis possibilities did not work out he offered me a new option and supported me in the realization of it. Secondly I wish to thank dr. ir. Remco Dijkman for both his intellectual and technical support and feedback during the process of writing this thesis.

This thesis would not have been possible without the support of my dear colleagues, Marcel and Frans, at Strukton who offered me the possibilities to take the extra time I required for finishing this thesis. I am also grateful for the initial opportunities offered to me by Robbert Aeyelts Averink.

For their support and much needed coffee breaks I want to thank my friends at the University and for their support and much needed Christmas break I want to thank my parents and brother.

Lastly I want to thank my girlfriend, Angélique, for without her support and aid this thesis would not have been possible, she stood by my side during the long weekends of endless labeling of message board data and offered me a home away from home in Brussels to work on this thesis.

Lennart Borst
April, 2013
Eindhoven

Table of content

1 Introduction

In this thesis, we will study the problem of analyzing the content of message boards for salespeople, in order to get a closer understanding into what salespeople discuss and how their sentiment is in relation to the subjects they discuss. Salespeople, their methods, and platforms of communication have remained subjects that require additional research.

In Section 1.1 we will delineate why additional research is required and the relation of message board content with the broader research theme of multi-channel marketing. In Section 1.2 we will formulate a problem definition based on the earlier stated research interest. For the readers' ease of reading, Section 1.3 will provide the outline of this thesis.

1.1 Motivation

Within the fields of marketing and sales, an important topic is multi-channel marketing. A major trend within this topic is social media. Like other channels within multi-channel marketing such as brick-and-mortar store or call-centers, social media can be used by salespeople. Salespeople have to interact, inform, and sell to and with customers through a wide variety of channels. More specifically, the Internet channels get more elaborate all the time with social media platforms and the integration of these platforms with various other channels and services (Shankar, Inman, Mantrala, Kelley, & Rizley, 2011). For instance, the way in which a customer earning badges with Foursquare can earn a free coffee at Starbucks (4SquareBadges.com, n.d.). Some level of data integration should be achieved if companies want social media to play a significant role as a channel for multi-channel marketing (Neslin et al., 2006), which sparks a need for quantitative data and insights into social media interaction and discussions. Social media channels are however not only reserved for use in interactions with the customer, but can also be used by salespeople themselves to interact with each other. By deploying social media exclusively for salespeople, information can be shared rapidly across different regions, entities, or organizations. An example of this is message boards. Through this application, message boards are a supporting platform for salespeople to perform their tasks within the different channels that they use for selling their products. This presents the message boards as a research subject within the field of multi-channel marketing in a different way than has been studied up until now. Thereby, addressing its information sharing aspects in much greater detail as opposed to its role as a purchasing channel, and showing a different opportunity of how organizations can leverage social media. This role of social media and message boards in particular, will thereby be the study object of this thesis.

For managers quantitative data on online salespeople interactions enables them to acquire further insights into how new products, sales strategies, organizational changes and work-related contacts are perceived by salespeople. If this information is related to performance measurements there are possibilities to further study and understand behavior of salespeople, thereby not only giving managers new information and data, but possibly also providing them with new tools to manage their sales force.

Within extant research, the marketing message boards have remained unstudied. The main research directions have a predominant B2C marketing focus where the internet channel is mostly studied as a collection of web shops. By looking at salespeople message boards the focus shifts to a sales perspective with a stronger focus on the information sharing function of the internet channel as opposed to the research shopping or selling function. The exact current state of research on multi-channel marketing will be further discussed in Sections 2.2 and 2.3.

1.2 Problem formulation

The context of this research will be Cafepharma, a message board for sales people in the pharmaceutical industry. This message board is not moderated by a pharmaceutical company, even though most of it is organized per company. This enables colleagues to easily find each other. On a message board, interactions take place by users placing posts in 'topics'; these topics form conversations or discussions. These topics will be the level at which the message board will be studied. This is because a single post cannot be studied in the context in which it was placed by the user. An interesting aspect to understand is the content of the topics and ratio of topics in relation to each other. To acquire additional insights, we could see how the content of topics relates to the sentiment of the topics. In order for us to do this, we would of course need to know the sentiment of a topic. It is however a challenging job to understand the content of each topic by reading every forum post manually, so this analysis of the content of message boards has to be done in a different way. But how does one analyze this much data without manually reading each and every one of them?

For analyzing large amounts of textual data, we argue to apply text mining techniques. Within text mining, a multitude options are available for analyzing text (Mitkov, 2005). When keeping the above in mind, we only need two dimensions to understand the content of topics in order to determine ratios in the relationships between them or relationships with external performance indices, like stock prices.

The first dimension is the category of the topic, which can be seen as the subject of a topic, and this is a key process in text mining and one of the applications often found when text mining (Apte, Damerau, & Weiss, 1998; Mitkov, 2005). The second dimension is sentiment. Sentiment is whether the text is positive, objective or negative (Pang & Lee, 2008). Through the category, it will be possible to understand what salespeople are talking about and through the sentiment it will be possible to measure how they feel about a certain topic. The reason to use only these two options from the tool set of text mining is that it reduces the enormous clutter of text to only a few variables of which we can immediately measure the frequency. By taking categories and sentiment as our point of view, we can then analyze words determine category and sentiment through text association rules (Mitkov, 2005).

The techniques we will apply are not new; our goal is to acquire useful quantitative data from salespeople interactions on Cafepharma through readily available tools that might enable managers to relatively easy acquire insights into salespeople behavior and opinions. Therefore the novelty is not in the techniques used but in the insights into online salespeople interactions, as this has remained unstudied.

For thorough insights into the relations between message board conversations and behavior performance measurements on salespeople is required. Since these measurements are not openly available and we wish to limit the scope of this study to a fitting scope for a thesis, we will focus on acquiring quantitative data from Cafepharma and explorative research on the relations between the content of conversations and an external performance measurement: stock prices.

The problem definition is as follows:
'How to analyze the content of conversations by sales representatives on online message boards? In the beginning, through understanding the category and sentiment of topics.'

To understand the position of this study within the field of multi-channel research, an answer will need to be found for the following questions:

1. What are the main concepts within the field of multi-channel research?
2. What are the main themes within extant multi-channel research?
3. What major gaps are identifiable within extant multi-channel research?

In order to find an answer for the problem definition, the following research questions need to be answered:

1. What kind of methods can be applied to identify the category and sentiment of a conversation?
2. How can these methods be applied?
3. Are there relations between the categories and sentiment?
4. What is the relation between category/sentiment ratios and the stock price index?

In light of research question 2, we will propose a roadmap for analyzing and understand the content of salespeople message boards. The benefit of the roadmap is that executing this roadmap will in practice provide answers to research question 4 to 6. The foundations of the roadmap design are the answers to research question 1 and 2.

1.3 Thesis outline

This thesis will start with some theoretical background behind choosing this subject; this will be done in Section 2. In Section 3 we will explain the research setting. The methodological approach will be explained based on a roadmap that will be proposed in the same section; Section 4. Some techniques require extra explanation for actual execution; this will be done in Section 5. Models and results will be presented and discussed in Section 6. In Section 7, we will draw the final conclusions and review to what extent we have been able to answer the research questions.

2 Theoretical background

Over the last decade, the number of channels to sell and market products has considerably increased. Especially with the growing popularity of social media and smartphone apps (Shankar et al., 2011) which have enlarged the available channel options. As a result, companies and their customers have more options to gather information and more sales points at which to purchase. The growing number of channels provides opportunities for companies, as customers who buy through multiple channels are more profitable (Kumar & Venkatesan, 2005). However, within this multi-channel environment, organizations face challenges on how to setup and manage their sales (Neslin et al., 2006; Schoenbachler & Gordon, 2002). In the following sections we will elaborate more on the theoretical background of the relevance of this study.

Firstly, we will look at the general concepts of multi-channel research, like channels and multi-channel marketing, in Section 2.1

Secondly we will discuss the current state of research on the application and control of channels in a marketing and sales environment in Section 2.2.

The review of literature will be concluded in Section 2.3, where the gaps in the current literature will be analyzed.

2.1 Multi-channel concepts

To be able to understand what the major multi-channel research themes are, it is essential that the main concepts are defined. This will be addressed in the following sections.

In recent years, a lot of research has been done on multi. Most of these papers however, do not formulate a set definition of what channels and multi-channel sales are.
To gain clear understanding of what the topics of these papers are, definitions need to be formulated for each of them.

2.1.1 Channels

The word channel on its own has very broad usage. The New Oxford American Dictionary (*New Oxford American Dictionary,* 2010) defines a channel as "a medium for communication or the passage of information". This general definition will not suffice, and therefore a specific definition needs to be determined. The reason for not using either 'sales channel' or 'marketing channel' is that this difference is not made in multi-channel related literature. For instance, Neslin et al. (2006) uses the term 'channels,' whereas Rosenbloom (2011) uses the term 'marketing channels'. The term 'sales channels' is not broadly used amongst academic authors as ABI/INFORM returns 881 results for 'sales channels' in scholarly journals, whereas 'marketing channels' return 4446 results in scholarly journals[1]. Also, papers focusing on 'sales channels' were not found to be closely related to multi-channel sales.

Hence, what is a channel in relation to sales? If we define it as a 'marketing channel', it can be seen as an "external contractual organization that management operates to achieve its distribution objectives" (Rosenbloom, 2011). However, this definition conflicts with other definitions used in the literature. First this definition stresses that a channel is by definition an external entity. As often seen in studies, channels are often not external organizations, but part of the organization (Kollmann, Kuckertz, & Kayser, 2012; Montoya-Weiss, Voss, & Grewal, 2003; van Birgelen, de Jong, & de Ruyter, 2006). Furthermore, with this definition the goal of a marketing channel is "to achieve … distribution objectives". The aforementioned may be so in respect to the wider operation of an organization, but it does not address what a channel does and in what phase it is of importance (search, purchase and/or after-sales (Neslin et al., 2006)). Furthermore, it doesn't articulate the sales function in this process.

A definition of channels that might be more applicable is the one used in relation to multi-channel management (more on that later): "a customer contact point, or a medium through which the

[1] As consulted on June 13th 2012 through the ProQuest database. The ABI/INFORM database goes back until 1971. (*ABI/INFORM,* n.d.)

firm and the customer interact" (Neslin et al., 2006). Neslin et al. (2006) emphasize that this means that a channel is about two-way communication, thus mass media like television advertisements are excluded, and it includes more of the sales function in this definition. Other research in the same field does however include channels that are one-way (Gensler, Dekimpe, & Skiera, 2007; Konuş, Verhoef, & Neslin, 2008; Venkatesan, Kumar, & Ravishanker, 2007). This implies that a channel can be both one-directional and two-directional. Thus, a channel can perform the role of an information source as well, which gives it a role in each of the phases. This definition also allows channels to be part of the organization itself.

In conclusion, the definition of a channel by Neslin et al. (2006) seems to more closely match other research than the definition of marketing channels by Rosenbloom (2011). Nevertheless, this does not disqualify the term marketing channel as it is often used in a way similar to Neslins (2006) channels; the term is just not defined in other research. To conclude on what the construct of channel is we will use Neslins (2006) definition together with the organizationally related goal that Rosenbloom (2011) defines: "a customer contact point, or a medium trough which the firm and the customer interact, to ultimately achieve the firms distribution objectives".

In literature, four "traditional" channels are the subjects of research. These four channels are the *brick-and-mortar stores*, *catalogues*, *call-centers* and the *Internet* (Gensler et al., 2007; Kollmann et al., 2012; Konuş et al., 2008; van Birgelen et al., 2006; Venkatesan et al., 2007; Verhoef, Neslin, & Vroomen, 2007) (Konuş (2008) actually compared all four of these channels). At first glance, one could argue that these channels seem like a simplification of the different channels that exist. To understand the scope of the channel definitions, each of the channels will discussed in more detail.

Brick-and-mortar

Brick-and-mortar stores vary depending on format and goods sold. For retail format we can distinguish small general stores, urban retail specialists, general department stores, category killers (e.g. Toys 'r us or IKEA) and discount department stores/general merchandisers (Hollander, 1966). Additionally, different specialty formats exist like 'non-specialized (mainly food)', 'pharmacy & medical', 'textiles', 'clothing', 'furniture, lighting & household', 'second hand', 'specialty food', 'footwear & leather', 'electric, household & TV', 'hardware & paint' and 'Books' (Reynolds, Howard, Cuthbertson, & Hristov, 2007). This overview does not take retailers with service product – e.g. banks – into consideration (object of the study by van Birgelen et al. (2006)).
In general, one aspect seen in all literature is that brick-and-mortar stores are physical points of contact for a customer, whereas live interaction is possible with personnel/sales-people and – if the product or retail formula allows it – the products.

Internet

The Internet is also considered to be one channel from a multi-channel research perspective, even though authors describe different formats. For instance, van Birgelen (2006) describes online-banking as a form of online service provision. There is also a difference in selling your own products (Kumar & Venkatesan, 2005) versus reselling other products such as an online department store (Konuş et al., 2008; Verhoef et al., 2007). A special channel are the auction websites like eBay where products are being resold (Gopal, Pathak, Tripathi, & Yin, 2006), and a lot of this reselling of (new) products is done by private consumers. These are sales outside of the promotion and marketing by the original manufacturer or retailer (Gopal et al., 2006).

A channel that can be used for either information exchange or information exchange and sales are the online comparison websites – e.g. kieskeurig.nl – (Huang, Lurie, & Mitra, 2009; Xu & Kim, 2008). A very new sort of channel is the mobile channel – or m-commerce –, including the use of apps (Balasubramanian, Peterson, & Jarvenpaa, 2002; Shankar et al., 2011). Of course, the Internet is often used for reaching out to customer, either through "old fashion" email or through weblogs and new social media outlets like Twitter or Facebook (Rickman & Cosenza, 2007; Shankar et al., 2011). Lastly, it is important to note that in the thorough empirical studies only one overall Internet channel is studied, which generally is described as more of web shop or sales access point (Gensler et al., 2007; Konuş et al., 2008; van Birgelen et al., 2006).

Catalogues

Catalogues have quite often been the subject of research (Konuş et al., 2008; Venkatesan et al., 2007; Verhoef et al., 2007). This is a channel where products are being offered through a catalogue (e.g. Wehkamp, Otto, and originally Sears), and this kind of channel is also sometimes – in daily use – referred to as 'mail order'. Mail order is defined by the New Oxford American Dictionary (*New Oxford American Dictionary,* 2010) as: "the selling of goods to customers by mail, generally involving selection from a special catalogue". This definition shows the role and function of a catalogue within the information and purchase phases.

Call-center

The last channel to be discussed is the call-center channel; it is again often used for studies (Gensler et al., 2007; Konuş et al., 2008; Venkatesan et al., 2007; Verhoef et al., 2007). In the case of Gensler et al. (2007) the call-center is the purchase (and possibly the additional information) channel for a TV home shopping channel. The definition of a call center given by the New Oxford American Dictionary (*New Oxford American Dictionary,* 2010) is: "an office set up to handle a large volume of telephone calls, especially for taking orders and providing customer service". This definition tells us that a call-center can fulfill both an informative role as a purchase function, and it is able to interact with large amounts of customers without the need of large amounts of contact points (shops). It offers personal contact, to a slightly limited degree, but it does not offer interaction with the product.

2.1.2 Multi-channel marketing

Similar to the term channel, the term multi-channel is left undefined in most recent literature. Various multi-channel terms are used like multi-channel marketing, multi-channel sales, multi-channel retailing, multi-channel shopping, multi-channel shoppers, multi-channel customers, multi-channel management, multi-channel customer management, etc. Papers on multi-channel shopping or shoppers are all on customer behavior, so they do not describe the core concept of multi-channel. Instead, they describe how customers act when in a multi-channel situation (Venkatesan et al., 2007).

Though recent studies are not very strict on their usage of adjectives, some earlier works (Montoya-Weiss et al., 2003; Schoenbachler & Gordon, 2002) do have the tendency to stick to the term multi-channel marketing. Because this term is often used and it also links closely to innovative retailing trends (Shankar et al., 2011), this term will be used throughout this thesis. To show that it is broadly applicable, even if other studies have used different definitions, we have to find a definition that proves that it is the key concept.

A concise definition of multi-channel marketing is given by Brassington and Pettitt (2006), as they state that multi-channel marketing is making products and services available to customers through linking a group of channels. Or to offer products to consumers through one or more channels (Schoenbachler & Gordon, 2002). This definition is appropriate, as it defines the core of multi-channel marketing, but the scope of the definition can be broadened to include the aspects we defined in the channel definition. To make the definition of multi-channel marketing more consistent with that of channels, we define that multi-channel marketing uses a mix of channels to reach customers, and the objectives are to distribute resources across this channel mix with the goal to satisfy customers and maximize profits (Montoya-Weiss et al., 2003; Moriarty & Moran, 1990). What this definition shows – additionally to being more consistent with the definition of channels – is the challenge that lies within multi-channel marketing; how to distribute resource across different channels to reach the goal of maximizing profit. Another important aspect of this definition is that it is in line with the modern marketing view that is more customer centric. Once again, this again is in line with the modern marketing trends (Shankar et al., 2011). A key aspect of this definition is that the mix of channels is used 'to reach customers'. This implies not only selling a product but also to inform, making multi-channel marketing not only a sales instrument but also a wider marketing instrument.

An additional definition needed when talking about multi-channel marketing is 'channel adoption'. This is when customers start to use additional channels beside the one(s) they already use (Venkatesan et al., 2007). This is a definition that we need, because it is relevant in the process of customers becoming multi-channel shoppers.

2.1.3 Concepts conclusion

In this chapter, the main concepts within the field of multi-channel have been defined. What can be said is that four main channels are the subject of studies; brick-and-mortar stores, internet, catalogs and call-centers. Where catalog retailers have been making a shift into other channels and the Internet channel has a wide variety of aspects that make it a good candidate for channel integration with brick-and-mortar stores. However, the aspects of the Internet that makes it a good candidate for channel integration – namely social media and mobile internet – are often not specified in thorough, empirical multi-channel research.

We have also been able to find a general concept for multi-channel which is called multi-channel marketing. Even though this concept has a possible broad application, the sales function only fits in the margins of the scope, thereby still neglecting the sales function for the most part. A conceptual integration of the marketing and the sales function in multi-channel research was not found in the literature, and a first hint of a gap has thus been found.

Relevant interesting themes that draw attention when looking at the current state of multi-channel research are: the Internet as a channel in a multi-channel environment, the marketing versus sales perspective of multi-channel and B2B versus B2C. These themes will be further discussed in the next section.

2.2 Multi-channel research

After having defined the key concepts of multi-channel research in the previous chapter, this section will discuss what has and what has not been studied. This will be discussed through identifying the major themes across these studies. Across these studies are the following themes that draw attention:

1. Internet channel
2. Marketing versus sales
3. B2B versus B2C

These themes will be further discussed in the following sections. Thereby not only illustrating what has been studied but also which subjects have remained unstudied, as this is the basis of determining the status quo and the gaps within multi-channel research.

Through a comparison table, an overview is presented of what has and has not been. This table can be found in Appendix I. In this table, characteristics of studies are given such as the variables, concepts, context, multi-channel setting, channels, and kind of study and sample size

2.2.1 Internet channel

When looking at retailing trends (Shankar et al., 2011), the importance of the internet channel becomes evident. Even more so, this internet channels is everywhere and does everything through mobile platforms and a wide variety of online services. This raises the question, to what extend has the internet been studied like this? Or has it just been studied as a collection of web shops?

As just shortly pointed out (Section 2.1.1); most authors only look at web shops when referring to internet channels. Only van Birgelen (2006) and Montoya-Weiss et al. (2003) describe internet channels that are not web shops, but also service selling, service delivery and sales support. Internet as just a web shop would actually disqualify internet as a channel (Section 2.1.1) and would go against the origins of the internet (Edosomwan, Prakasan, Kouame, Watson, & Seymour, 2011). Internet has always been about the two-way communication. To have an even more complete look at internet as a channel – than van Birgelen et al. (2006) and Montoyas-Weiss et al. (2003) already do – also means that social media should be added to the scope of research (Edosomwan et al., 2011; Shankar et al., 2011). This relatively new side of the internet goes back to the origins of the internet where it was a meeting place for people.

Additionally, the Internet offers a great deal of possibilities for the sales function both in interacting with customers and amongst sales peoples for sharing information. Information sharing – or data integration – has not been studied in great detail (Neslin et al., 2006) and could benefit from the application of the Internet within an organization for internal information sharing.

In conclusion, in all modern multi-channel research the Internet as a channel has been part of the scope. But the application of the Internet has been shallow, either purely as a web shop or in some cases as a service access point. A more in depth application of the Internet would incorporate more facets of these channels such as the mobile Internet, social media and internal information sharing. This could also have a significant influence on the dynamics of the Internet channel and its effects on the other channels in the channel mix.

When looking at trends about channel integration, it is important to note that these trends are often founded in a broader definition of the Internet channel, with social media and mobile apps as a strong component (Shankar et al., 2011). The Internet has made its first big step into the multi-channel domain, but it has yet to be studied in a broader perspective.

2.2.2 Marketing versus sales

As the multi-channel marketing definition (Section 2.1.2), informing and purchasing are important phases within the multi-channel domain (Blattberg, Kim, Kim, & Neslin, 2008; Neslin et al., 2006). These phases could be identified as sales functions, but what is the predominant perspective in multi-channel studies?

The perspective – used by the authors of the articles in the comparison table in Appendix I– is a marketing perspective. A sales perspective is largely overlooked, only van Birgelen et al. (2006) and Verhoef et al. (2007) have – in varying degrees – a more sales oriented perspective. However clear research in the added value of multi-channel marketing as a sales tool or technique has remained a topic that has yet to be researched. This is evident through focus on resource allocation on a strategic and tactical distribution level (Neslin et al., 2006), where the translation to and the impact on the sales function are left out of the scope. For instance, selling through multiple channels could require different sales techniques, sales tools and a different composition or management of the sales force. This impact of the multi-channels marketing strategy on the sales strategy has hardly been studied. An interesting perspective on the application of multi-channel marketing would be the use of these channels by sales people, not only in their interaction with customers but also the sales information it might create for them – for instance how social media gives insights into word-to-mouth communication amongst customers. Another option would be to take a completely other perspective on multi-channel sales and look at the internal workings of it, for instance how sales people could share information about customers amongst each other. This could possibly aid information exchange amongst entities or regions within a sales organization.

The literature overview by Neslin et al. (2006) shows a strong focus on marketing related research topics like *data integration across channels*, *understanding customer behavior*, *channel evaluation*, *allocating resources across channels* and *coordinating channels strategies*.
The step to take the marketing insights and providing them for sales application is however not addressed.

To conclude, the research field of multi-channel has had a predominant marketing focus. Nevertheless the strong insights in customer behavior might be useful insights if and when they are translated, tested and studied for sales application. These marketing insights combined with the possibilities for the sales function – for instance through the application of the internet and information exchange amongst sales people – could show interesting new sales dynamics.
In a more general sense the impact of multi-channel marketing on sales has not been studied, these insights could be vital for successful multi-channel customer management.

2.2.3 B2B versus B2C

Sales and marketing are not limited to business to consumer (B2C), before products or services end up in the hands of consumers it is often traded in a business to business (B2B) environment. The question is how well is this has been studied in the field of multi-channels research.

The focus of all papers from the comparison table in Appendix I focus on a B2C setting, or where it is not further specified what possible challenges might arise when dealing in a B2B setting. In B2B an entirely different set of channels might be applicable, but insight into this topic limits itself to parallels that can be drawn from the service selling orientation van Birgelen et al. (2006) offers us. For the B2C perspective a broad focus can be found between the different studies. From service side (Montoya-Weiss et al., 2003; van Birgelen et al., 2006) to home shopping & shop shopping (Gensler

et al., 2007; Kollmann et al., 2012; Konuş et al., 2008; Venkatesan et al., 2007; Verhoef et al., 2007) and understanding of behavior of customers (Kollmann et al., 2012; Montoya-Weiss et al., 2003; Neslin et al., 2006; Schoenbachler & Gordon, 2002; Venkatesan et al., 2007; Verhoef et al., 2007). A wide scope with a wide variety of channels, however within the limits discussed earlier.

To conclude, the studies within the field of B2C have been thorough. But since no parallels have been studied between B2C and B2B, the application of multi-channel marketing on a B2B environment is still unknown territory for multi-channels research.

2.3 Gaps within multi-channel research

The first big gap is the lack of studies on channel integration. Though a lot is known to this date on the processes and behaviors of multiple channels, the integration of these channels is mostly unstudied. However, when looking at the definition of the internet channel and retail trends the Internet has a very important role in channel integration. This cannot be studied right now, because the Internet has mostly been studied as a collection of web shops instead as a mobile and social platform. Until the research scope on internet has been widened, this theme cannot be further researched to its fullest.

Secondly, the internet as a channel has been quite thoroughly studied within the scope of a collection of web shops. Further research is, however, dependent on studies on the Internet as a channel in a multi-channel setting with a broad scope on what the Internet as a channel has to offer. More specifically, either its mobile aspects or its interacting aspects through social media.

Thirdly, continuing on the same track as channel integration and a wider Internet scope is the discussion around marketing versus sales. Marketing is only a part of the complete multi-channel picture. The question remains as to what is the impact of multi-channel strategy decisions at a marketing level on the sales strategy.

Topics such as sales strategy, sales force and sales techniques have mostly been left out of the research scope and should be researched in the future. Additionally, multi-channel sales and the application of the Internet channel give possibilities for a lot of different study objects, such as information sharing amongst sales people through social media

Finally, there is the B2B versus B2C scope within multi-channel studies. All of the multi-channel studies have a predominant focus on B2C and give no insights in the specific challenges that might arise when dealing with a B2B environment. The same applies for including sales within the research scope, as B2B should also be more often included in future research scopes.

With this thesis, we aim to offer insights within three gaps: the application of the internet channel, and the lack of a sales and the B2B perspective. This is done by studying the content of salespeople message boards. These message boards are reserved for salespeople in the pharmaceutical industry who perform a B2B sales function towards physicians.

These message boards are in itself not sales channels, as they perform the role of a supporting platform for sales people to perform their tasks within the different channels that they use for selling their products. In this way, the Internet channel is the research subject within the field of multi-channel in a different way than has been studied up until now, but thereby addressing its information sharing aspects in much greater detail as opposed to its role as a purchasing channel. So far, however, little is known about the content and impact of these salespeople message boards.

3 Research setting

The data for this study will be acquired from Cafepharma. Cafepharma (www.cafepharma.com) is a website for pharmaceutical and medical sales professionals. Besides different news feeds, blog aggregators and a jobs section it has a dedicated message board for sales representatives. The message board of Cafepharma is organized into different boards with most of these boards themselves again sorted by company. Cafepharma also contains some smaller boards for other people employed by pharmaceutical companies, but the bulk of the boards are for sales representatives. The message boards on Cafepharma are not moderated by any pharmaceutical company itself.

The reason why Cafepharma was selected is because it is a publicly accessible and independent company. This means that outsiders can access the message board, and it is not moderated by one of the companies that are being discussed. A second reason is the technical ease of extracting data. Salespeople are not just a speck of dust amongst the millions of others as would be the case with Twitter. This also reduces the amount of data that is needed.

4 Methodological approach to analyze content of message boards

As defined earlier, the problem definition is about how to analyze the content of topics on Cafepharma. Topic categories and sentiment have a key role in this and so it is essential to acquire this information from the topics on Cafepharma. Since Cafepharma does not offer categories or sentiment of topics we will have to "create" or generate this information. The topic itself contains the category and sentiment, since a person would be able to determine these when reading the topics. Because we aim to extract this data on a larger scale, we step into the field of *text mining* or *semantic text analysis*. The choice to apply text mining and how it will be applied will be explained and defended in this section on the basis of a roadmap (Figure 2). Each facet of this study will be explained in relation to this roadmap.

The first step, to get the categories and sentiment of a lot of topics, is a manual process that will be input for self-learning models. These models will then be able to further categorize additional data. (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006; Mitkov, 2005).

Acquiring the information we need requires a stepwise sequential procedure. This a key element of text mining, as it requires that text is prepared before it is analyzed and it increases the quality of self-learning model (Howland & Park, 2008; Mitkov, 2005).

To further understand the relations and general trends within the interactions on Cafepharma, an explorative study will be performed by looking at possible ratios between categories and sentiment and by looking at possible relations between these ratios and an external performance measure. For this, we have selected stock prices over time.

To enable the text mining application to learn how to categorize or even to load the data some pre-processing of the data is required. After pre-processing, the data can be analyzed for categories and sentiment. This will be a manual analysis. The categorized data will be learning data for self-learning models that – when applied – will be able to process more message board data and categorize it. The same process will be performed for the sentiment of topics.
Lastly, an explorative research will be done on the content of the topics to see what kind of relations are visible within a company and if any of these relations show a relation with external performance measures like the stock prices of a company.

An important last note is that the text techniques themselves are not the subject of the thesis. The subject is the application of these techniques to acquire data through which additional behavioral studies on salespeople can be conducted. Or in respect to a managerial perspective, to acquire information that aids them in the management of the sales force, new product launches or organizational change.

This approach is mainly based on the techniques and standard approaches available in text mining (Höppner, 2005; Manning & Schütze, 1999; Mitkov, 2005), combined with how the data mining application RapidMiner works and the limitations it might impose on the process(Mierswa et al., 2006).

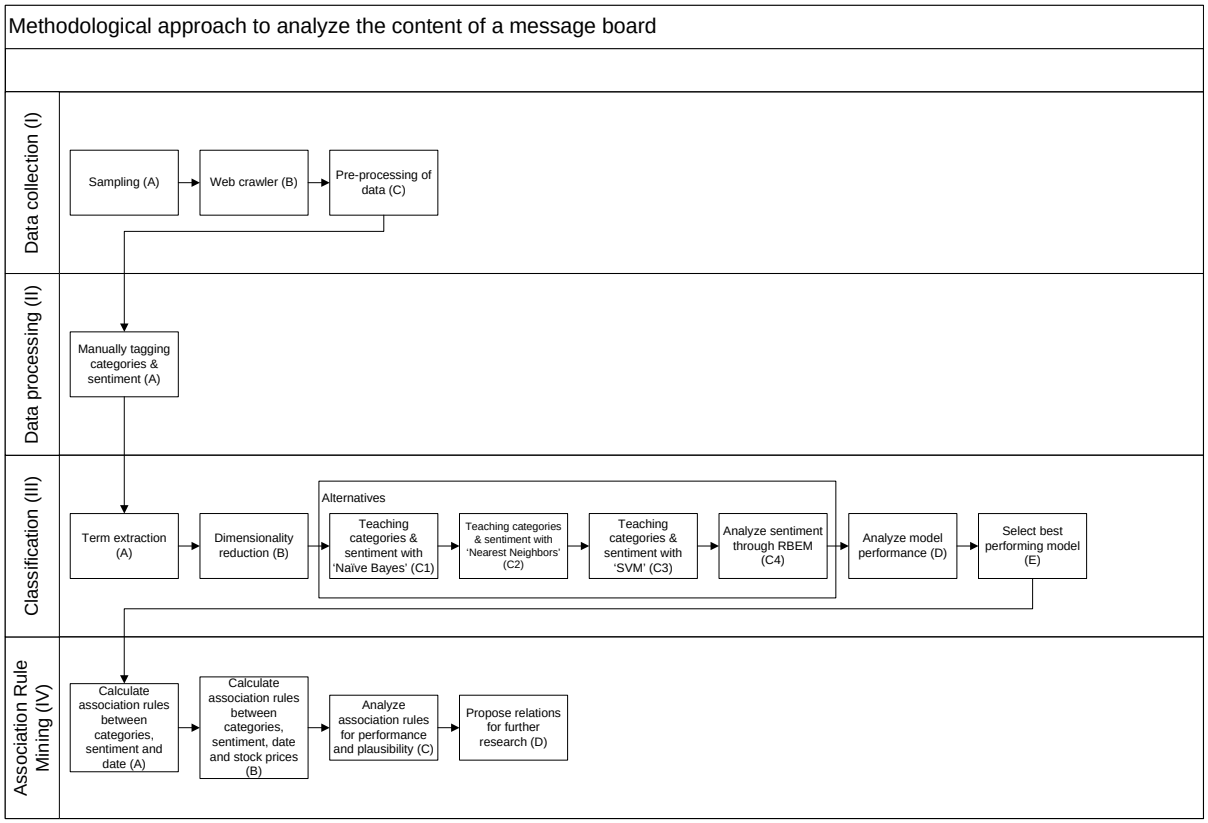

Figure 2 Research design roadmap, all steps are labeled by column and activity.

Each column of the roadmap (Figure 2) is labeled with a roman numeral from I to IV, and each activity is labeled with a letter. We will use these numbers and letters to relate to the roadmap throughout the thesis. For instance, I-A for sampling.

The roadmap is made up out of four steps (I to IV). The steps and activities of the roadmap describe the standard and accepted approach of text mining (Höppner, 2005; Manning & Schütze, 1999; Mitkov, 2005) where different text mining techniques are executed (Steps III and IV) with alternative activities for acquiring the highest quality of data (III-C1 through III-C4). Steps I and II are required to acquire data for Steps III and IV. Steps III and IV each describe a certain text mining technique that is required to gain the information and insights needed to give managers additional insights into the behavior of the sales force. The argumentation for the choice for these two techniques will be explained further in Section 4. Step III is called 'Classification' and is based text categorization (Manning & Schütze, 1999; Mitchell, 1997), within the limits of RapidMiner (Hsu, Chang, & Lin, 2003; Mierswa et al., 2006). Step IV is 'Association Rule Mining', this is where we look at the context of topics by looking for relationships between categories, sentiment and stock prices as an external measure (Höppner, 2005). Step IV is an explorative study of our data set.

As should be clearly pointed out, we do not propose a new text mining technique or approach. Text mining techniques are merely applies in order to acquire data through which additional behavioral studies on salespeople can be conducted. Or in respect to a managerial perspective, to acquire information that aids them in the management of the sales force, new product launches or organizational change.

For a better understanding of the roadmap we will argue for the applying text mining and discuss tools and application choices in Section 4.1.

4.1 Text mining, tools and applications selection

As of February 2nd 2013 Cafepharma contains over 3 million posts. If the aim was – for instance – finding topics on new products, manually reading every topic in order to find the relevant topics would require vast resources. Because we want to understand the content of a vast amount of data, manually reading every topic is thereby not an option. This is where we find an application for text mining techniques based on machine learning. Machine learning is a computational system that improves its performance on some tasks based on experience (Mitchell, 1997; Mooney, 2005). What this means is that a computational system is able to perform a certain task based on learned knowledge. Through this application, we only need to read and categorize a sample of the Cafepharma message boards. By feeding this "learned knowledge" into a machine learning model, it can perform the categorizing on new data.

In order to assure maximum practical applicability and ease of use for managers, we aim to perform the analysis on the content of topics within a single application where no knowledge of coding is required. The application that is most suitable for our analysis is RapidMiner (Mierswa et al., 2006). Two possible alternatives are Knime and Weka. However, Weka's machine learning library has been integrated into both RapidMiner and Knime. Additionally, both RapidMiner and Knime have the added benefit of having a comprehensible graphical user interface (GUI) (Berthold et al., 2007; Hall et al., 2009; Mierswa et al., 2006). The reason for choosing RapidMiner over Knime is due to a strong set of text mining techniques included with RapidMiner (Mierswa et al., 2006).

4.2 Data collection (I)

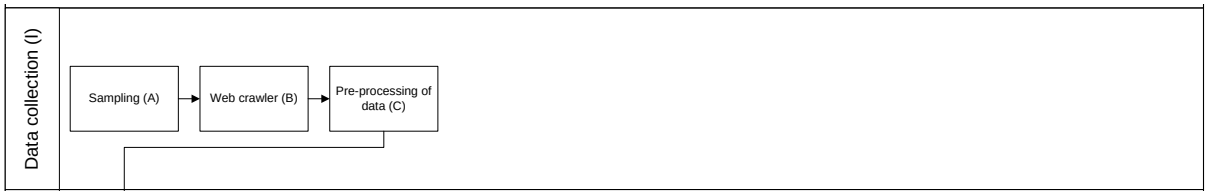

Figure 3 Roadmap Step I: Data collection

In this section sampling (I-A), web crawling (I-B) and pre-processing of data (I-C) will be discussed in general. The exact execution will be disused in Section 5.2 as especially this element of study is closely tied to the research setting: the message board Cafepharma.

Data collection (I) is made up out of the activities that acquire data and prepare the data for further analysis. These activities are sampling (I-A), web crawling (I-B) and pre-processing of the data (I-C).

Sampling (I-A) is a requirement because text categorization requires both a learning and a test set (Manning & Schütze, 1999; Mitchell, 1997). This sample will have to contain 1000 topics for text categorization models to perform their self learning tasks (Mitchell, 1997).

The data will be acquired through web crawling (I-B). This method is required because we have no direct access to the database of Cafepharma. A web crawler in general is a program that iteratively and automatically downloads web pages, follows URLs in these web pages and downloads these web pages as well (Thelwall, 2001). Because of precise structure of Cafepharma, a custom web crawler had to be built. This will be further explained in Section 5.2.1.

Because pre-processing (I-C) is a more general technique, we can explain it in greater detail. An issue with message board entries is that the natural way of typing text adds a lot of elements that can become "clutter" for a computer model trying to learn categories. These are often occurring words that are not distinguishing topics. Some of these words are absolutely vital for humans to understand text, but for a text analytic technique – that relies on pattern recognition – it deludes the elements we want the model pay attention for (Manning & Schütze, 1999).

4.3 Data processing (II)

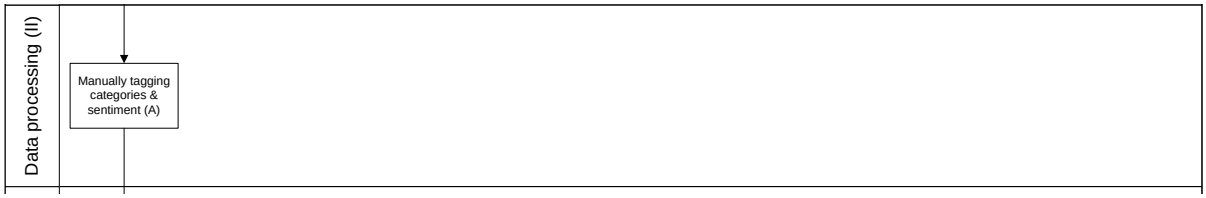

Figure 4 Roadmap Step II: Data Processing

This process is a manual one. The goal is create a learning set that will be used by the data mining application for tagging (II-A). The learning set will be created by manually tagging the sample of 1000 topics. For higher accuracy, it is better to have multiple people tag the data. However, due to restricted resource this technique is not part of the scope of this thesis. This can lead to *subjective* tagging of topics. This means that the tag does not necessarily reflect the actual content of a topic, but it possible just reflects the taggers opinion on the topic.

An issue might be inconsistency, especially when tagging sentiment due to a bias based on emotions. For instance, because over time the attitude of the tagger changes, he or she labels an almost identical topic with another label.

It is impossible to completely prevent wrong categories due to their subjective nature. However, to overcome this issue to some degree the category tagging process knows two layers: a general category and a detailed category of the topic. This way the performance of categorization models can be measured at different levels of detail to see what level of detail self-learning models can replicate within the set sample.

The categories that are being used are depicted in Table 1. The general categories are based on Wheeler and Reis (1991), as they proposed the categories *Task related* and *Social related* which have been used in further research on interactions at work (Tschan, Semmer, & Inversin, 2004). To cope with some 'noise' on the message board, the categories flame and off-topic have been added. Flame is for extreme swearing and off-topic has been reserved for topics that are not related to sales, the pharmaceutical industry in general, or outsiders posting on Cafepharma.
For the detailed categories, a more pragmatic solution was found due to the lack of relevant categories in the literature. The content of Cafepharma was therefore leading for creating the detailed categories within the general categories. Because of this process, the tagger will influence the choice of categories and their exact meaning. This influence might be larger due to the fact that only one tagger will be used for this study.

General categories	Detailed categories	Description
Off-topic	Off-topic	*Not related to the pharmaceutical industry or to the sales function in general, also includes job inquiries by externals*
Flame	Flame	*Discussions hijacked for the shear purpose of calling names and swearing ("flame war")*
Social related	Social: Performance	*Personal/sales performance (also includes bonuses)*
	Social: Organization	*Organizational changes, corporate news, corporate initiatives, layoffs and HR-related*
	Social: Person	*Discussions about certain individuals*
	Social: Personal effects	*Discussions about the job has personally on an individual (e.g. personal stories after an layoff)*
Task related	Task: Product	*Questions/information about current*

		products
	Task: New Product	*Questions/information about new products*
	Task: Sales Technique	*Discussions on sales approaches, clients, sales techniques, sales related trainings, etc.*
	Task: Market development	*Discussions about developments in the market*

Table 1 categories and sentiment

A third kind of tagging that will be done is sentiment tagging. A problem with judging the sentiment of a topic is that it might change during the conversation, and the influence the set of topics has on the sentiment of the tagger (Forgas, Bower, & Krantz, 1984; Schiffenbauer, 1974). The sentiment tags that will be used are: *Positive*, *Objective* and *Negative*. These sentiments are based on the kinds of sentiment used within a sentiment lexicon called SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010).

4.4 Classification (III)

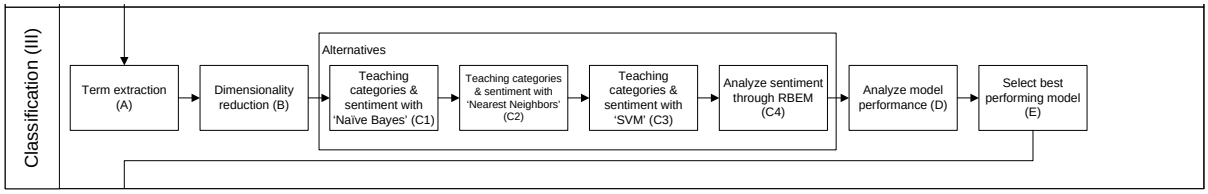

Figure 5 Roadmap Step III: Classification

Firstly an overview of all activities in step III will be given, due to their complexity we will elaborate on some of the activities for further understand on the workings of text mining

In order for text mining techniques to work with text terms have to extracted and a document space has to be build (III-A). This in itself is a text mining technique where elements of a text are separated as tokens and where for each token a column is made. The content of a text is represented by a row with values in token-columns is these tokens occur in the text (Manning & Schütze, 1999). An example is given in Table 2 and Table 3.

ID	Text
1.	We saw a bright sun in the blue sky
2.	The sky blue car was chased

Table 2 Document space example texts

ID	we	saw	a	bright	sun	in	the	blue	sky	car	was	chased
1.	1	1	1	1	1	1	1	1	1			
2.							1	1	1	1	1	1

Table 3 Document space example

To ensure that the data can be processed in an effective and efficient manner, the dimensionality has to be reduced. Through this process the number of columns in the document space will be reduced. However, this has to happen in such a way that it reflects the original input (Howland & Park, 2008) (III-B). Dimensionality reduction is further explained in Section 4.4.1.

Because the goal is to not tag the topics manually, the tagged set will be used to create a self-learning model, a more elaborate explanation of text categorization with self-learning models is given is Section 4.4.2. This is done by feeding the tagged data set to a cross-validating, self-learning model. To get the best possible result, three different cross validating techniques will be applied:

1. Naïve Bayes (III-C1)
2. Nearest Neighbors (III-C2)
3. Support Vector Machine (SVM) (III-C3)

Nearest neighbors is a key modeling technique for text categorization, it is an older technique that traditionally yields good results. It works by looking for a case that is most similar to the text X and applies the label of this "nearest neighbor" to the text X (Manning & Schütze, 1999; Mooney, 2005). Naïve Bayes is a simple to use modeling technique often used for text clustering and is considered an efficient modeling technique that can handle large data sets. It is a statistical method that takes the words around an ambiguous word in consideration to understand the sense in which this ambiguous word is used (Manning & Schütze, 1999). Finally, SVM is considered a robust and efficient modeling technique that can perform well without a lot of parameter tuning. SVM is still in development, but it often outperformance other categorization models. This techniques is based on risk minimization, the aim is to find a hypothesis for which the lowest true risk can be guaranteed. (Joachims, 1998). A full explanation of Naïve Bayes, Nearest Neighbor and SVM can be found in Appendix II.

Each of these three techniques will be both applied for labeling topics with a category, as labeling it with a sentiment tag. In essence a sentiment tag is the same as a category tag and these self-learning models are suitable to perform a sentiment analysis (Manning & Schütze, 1999). Over the years many dedicated sentiment analyses have been developed (Tromp, 2011). Beside a sentiment analysis in RapidMiner with self-learning models we will also apply RBEM (III-C4), a dedicated social media sentiment analysis that has yielded great results on social media like Facebook and Twitter (Tromp, 2011). The reason why we do not limit us to only performing the sentiment analysis with a dedicated solution is because the aim of this thesis is to offer relevant sales force data for managers in an easy practical process that they can execute themselves. The expectations are that a dedicated sentiment analysis solution yields better results. By comparing III-C1 through III-C3 with III-C4 (RBEM) we can review the performance of RapidMiner for sentiment analysis.

All models will be reviewed based on their performance (how well they can replicate the learning set), and the best will be selected and applied to the complete set (III-D and III-E). However, a bad model fit is also possible. A bad model fit could be caused by an inadequate learning set. This problem can be fixed in two ways which is either by creating a bigger learning set or by letting more people tag the same set of topics – this would create a weighted average in the categories.

4.4.1 Dimensionality reduction (III-B)

An initial step that increases the efficiency and performance of a text categorization model is dimensionality reduction (Howland & Park, 2008). With dimensionality reduction, a set of attributes that exist in a high dimensional space are represented in a low dimensional space, or dimensionality reduction represents an n-dimensional space onto a k-dimensional space where $n \gg k$ (Manning & Schütze, 1999). For our purpose, we will rely on singular value decomposition (SVD) as this technique is often applied in the field of text mining with good results (Howland & Park, 2008; Manning & Schütze, 1999).

An initial step for reducing the dimensions is done by stemming, removing stop words, creating n-grams and omitting the most and least occurring words through pruning. With stemming words with the same origin are transformed into a single token, for instance ‘conclusion’ and ‘concluding’ become ‘conclu*’. By removing stop words, standard English stop words are removed. N-grams are pairs of words that often occur together, an example of this could be ‘region manager’. Lastly pruning ensures that words that are too unique for a text or too common within the data set are omitted.

Within the application of dimensionality reduction SVD maps co-occurring terms onto the same dimension, thereby increasing the similarity between similar documents. For text mining, SVD is applied to document-by-term matrices by a technique called latent semantic indexing (LSI). Document-by-term matrices are matrices such as TF-IDF matrices. TF-IDF or turn ‘frequency, inverse document frequency’ are matrices where the value of words in a document is calculated through the inverse proportion of the frequency of words in a document to the overall percentage of documents the word is found in (Ramos, 2003). SVD represents such a matrix A as $\hat{A}$ in a lower dimensional space ensuring that the distance between these two matrices is minimized: $\Delta = \left\| A - \hat{A} \right\|_2$ (Manning & Schütze, 1999). This distance between matrices is measured by the 2-norm, which is equivalent to the Euclidean distance for vectors.

The representation is calculated by decomposing the original matrix $A_{t \times d}$ into the product of $T_{t \times n}$, $S_{n \times n}$, $D_{d \times n}$: $A_{t \times d} = T_{t \times n} \times S_{n \times n} (D_{d \times n})^T$ (Manning & Schütze, 1999). Where t are the terms, d the documents and n is defined as$n = \min(t, d)$. The first position of the subscript are the rows, with the second position – after the × – representing the columns. D^T is the transpose of the D, where the matrix is rotated around the diagonal as defined like $D_{ij} = (D^T)_{ji}$.

4.4.2 Text categorization (III-C1, III-C2 & III-C3)

The application of text mining that is relevant for our problem is text categorization. Categorization is “the task of assigning objects from a universe to two or more categories” (Manning & Schütze, 1999). An important aspect of text categorization is machine learning – as noted in the previous section. Machine learning in text categorization relies on a training set. A training set contains cases – in our case topics – which are labeled with one or more classes (Manning & Schütze,

1999) – categories and sentiment for this specific study. In this training set, each topic is represented as a vector of a word count. With this training set a classifier is trained. This is done with a training procedure or modeling technique like naïve Bayes or nearest neighbor. To test the performance of the trained classifier it has to be tested on a test set. A test set is similar to the training set but contains cases that are not included in the training set to see how the classifier performs when categorizing new data. Results can be presented in a contingency table as shown below in [X].

	CAR is correct	BIKE is correct
CAR was predicted	a	b
BIKE was predicted	c	d

Table 4 Contingency table for evaluating a classifier. In this example the classifier was trained to categorize document as either CAR or BIKE, *a*, *b*, *c* and *d* are values representing the number of documents that were assigned to a certain category and to what category they are actually belonging.

An important measure is the accuracy, or the proportion of correctly categorized cases. For a non-binary classification this is done by making a 2×2 contingency matrix for each category where the accuracy is computed according to the definition $\frac{a+b}{a+b+c+d}$ after which an average is computed over the categories to calculate an overall accuracy (Manning & Schütze, 1999). The accuracy gives important insights into the performance of a trained classifier, by showing how well the classifier performs compared to the actual categories in the test set.

A trained classifier can be applied to categorize new data. This illustrates a key benefit of machine learning; the classifier is able to process large quantities of data based on earlier experience. This is something humans would not be able to do at the same speed. An important note to keep in mind is that the quality of the classifier and its accuracy are highly depended on the quality of training and test set. Depending on the source of the data, this method can thus still be dependent on human qualities like initial categorization.

4.5 Association rule mining (IV)

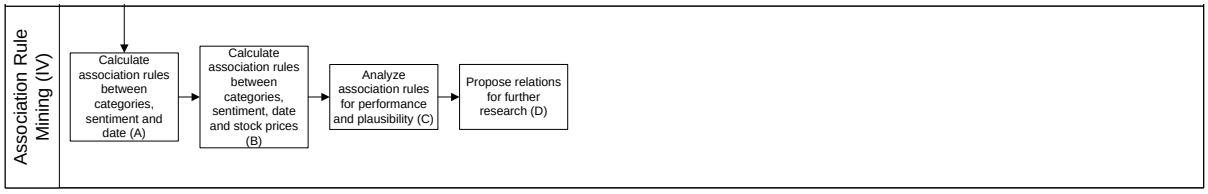

Table 5 Roadmap IV: Association Rule Mining

The analysis of topics will be an explorative process through the use of an unsupervised data mining technique called association rule learning. This technique was originally used for market basket analysis, where it was applied to find what products a customer bought in combination with other products. The result of this analysis is an association rule in the form of "70% of the customers who buy vine and cheese also buy grapes" or $\{vine, cheese\} \rightarrow \{grapes\}$ with a confidence of 0,70 (Höppner, 2005). The unsupervised nature of this technique makes it an applicable technique for this explorative study.

Association rule learning employs a scheme with binary attributes (Höppner, 2005). Therefore, a category like color would be split in such a way for each color an attribute is created that can be either 0 or 1. An association rule could therefore be formulated like $\{A_2, A_{33}\} \rightarrow \{A_5, A_{12}, A_{54}\}$, this means that when a case has the attributes A_2 andA_{33}, it will also have the attributes A_5, A_{12} and A_{54}.

The evaluation of association rules will be done based on three criteria which are the support, confidence and lift. Support is defined as $supp(X \rightarrow Y) = supp(X \cup Y)$ (Höppner, 2005). With support we can evaluate for how much of the total cases a rule is true, or in short if the rule occurs often or not. To get better insight into the quality of a rule, the confidence is a better measurement. Confidence is defined as $conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$ (Höppner, 2005). With this measurement we can evaluate in how many cases that contain X it also holds that the case contains Y.

When dealing with a lot of attributes the lift measurement is good measurement to identify rules that are strong compared to other rules. Lift is defined as $lift(X \rightarrow Y) = supp(X \cup Y)/(supp(X) \times supp(Y))$ (Höppner, 2005). In short, lift can be explained as the number of times a case that contains X is more likely to contain Y compared to all other cases.

In the context of this study, association rule learning will be applied to discover relations between topic elements (IV-A) – like category, sentiment and date – and an external measurement: stock prices (IV-B).

For the first application of association rule learning, no additional data is required as the data set will already contain all required data. Some recoding of the data is required to filter out the dates which are stored in the file names.

The second application requires an external data source. Historical stock prices are widely and freely available. Because of its practical CSV export, *Yahoo! Finance* has been selected as the data source for these stock prices. Individual opening and closing stock prices vary a lot. They will, therefore, have to be recoded in order to see the more general trends of closing higher, lower or equal to the opening of the stock markets.

These association rules will be analyzed based on their support, confidence and lift and their general plausibility (IV-C). With these association rules we hope to find possible relations that can be further studied in future research (IV-D).

5 Execution of the Roadmap

In this section the translation is explained from roadmap to the tools and applications that were used. Therefore step II (Data processing) is not further discussed in this section, as it was a manual step that was explained in Section 4.3 and required no further translation in order to execute it.

5.1 Sampling (I-A)

Figure 6 Roadmap Step I: Data collection

First a sample size and sample origin was determined. This was a key decision because sampling is strongly connected to the research design and activities and influences its outcomes.

Sampling the data had two reasons. Firstly it would create redundant work to automatically categorizing data if all the data has to be manually tagged. Successfully categorizing a model requires a learning set, and thereby not all the data Cafepharma contains. Through a learning set a self-learning model is able to learn to replicate the tagging of topics. These techniques require a rather large sample set. For this research a sample set of 1000 was selected, since a sample of around a 1000 has been tested and shown to work with different algorithms (Mitchell, 1997).
Secondly, due to system requirements when dealing with large textual datasets it would take too much time to process all the raw data, so the technique needs to be tried with a sample. A sample of 1000 topics is however possible with proper dimensionality reduction.

The question arose from what population of data the 1000 topics would have to be sampled. Crawling all the company boards available at Cafepharma.com would create a complete overview, but it would require a lot of time and generate too much data, which might be hard to handle for a desktop computer. Additionally, with a more compact set of data, quick iterations were easier to execute. Therefore, a selection had to be made of a couple of companies. The most important selection criterion was that the company should be publicly listed; this way stock price information could be collected that was required for association rule mining. To get a balanced set of data companies, varying amounts of posts had been selected.
The following companies were selected:

- Pfizer (USA)
    - 17.500+ topics
    - 191.400+ posts
- Merck & Co., (USA) known as Merck Sharp & Dohme (MSD) outside of USA and Canada
    - 7.000+ topics
    - 92.000+ posts
- AstraZeneca (UK/Sweden)
    - 6.500+ topics
    - 80.000+ posts
- Johnson & Johnson (USA)
    - 1.900+ topics
    - 18.000+ posts
- Daiichi Sankyo (Japan)

- 2.500+ topics
- 19.000+ posts

Pfizer was selected due to the large amount of topics the Pfizer company boards contain, due to the fact that it is the biggest company board on Cafepharma.com and the biggest pharmaceutical company in the world[2]. The others are boards with varying numbers of topics that are large and varying enough to prevent overrepresentation of a single company board. These company boards were still some of the biggest company boards with the exception of Johnson & Johnson. Together this list of companies gave a decent cross section of the message board.

The proportions amongst the different company boards have been maintained when drawing the sample of 1000 topics. This gave better insights in model performance, as the learning set would contain fewer topics from the smaller company boards. If a model was still able to correctly categorize these topics, the model had proven to a certain degree that it was more widely applicable.

5.2 Data collection (I-B & I-C)

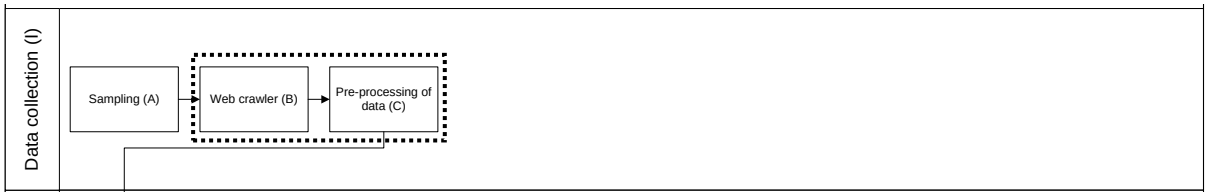

Figure 7 Roadmap Step I: Data collection

This step was intertwined with research objective and object (the data source). For "simple" crawls – e.g. every subject or topic has just one child page – most built in web crawl plugins of data mine applications would suffice. However, for more flexibility a custom crawler was more adequate. This way less redundant information would be crawled, and the data could be better organized immediately while crawling. Another advantage is that the data mining tool was relieved of web crawling, and this way text mining could continue while crawling. For this crawl the second option was chosen and a custom Java web crawler was built. All this data has been saved to separate files per topic. Topics are discussions or conversations that have a subject. A topic is built up out of several posts, which are contributions by the message boards' members.

For this study, topics were mined from Cafepharma's so-called company boards[3], as these were easiest to relate to a certain organization. Cafepharma also has more specific boards aimed at certain product groups. However, using these boards– for instance – could have led to a categorizing model only working for dental products. Using the product related boards could have meant that the topics contain more product and task related topics, where the company boards might mostly be aimed at organizational related topics. However, for the sake of setting a scope the company boards have been selected to be the research object of this study.

5.2.1 Web crawler (I-B)

As stated in Section 5.2 a web custom crawler was made for collecting the data from the Cafepharma company boards. A web crawler is a program that iteratively and automatically downloads web pages, follows URLs in these web pages and downloads these web pages as well (Thelwall, 2001). The web crawling function within RapidMiner did not suffice for this web crawl, since we wanted absolute control over what data was being crawled and in what format it was stored.

The web crawler followed the hierarchy of the message board. After manually inputting a company board, the web crawler would index each post per topic. Because our goal is to understand the content of each topic, the data was stored per topic. The last actions of the web crawler were to apply the sampling of a 1000 cases. This was done as an extension of the web crawler to maintain the correct hierarchy and proportions of the web crawl.

[2] According to Forbes in 2012

[3] http://www.Cafepharma.com/boards/forumdisplay.php?f=4

5.2.2 Data clean up (I-C)

Remnants of HTML code were often still found in posts, and this HTML code had to be removed. This was done to prevent possible errors when analyzing the data and to make sure that the data had less “foreign objects” in it. The HTML code was a “foreign object” since the author of the post did not put this code in as part of the message he or she was trying to convey.

Even though a custom crawler ensured a better quality of raw data, the build-up of the message board still affected the actual organization of posts. Due to the construction of the message board, data was collected at the post level. To ensure the data was usable for this study it needed to be clustered to topic level.

Both steps could have been done in a data mining application. However, for the purpose of time saving and the amount of data, the data clean-up process was appended to the process of crawling and a custom Java solution was be build for this as well. The output consisted of separate text files for each topic nested within company folders.

5.3 Classification model design (III)

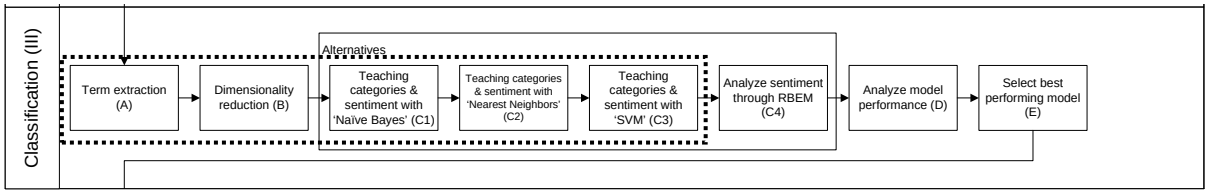

Figure 8 Roadmap Step III: Classification

In this section, we will explain the design of the three self-learning text categorization models based on naïve Bayes, nearest neighbor and SVM modeling techniques. The full RapidMiner model can be found in Appendix III. The RBEM sentiment analysis was executed by the company currently in charge of further developing RBEM into a commercial product and is therefore not further explained in this section.

5.3.1 Dimensionality reduction (III-B)

The resultant text categorization might be more efficient with dimensionality (Howland & Park, 2008), its execution within RapidMiner however is not. Due to the heavy computational demands of the SVD operator, the following process was only performed on the best performing models as selected in Section 6. The model is mostly identical to the other models except for a SVD operator nested within the *optimize parameter* operator, but not nested within the *cross validation* operator (see Section 5.3.2, 5.3.3 and 5.3.4). This is depicted in Figure 9. For the SVD operator the number of dimensions was optimized.

Executing dimensionality reduction exclusively on the best performing model was a deviation from the roadmap.

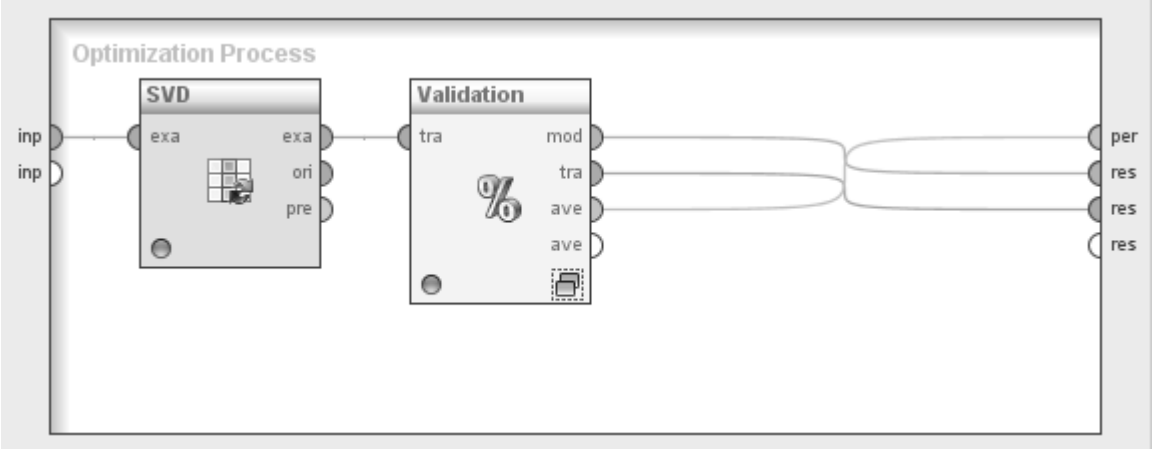

Figure 9 SVD dimensionality reduction

5.3.2 Naïve Bayes (III-C1)

For the self learning model based on a naïve bayes cross validation, the following process was built:

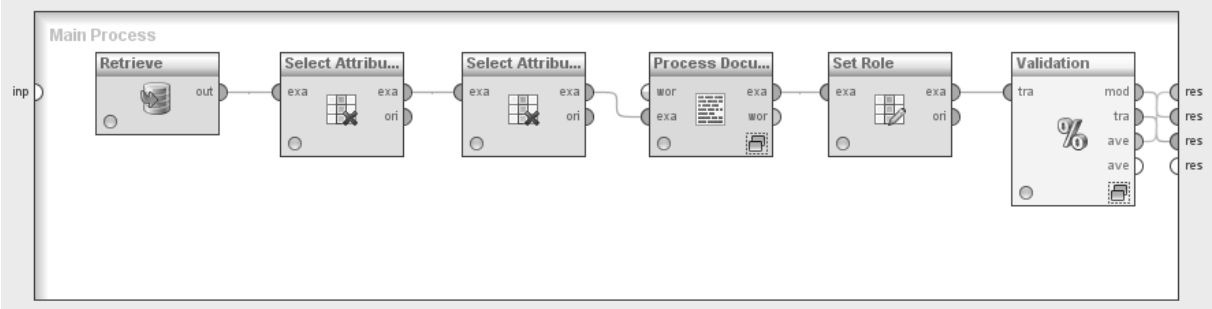

Figure 10 Nearest Neighbor self-learning model

This process reads the data, removes any column with missing values and selects the columns that will be needed for the learning model. This will always be the ID and the text, and alternately the general category, the detailed category and the sentiment. The *set role* operator will ensure that RapidMiner knows that is has to learn how to categorize or determine sentiment, by defining these as the *label*.

For better processing, the topic text requires some clean up. This is achieved by running the data through a *process document from data* operator, this operator execute the functions of step term extraction (III-A). This process can hold different operators within itself (a nested operator) for further data optimization. In this particular case that was achieved through the operators that transform everything to lowercase, replace some words (like *I am* for *I'm*), tokenize the entire text, filter out common English stop words, stem tokens, filter out one letter tokens and produce n-grams.

Additionally, the *process documents from data* operator also creates an TF-IDF table and prunes it so that it leaves out any words that occur in less than 1% and in more than 90% of the documents. This last step is part of the dimensionality reduction (III-B), but in RapidMiner it is build in into this pre-processing operator. The entire process looks like this:

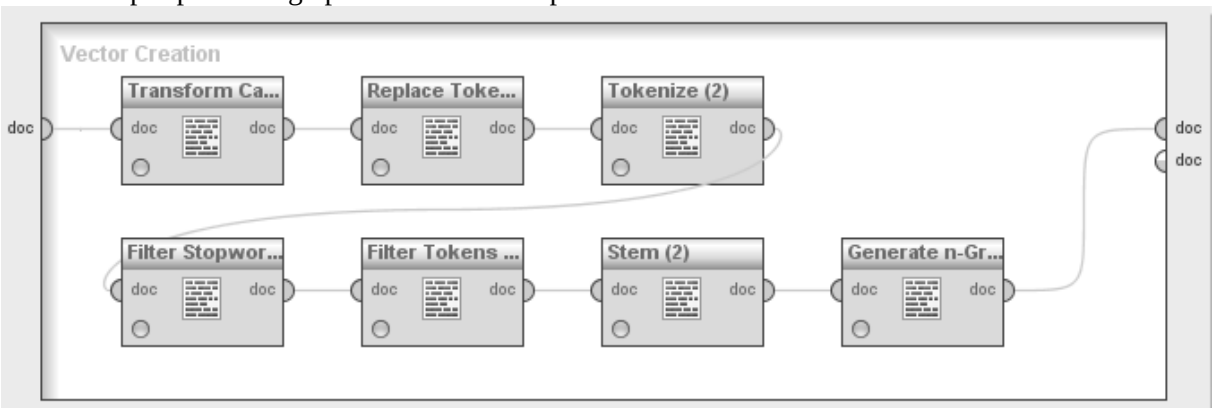

Figure 11 Process documents from data (III-A)

There is no additional need for an optimized operator, because this implementation of the naïve bayes model does not have any parameters that can be optimized.

The learning itself happens within the *cross validation* operator. The cross validation splits the data set in two parts with one part being the learning set and the other being the test set. The training set is used by the modeling operator to learn how to categorize and the test set is used to apply this learned model on. Of course, the important aspect here is that the cross validation will test the performance of the model and rerun the learning 10 times. To evaluate the performance of the learned model, the *performance* operator will summarize the model performance.

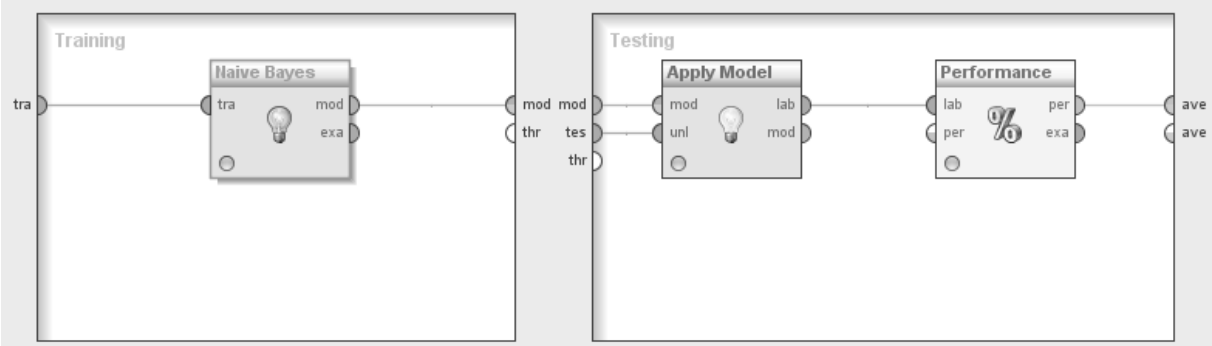

Figure 12 Cross validation

5.3.3 K Nearest Neighbor (III-C2)

The basis of this model is the same as the naïve Bayes model Figure 10, with the exception that within the cross validation the training operator is a K nearest neighbor (k-NN) operator instead of a naïve Bayes operator. This model also has an *optimized parameter* operator, since the k-NN has a parameter k on which the performance of the model depends. The *optimized parameter* operator tests different values for k until it finds the best performing model. This operator is a nested operator like the *cross validation* operator. The *cross validation* operator is nested within the *optimize parameter* operator as the *cross validation* operator contains the k-NN operator that has to be optimized in the context of an optimal model performance, hence the entire *cross validation* operator has to be nested within the *optimize parameter* operator.
The inside of the *cross validation* operator is almost identical to the one for the naïve Bayes model Figure 12, with the exception that the naïve Bayes operator is replaced by the k-NN operator.

5.3.4 Support Vector Machine (SVM) (III-C3)

This mode is very similar to the k-NN model and also has an *optimize parameter* operator. The main difference is that the SVM operator has two important parameters that need to optimized, C and gamma. This does not change much to the design, but it does require more time and computation power when executing.

5.4 Association Rule Mining (IV)

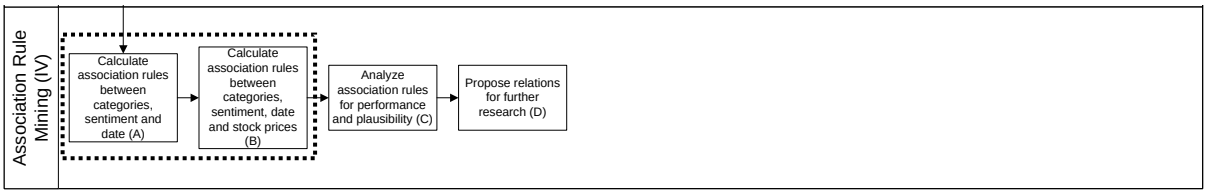

Table 6 Roadmap IV: Association Rule Mining

Due to the limitation imposed by RapidMiner (Mierswa et al., 2006) association rule mining requires a lot of operators that ensure that the data is of the right type and to select only the relevant data. Since the association rule mining operators require binominal data this imposed varying challenges depending on the data. Since these steps were highly depending on the data sources and encoding they will not be further explained in this section. Detailed information can be found in Appendix III.

The *FP-growth* operator calculates frequent item sets, the *Create Association Rule* operator creates association rules form the frequent item sets (Figure 13).

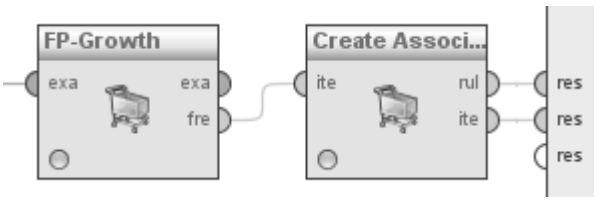

Figure 13 Association rule mining

6 Research results

The results are reported in two sections. Section 6.1 focuses on model performance for each model and in Section 6.1.4 we will select the best performing model. In Section **Fout! Verwijzingsbron niet gevonden.** we will discuss the insights we have gained through the explorative research of the topic content.

The tables with the model performance can be found in Appendix IV.

6.1 Classification model performance

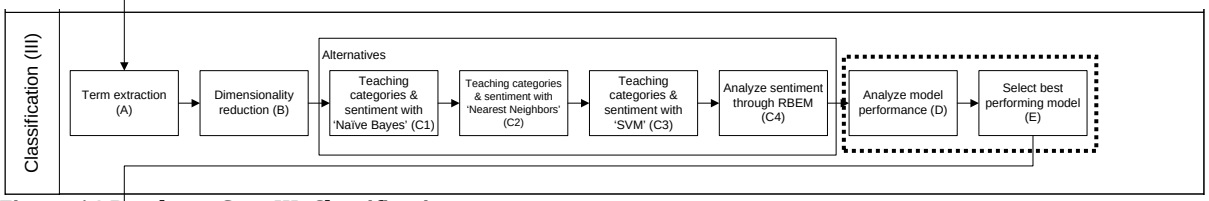

Figure 14 Roadmap Step III: Classification

The performance will be discussed per modeling technique. For each modeling technique, the model accuracy is presented for the category, the second level category and for the sentiment. This is done with both an unbalanced data set (all the data) and a balanced data set. The balanced data set excludes the over representation of certain categories or sentiments. Through this over representation the accuracy of a model might seem to perform acceptable, closer investigation can lead to the discovery that the model is only able to predict a limited amount of categories or sentiments.

Lastly ,we will also look at the performance of the alternative sentiment analysis proposed by Tromp (2011) and how it compares to the other sentiment analysis.

All results can be found in larger print in Appendix IV.

6.1.1 Naïve Bayes

The overall accuracy is higher for the model with the unbalanced data as opposed to the balanced data. With the first level category, a further inspection leads us to the conclusion that this higher performance is due to a higher accuracy for *Social related*. Since this data set contains more cases with the category *Social related*, the model will have a high accuracy score, even if it would label all cases as *Social related*. With the balanced data set a decrease in class precision for *Social related* and *Off topic* is observed. For *Task related* the contrary is observed. Due to the over representation of *Social related* and *Off topic*, the first model's accuracy does not give an actual better performing model. The same effects occur as well for both the second level categories and the sentiment. The models on the balanced data set are better models, because these models provide a higher accuracy for the underrepresented categories or sentiments.

Accuracy: 55,91% ± 4,08% (mikro: 55,89%)					
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	5	2	10	0,00%
pred. Social related	15	461	115	166	60,90%
pred. Task related	0	42	30	32	28,85%
pred. Off topic	1	36	10	59	55,66%
class recall	0,00%	84,74%	19,11%	22,10%	

Table 7 Naïve Bayes for first level category (unbalanced)

Accuracy: 42,76% ± 5,19% (mikro: 42,76%)					
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	4	1	9	0,00%
pred. Social related	10	115	57	92	41,97%
pred. Task related	4	55	90	59	43,27%
pred. Off topic	2	9	26	40	51,95%
class recall	0,00%	57,50%	57,32%	20,00%	

Table 8 Naïve Bayes for first level category (balanced)

Because we established that the balanced models give a more accurate representation, we will discuss this model in greater detail, and we will discuss each level of categorization and sentiment separately. With the first level category (Table 8) both the class recall and class precision of *Flame* is 0,00%. This is actually observed with all models. We can reason that this is due to the fact that these topics are – or were – normal topics on one of the other categories that contain a lot of swear words or (personal) insults. Maybe this is quite evident for the human tagger, but for the model it is just a nuance because the model might see more patterns that match with other categories. This action is quite logical since flame topics can sprout from any conversation whether it was originally *Social related* or *Task related*. The precisions of the naïve Bayes model for *Social related* and *Task related* are 41,97% and 43,27% respectively, which means that in neither of these cases less than half of the cases actually belonged to these categories. In the end, the mistakes made other categories lead to a class recall of almost 60% for both *Social related* and *Task related*, but the low precision makes it so that this models performance is too low to be usable for us.

An interesting phenomenon can be observed with *Off topic*. Even in the balanced data set, this was a small category. This was done to ensure that the data set did not get too small, which would have decreased overall accuracy even more (Mitchell, 1997). Compared to the other categories the precision is quite high with 51,95%, which means that about half of the cases that were labeled as being *Off topic* actually belonged to this category. When we look at the class recall, it becomes clear that most of the cases that should have been labeled as *Off topic* were actually labeled as something else. In short, if the model decided to label a case as *Off topics* it was right in about half of the cases, but the model as a whole was unable to ensure more than 20% of the cases that were actually *Off topic* got labeled as such.

Accuracy: 22,91% ± 5,55% (mikro: 22,91%)											
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	1	0	0	0	2	2	0	0	0	0,00%
pred. Social: Person	0	8	2	0	0	4	2	2	1	1	40,00%
pred. Task: product	0	5	11	1	6	2	7	1	1	5	28,21%
pred. Social: Performance	1	4	3	2	1	1	3	1	4	0	10,00%
pred. Task: New product	0	1	6	0	0	4	2	0	0	1	0,00%
pred. Social: Organization	3	9	3	3	2	12	8	3	8	2	22,64%
pred. Off topic	2	5	3	0	1	1	9	2	1	1	36,00%
pred. Social: Personal effects	10	30	15	17	7	34	28	54	23	13	23,38%
pred. Task: Sales Technique	0	10	4	6	1	10	10	10	11	3	16,92%
pred. Task: Market developments	0	2	5	0	3	5	4	2	1	8	26,67%
class recall	0,00 %	10,67 %	21,15 %	6,90%	0,00%	16,00%	12,00 %	71,00 %	22,00%	23,53%	

Table 9 Naïve Bayes for second level category (balanced)

The discussion of the second level category will again be limited to only the balanced data set (Table 9) since this a more accurate representation and this model has a more even spread performance instead of just being able to predict a very limited number of categories. Even though, in that respect, the precision was poor for any of the categories in the unbalanced data set. In class recall, we observe a slight shift in performance from *Social: Organization* to *Social: Personal effects*.

With an accuracy of only 22,91% and no class precisions exceeding 40%, the model performance is too low to be considered useful.

Accuracy: 51,03% ± 4,09% (mikro: 51,04%)				
	true Negative	true Objective	true Positive	class precision
pred. Negative	275	187	74	51,31%
pred. Objective	20	64	8	69,67%
pred. Positive	5	35	4	9,09%
class recall	91,67%	22,38%	4,65%	

Table 10 Naïve Bayes for sentiment (balanced)

As with the other two discussions of the result, only the balanced data set will be discussed (Table 10). For the sentiment this model performance is better than for the two levels of categories. With an accuracy of 51,03% the sentiment is the only level that exceeds 50%. This indicates a decent level of performance. The precision for both *Negative* and *Objective* is decent with 51,31% and 69,67% respectively. The class recall however is skewed strongly in favor of *Negative*. Closer inspection explains us that this is due to large amount of cases that are assigned to the *Negative* sentiment. This means the relative large accuracy is due to model being able to predict the larger set of cases that are *Negative*. The model is therefore not very fit to label sentiment.

6.1.2 K Nearest Neighbor

The overall accuracy is higher for the model with the unbalanced data as opposed to the balanced data. With the first level category, a further inspection leads us to the conclusion that this higher performance is due to a higher accuracy for *Social related*. Since this data set contains more cases with the category *Social related,* the model will have a high accuracy score even if it would label all cases as *Social related*. The class precision is increased for each of the first level categories with the balanced data set. For the second level categories and the sentiment, the class precision shifted only slightly between different categories and sentiments. For all category levels and sentiment, the class recall is more spread out with the balanced data set, thereby giving a better indication whether the model is actually performing well. For the first level categories and the sentiment, the class recall gives much better results when using the balanced data set. The same increase in class recall is not observed with the second level categories. This model is only able to get decent class recalls for *Social: Personal effects* and *Task: Sales techniques*, however the class precision is still low with values around 30%.

Accuracy: 64,63% ± 3,58% (mikro: 64,63%)			**K = 15**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	12	508	82	190	64,14%
pred. Task related	2	15	63	12	68,48%
pred. Off topic	2	21	12	65	65,00%
class recall	0,00%	93,38%	40,13%	24,24%	

Table 11 k-NN for first level category, k=13 (unbalanced)

Accuracy: 57,41% ± 6,22% (mikro: 57,42%)			**K = 18**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	6	134	27	87	52,76%
pred. Task related	8	42	118	36	57,84%
pred. Off topic	2	24	12	77	66,96%
class recall	0,00%	67,00%	75,16%	38,50%	

Table 12 k-NN for first level category, k=18 (balanced)

As in the previous section, we will only discuss the model performance of the balanced data set in detail. The first one is the model performance of the first level category (Table 12).

Flame does not get predicted correctly at all, as was the case with the naïve Bayes model (Table 8, Section 6.1.1). On other aspects the K nearest neighbor model performs better. The accuracy is 57,41% compared to 42,76%. The categories other than *Flame* all have a precision larger than 50%. Even though the class recall for *Off topic* is still rather low with 38,50% the class recalls for *Social related* and *Task related* are good with percentages around 70%.

The performance of the k nearest neighbor model for the first level category makes it a good candidate for future categorization.

Accuracy: 36,67% ± 6,55% (mikro: 36,65%)							K = 18				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Techniqu e	true Task: Market developments	class precision
pred. Flame	0	0	1	0	1	0	1	0	0	0	0,00%
pred. Social: Person	1	15	1	1	0	1	2	2	0	0	65,22%
pred. Task: product	3	5	23	1	7	2	5	1	3	10	38,33%
pred. Social: Performance	2	4	0	3	0	5	1	3	2	1	14,29%
pred. Task: New product	1	1	4	0	6	0	0	0	0	1	46,15%
pred. Social: Organization	0	9	6	2	1	18	7	7	5	1	32,14%
pred. Off topic	3	9	3	2	0	6	22	2	2	1	44,00%
pred. Social: Personal effects	4	21	5	15	2	29	23	53	7	5	32,32%
pred. Task: Sales Technique	2	10	4	5	1	12	10	6	31	2	37,35%
pred. Task: Market developments	0	1	5	0	3	2	4	1	0	13	44,83%
class recall	0,00 %	20,00 %	44,23 %	10,34%	28,57 %	24,00%	29,33 %	70,67 %	62,00%	38,24%	

Table 13 k-NN for second level category, k=18 (balanced)

The performance of the second level category K nearest neighbor model (Table 13) is better than the performance of the same category level with a naïve Bayes model (Table 9, Section 6.1.1). The accuracy for the K nearest neighbor is 36,67% compared to 22,91% for the naïve Bayes model. Precision is better than the naïve Bayes model. Where that model had a maximum precision of 40% for a single category (*Social: Person*), this model has +40% for 4 categories (*Social: Person*, *Task: New product*, *Off topic* and *Task: Market development*), where *Social: Person* has a precision of 65,22%.

With the K nearest neighbor model two categories have a class recall larger than 50% (*Social: Personal effects* and *Task: Sales Techniques*), where the naïve Bayes model had this for one category.

In general, even the K nearest neighbor model – which performed well for the first level category – disappoints for the second level category. No strong precision and class recall combination can be observed in the performance table (Table 13), where both are at least 50%. This fact makes this model useless for our goal of being able to categorize topics from Cafepharma.

Accuracy: 54,48% ± 5,74% (mikro: 54,46%)			K = 6	
	true Negative	true Objective	true Positive	class precision
pred. Negative	226	139	55	53,81%
pred. Objective	70	134	25	58,52%
pred. Positive	4	13	6	26,09%
class recall	75,33%	56,85%	6,98%	

Table 14 k-NN for sentiment, k=6 (balanced)

The K nearest neighbor model has a weak performance in both precision and recall when it comes to the *Positive* sentiment (Table 14). In this sense, it does not perform significantly better than the naïve Bayes model since neither of them have a precision or recall of at least 50%, when looking at the models based on the balanced data set.

Even though the accuracy compared to the naïve Bayes model has only increased by 3,45% to 54,48% we observe a more even performance between *Negative* and *Objective*. The precision of *Objective* lower with 58,52% compared to 69,67%, but the class recall has improved greatly from just 22,38% to 56,85%. Compared to the naïve Bayes model the class recall for *Negative* has dropped from just over 90% to 75,33%, but the precision increased by 2,5% to 53,81%. Even without a great performance increase for *Positive*, this model still performance better than the naïve Bayes model due to its greater accuracy for *Objective*.

6.1.3 SVM

Unlike the other models, the performance of model does increase through the usage of a balanced data set. Of course with the other models the accuracy would often decrease, but when looking into the class recall and class precision it can be observed that the models are performing better for the less represented categories and sentiments. The same observation is not possible with the SVM. For the first level categories, the model with unbalanced data could only predict Social *related*. The model with balanced data has major decrease in class recall for *Social related*, this comes with an increase in class recall for *Off-topic*. But since neither of them have a class precision of at least 50%, this model will not lead to successful labeling. For the second level category, the performance of the SVM was already disappointing; using a balanced data did not improve this outcome. The SVM model was not performing well at all with the unbalanced data set. The accuracy of 62,60% was almost completely due to the unbalanced nature of the data set. The balanced data set greatly improved this, even though it was still not able to accurately predict any *Positive* cases.

Accuracy: 55,59% ± 0,50% (mikro: 55,59%)			γ = 6,0 & C = 3,0		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	16	543	155	265	55,46%
pred. Task related	0	0	2	0	100,00%
pred. Off topic	0	1	0	2	66,67%
class recall	0,00%	99,82%	1,27%	0,75%	

Table 15 SVM for first level category, γ=6,0 & C=3,0 (unbalanced)

Accuracy: 45,38% ± 3,19% (mikro: 45,38%)			γ = 3,0 & C = 9,0		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	5	110	77	54	44,72%
pred. Task related	0	1	6	2	66,67%
pred. Off topic	11	89	74	144	45,28%
class recall	0,00%	55,00%	3,82%	72,00%	

Table 16 SVM for first level category, γ=3,0 & C=9,0 (balanced)

As for both previous models, we will limit the detailed discussion of the performance results to the balanced data set.

Table 16 contains the performance results of the SVM model for the first level category. As with the previous two models, *Flame* has a precision and a class recall of 0,00%. The SVM model does not perform very well for the other categories unlike the K nearest neighbor model. None of the

categories have a combination where both precision and class recall exceed 50%. This fact alone excludes this model from being applied for further use.

Accuracy: 29,29% ± 4,57% (mikro: 29,28%)							**γ = 3,0 & C = 12,0**				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Person	7	36	13	5	3	13	17	14	6	4	30,51%
pred. Task: product	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Performance	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: New product	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Organization	2	18	17	8	9	30	16	10	18	14	21,13%
pred. Off topic	4	11	16	3	7	21	33	5	13	10	26,83%
pred. Social: Personal effects	3	10	6	13	2	11	9	46	13	4	39,32%
pred. Task: Sales Technique	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: Market developments	0	0	0	0	0	0	0	0	0	2	100,00 %
class recall	0,00 %	48,00 %	0,00%	0,00%	0,00%	40,00%	44,00 %	61,33 %	0,00%	5,88%	

Table 17 SVM for second level category, γ=3,0 & C=12,0 (balanced)

The SVM model for the second level category (Table 17) has a slightly higher accuracy (29,29%) compared to the worst performing model, the naïve Bayes model (22,91). But it is not able to outperform the K nearest neighbor model that has an accuracy of 36,67%. Closer inspection of the precision and class recall leads us to the conclusion that this model is not useful for our application. Expect for the outlier *Task: Market development* none of the other categories have a precision close to 50%. *Task: Market development* has a class recall of just 5,88%, therefore the high precision is meaningless. As with the precision the class recall has many categories performing at 0,00% and besides *Social: Personal effects* none of the categories have a class recall of at least 50%. This model cannot meet the class recall of *Social: Personal effects* (61,33%) with a precision that comes close to that, leading us to earlier stated conclusion that this model is not useful for us.

Accuracy: 61,91% ± 5,76% (mikro: 61,90%)			**γ = 3,0 & C = 12,0**	
	true Negative	true Objective	true Positive	class precision
pred. Negative	243	113	55	59,12%
pred. Objective	57	173	31	66,28%
pred. Positive	0	0	0	0,00%
class recall	81,00%	60,49%	0,00%	

Table 18 SVM for sentiment, γ=6,0 & C=6,0 (balanced)

For the sentiment labeling the SVM model (Table 18) has the highest accuracy of all three modeling techniques (61,91%). A closer inspection of the precision and class recall leads us to conclude that this model outperforms the K nearest neighbor model for *Negative* and *Objective*. Both the precision and class recall for both *Negative* and *Objective* are higher. However the precision and recall for

Positive have dropped to 0,00%. The performance for both these performance measurements were low in either of the two other models, which made them unable to correctly label *Positive* as well.

6.1.4 Model selection

Based on the assessments of each model in the previous sections the following models are selected as best performing option for each level of category and sentiment:

1. First level categories: k-NN
2. Second level categories: k-NN
3. Sentiment: SVM

These models are selected based on their accuracy and class precision and recall compared to the other models. It should be noted that the k nearest neighbor model is the best performing model for the second level category in comparison to the other models, but still a bad performing in its own sense. These models will be used with dimensionality reduction to test if higher performance is achievable.

6.1.5 Model performance with dimensionality reduction

In the previous section, we stated that we would test three models with dimensionality reduction. These three will be presented first, however we have also experimented with applying SVM and SVD to the first and second categories. Due to some interesting findings, we will also present the results of these models. Extra models based on naïve Bayes will not be presented as none of these model came close to the performance of either the K nearest neighbor or the SVM models. The tables with performance results can also be found in Appendix IV. All models only used the balanced data sets, as these present more realistic performance measurements.

Accuracy: 60,37% ± 6,79% (mikro: 60,38%)			**K = 15, Dimensions=40**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	6	129	31	70	54,66%
pred. Task related	3	24	110	23	68,75%
pred. Off topic	7	47	16	107	60,45%
class recall	0,00%	64,50%	70,06%	53,50%	

Table 19 k-NN for first level category with dimensionality reduction through SVD, k=15, number of dimensions=40(balanced)

The class precisions of the model in Table 19 are lower when compared to the model without dimensionality reduction Table 12. The overall accuracy is higher by 2,96% and the class recall is more spread with a better recall for *Off topic*, but lower recall for *Social related* and *Task related*. Even though the accuracy is higher for this model, closer inspection does not lead to the conclusion that this model actually performed better than the model without dimensionality reduction.

Accuracy: 38,67% ± 5,58% (mikro: 38,66%)							K = 20, Dimensions=20				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	2	0	1	0	0	1	0	0	0	0,00%
pred. Social: Person	7	35	8	2	0	9	19	10	5	2	36,08%
pred. Task: product	1	4	22	1	11	5	1	0	5	8	37,93%
pred. Social: Performance	0	2	1	2	0	6	0	4	2	1	11,11%
pred. Task: New product	0	0	1	0	0	1	0	0	0	0	0,00%
pred. Social: Organization	0	10	8	6	3	21	5	3	5	3	32,31%
pred. Off topic	4	6	5	2	3	11	29	4	2	4	41,43%
pred. Social: Personal effects	2	9	3	12	1	17	8	47	4	3	44,34%
pred. Task: Sales Technique	2	6	2	3	1	4	9	5	27	1	45,00%
pred. Task: Market developments	0	1	2	0	2	1	3	2	0	11	50,00%
class recall	0,00 %	46,67 %	42,31 %	6,90%	0,00%	28,00%	38,67 %	62,67 %	54,00%	32,35%	

Table 20 k-NN for second level category with dimensionality reduction through SVD, k=20, number of dimensions=20 (balanced)

The accuracy of the model in Table 20 compared to the original model (Table 13) increased by 2%. Both the precision and class recall went down for many categories. Some spread amongst the recall and precision boosted the accuracy, but like the model for the first level categories a closer inspection of the precision and class recall leads us to the conclusion that model is out performed by the original.

Accuracy: 61,59% ± 4,95% (mikro: 61,61%)			γ = 15, C = 8 and Dimensions=14	
	true Negative	true Objective	true Positive	class precision
pred. Negative	193	69	40	63,91%
pred. Objective	104	211	36	60,11%
pred. Positive	3	6	10	52,63%
class recall	64,33%	73,78%	11,63%	

Table 21 SVM for sentiment with dimensionality reduction through SVD, γ=15,0, C=8,0 and number of dimensions=14 (balanced)

The model in Table 21 does not show any improvement when it comes to the accuracy, but the precision and class recall are better spread and even perform to some extent for the *Positive* sentiment. For *Negative* the class recall is 16,67% lower compared to the original model (Table 18), but the precision increased by 4,79%. The reverse in observable for *Objective* where the class recall increased by 13,29%, but the precision decreased by 6,17%.

Even with a slightly lower accuracy, this model outperforms the original and is the best suitable model for labeling topics with sentiments.

Accuracy: 61,43% ± 5,65% (mikro: 61,43%)			γ = 1, C = 15 and Dimensions=40		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	3	123	29	58	57,75%
pred. Task related	3	22	103	16	71,53%
pred. Off topic	10	55	25	126	58,33%
class recall	0,00%	61,50%	65,61%	63,00%	

Table 22 SVM for first level category with dimensionality reduction through SVD, γ=1, C=15 and number of dimensions=40 (balanced)

The model in Table 22 was computed after finding the results from applying the combination of SVM and SVD for sentiment. The accuracy comes close to that of the unbalanced k-NN model for the first level categories, but with a much better performance for *Task related* and *Off topic*, and this model out performs the model form Table 19 for *Social related* and *Off topic*. Therefore, this model is the best performing model for the first level category.

Accuracy: 41,41% ± 7,48% (mikro: 41,43%)							γ = 1, C = 15 and Dimensions=40				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	1	2	0	0	0	0	0	0	1	0	20,00%
pred. Social: Person	5	41	6	1	1	11	14	6	2	0	47,13%
pred. Task: product	3	2	27	2	9	3	2	2	6	6	43,55%
pred. Social: Performance	0	3	0	4	0	5	0	2	4	0	22,22%
pred. Task: New product	0	0	3	0	1	0	1	0	0	2	14,29%
pred. Social: Organization	1	9	8	9	3	25	9	10	8	5	28,74%
pred. Off topic	2	13	2	3	1	11	34	8	4	3	41,98%
pred. Social: Personal effects	2	3	2	6	1	11	11	41	4	3	48,81%
pred. Task: Sales Technique	2	2	2	4	1	8	2	3	21	2	44,68%
pred. Task: Market developments	0	0	2	0	4	1	2	2	0	13	54,17%
class recall	6,25 %	54,67 %	51,92 %	13,79%	4,76%	33,33%	45,33 %	54,67 %	42,00%	38,24%	

Table 23 SVM for second level category with dimensionality reduction through SVD, γ=1, C=15 and number of dimensions=40 (balanced)

Both in accuracy and in overall performance per category, this model (Table 23) outperforms any other model for the second level category. It is even the only model that has any performance for *Flame*. Although 41,41% is the best performance, it is still not a very good accuracy. We must therefore conclude that we were unable to find a good performing model for the second level category.

6.1.6 Alternative Sentiment analysis

As noted earlier, an issue with manually determining the sentiment of a topic – especially multiple topics in a row – is an emotional bias of the person performing this labeling of sentiment. To remove this emotional bias, we also applied a different sentiment analysis as proposed by Tromp (2011). Table 24 shows a comparison between the sentiment analysis by Tromp (2011) and the manual sentiment labeling, where the manual sentiment is meant as input for the self-learning modeling techniques within RapidMiner.

	Sentiment with RBEM			
Manual sentiment	negative	neutral	positive	**Total**
Negative	294	28	290	**612**
Objective	63	75	148	**286**
Positive	22	7	57	**86**
Total	**379**	**110**	**495**	**984**

Table 24 Comparison table between manual sentiment and sentiment analysis

According to the manual sentiment labeling, the majority of cases were negative, while the sentiment analysis illustrates that a slight majority of the cases is positive. Another striking observation is the spread of cases amongst the different sentiments. The manual tagging leans mostly to the negative sentiment with hardly any cases on the positive end of the sentiment spectrum. The objective sentiment has a lot of cases, especially when comparing it to the sentiment analysis. The sentiment analysis allocated the bulk of all cases on one of the extremes of the sentiment analysis. In this situation, the objective sentiment looks to more or less a bridge between negative and positive.

An issue that might have occurred with the manual tagging of the sentiment is the influence of the emotional state of the tagger. Studies have shown that when observers have to judge the sentiment of social interactions or judge the emotional state of others they are significantly influenced by their own emotional state (Forgas et al., 1984; Schiffenbauer, 1974). This could lead to taking sarcasm too literally and therefore interpreting a topic as negative instead of objective or possibly even positive. Related to this emotional bias is that the sample data contains a lot of topics on layoffs and organizational changes. These messages might be interpreted overly negative as they occur plentiful and thereby might influence the mood of the tagger thereby triggering a vicious circle. We can also reason that salespeople are actually talking in a negative sense about these layoffs. As is being noted by users on Cafepharma, big pharmaceutical companies have had many rounds of layoffs during the last years due to competition of generic products[4]. Layoffs create a sense of job uncertainty, a decrease of personal control and job satisfaction throughout an organization. This effect is even stronger for the 'victims' of a layoff round (Paulsen et al., 2005), thereby the sentiment might actually be overly negative. It has been observed earlier that a negative sentiment boosts user activity and keeps topics active longer than positive topics (Chmiel et al., 2011).

The study performed by Chmiel et al. (2011) was performed by an automated sentiment analysis. The results show insights that are closer related to the manual sentiment tagging than to the sentiment analysis by Tromp (2011). An important reason for this could be that the latter was developed and tested on short social media posts on Twitter, Facebook and Hyves (Tromp, 2011). As this technique determines the sentiment by summing the sentiment of all elements within a case we might not get an accurate representation of the sentiment within a conversation. We can argue that it might a better strategy to determine the sentiment of different posts within a topic separately and then summing these. This gives each post an equal weight in the sentiment of a conversation and enables us to see whether the majority of the participants of a topic share a positive, objective or negative sentiment.

[4] Layoff rounds within Pfizer being discussed on Cafepharma:
http://www.Cafepharma.com/boards/showthread.php?t=518922
In late 2012 additional layoffs were announced to the sales force (NASDAQ Dow Jones Business News, 2012)

From sections 6.1.1 to 6.1.3 we know that the largest category is on the organization. Since pharmaceutical organizations have had many layoff rounds the last decade, as well as a lot of important products losing their protected status, the observation of a widespread negative sentiment seems acceptable in a sentiment analysis. However, the manual tagging of sentiment has some major drawbacks. The sentiment analysis of Tromp (2011) has been proven to work very well with social media like Twitter and Facebook. It should be further studied whether these results can be replicated when applying this technique in a manner as we proposed earlier on. Based on earlier research on both message board activity and the influence of layoffs (Chmiel et al., 2011; Paulsen et al., 2005) we conclude that the manual sentiment analysis approximates the actual sentiment closer that the method proposed by Tromp (2011).

6.1.7 Classification conclusion (III)

The order of operation did not conform to the earlier proposed roadmap. As the literature on which the roadmap is based clearly points out, dimensionality reduction is essential for model performance. Due to the computational demand of dimensionality reduction in RapidMiner this step was postponed. As the results expose, altering the order of executing was a mistake and led to incongruous conclusions on performance of models.

The best performing models were SVM models with dimensionality reduction through SVD. For sentiment and the first level categories the performance was around 61%. This is a decent performance but not yet very usable. Excluding *Flame* and *Positive* might lead to a more useful model in a technical sense but it would not lead to a model that meets practical needs. No models were found for the second level category that had adequate performance.

The manual sentiment tagging as input for the sentiment analysis with RapidMiner is more similar to another study related to message boards (Chmiel et al., 2011). Based on that study, the lay-offs in recent years and the fact that RBEM was not executed in respect to the message board conversation layout, we estimate that the manual tagging was more accurate. For future application it is advised to test RBEM in a way that it determines the sentiment of different posts within a topic separately and then sums these sentiments.

6.2 Association Rule Mining (IV)

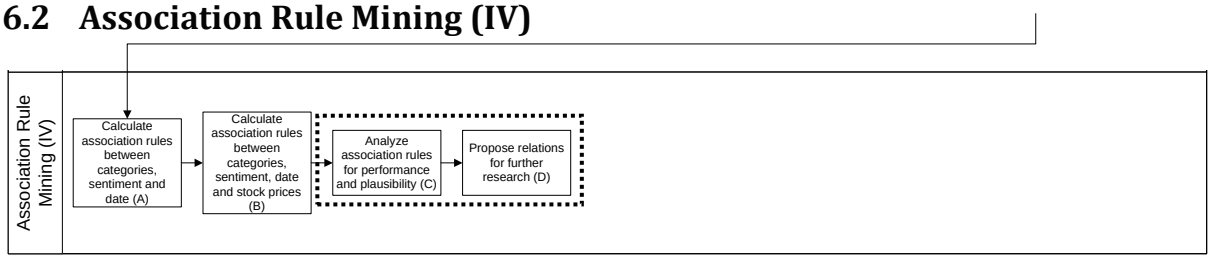

Table 25 Roadmap IV: Association Rule Mining

For closer understanding of the content of topics on Cafepharma we let RapidMiner, look for association rules between different attributes. The relations that have been explored with association rules are the relation between categories and sentiment, between categories and the year of posting, between the sentiment and the year of posting, between the change in the stock price and the category, between the stock price and the sentiment and between the change in stock price and the year of posting. The following figures and tables summarize the results.

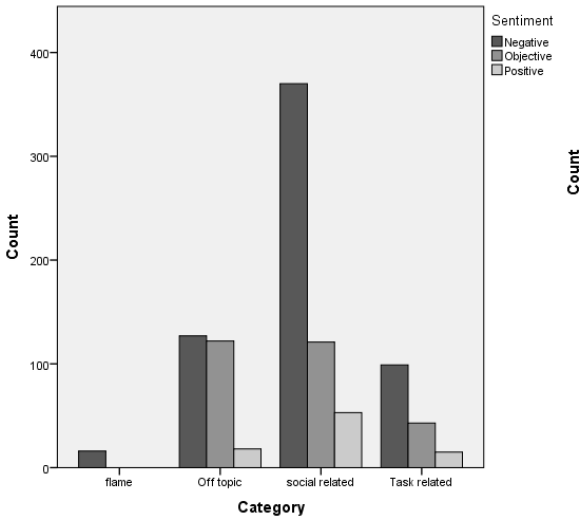

Figure 15 number of cases per sentiment per category

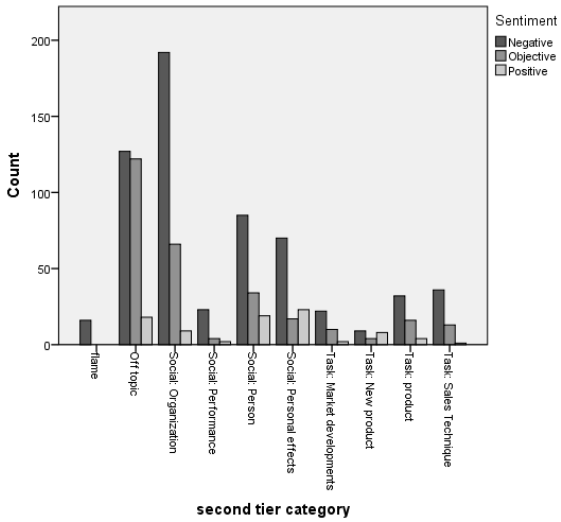

Figure 16 number of cases per sentiment per second level category

Figure 15 and Figure 16 illustrate that category and sentiment are to a certain degree related to each other. From these graphs we cannot deduce the actual relations or their strength. More thorough insights into possible relations can be acquired through association rules. Association rules with a confidence lower than 0,5 have been omitted as these rules do not propose very strong relations. Not all tables with results have been included in this section. Additional results can be found in Appendix V.

Categories	Number of cases
flame	16
Off topic	267
social related	544
Task related	157

Sentiments	Number of cases
Negative	612
Objective	286
Positive	86

Categories	Number of cases
flame	16
Off topic	267
Social: Organization	267
Social: Performance	29
Social: Person	138
Social: Personal effects	110
Task: Market developments	34
Task: New product	21
Task: product	52
Task: Sales Technique	50

Table 26 Frequency of categories and sentiments in the Cafepharma data set

6.2.1 Association rules for category, sentiment and year of posting

No.	Premises	Conclusion	Support	Confidence	Lift
1	Sentiment = Negative	Category = social related	0,38	0,60	1,094
2	Sentiment = Positive	Category = social related	0,05	0,62	1,115
3	Category = Task related	Sentiment = Negative	0,10	0,63	1,014
4	Category = social related	Sentiment = Negative	0,38	0,68	1,094
5	Category = flame	Sentiment = Negative	0,02	1,00	1,608

Table 27 Association rules for first level categories and sentiment.

An interesting combination of rules are $\{Sentiment = Negative\} \rightarrow \{Category = social\ related\}$ and $\{Category = social\ related\} \rightarrow \{Sentiment = Negative\}$.We can sum the support because this is a fraction of the complete data set, this leads us to the conclusion that these two rules combined are found in 76% of the cases. The confidence of these two rules is decent with 0,60 and 0,68 respectively. The lift for both of these rules is around 1 which makes neither of these rules particularly strong as this illustrates that these rules do not propose a very unique relation between two attributes. The data contain a lot of cases that are labeled as *Social related* and/or *Negative* (Table 26), therefore these association rules were to be expected.

Another obvious rule is $\{Category = flame\} \rightarrow \{Sentiment = Negative\}$. This relation does not occur a lot in the data (support=0,02) since *Flame* does not occur often in the data (Table 26). The confidence is very strong because every case that was labeled *Flame* was also labeled *Negative*. The lift of 1,6 also illustrates the strong relation between *Flame* and *Negative*.

Given the rule $\{Sentiment = Negative\} \rightarrow \{Category = social\ related\}$, the rule $\{Sentiment = Positive\} \rightarrow \{Category = social\ related\}$ might seem contra intuitive. When keeping in mind that *Social related* is an often occurring category and *Positive* a rarely occurring sentiment we can see the logic behind this rule. The support is very low with only 0,05, because of the limited number of cases with a *Positive* sentiment. The confidence could be explained due to large amount of cases with the label *Social related*. The same explanation is applicable to $\{Category = Task\ related\} \rightarrow \{Sentiment = Negative\}$ since *Task related* is a relatively small category and *Negative* a relatively large sentiment category.

Because the data is skewed towards *Social related* and *Negative*, and the number of attributes are limited these rules do not provide strong and new insights.

Premises	Conclusion	Support	Confidence	Lift
second tier category = Task: product	Sentiment = Negative	0,03	0,62	0,989
second tier category = Social: Person	Sentiment = Negative	0,09	0,62	0,990
second tier category = Social: Personal effects	Sentiment = Negative	0,07	0,64	1,023
second tier category = Task: Market developments	Sentiment = Negative	0,02	0,65	1,040
second tier category = Social: Organization	Sentiment = Negative	0,20	0,72	1,156
second tier category = Task: Sales Technique	Sentiment = Negative	0,04	0,72	1,158
second tier category = Social: Performance	Sentiment = Negative	0,02	0,79	1,275
second tier category = flame	Sentiment = Negative	0,20	1,00	1,608

Table 28 Association rules for second level categories and sentiment.

The association rules for the second level categories and the sentiment (Table 28) all have limited support and additionally most have limited lift. These factors combined make these rather weak rules.

The stronger rules are $\{second\ tier\ category = Social: Organization\} \rightarrow \{Sentiment = Negative\}$, $\{second\ tier\ category = Task: Sales\ Technique\} \rightarrow \{Sentiment = Negative\}$, $\{second\ tier\ category = Social: Performance\} \rightarrow \{Sentiment = Negative\}$ and $\{second\ tier\ category = flame\} \rightarrow \{Sentiment = Negative\}$. Especially the first and the last rule are strong compared to the other rules due to a support of 0,20 and a lift of 1,608 for the last rule.

All of the rules in Table 28 give a clear indication that when looking for a strong relation it is bound to the *Negative* sentiment.

The association rules for years of posting and the first level category and sentiment only illustrated in what year more than half of the topics were either belonging to the category *Social related* or had the sentiment *Negative*. Table 26 already gave us the insight that both *Social related* and *Negative* are often occurring, therefore association rules were expected but do not give new interesting insights.

A trend was observed of increased negativity over the years, with a small recovery in 2008. Especially in 2012 where 75% of all posts were negative.

No association rules were found for the second level categories.

6.2.2 Association rules for category, sentiment, year of posting and stock prices

Premises	Conclusion	Support	Confidence	Lift
Category = Task related	STOCK CHANGE = increase	0,08	0,51	1,038
STOCK CHANGE = increase	Category = social related	0,28	0,57	0,988
Category = flame	STOCK CHANGE = increase	0,01	0,57	1,153
STOCK CHANGE = decrease	Category = social related	0,28	0,58	1,014

Table 29 Association rules for change in stock and the first level category.

For the association rules with the change in stock prices (Table 29 and Table 30) the opening and closing stock prices of AstraZeneca, Daiichi, Johnson & Johnson, Merck and Pfizer were linked to topics that started on the same day. *STOCK CHANGE* is a measurement of whether the stock prices closed higher (*increase*), lower (*decrease*) or ended on the same price as it had opened (*level*). All data was acquired through *Yahoo! Finance*[5]. A strong limitation that should be considered is that topics may sometimes last for days and that *STOCK CHANGE* is not a measurement of the strength of the change. We can use *STOCK CHANGE* to find some initial relation that have to be further tested in additional research.

None of the rules in Table 29 have a particularly high confidence or support, support or lift. $\{STOCK\ CHANGE = increase\} \rightarrow \{Category = social\ related\}$ and $\{STOCK\ CHANGE = decrease\} \rightarrow \{Category = social\ related\}$ could both be due to the large amount of cases with the category *Social related*. For the latter we could reason that people might be inclined to discuss their job security or company performance during difficult times, but this should be further investigated.

$\{Category = flame\} \rightarrow \{STOCK\ CHANGE = increase\}$ look very counter intuitive. The very low support of 0,01 and the limited amount of cases that are labeled as *Flame* make this rule unreliable.

The same reasoning could apply for $\{Category = Task\ related\} \rightarrow \{STOCK\ CHANGE = increase\}$. It could however be argued that this rule makes more sense. If the company is performing well sales people might talk more about task related topics. However in that case the premises and the conclusion are in the wrong order, therefore we have to conclude that this reasoning is pure speculation and should be further researched.

Association rule mining for rules between second level category and stock prices led to no explanation for $\{Category = Task\ related\} \rightarrow \{STOCK\ CHANGE = increase\}$ from Table 29. Beside that we must conclude that any of the rules for second level category and stock prices were weak with very low support and confidence. We cannot draw any solid conclusion from these rules.

[5] http://finance.yahoo.com/q/hp?s=YHOO

Premises	Conclusion	Support	Confidence	Lift
Sentiment = Objective	STOCK CHANGE = increase	0,16	0,51	1,024
Sentiment = Positive	STOCK CHANGE = increase	0,05	0,54	1,087
STOCK CHANGE = increase	Sentiment = Negative	0,29	0,58	0,974
STOCK CHANGE = decrease	Sentiment = Negative	0,30	0,61	1,019
STOCK CHANGE = level	Sentiment = Negative	0,01	0,73	1,221

Table 30 Association rules for change in stock and the sentiment.

The rules of Table 30 are not very strong. Closer inspection of the rules leads to the conclusion that rules with a confidence higher than 0,55 are not useful at all. With any situation on the stock market the most often occurring sentiment is *Negative*. This might be due to the overrepresentation of the sentiment *Negative*, or it might be that there is no relation between sentiment and stock prices.

6.2.3 Association rules for stock prices with time delay

The structure of the association rules in the previous section imposes a direct relation between category and sentiment on one side and stock price on the other. We can argue that stock prices might not affect category and sentiment on the same day, therefore we have to look for a possible delay between these. We limit ourselves to the sentiment as the sentiment can be easily visualized and has the least amount of possible values. The following plots (Figure 17 and Figure 18) are presenting both daily closing stock prices versus the percentage of positives post on the same day between early 2006 and mid 2011. Figure 17 is based on the sentiment analysis with RapidMiner, whereas Figure 18 is based on the RBEM sentiment analysis. The sentiment data has been plotted as a scatter plot due to its nominal nature.

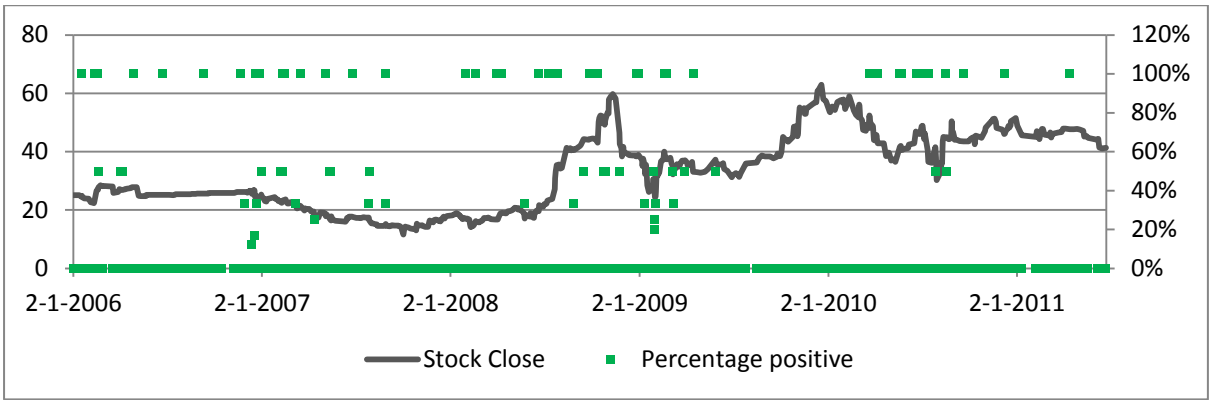

Figure 17 Percentage of topics with a positive sentiment (RapidMiner) versus the closing stock price (x=time(day), y1=stock, y2=%positive)

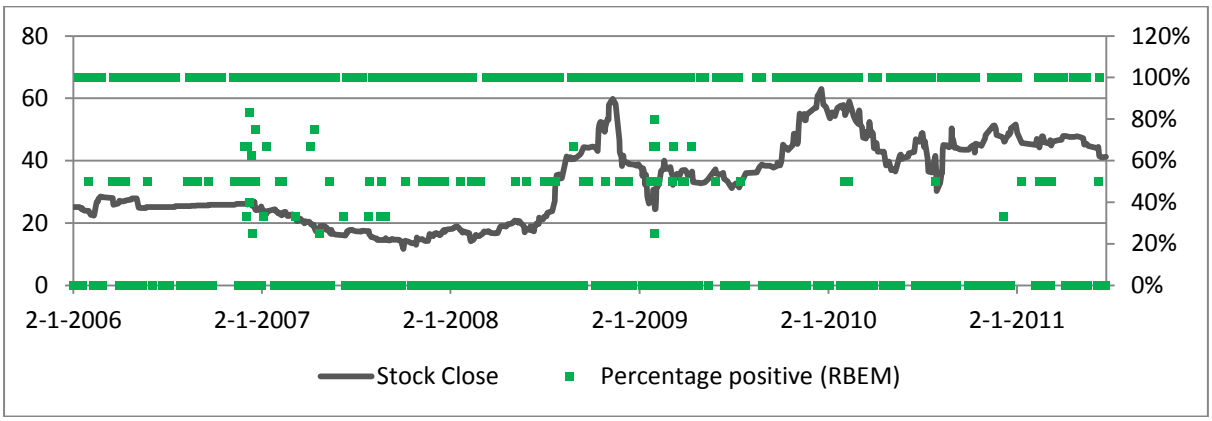

Figure 18 Percentage of topics with a positive sentiment (RBEM) versus the closing stock price (x=time(day), y1=stock, y2=%positive)

Figure 17 contains some concentration of data points larger that 0% around the two spikes in stock prices. This observation goes beyond the fact that we observer the same kind of concentration for the first 2 years or the last year where the stock prices did not show significant increases or decreases.

The data points of the percentage of topics with a positive sentiment in Figure 18 can be interpreted as an even plot of sentiment that does not have a relation with the stock price.

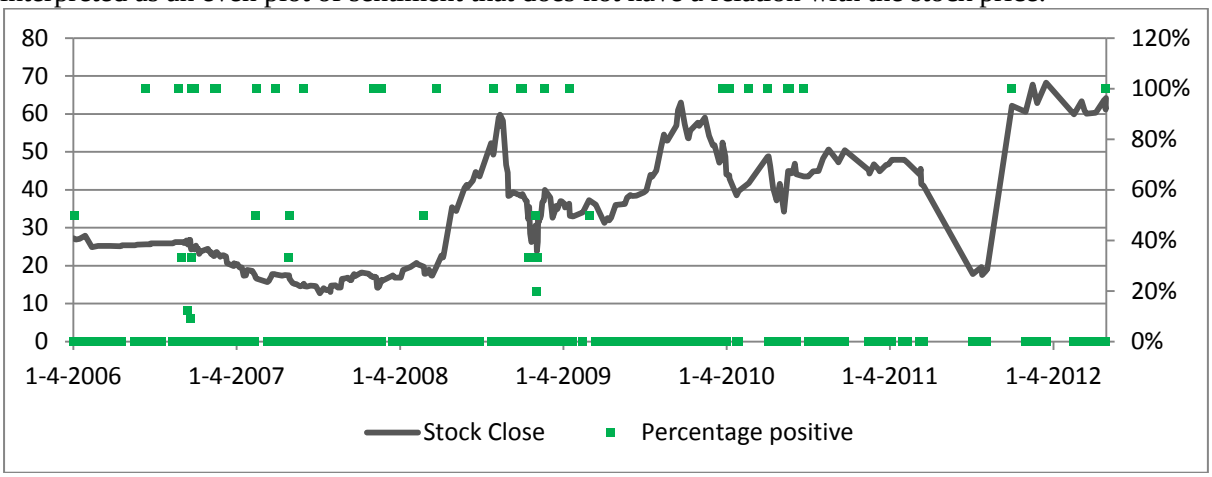

Figure 19 Percentage of topics with a positive sentiment (RapidMiner) versus the closing stock price (Pfizer) (x=time(day), y1=stock, y2=%positive)

Figure 19 is a plot of the stock prices versus the percentage of the positive sentiment for Pfizer. With this plot we hoped to observe a less cluttered plot. This plot has again some vague concentrations of data points larger than 0%. On this basis we must conclude that the data we collected does not contain a relation between sentiment and the stock price.

As some data point around 100% seem to follow after a spike in stock prices we used RapidMiner to find association rules, none of these rules indicated a relation (Appendix V).

6.2.4 Association rules mining (IV)

With respect to insights gained in Section 6.2.3 we must conclude that in our data there is no indication for a relation between any of the variable in our data and stock prices. The only relation that was found in the data and that could be explained is the negative trend over the last years.

7 Conclusions

In the conclusions, we will firstly reflect on the roadmap before we review to what extent the research questions have been answered. This will be followed by the discussing of the practical implications and finally a general discussion and possibilities for future research.

7.1 Roadmap

The roadmap (Figure 2) is made up of 4 steps, *Data collection* (I), *Data processing* (II), *Model teaching* (III) and *Topic analysis* (IV). In practice step I is very dependable on the data source. Factors like website formatting and hierarchy heavily influence the crawling. Even though data mining applications like RapidMiner provide options for web crawling, it is advised to build a custom miner for the best results.

Based on the RBEM sentiment analysis by Tromp (2011) and studies on emotional influences (Forgas et al., 1984; Schiffenbauer, 1974) we conclude that the manual sentiment analysis (II-A) was more accurate than the RBEM sentiment analysis (III-C4) due to the construction of topics where these topics are made up of posts by possibly different people. If the RBEM analysis is applied in respect to this structure it might outperform the manual sentiment analysis. Another possibility is to let the manual sentiment analysis be performed by multiple people. This way the inter rater reliability could be calculated ensuring more rigor into the tagging of topics (Fleiss, Cohen, & Everitt, 1969; Fleiss, 1971). The same technique should also be applied to the labeling of topics with categories, as this would ensure a higher level of rigor and thus make the data more suitable for quantitative research. For this study, the means were not available to perform a labeling of this kind.

As noted earlier Step III was not executed according to the chronology as proposed by the roadmap and literature (Howland & Park, 2008; Manning & Schütze, 1999). Major aspects of the dimensionality reduction (III-B) were executed after the categorization (III-C). The results of this have become evident in Section 6.1.5, as the models we thought had the best performance were outperformed by SVM models. This deviation from the roadmap was initialized due to the higher computational demands SVD models have in RapidMiner. In hindsight we conclude that this deviation only led to rework. Based on the results of Section 6.1.5 we conclude that Step III offers a correct approach to selecting a good categorization model. For both the first level category and the sentiment a SVM model was selected as the best performing model. No adequate model was found for the second level category. We proved the applicability of association rules for this kind of explorative research. The quality of the rules is debatable, but we will be going more in depth into discussing these in Section 7.5.

In conclusion, we found that the roadmap (Figure 2) offers good guide lines for building a categorization model and exploring the data for possible relations. Two problems have arisen. The first is that Step II requires more emphasis on the process of manually categorizing and labeling sentiment. A more thorough approach in this phase could boost the model performance in Step III and the number of valuable rules in Step IV. A second issue was in the execution of the roadmap. In our execution, we performed the dimensionality reduction at the very end due to the time consuming nature of computing an optimization of variables within RapidMiner. This resulted in selecting the wrong models as best performing models. Therefore it is advised to execute the roadmap in the order it was presented in and not performing the dimensionality reduction as a final model optimization step.

7.2 Roadmap improvements

The overall quality of the models from Step III and the association rules from Step IV we not high enough to be useful for direct practical usage. For this we propose some improvements to the roadmap. Additionally we propose some execution improvements to II-A, III-C4 and Step IV.

It became evident in Section 6.2.3 that there was no clear relation between the categories and sentiment on the one hand and the stock prices on the other (Step IV). Due to this conclusion we propose to use a different data source that is more closely related to the sales force, for instance sales KPIs or turnover. Since this data was not available for this study we were not able to test this.

For Step IV we proposed textbook approach, this left out a creative aspect of text mining in the pre-processing phase. We propose an extra activity between III-A and III-B called 'Word list creation'. The problem of self-learning model is that they only have the text that was given as input to understand the relation between words and categories. A human being would label a text 'new product' if it saw a product name of a new product often in a single text. A self-learning model might not have the data to relate this product name to the category 'new product'. A solution would be to provide all product names of new products. If words from a certain wordlist occur this would increase the weight of a text belonging to a certain category. The current problem with RapidMiner is that it cannot deal with word lists in this manner (Mierswa et al., 2006).

For II-A the quality of the manual categorization and sentiment labeling is debatable. The quality and rigor of this data would greatly benefit from multiple labelers (Fleiss et al., 1969; Fleiss, 1971).

Beside improving the quality of the sentiment analysis through multiple labelers we could also improve the quality of the RBEM sentiment analysis by applying this technique with respect to the construction of a topic on Cafepharma, this would lead to an overall sentiment of a topic based on the sentiment of each post within this topics.

Lastly, for better insights into message board conversations in relation to performance (Step IV) we propose the use of a data source that is more closely related to sales, for instance sales KPIs or turnover.

7.3 Research questions

To conclude the results of this thesis we will link the findings of this study to the research questions as they were formulated in the first Section. We will firstly answer the research questions related to the extant research of multi-channel and secondly the research question in relation to the problem definition.

Research question 1: 'What are the main concepts within the field of multi-channel research?' These are the channels that are subject of extant studies (brick-and-mortar stores, internet, catalogue and call-centers) and multi-channel marketing.

Research question 2: 'What are the main themes within extant multi-channel research?' These are the usage of the internet channel, the focus in research on marketing versus sales and B2C versus B2C.

Research question 3: 'What major gaps are identifiable within extant multi-channel research?' Firstly the internet channel has only been studied as a collection of web shops. Secondly and thirdly the focus of research has been almost exclusively on marketing and B2C.

The next 4 research question will be in relation to the problem definition.

Research question 1: 'What kind of methods can be applied to identify the category and sentiment of a conversation?' For text categorization the SVM training operator returned the best performance in combination with SVD. With respect to the sentiment analysis no definitive answer was found for this question. Either applying the RBEM sentiment analysis with respect to the structures of topics and the posts that they are made up of, or by having the manual sentiment analysis be performed by multiple people. Additional testing of the sentiment analysis in this manner is required to get a definitive answer to research question 1 for determining sentiment.

Research question 3: ‘How can these methods be applied?’ The roadmap Figure 2 presented a good approach. It is of key importance that the order of performing the steps and activities is followed. As results have shown in this study, it could lead to suboptimal outcomes if the order of activities is altered. To increase the reliability of the categorization models we propose a larger sample set, multiple people to perform the tagging procedure and when the technology allows it, word lists. To increase the usability of the association rules we propose the use of internal sales related data.

Research question 4: ‘What is the relation between category/sentiment ratios and the stock price index?’A strong relation was found between the category *Flame* and the sentiment *Negative* and that the sentiment has become more negative over the last years with a peak in 2012 of 75% of all posts having a negative sentiment. Other relations were not very strong and mostly indicated a strong overrepresentation of the sentiment *Negative* and the category *Social related*. No reliable rules were found for relations between the message board topics and stock prices.

7.4 Practical impact

In this study, we made use of a publicly available performance measure of companies: stock prices. Companies themselves have much more accurate and specific performance measurements down to sales groups and products. Through these performance measurements and – for instance – the category *Task: New Product* combined with the sentiment an organization can get insights into the reception of its salespeople of new products. Without categorization it would be very hard to find related topics on a message board and to get quantitative data on the sentiment towards a certain category. With this data it becomes easier for companies to find – for instance – criticism of salespeople on certain products or measure the effects of organizational performance on salespeople sentiment. Thereby managers are able to gain quantitative insights into salespeople behavior and opinions through using this roadmap with readily available PC applications. The current limit of these readily available applications is that they are not yet able to categorize text with word lists.

This application is also usable for academic purposes. An additional application can be found in the explorative nature of association rules. Through association rules possible relations can be found on how people react to organizational performance. This helps in the formulation of hypotheses on behavioral studies of salespeople in a social media context.

7.5 Discussion and future research suggestions

An issue that exists – and the reason why we have used balanced data sets – is the over representation of certain categories in the data. To overcome the possible negative effects of these larger categories, some cases should be randomly deleted. This can however lead to a data set that does not contain enough cases for successful categorization. It is therefore advised to use a larger initial sample set to ensure that after deleting over represented case the data set still contains 1000+ cases.

Beside the in Section 7.4 mentioned broad academic possibilities within the field of behavioral studies related to social media and sales, we also propose some more concrete research suggestions. We found in Section 6.2 that the sentiment has become more negative over the past years. In Section 6.1.6 we hinted to the large layoffs in the pharmaceutical industry. Therefore, an interesting topic might be the job satisfaction amongst pharmaceutical salespeople, and as an extension on that topic, the influence this job satisfaction might have had on the organizational performance of these pharmaceutical companies. Combined with Cafepharma data, this research could lead to an answer on the question of whether the sentiment on Cafepharma is related to the actual sentiment of salespeople towards their jobs.

In a more technical sense the creation of an application that can implement the creative aspects of text categorization through word lists would mean a leap forwards into the practical usage of text mining for managers.

Lastly, the causality between organizational performance and category or sentiment could be studied. The only indication we found of some relation is between declining stock markets and the topics within the *Social related* category.

References

4SquareBadges.com. (n.d.). How to get the Barista (Starbucks) badge? Retrieved from http://www.4squarebadges.com/foursquare-badge-list/barista-starbucks-badge/

ABI/INFORM. (n.d.). ProQuest.

Apte, C., Damerau, F., & Weiss, S. (1998). *Text mining with decision rules and decision trees*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.6018&rep=rep1&type=pdf

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *in Proc. of LREC*.

Balasubramanian, S., Peterson, R., & Jarvenpaa, S. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science*, *30*(4), 348–361. doi:10.1177/009207002236910

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., … Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.

Blattberg, R. C., Kim, P., Kim, B.-D., & Neslin, S. A. (2008). *Database Marketing: Analyzing and Managing Customers*. Springer.

Brassington, F., & Pettitt, S. (2006). *Principles of Marketing*. Pearson Education.

Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., & Hołyst, J. A. (2011). Negative emotions boost user activity at BBC forum. *Physica A: Statistical Mechanics and its Applications*, *390*(16), 2936–2944. doi:10.1016/j.physa.2011.03.040

Colombo, R. A., & Morrison, D. G. (1989). A Brand Switching Model with Implications for Marketing Strategies. *Marketing Science*, *8*(1), 89–99.

Duda, R. O., & Hart, P. E. (1973). *Pattern recognition and scene analysis*. Wiley, New York.

Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J., & Seymour, T. (2011). The History of Social Media and its Impact on Business. *Journal of Applied Management and Entrepreneurship*, *16*(3), 79–91.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), 323–327. doi:10.1037/h0028106

Forgas, J. P., Bower, G. H., & Krantz, S. E. (1984). The influence of mood on perceptions of social interactions. *Journal of Experimental Social Psychology*, *20*(6), 497–513. doi:10.1016/0022-1031(84)90040-4

Freund, Y., & Schapire, R. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory* (pp. 23–37). Retrieved from http://www.springerlink.com/index/7QP818392282L161.pdf

Gale, W. A., Church, K. W., & Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, *26*(5), 415–439.

Gensler, S., Dekimpe, M. G., & Skiera, B. (2007). Evaluating channel performance in multi-channel environments. *Journal of Retailing and Consumer Services*, *14*(1), 17–23. doi:10.1016/j.jretconser.2006.02.001

Gopal, R. D., Pathak, B., Tripathi, A. K., & Yin, F. (2006). From Fatwallet to eBay: An investigation of online deal-forums and sales promotions. *Journal of Retailing*, *82*(2), 155–164. doi:10.1016/j.jretai.2006.02.002

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1).

Hollander, S. (1966). Notes on the retail accordion. *Journal of Retailing*, *42*, 29–40.

Höppner, F. (2005). Association Rules. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 353–376). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/0-387-25465-X_16

Howland, P., & Park, H. (2008, January 1). Cluster-Preserving Dimension Reduction Methods for Document Classification. Retrieved December 8, 2012, from http://link.springer.com/chapter/10.1007/978-1-84800-046-9_1

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A practical guide to support vector classification*. Retrieved from https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf

Huang, P., Lurie, N. H., & Mitra, S. (2009). Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods. *Journal of Marketing*, *73*(2), 55–69. doi:10.1509/jmkg.73.2.55

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, 137–142.

Kollmann, T., Kuckertz, A., & Kayser, I. (2012). Cannibalization or synergy? Consumers' channel selection in online–offline multichannel systems. *Journal of Retailing and Consumer Services*, *19*(2), 186–194. doi:10.1016/j.jretconser.2011.11.008

Konuş, U., Verhoef, P. C., & Neslin, S. A. (2008). Multichannel Shopper Segments and Their Covariates. *Journal of Retailing*, *84*(4), 398–413. doi:10.1016/j.jretai.2008.09.002

Kumar, V., & Venkatesan, R. (2005). Who are the multichannel shoppers and how do they perform?: Correlates of multichannel shopping behavior. *Journal of Interactive Marketing*, *19*(2), 44–62. doi:10.1002/dir.20034

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press. Retrieved from http://books.google.nl/books?hl=nl&lr=&id=YiFDxbEX3SUC&oi=fnd&pg=PR16&dq=Manning+%22Foundations+of+statistical%22&ots=vYvmsvdLLP&sig=DYuBt2h63mhVvzCSNIyehrmgi_s

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In L. Ungar, M. Craven, D. Gunopulos, & T. Eliassi-Rad (Eds.), *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935–940). New York, NY, USA: ACM. Retrieved from http://rapid-i.com/component/option,com_docman/task,doc_download/gid,25/Itemid,62/

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill New York:

Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press.

Montoya-Weiss, M. M., Voss, G. B., & Grewal, D. (2003). Determinants of Online Channel Use and Overall Satisfaction with a Relational, Multichannel Service Provider. *Journal of the Academy of Marketing Science*, *31*(4), 448–458. doi:10.1177/0092070303254408

Mooney, R. J. (2005). Machine learning. In *The Oxford Handbook of Computational Linguistics* (pp. 376–394). Oxford: Oxford University Press.

Moriarty, R. T., & Moran, U. (1990, December). Managing Hybrid Marketing Systems. *Harvard Business Review*, *68*(6), 146.

NASDAQ Dow Jones Business News. (2012, November 30). Pfizer to Shrink US Sales Force, Cites "Future Needs" of Business. *NASDAQ.com*. Financial news. Retrieved February 19, 2013, from http://www.nasdaq.com/article/pfizer-to-shrink-us-sales-force-cites-future-needs-of-business-20121130-00845

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., & Verhoef, P. C. (2006). Challenges and Opportunities in Multichannel Customer Management. *Journal of Service Research*, *9*(2), 95–112. doi:10.1177/1094670506293559

New Oxford American Dictionary. (2010) (third.). Oxford University Press.

Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Now Pub. Retrieved from http://books.google.nl/books?hl=nl&lr=&id=XQswsqLLKrEC&oi=fnd&pg=PA1&dq=%22sentiment+analysis%22&ots=FM17BLu4l1&sig=74hNeGjS1jho0MR-o3H47qocaEA

Paulsen, N., Callan, V. J., Grice, T. A., Rooney, D., Gallois, C., Jones, E., … Bordia, P. (2005). Job uncertainty and personal control during downsizing: A comparison of survivors and victims. *Human Relations*, *58*(4), 463–496. doi:10.1177/0018726705055033

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*. Retrieved from https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf

Reynolds, J., Howard, E., Cuthbertson, C., & Hristov, L. (2007). Perspectives on Retail Format Innovation: Relating Theory and Practice. *International Journal of Retail & Distribution Management*, *35*(8), 647–660. doi:10.1108/09590550710758630

Rickman, T. A., & Cosenza, R. M. (2007). The Changing Digital Dynamics of Multichannel Marketing: The Feasibility of the Weblog: Text Mining Approach for Fast Fashion Trending. *Journal of Fashion Marketing and Management*, *11*(4), 604–621. doi:10.1108/13612020710824634

Rosenbloom, B. (2011). *Marketing Channels*. Cengage Learning.

Schachter, P. (1985). Parts-of-speech systems. In *Language Typology and syntactic Description*. cambridge: Cambridge University Press.

Schiffenbauer, A. (1974). Effect of observer's emotional state on judgments of the emotional state of others. *Journal of Personality and Social Psychology*, *30*(1), 31.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing* (Vol. 12, pp. 44–49). Retrieved from http://www.stttelkom.ac.id/staf/imd/Riset/POS%20Tagging/Using%20Decision%20Tree.pdf

Schoenbachler, D. D., & Gordon, G. L. (2002). Multi-channel shopping: understanding what drives channel choice. *Journal of Consumer Marketing*, *19*(1), 42–53. doi:10.1108/07363760210414943

Shankar, V., Inman, J. J., Mantrala, M., Kelley, E., & Rizley, R. (2011). Innovations in Shopper Marketing: Current Insights and Future Research Issues. *Journal of Retailing*, *87, Supplement 1*, S29–S42. doi:10.1016/j.jretai.2011.04.007

Thelwall, M. (2001). A web crawler design for data mining. *Journal of Information Science*, *27*(5), 319–325. doi:10.1177/016555150102700503

Thrax, D. (1883). Ars grammatica. *Grammatici Graeci*, *1*, 5–100.

Tromp, E. (2011). *Multilingual Sentiment Analysis on Social Media*. Eindhoven University of Technology, Eindhoven.

Tschan, F., Semmer, N. K., & Inversin, L. (2004). Work Related and "private" Social Interactions at Work. *Social Indicators Research*, *67*(1/2), 145–182.

Van Birgelen, M., De Jong, A., & De Ruyter, K. (2006). Multi-channel service retailing: The effects of channel performance satisfaction on behavioral intentions. *Journal of Retailing, 82*(4), 367–377. doi:10.1016/j.jretai.2006.08.010

Vapnik, V. (1979). Estimation of dependences based on empirical data. *Nauka*. Retrieved from http://www.citeulike.org/group/1938/article/1055256

Vapnik, V. (1995). *The nature of statistical learning theory*. springer. Retrieved from http://books.google.nl/books?hl=nl&lr=&id=sna9BaxVbj8C&oi=fnd&pg=PR7&dq=%22The+Nature+of+Statistical+Learning+Theory%22+Vapnik+1995&ots=onM7IVoka8&sig=wvVBtk1CgkSPVPQY4kls2yWxNqU

Venkatesan, R., Kumar, V., & Ravishanker, N. (2007). Multichannel Shopping: Causes and Consequences. *Journal of Marketing, 71*(2), 114–132. doi:10.1509/jmkg.71.2.114

Verhoef, P. C., Neslin, S. A., & Vroomen, B. (2007). Multichannel customer management: Understanding the research-shopper phenomenon. *International Journal of Research in Marketing, 24*(2), 129–148. doi:10.1016/j.ijresmar.2006.11.002

Voutilainen, A. (2005). Parts of speech, tagging. In *The Oxford Handbook of Computational Linguistics* (pp. 219–232). Oxford: Oxford University Press.

Wheeler, L., & Reis, H. T. (1991). Self-Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of Personality, 59*(3), 339–354. doi:10.1111/j.1467-6494.1991.tb00252.x

Xu, Y. (Calvin), & Kim, H.-W. (2008). Order Effect and Vendor Inspection in Online Comparison Shopping. *Journal of Retailing, 84*(4), 477–486. doi:10.1016/j.jretai.2008.09.007

Appendix I Comparison table

No.	Study	Independent variable	Dependent variable	Key concepts	Industry & Context	Multi-channel setting & Channels	Empirical/ theoretical	Sample size
1	**Birgelen, de Jong & de Ruyter (2006)** ***Journal of Retailing***	Channel performance (traditional service channel & technology-mediated channels	Behavioral intentions	– Multi-channel marketing – Customer behavior based on channel performance – Channel interaction performance	Retail banking (the Netherlands) B2C	Multi-channel behavior (Brick-and-mortar, internet)	Empirical	n=809 (customers of 37 non-routine service branch units) n=567 (customer of 25 routine service branch units)
2	**Gensler, Dekimpe & Skiera (2007)** ***Journal of Retailing and Consumer services***	Customer group: Hard-core loyals and potential switchers (according to the model by Colombo and Morrison (1989))	Customers' intrinsic loyalty to a particular channel Channel's ability to attract switching customers	– Multi-channel marketing channel performance measurement	Home-shopping (Europe) B2C	Multi-channel performance (Call-center, internet)	Empirical	Data available on ± 1,5 million customers for 15 consecutive years
3	**Kollman, Kuckertz, Kayser (2012)** ***Journal of Retailing and Consumer services***	– Convenience orientation – Risk aversion – Service orientation	– Channel selection – Channel of initially getting information – Customer type – Whether the transaction could be completed in a competing channel	– Multi-channel marketing – Multi-channel customer segmentation – Cannibalization/Synergy	Telecom (Germany) B2C	Multi-channel behavior (Brick-and-mortar, internet)	Empirical	– Offline customers: n=163 – Online customers: n=1075
4	**Montoya-Weiss, Voss & Grewal (2003)** ***Journal of the Academy of Marketing Science***	– Perceived quality of web site's information content – Perceived attractiveness of web site's graphic style – Perceived online channel service quality – Perceived service quality of primary alternative channel – Perceived channel risk – Level of internet expertise	– Perceived online channel service quality – Perceived risk of online channel – Overall satisfaction with service provider	– Multi-channel marketing – Customer behavior to using an internet channel	Financial services (study 1) (USA) University course registration (study 2) (USA) B2C	Channel behavior (internet)	Empirical	Study 1 1. Pretest: n=600 2. Main study: n=1137 Study 2 n=493
5	**Schoenbachler & Gordon (2002)** ***Journal of Consumer Marketing***			– Multi-channel marketing – Multi-channel customer behavior		Multi-channel behavior	Conceptual	

6	**Konuş, Verhoef & Neslin (2008)** ***Journal of Retailing***	Consumer attitudes toward various channels: – Innovativeness – Loyalty – Shopping enjoyment – Price conciousness	Multi-channel customer segments	– Multi-channel customer management – Multi-channel customer segmentation	Consumer (mortgage, health insurance, holidays, books, computers, electronics and clothing) B2C	Multi-channel behavior (brick-and-mortar, internet, catalogue)	Empirical & academic research review	n=364
7	**Venkatesan, Kumar & Ravishanker (2007)** ***Journal of Marketing***	Longitudinal analysis – Number of transactions – Lagged profits – Lagged multi-channel shopping Customer channel adoption duration – Basket size – Cross-buying – Level of price discounts – Proportion of returns – Purchase frequency – Frequency of marketing communications – Travel cost proportion – IPA proportion (immediate product availability)	Longitudinal analysis – Total customer profit (Longitudinal customer profitability model) – Indicator of multi-channel shopping (Longitudinal multi-channel shopping model) Customer channel adoption duration – Duration to adopt second channel – Duration to adopt third channel	– Multi-channel marketing – Multi-channel performance measurement – Channel adoption	Apparel B2C	Multi-channel behavior (brick-and-mortar, internet, catalogue	Empirical (longitudinal)	Longitudinal analysis – N=8882 Customer channel adoption – Calibration sample: n=1165 – Holdout sample: n=379
8	**Neslin, Grewal, Leghorn, Shankar, Teerling, Thomas & Verhoef (2006)** ***Journal of Service Research***			– Multi-channel customer management – Data integration – Multi-channel customer behavior – Multi-channel marketing channel performance measurement – Resource allocation – Channel coordination strategies		Multi-channel customer management & challenges for research.	Conceptual & academic research review	

9	**Verhoef, Neslin & Vroomen (2007)** ***International Journal of Research in Marketing***	– Information availability – Search convenience – Search effort – Service quality – After sales service – Purchase convenience – Negotiation possibilities – Purchase effort – Purchase risk – Enjoyment – Assortment	– Search attractiveness – Purchase attractiveness	– Multi-channel customer management – Research-shopper	Consumer (loans, vacations, books, computers, clothing and electronic appliances) B2C	Multi-channel behavior (brick-and-mortar, internet, catalogue)	Conceptual & Empirical	N=396		

Table 31 Comparison table

Appendix II Modeling techniques & RBEM

In the following appendix, we will explain the modeling techniques used to train categorization models. The modeling techniques are naïve bayes, k nearest neighbor and support vector machine (SVM).

Naïve Bayes

Naïve Bayes is a supervised disambiguation modeling technique proposed by Gale, Church and Yarowsky (1992). Supervised disambiguation is a statistical classification that relies on a disambiguated corpus for training where a training set where each occurrence of the word w is annotated with a semantic label (Manning & Schütze, 1999).

A Bayes classifier looks at the words around an ambiguous word. Each of these words gives potentially useful information about the sense in which the ambiguous word is used (Manning & Schütze, 1999).

These classifiers rely on the Bayes decision rule for choosing a category c (Duda & Hart, 1973). With the Bayes decision rule we want to decide s' if $P(s'|c) > P(s_k|c)$ for $s_k \neq s'$. s_k is the contextually appropriate sense of a semantic label. Because $P(s_k|c)$ is often unknown it has to be calculated using Bayes' rules, which is formulated as $P(s_k|c) = \frac{P(c|s_k)}{P(c)} P(s_k)$.

For the classification the Naïve Bayes classifier will be used. It is a popular classifier in machine learning in general due to its efficiency and its ability to handle large sets of features (Mitchell, 1997). The Naïve Bayes assumption is formulated as follows: $P(c|s_k) = P(\{v_j|v_j \text{ in } c\}|s_k) = \prod_{v_j \, in \, c} P(v_j|s_k)$, where v_j are the words that are in the context of w (Gale et al., 1992; Manning & Schütze, 1999). The Naïve Bayes assumption lead to the decision rule where have to decide s' if $s' = \arg max_{s_k}[\log P(s_k) + \sum_{v_j \, in \, c} \log P(v_j|s_k)]$. $P(v_j|s_k)$ and $P(s_k)$ can be calculated through a Maximum-likelihood estimation: $P(v_j|s_k) = \frac{C(v_j, s_k)}{C(s_k)}$ and $P(s_k) = \frac{C(s_k)}{C(w)}$.

K Nearest Neighbor

The nearest neighbor modeling technique working according to the principal of finding a case which is most similar and assigning the category of this neighbor to the case itself (Manning & Schütze, 1999). The technique we applied is known as k nearest neighbor. This technique does not rely on just one neighbor but on k neighbors, where $k > 1$. The k nearest neighbor is more robust than the single nearest neighbor thanks to this multiple neighbor aspect.

A good similarity measure for text categorization is cosine similarity (Manning & Schütze, 1999). Cosine similarity is defined as $\frac{|X \cap Y|}{\sqrt{|X| \times |Y|}}$. When using this similarity measure in a binary categorization where $k = 1$, the categorization looks as follows. The goal is to categorize our case $\vec{y}$ based on the training set X. For this we have to know the largest similarity for $\vec{y}$ with any case in the training set, defined as $sim_{max}(\vec{y}) = max_{\vec{x}} sim(\vec{x}, \vec{y})$. We must then find a subset of X with the largest similarity with $\vec{y}$. This is defined as: $A = \{\vec{x} \in X | sim(\vec{x}, \vec{y}) = sim_{max}(\vec{y})\}$. If we then state that n_1 and n_2 are the number of cases in A that belong to either one of the categories c_1 or c_2 we can estimate the probability of a case belonging to one of the categories as follows: $P(c_1|\vec{y}) = \frac{n_1}{n_1 + n_2}$ and $P(c_2|\vec{y}) = \frac{n_2}{n_1 + n_2}$. Solve $P(c_1|\vec{y}) > P(c_2|\vec{y})$ for c_1 or $P(c_2|\vec{y}) > P(c_1|\vec{y})$ for c_2.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a technique based on structural risk minimization. This means that SVM tries to find a hypothesis h for which it can guarantee the lowest possible true error. This true error is the probability that h will make an error on a test example (Joachims, 1998).

Through the following upper bound the true error of h is connected with both the error of h on a training set and the complexity of h (Vapnik, 1995): $P\big(error(h)\big) \leq train\_error(h) + 2\sqrt{\frac{d(\ln\frac{2n}{d}+1) - \ln\frac{n}{4}}{n}}$. In this equation n is the number of training examples and d is what Vapnik (1995)

defined as VC-Dimension or VCdim. VCdim is the expressiveness of hypothesis space. Because a small VCdim leads to a big error and a large VCdim leads to over fitting it is important to find the best VCdim. Finding the right VCdim is done by defining a hypothesis space structure H_i so that their VCdim d_i increases (Joachims, 1998): $H_1 \subset H_2 \subset \cdots \subset H_i$ and $\forall i: d_i \leq d_{i+1}$. The goal is to an i where the upper bound minimum.

The goal of a SVM is to find a hyperplane that separates the training data with the shortest weight factor as. The relation between VCdim and hyperplanes is defined as follows (Vapnik, 1979): as a hypothesis we consider $hyperplanes\ h(\vec{d}) = sign\{\vec{w} \times \vec{d} + b\}$. If all $\overrightarrow{d_i}$ are contained in a ball radius R and for all $\overrightarrow{d_i}$ it is required that $|\vec{w} \times \overrightarrow{d_i} + b| \geq 1$ with $||\vec{w}|| = A$, it follows that the hypeplane has a VCdim bound by $d \leq \min([R^2A^2], n) + 1$. With this is mind the hyperplane can be found by solving the optimization problem where we want to minimize $||\vec{w}||$, so that $\forall: y_i[\vec{w} \times \overrightarrow{d_i} + b] \geq 1$.

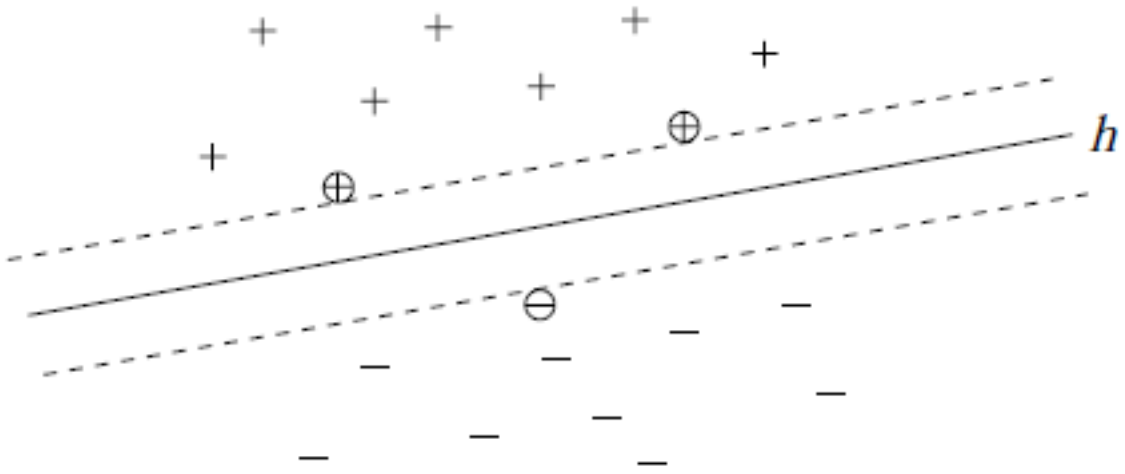

Figure 20 Hyperplane h which separates positive from negative training cases. The examples closest to the hyperplane are Support Vectors (marked with circles in this figure) (Joachims, 1998)

RBEM (Sentiment analysis)

As an alternative method for sentiment analysis, we will apply a technique specifically tailored for social media. This technique was developed by Tromp (2011).

This sentiment analysis was developed for multilingual social media data. Since our data is exclusively in English, the multilingual aspect is of less importance to us. This technique has been selected due to its performance compared to other sentiment analysis (Tromp, 2011). The sentiment analysis by Tromp (2011) is a four step approach.

The first step is to determine the language, since this is not applicable to our study we will not elaborate on this.

The second step is to parts of speech tagging (POS). With POS tagging every word is labeled with a 'tag'. Parts of speech was introduced by Thrax (1883, original c. 100 BC). He distinguished eight word classes in parts of speech:

1. Noun
2. Verb
3. Participle
4. Article (incl. relative pronoun)
5. Pronoun
6. Preposition

7. Adverb
8. Conjunction

Schachter (1985) defined that for part of speech grammatical aspects are more important than semantics:

1. Syntactic distribution;
2. Syntactic function;
3. Morphological and syntactical classes to which different parts of speech can be assigned.

Tagging is the automatically assignment of 'tags' to input tokens (Voutilainen, 2005).
To clarify what happens with POS tagging an example by Voutilainen (2005) is given in Table 32.

Input	*However, if two contiguous words are both unambiguous, the time-slice corresponding to those words will contain only a single state with non-zero probability.*
Output[6]	However_ADVwh , if_Cs two_Ncard contiguous_A words_Npl are_Vpres both_P unambiguous_A , the_DET time-slice_N corresponding_ING to_PREP those_DET words_Npl will_Vmod contain_Vinf only_ADV a_DET single_A stat_N with_PREP non-zero_N probability_N . <p>

Table 32 Example of parts of speech tagging (POS) (Voutilainen, 2005)

The tags behind each word tells us what kind of word it is. For instance are_Vpres tells us that are is a verb in the present tense.
This tagged output can be filtered so that other analytic techniques further down the line only look for nouns and adjectives for instance.

The approach of a POS tagger is as follows:

1. Tokenization
    - This part of the architecture separates the different objects of the input text. These can for instance be objects that look like words or punctuation marks. This step is necessary for further analysis.
2. Ambiguity look-up
    - The first tool to give tokens a tag is a *lexicon*, this can be a list of word forms and their possible parts of speech.
    - The second tool is a *guesser*, these are often built around what is known about the lexicon; this way the guesser can estimate what a unknown token will probably be.
    - Finally a POS tagger has a compiler/interpreter, and this will provide the alternative tags for each token given the results of the lexicon and guesser.
3. Ambiguity resolution of disambiguation
    - This step determines which of the alternative tags the *ambiguity look-up* proposed is the "right" one. It does this both with information about the word itself as with information about the sequence of the words.

The tagger applied by Tromp (2011) is the TreeTagger (Schmid, 1994).

[6] POS tagger used: EngCG-2 (Voutilainen, 2005)

The third step is to determine subjectivity. The aim is to figure out whether a case contains objective information or conveys a subjective message. Subsequently, only the subjective messages are further analyzed in step four. Tromp (2011) attains this by using AdaBoost (Freund & Schapire, 1995). AdaBoost uses weak learners to create a strong learner. Tromp (2011) uses decision stomp for this. Decision stomps are decision trees with a tree length of one (Tromp, 2011) through which the learner can either assign a case to be subject or objective. The subjectivity is then determined by applying AdaBoost with a polarity lexicon on cases that have been processed by the POS Tagger. A polarity lexicon contains polarity information (e.g. positive, negative and objective) on words (Baccianella et al., 2010).

The fourth step is to determine the polarity of a case. The polarity is determined by an algorithm proposed by Tromp, called Rule-Based Emission Model (RBEM) (Tromp, 2011). This technique determines the polarity based on a set of rules. These rules are derived from a set of eight patterns that elements in a case can emit (Tromp, 2011). These patterns are:

1. Positive: positive elements when taken out of context. E.g. *good* and *well done*.
2. Negative: negative elements when taken out of context. E.g. *bad* and *terrible*.
3. Amplifier: amplifying the polarity, positive or negative. E.g. *very much* and *a lot*.
4. Attenuator: Weakening the polarity, positive or negative. E.g. *a little* and a *tiny bit*.
5. Right flip: Flips the polarity of n elements to its right. E.g. *not* and no.
6. Left flip: Flips the polarity of n elements to its left. E.g. *but* and *however*.
7. Continuator: Continues the emission of polarity of elements. E.g. *and* and *also*.
8. Stop: Interrupts the continuous emission of polarity. E.g. full stops and exclamation marks.

The RBEM algorithm applies a set of sequential rules based on these rules to determine the polarity emission of each element in a case. The polarity of a case is determined by computing the sum of all emission values within a case (Tromp, 2011).

Appendix III Model design

Naïve Bayes

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="193" width="835">
      <operator activated="true" class="retrieve" compatibility="5.2.008" expanded="true" height="60"
name="Retrieve" width="90" x="45" y="30">
        <parameter key="repository_entry" value="../Cafepharma_sampleset"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes" width="90" x="180" y="30">
        <parameter key="attribute_filter_type" value="no_missing_values"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes (2)" width="90" x="313" y="30">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="id|text|Sentiment"/>
      </operator>
      <operator activated="true" class="text:process_document_from_data" compatibility="5.2.004"
expanded="true" height="76" name="Process Documents from Data" width="90" x="450" y="30">
        <parameter key="keep_text" value="true"/>
        <parameter key="prune_method" value="percentual"/>
        <parameter key="prunde_below_percent" value="1.0"/>
        <parameter key="prune_above_percent" value="90.0"/>
        <parameter key="prune_below_absolute" value="2"/>
        <parameter key="prune_above_absolute" value="999"/>
        <list key="specify_weights"/>
        <process expanded="true" height="210" width="567">
          <operator activated="true" class="text:transform_cases" compatibility="5.2.004" expanded="true"
height="60" name="Transform Cases (3)" width="90" x="45" y="30"/>
          <operator activated="true" class="text:replace_tokens" compatibility="5.2.004" expanded="true"
height="60" name="Replace Tokens (2)" width="90" x="179" y="30">
            <list key="replace_dictionary">
              <parameter key="i'm" value="I am"/>
              <parameter key="you're" value="you are"/>
              <parameter key="won't" value="will not"/>
              <parameter key="can't" value="cannot"/>
              <parameter key="i've" value="i have"/>
              <parameter key="haven't" value="have not"/>
              <parameter key="you've" value="you have"/>
            </list>
          </operator>
          <operator activated="true" class="text:tokenize" compatibility="5.2.004" expanded="true"
height="60" name="Tokenize (2)" width="90" x="313" y="30"/>
          <operator activated="true" class="text:filter_stopwords_english" compatibility="5.2.004"
expanded="true" height="60" name="Filter Stopwords (2)" width="90" x="45" y="120"/>
          <operator activated="true" class="text:filter_by_length" compatibility="5.2.004" expanded="true"
height="60" name="Filter Tokens (2)" width="90" x="179" y="120">
            <parameter key="min_chars" value="2"/>
            <parameter key="max_chars" value="40"/>
          </operator>
          <operator activated="true" class="text:stem_snowball" compatibility="5.2.004" expanded="true"
height="60" name="Stem (2)" width="90" x="313" y="120"/>
          <operator activated="true" class="text:generate_n_grams_terms" compatibility="5.2.004"
expanded="true" height="60" name="Generate n-Grams (Terms)" width="90" x="447" y="120"/>
          <connect from_port="document" to_op="Transform Cases (3)" to_port="document"/>
          <connect from_op="Transform Cases (3)" from_port="document" to_op="Replace Tokens (2)"
to_port="document"/>
          <connect from_op="Replace Tokens (2)" from_port="document" to_op="Tokenize (2)"
to_port="document"/>
```

```
        <connect from_op="Tokenize (2)" from_port="document" to_op="Filter Stopwords (2)"
to_port="document" />
        <connect from_op="Filter Stopwords (2)" from_port="document" to_op="Filter Tokens (2)"
to_port="document" />
        <connect from_op="Filter Tokens (2)" from_port="document" to_op="Stem (2)"
to_port="document" />
        <connect from_op="Stem (2)" from_port="document" to_op="Generate n-Grams (Terms)"
to_port="document" />
        <connect from_op="Generate n-Grams (Terms)" from_port="document" to_port="document 1" />
        <portSpacing port="source_document" spacing="0" />
        <portSpacing port="sink_document 1" spacing="0" />
        <portSpacing port="sink_document 2" spacing="0" />
      </process>
    </operator>
    <operator activated="true" class="set_role" compatibility="5.2.008" expanded="true" height="76"
name="Set Role" width="90" x="581" y="30">
      <parameter key="name" value="Sentiment" />
      <parameter key="target_role" value="label" />
      <list key="set_additional_roles" />
    </operator>
    <operator activated="true" class="x_validation" compatibility="5.2.008" expanded="true" height="112"
name="Validation" width="90" x="715" y="30">
      <process expanded="true" height="193" width="275">
        <operator activated="true" class="naive_bayes" compatibility="5.2.008" expanded="true"
height="76" name="Naive Bayes" width="90" x="112" y="30" />
        <connect from_port="training" to_op="Naive Bayes" to_port="training set" />
        <connect from_op="Naive Bayes" from_port="model" to_port="model" />
        <portSpacing port="source_training" spacing="0" />
        <portSpacing port="sink_model" spacing="0" />
        <portSpacing port="sink_through 1" spacing="0" />
      </process>
      <process expanded="true" height="193" width="342">
        <operator activated="true" class="apply_model" compatibility="5.2.008" expanded="true"
height="76" name="Apply Model" width="90" x="45" y="30">
          <list key="application_parameters" />
        </operator>
        <operator activated="true" class="performance" compatibility="5.2.008" expanded="true"
height="76" name="Performance" width="90" x="222" y="30" />
        <connect from_port="model" to_op="Apply Model" to_port="model" />
        <connect from_port="test set" to_op="Apply Model" to_port="unlabelled data" />
        <connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled
data" />
        <connect from_op="Performance" from_port="performance" to_port="averagable 1" />
        <portSpacing port="source_model" spacing="0" />
        <portSpacing port="source_test set" spacing="0" />
        <portSpacing port="source_through 1" spacing="0" />
        <portSpacing port="sink_averagable 1" spacing="0" />
        <portSpacing port="sink_averagable 2" spacing="0" />
      </process>
    </operator>
    <connect from_op="Retrieve" from_port="output" to_op="Select Attributes" to_port="example set
input" />
    <connect from_op="Select Attributes" from_port="example set output" to_op="Select Attributes (2)"
to_port="example set input" />
    <connect from_op="Select Attributes (2)" from_port="example set output" to_op="Process Documents
from Data" to_port="example set" />
    <connect from_op="Process Documents from Data" from_port="example set" to_op="Set Role"
to_port="example set input" />
    <connect from_op="Set Role" from_port="example set output" to_op="Validation" to_port="training" />
    <connect from_op="Validation" from_port="model" to_port="result 2" />
    <connect from_op="Validation" from_port="training" to_port="result 3" />
    <connect from_op="Validation" from_port="averagable 1" to_port="result 1" />
    <portSpacing port="source_input 1" spacing="0" />
    <portSpacing port="sink_result 1" spacing="0" />
    <portSpacing port="sink_result 2" spacing="0" />
    <portSpacing port="sink_result 3" spacing="0" />
    <portSpacing port="sink_result 4" spacing="0" />
  </process>
 </operator>
</process>
```

Nearest Neighbor

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="190" width="835">
      <operator activated="true" class="retrieve" compatibility="5.2.008" expanded="true" height="60"
name="Retrieve" width="90" x="45" y="30">
        <parameter key="repository_entry" value="../Cafepharma_sampleset"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes" width="90" x="180" y="30">
        <parameter key="attribute_filter_type" value="no_missing_values"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes (2)" width="90" x="315" y="30">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="id|text|Sentiment"/>
      </operator>
      <operator activated="true" class="text:process_document_from_data" compatibility="5.2.004"
expanded="true" height="76" name="Process Documents from Data" width="90" x="450" y="30">
        <parameter key="keep_text" value="true"/>
        <parameter key="prune_method" value="percentual"/>
        <parameter key="prunde_below_percent" value="1.0"/>
        <parameter key="prune_above_percent" value="90.0"/>
        <parameter key="prune_below_absolute" value="2"/>
        <parameter key="prune_above_absolute" value="999"/>
        <list key="specify_weights"/>
        <process expanded="true" height="210" width="567">
          <operator activated="true" class="text:transform_cases" compatibility="5.2.004" expanded="true"
height="60" name="Transform Cases (3)" width="90" x="45" y="30"/>
          <operator activated="true" class="text:replace_tokens" compatibility="5.2.004" expanded="true"
height="60" name="Replace Tokens (2)" width="90" x="179" y="30">
            <list key="replace_dictionary">
              <parameter key="i'm" value="I am"/>
              <parameter key="you're" value="you are"/>
              <parameter key="won't" value="will not"/>
              <parameter key="can't" value="cannot"/>
              <parameter key="i've" value="i have"/>
              <parameter key="haven't" value="have not"/>
              <parameter key="you've" value="you have"/>
            </list>
          </operator>
          <operator activated="true" class="text:tokenize" compatibility="5.2.004" expanded="true"
height="60" name="Tokenize (2)" width="90" x="313" y="30"/>
          <operator activated="true" class="text:filter_stopwords_english" compatibility="5.2.004"
expanded="true" height="60" name="Filter Stopwords (2)" width="90" x="45" y="120"/>
          <operator activated="true" class="text:filter_by_length" compatibility="5.2.004" expanded="true"
height="60" name="Filter Tokens (2)" width="90" x="179" y="120">
            <parameter key="min_chars" value="2"/>
            <parameter key="max_chars" value="40"/>
          </operator>
          <operator activated="true" class="text:stem_snowball" compatibility="5.2.004" expanded="true"
height="60" name="Stem (2)" width="90" x="313" y="120"/>
          <operator activated="true" class="text:generate_n_grams_terms" compatibility="5.2.004"
expanded="true" height="60" name="Generate n-Grams (Terms)" width="90" x="447" y="120"/>
          <connect from_port="document" to_op="Transform Cases (3)" to_port="document"/>
          <connect from_op="Transform Cases (3)" from_port="document" to_op="Replace Tokens (2)"
to_port="document"/>
          <connect from_op="Replace Tokens (2)" from_port="document" to_op="Tokenize (2)"
to_port="document"/>
          <connect from_op="Tokenize (2)" from_port="document" to_op="Filter Stopwords (2)"
to_port="document"/>
          <connect from_op="Filter Stopwords (2)" from_port="document" to_op="Filter Tokens (2)"
to_port="document"/>
```

```
      <connect from_op="Filter Tokens (2)" from_port="document" to_op="Stem (2)"
to_port="document" />
      <connect from_op="Stem (2)" from_port="document" to_op="Generate n-Grams (Terms)"
to_port="document" />
      <connect from_op="Generate n-Grams (Terms)" from_port="document" to_port="document 1" />
      <portSpacing port="source_document" spacing="0" />
      <portSpacing port="sink_document 1" spacing="0" />
      <portSpacing port="sink_document 2" spacing="0" />
     </process>
    </operator>
    <operator activated="true" class="set_role" compatibility="5.2.008" expanded="true" height="76"
name="Set Role" width="90" x="581" y="30">
     <parameter key="name" value="Sentiment" />
     <parameter key="target_role" value="label" />
     <list key="set_additional_roles" />
    </operator>
    <operator activated="true" class="optimize_parameters_grid" compatibility="5.2.008" expanded="true"
height="130" name="Optimize Parameters (Grid)" width="90" x="715" y="30">
     <list key="parameters">
      <parameter key="k-NN.k" value="[1.0;20;8;linear]" />
     </list>
     <process expanded="true" height="405" width="850">
      <operator activated="true" class="x_validation" compatibility="5.2.008" expanded="true"
height="112" name="Validation" width="90" x="338" y="198">
       <process expanded="true" height="405" width="400">
        <operator activated="true" class="k_nn" compatibility="5.2.008" expanded="true" height="76"
name="k-NN" width="90" x="155" y="30">
         <parameter key="k" value="20" />
         <parameter key="measure_types" value="NumericalMeasures" />
         <parameter key="numerical_measure" value="CosineSimilarity" />
        </operator>
        <connect from_port="training" to_op="k-NN" to_port="training set" />
        <connect from_op="k-NN" from_port="model" to_port="model" />
        <portSpacing port="source_training" spacing="0" />
        <portSpacing port="sink_model" spacing="0" />
        <portSpacing port="sink_through 1" spacing="0" />
       </process>
       <process expanded="true" height="405" width="400">
        <operator activated="true" class="apply_model" compatibility="5.2.008" expanded="true"
height="76" name="Apply Model" width="90" x="112" y="30">
         <list key="application_parameters" />
        </operator>
        <operator activated="true" class="performance" compatibility="5.2.008" expanded="true"
height="76" name="Performance" width="90" x="222" y="30" />
        <connect from_port="model" to_op="Apply Model" to_port="model" />
        <connect from_port="test set" to_op="Apply Model" to_port="unlabelled data" />
        <connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled
data" />
        <connect from_op="Performance" from_port="performance" to_port="averagable 1" />
        <portSpacing port="source_model" spacing="0" />
        <portSpacing port="source_test set" spacing="0" />
        <portSpacing port="source_through 1" spacing="0" />
        <portSpacing port="sink_averagable 1" spacing="0" />
        <portSpacing port="sink_averagable 2" spacing="0" />
       </process>
      </operator>
      <connect from_port="input 1" to_op="Validation" to_port="training" />
      <connect from_op="Validation" from_port="model" to_port="result 1" />
      <connect from_op="Validation" from_port="training" to_port="result 2" />
      <connect from_op="Validation" from_port="averagable 1" to_port="performance" />
      <portSpacing port="source_input 1" spacing="0" />
      <portSpacing port="source_input 2" spacing="0" />
      <portSpacing port="sink_performance" spacing="0" />
      <portSpacing port="sink_result 1" spacing="0" />
      <portSpacing port="sink_result 2" spacing="0" />
      <portSpacing port="sink_result 3" spacing="0" />
     </process>
    </operator>
    <connect from_op="Retrieve" from_port="output" to_op="Select Attributes" to_port="example set
input" />
```

```
    <connect from_op="Select Attributes" from_port="example set output" to_op="Select Attributes (2)"
to_port="example set input" />
    <connect from_op="Select Attributes (2)" from_port="example set output" to_op="Process Documents
from Data" to_port="example set" />
    <connect from_op="Process Documents from Data" from_port="example set" to_op="Set Role"
to_port="example set input" />
    <connect from_op="Set Role" from_port="example set output" to_op="Optimize Parameters (Grid)"
to_port="input 1" />
    <connect from_op="Optimize Parameters (Grid)" from_port="performance" to_port="result 2" />
    <connect from_op="Optimize Parameters (Grid)" from_port="parameter" to_port="result 1" />
    <connect from_op="Optimize Parameters (Grid)" from_port="result 1" to_port="result 3" />
    <portSpacing port="source_input 1" spacing="0" />
    <portSpacing port="sink_result 1" spacing="0" />
    <portSpacing port="sink_result 2" spacing="0" />
    <portSpacing port="sink_result 3" spacing="0" />
    <portSpacing port="sink_result 4" spacing="0" />
  </process>
 </operator>
</process>
```

Support Vector Machine (SVM)

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="449" width="835">
      <operator activated="true" class="retrieve" compatibility="5.2.008" expanded="true" height="60"
name="Retrieve" width="90" x="45" y="30">
        <parameter key="repository_entry" value="../Cafepharma_sampleset"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes" width="90" x="180" y="30">
        <parameter key="attribute_filter_type" value="no_missing_values"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes (2)" width="90" x="315" y="30">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="id|text|Sentiment"/>
      </operator>
      <operator activated="true" class="text:process_document_from_data" compatibility="5.2.004"
expanded="true" height="76" name="Process Documents from Data" width="90" x="450" y="30">
        <parameter key="keep_text" value="true"/>
        <parameter key="prune_method" value="percentual"/>
        <parameter key="prunde_below_percent" value="1.0"/>
        <parameter key="prune_above_percent" value="90.0"/>
        <parameter key="prune_below_absolute" value="2"/>
        <parameter key="prune_above_absolute" value="999"/>
        <list key="specify_weights"/>
        <process expanded="true" height="210" width="567">
          <operator activated="true" class="text:transform_cases" compatibility="5.2.004" expanded="true"
height="60" name="Transform Cases (3)" width="90" x="45" y="30"/>
          <operator activated="true" class="text:replace_tokens" compatibility="5.2.004" expanded="true"
height="60" name="Replace Tokens (2)" width="90" x="179" y="30">
            <list key="replace_dictionary">
              <parameter key="i'm" value="I am"/>
              <parameter key="you're" value="you are"/>
              <parameter key="won't" value="will not"/>
              <parameter key="can't" value="cannot"/>
              <parameter key="i've" value="i have"/>
              <parameter key="haven't" value="have not"/>
              <parameter key="you've" value="you have"/>
            </list>
          </operator>
          <operator activated="true" class="text:tokenize" compatibility="5.2.004" expanded="true"
height="60" name="Tokenize (2)" width="90" x="313" y="30"/>
          <operator activated="true" class="text:filter_stopwords_english" compatibility="5.2.004"
expanded="true" height="60" name="Filter Stopwords (2)" width="90" x="45" y="120"/>
          <operator activated="true" class="text:filter_by_length" compatibility="5.2.004" expanded="true"
height="60" name="Filter Tokens (2)" width="90" x="179" y="120">
            <parameter key="min_chars" value="2"/>
            <parameter key="max_chars" value="40"/>
          </operator>
          <operator activated="true" class="text:stem_snowball" compatibility="5.2.004" expanded="true"
height="60" name="Stem (2)" width="90" x="313" y="120"/>
          <operator activated="true" class="text:generate_n_grams_terms" compatibility="5.2.004"
expanded="true" height="60" name="Generate n-Grams (Terms)" width="90" x="447" y="120"/>
          <connect from_port="document" to_op="Transform Cases (3)" to_port="document"/>
          <connect from_op="Transform Cases (3)" from_port="document" to_op="Replace Tokens (2)"
to_port="document"/>
          <connect from_op="Replace Tokens (2)" from_port="document" to_op="Tokenize (2)"
to_port="document"/>
          <connect from_op="Tokenize (2)" from_port="document" to_op="Filter Stopwords (2)"
to_port="document"/>
          <connect from_op="Filter Stopwords (2)" from_port="document" to_op="Filter Tokens (2)"
to_port="document"/>
          <connect from_op="Filter Tokens (2)" from_port="document" to_op="Stem (2)"
to_port="document"/>
```

```
        <connect from_op="Stem (2)" from_port="document" to_op="Generate n-Grams (Terms)"
to_port="document" />
        <connect from_op="Generate n-Grams (Terms)" from_port="document" to_port="document 1" />
        <portSpacing port="source_document" spacing="0" />
        <portSpacing port="sink_document 1" spacing="0" />
        <portSpacing port="sink_document 2" spacing="0" />
      </process>
    </operator>
    <operator activated="true" class="set_role" compatibility="5.2.008" expanded="true" height="76"
name="Set Role" width="90" x="581" y="30">
      <parameter key="name" value="Sentiment" />
      <parameter key="target_role" value="label" />
      <list key="set_additional_roles" />
    </operator>
    <operator activated="true" class="optimize_parameters_grid" compatibility="5.2.008" expanded="true"
height="130" name="Optimize Parameters (Grid)" width="90" x="715" y="30">
      <list key="parameters">
        <parameter key="SVM.gamma" value="[0.0;15;5;linear]" />
        <parameter key="SVM.C" value="[0.0;15;5;linear]" />
      </list>
      <parameter key="parallelize_optimization_process" value="true" />
      <process expanded="true" height="405" width="850">
        <operator activated="true" class="x_validation" compatibility="5.2.008" expanded="true"
height="112" name="Validation" width="90" x="338" y="198">
          <process expanded="true" height="405" width="400">
            <operator activated="true" class="support_vector_machine_libsvm" compatibility="5.2.008"
expanded="true" height="76" name="SVM" width="90" x="179" y="30">
              <parameter key="gamma" value="15.0" />
              <parameter key="C" value="15.0" />
              <list key="class_weights" />
            </operator>
            <connect from_port="training" to_op="SVM" to_port="training set" />
            <connect from_op="SVM" from_port="model" to_port="model" />
            <portSpacing port="source_training" spacing="0" />
            <portSpacing port="sink_model" spacing="0" />
            <portSpacing port="sink_through 1" spacing="0" />
          </process>
          <process expanded="true" height="405" width="400">
            <operator activated="true" class="apply_model" compatibility="5.2.008" expanded="true"
height="76" name="Apply Model" width="90" x="112" y="30">
              <list key="application_parameters" />
            </operator>
            <operator activated="true" class="performance" compatibility="5.2.008" expanded="true"
height="76" name="Performance" width="90" x="222" y="30" />
            <connect from_port="model" to_op="Apply Model" to_port="model" />
            <connect from_port="test set" to_op="Apply Model" to_port="unlabelled data" />
            <connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled
data" />
            <connect from_op="Performance" from_port="performance" to_port="averagable 1" />
            <portSpacing port="source_model" spacing="0" />
            <portSpacing port="source_test set" spacing="0" />
            <portSpacing port="source_through 1" spacing="0" />
            <portSpacing port="sink_averagable 1" spacing="0" />
            <portSpacing port="sink_averagable 2" spacing="0" />
          </process>
        </operator>
        <connect from_port="input 1" to_op="Validation" to_port="training" />
        <connect from_op="Validation" from_port="model" to_port="result 1" />
        <connect from_op="Validation" from_port="training" to_port="result 2" />
        <connect from_op="Validation" from_port="averagable 1" to_port="performance" />
        <portSpacing port="source_input 1" spacing="0" />
        <portSpacing port="source_input 2" spacing="0" />
        <portSpacing port="sink_performance" spacing="0" />
        <portSpacing port="sink_result 1" spacing="0" />
        <portSpacing port="sink_result 2" spacing="0" />
        <portSpacing port="sink_result 3" spacing="0" />
      </process>
    </operator>
    <connect from_op="Retrieve" from_port="output" to_op="Select Attributes" to_port="example set
input" />
```

```
    <connect from_op="Select Attributes" from_port="example set output" to_op="Select Attributes (2)"
to_port="example set input" />
    <connect from_op="Select Attributes (2)" from_port="example set output" to_op="Process Documents
from Data" to_port="example set" />
    <connect from_op="Process Documents from Data" from_port="example set" to_op="Set Role"
to_port="example set input" />
    <connect from_op="Set Role" from_port="example set output" to_op="Optimize Parameters (Grid)"
to_port="input 1" />
    <connect from_op="Optimize Parameters (Grid)" from_port="performance" to_port="result 2" />
    <connect from_op="Optimize Parameters (Grid)" from_port="parameter" to_port="result 1" />
    <connect from_op="Optimize Parameters (Grid)" from_port="result 1" to_port="result 3" />
    <portSpacing port="source_input 1" spacing="0" />
    <portSpacing port="sink_result 1" spacing="0" />
    <portSpacing port="sink_result 2" spacing="0" />
    <portSpacing port="sink_result 3" spacing="0" />
    <portSpacing port="sink_result 4" spacing="0" />
  </process>
 </operator>
</process>
```

Dimensionality reduction through Singular Value Decomposition (SVD)

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="449" width="835">
      <operator activated="true" class="retrieve" compatibility="5.2.008" expanded="true" height="60"
name="Retrieve" width="90" x="45" y="30">
        <parameter key="repository_entry" value="../Cafepharma_balanced_cat2"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes" width="90" x="180" y="30">
        <parameter key="attribute_filter_type" value="no_missing_values"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes (2)" width="90" x="313" y="30">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="id|text|second tier category"/>
      </operator>
      <operator activated="true" class="text:process_document_from_data" compatibility="5.2.004"
expanded="true" height="76" name="Process Documents from Data" width="90" x="450" y="30">
        <parameter key="keep_text" value="true"/>
        <parameter key="prune_method" value="percentual"/>
        <parameter key="prunde_below_percent" value="1.0"/>
        <parameter key="prune_above_percent" value="90.0"/>
        <parameter key="prune_below_absolute" value="2"/>
        <parameter key="prune_above_absolute" value="999"/>
        <list key="specify_weights"/>
        <process expanded="true" height="210" width="567">
          <operator activated="true" class="text:transform_cases" compatibility="5.2.004" expanded="true"
height="60" name="Transform Cases (3)" width="90" x="45" y="30"/>
          <operator activated="true" class="text:replace_tokens" compatibility="5.2.004" expanded="true"
height="60" name="Replace Tokens (2)" width="90" x="179" y="30">
            <list key="replace_dictionary">
              <parameter key="i'm" value="I am"/>
              <parameter key="you're" value="you are"/>
              <parameter key="won't" value="will not"/>
              <parameter key="can't" value="cannot"/>
              <parameter key="i've" value="i have"/>
              <parameter key="haven't" value="have not"/>
              <parameter key="you've" value="you have"/>
            </list>
          </operator>
          <operator activated="true" class="text:tokenize" compatibility="5.2.004" expanded="true"
height="60" name="Tokenize (2)" width="90" x="313" y="30"/>
          <operator activated="true" class="text:filter_stopwords_english" compatibility="5.2.004"
expanded="true" height="60" name="Filter Stopwords (2)" width="90" x="45" y="120"/>
          <operator activated="true" class="text:filter_by_length" compatibility="5.2.004" expanded="true"
height="60" name="Filter Tokens (2)" width="90" x="179" y="120">
            <parameter key="min_chars" value="2"/>
            <parameter key="max_chars" value="40"/>
          </operator>
          <operator activated="true" class="text:stem_snowball" compatibility="5.2.004" expanded="true"
height="60" name="Stem (2)" width="90" x="313" y="120"/>
          <operator activated="true" class="text:generate_n_grams_terms" compatibility="5.2.004"
expanded="true" height="60" name="Generate n-Grams (Terms)" width="90" x="447" y="120"/>
          <connect from_port="document" to_op="Transform Cases (3)" to_port="document"/>
          <connect from_op="Transform Cases (3)" from_port="document" to_op="Replace Tokens (2)"
to_port="document"/>
          <connect from_op="Replace Tokens (2)" from_port="document" to_op="Tokenize (2)"
to_port="document"/>
          <connect from_op="Tokenize (2)" from_port="document" to_op="Filter Stopwords (2)"
to_port="document"/>
          <connect from_op="Filter Stopwords (2)" from_port="document" to_op="Filter Tokens (2)"
to_port="document"/>
          <connect from_op="Filter Tokens (2)" from_port="document" to_op="Stem (2)"
to_port="document"/>
```

```
      <connect from_op="Stem (2)" from_port="document" to_op="Generate n-Grams (Terms)"
to_port="document" />
      <connect from_op="Generate n-Grams (Terms)" from_port="document" to_port="document 1" />
      <portSpacing port="source_document" spacing="0" />
      <portSpacing port="sink_document 1" spacing="0" />
      <portSpacing port="sink_document 2" spacing="0" />
     </process>
    </operator>
    <operator activated="true" class="set_role" compatibility="5.2.008" expanded="true" height="76"
name="Set Role" width="90" x="581" y="30">
     <parameter key="name" value="second tier category" />
     <parameter key="target_role" value="label" />
     <list key="set_additional_roles" />
    </operator>
    <operator activated="true" class="optimize_parameters_grid" compatibility="5.2.008" expanded="true"
height="130" name="Optimize Parameters (Grid)" width="90" x="514" y="210">
     <list key="parameters">
      <parameter key="SVM.gamma" value="[1;15;2;linear]" />
      <parameter key="SVM.C" value="[1;15;2;linear]" />
      <parameter key="SVD.dimensions" value="[1;40;3;linear]" />
     </list>
     <parameter key="parallelize_optimization_process" value="true" />
     <process expanded="true" height="405" width="850">
      <operator activated="true" class="singular_value_decomposition" compatibility="5.2.008"
expanded="true" height="94" name="SVD" width="90" x="137" y="229">
       <parameter key="dimensions" value="40" />
      </operator>
      <operator activated="true" class="x_validation" compatibility="5.2.008" expanded="true"
height="112" name="Validation" width="90" x="338" y="198">
       <process expanded="true" height="405" width="400">
        <operator activated="true" class="support_vector_machine_libsvm" compatibility="5.2.008"
expanded="true" height="76" name="SVM" width="90" x="179" y="30">
         <parameter key="gamma" value="15.0" />
         <parameter key="C" value="15.0" />
         <list key="class_weights" />
        </operator>
        <connect from_port="training" to_op="SVM" to_port="training set" />
        <connect from_op="SVM" from_port="model" to_port="model" />
        <portSpacing port="source_training" spacing="0" />
        <portSpacing port="sink_model" spacing="0" />
        <portSpacing port="sink_through 1" spacing="0" />
       </process>
       <process expanded="true" height="405" width="400">
        <operator activated="true" class="apply_model" compatibility="5.2.008" expanded="true"
height="76" name="Apply Model" width="90" x="45" y="30">
         <list key="application_parameters" />
        </operator>
        <operator activated="true" class="performance" compatibility="5.2.008" expanded="true"
height="76" name="Performance" width="90" x="222" y="30" />
        <connect from_port="model" to_op="Apply Model" to_port="model" />
        <connect from_port="test set" to_op="Apply Model" to_port="unlabelled data" />
        <connect from_op="Apply Model" from_port="labelled data" to_op="Performance" to_port="labelled
data" />
        <connect from_op="Performance" from_port="performance" to_port="averagable 1" />
        <portSpacing port="source_model" spacing="0" />
        <portSpacing port="source_test set" spacing="0" />
        <portSpacing port="source_through 1" spacing="0" />
        <portSpacing port="sink_averagable 1" spacing="0" />
        <portSpacing port="sink_averagable 2" spacing="0" />
       </process>
      </operator>
      <connect from_port="input 1" to_op="SVD" to_port="example set input" />
      <connect from_op="SVD" from_port="example set output" to_op="Validation" to_port="training" />
      <connect from_op="Validation" from_port="model" to_port="result 1" />
      <connect from_op="Validation" from_port="training" to_port="result 2" />
      <connect from_op="Validation" from_port="averagable 1" to_port="performance" />
      <portSpacing port="source_input 1" spacing="0" />
      <portSpacing port="source_input 2" spacing="0" />
      <portSpacing port="sink_performance" spacing="0" />
      <portSpacing port="sink_result 1" spacing="0" />
      <portSpacing port="sink_result 2" spacing="0" />
```

```
        <portSpacing port="sink_result 3" spacing="0" />
      </process>
    </operator>
    <connect from_op="Retrieve" from_port="output" to_op="Select Attributes" to_port="example set
input" />
    <connect from_op="Select Attributes" from_port="example set output" to_op="Select Attributes (2)"
to_port="example set input" />
    <connect from_op="Select Attributes (2)" from_port="example set output" to_op="Process Documents
from Data" to_port="example set" />
    <connect from_op="Process Documents from Data" from_port="example set" to_op="Set Role"
to_port="example set input" />
    <connect from_op="Set Role" from_port="example set output" to_op="Optimize Parameters (Grid)"
to_port="input 1" />
    <connect from_op="Optimize Parameters (Grid)" from_port="performance" to_port="result 2" />
    <connect from_op="Optimize Parameters (Grid)" from_port="parameter" to_port="result 1" />
    <connect from_op="Optimize Parameters (Grid)" from_port="result 1" to_port="result 3" />
    <portSpacing port="source_input 1" spacing="0" />
    <portSpacing port="sink_result 1" spacing="0" />
    <portSpacing port="sink_result 2" spacing="0" />
    <portSpacing port="sink_result 3" spacing="0" />
    <portSpacing port="sink_result 4" spacing="0" />
  </process>
 </operator>
</process>
```

Association Rule Mining

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="475" width="832">
      <operator activated="true" class="read_excel" compatibility="5.2.008" expanded="true" height="60"
name="Read Excel" width="90" x="45" y="75">
        <parameter key="excel_file"
value="C:\Users\lennart.borst\Dropbox\Thesis\Crawler\Sample\sample.xlsx"/>
        <parameter key="imported_cell_range" value="A1:I985"/>
        <parameter key="first_row_as_names" value="false"/>
        <list key="annotations">
          <parameter key="0" value="Name"/>
        </list>
        <list key="data_set_meta_data_information">
          <parameter key="0" value="id.true.integer.id"/>
          <parameter key="1" value="text.true.text.attribute"/>
          <parameter key="2" value="label.true.polynominal.attribute"/>
          <parameter key="3" value="metadata_file.true.polynominal.attribute"/>
          <parameter key="4" value="metadata_path.false.polynominal.attribute"/>
          <parameter key="5" value="metadata_date.false.date_time.attribute"/>
          <parameter key="6" value="Category.true.nominal.attribute"/>
          <parameter key="7" value="second tier category.true.nominal.attribute"/>
          <parameter key="8" value="Sentiment.true.nominal.attribute"/>
        </list>
      </operator>
      <operator activated="true" class="generate_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Generate Attributes" width="90" x="179" y="75">
        <list key="function_descriptions">
          <parameter key="date" value="cut(metadata_file,0,10)"/>
          <parameter key="year" value="cut(date,6,4)"/>
        </list>
      </operator>
      <operator activated="true" class="nominal_to_date" compatibility="5.2.008" expanded="true"
height="76" name="Nominal to Date" width="90" x="313" y="75">
        <parameter key="attribute_name" value="date"/>
        <parameter key="date_format" value="MM-dd-yyyy"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes" width="90" x="447" y="75">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="Sentiment|second tier category"/>
      </operator>
      <operator activated="true" class="nominal_to_binominal" compatibility="5.2.008" expanded="true"
height="94" name="Nominal to Binominal" width="90" x="514" y="255"/>
      <operator activated="true" class="fp_growth" compatibility="5.2.008" expanded="true" height="76"
name="FP-Growth" width="90" x="581" y="120">
        <parameter key="min_support" value="0.05"/>
      </operator>
      <operator activated="true" class="create_association_rules" compatibility="5.2.008" expanded="true"
height="76" name="Create Association Rules" width="90" x="715" y="165">
        <parameter key="min_confidence" value="0.5"/>
      </operator>
      <connect from_op="Read Excel" from_port="output" to_op="Generate Attributes" to_port="example set
input"/>
      <connect from_op="Generate Attributes" from_port="example set output" to_op="Nominal to Date"
to_port="example set input"/>
      <connect from_op="Nominal to Date" from_port="example set output" to_op="Select Attributes"
to_port="example set input"/>
      <connect from_op="Select Attributes" from_port="example set output" to_op="Nominal to Binominal"
to_port="example set input"/>
      <connect from_op="Nominal to Binominal" from_port="example set output" to_op="FP-Growth"
to_port="example set"/>
      <connect from_op="FP-Growth" from_port="frequent sets" to_op="Create Association Rules"
to_port="item sets"/>
      <connect from_op="Create Association Rules" from_port="rules" to_port="result 1"/>
```

```
    <connect from_op="Create Association Rules" from_port="item sets" to_port="result 2" />
    <portSpacing port="source_input 1" spacing="0" />
    <portSpacing port="sink_result 1" spacing="0" />
    <portSpacing port="sink_result 2" spacing="0" />
    <portSpacing port="sink_result 3" spacing="0" />
  </process>
 </operator>
</process>
```

Association Rule Mining (with stock prices)

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="494" width="882">
      <operator activated="true" class="read_excel" compatibility="5.2.008" expanded="true" height="60"
name="Read Excel" width="90" x="45" y="120">
        <parameter key="excel_file"
value="C:\Users\lennart.borst\Dropbox\Thesis\Crawler\Sample\STOCKPRICE_DATA\complete_stock.xlsx"/>
        <parameter key="imported_cell_range" value="A1:J985"/>
        <parameter key="first_row_as_names" value="false"/>
        <list key="annotations">
          <parameter key="0" value="Name"/>
        </list>
        <list key="data_set_meta_data_information">
          <parameter key="0" value="id.true.integer.id"/>
          <parameter key="1" value="label.true.binominal.attribute"/>
          <parameter key="2" value="metadata_file.true.polynominal.attribute"/>
          <parameter key="3" value="Date.false.date_time.attribute"/>
          <parameter key="4" value="Category.true.nominal.attribute"/>
          <parameter key="5" value="second tier category.true.nominal.attribute"/>
          <parameter key="6" value="Sentiment.true.nominal.attribute"/>
          <parameter key="7" value="STOCK OPEN.true.numeric.attribute"/>
          <parameter key="8" value="STOCK CLOSE.true.numeric.attribute"/>
          <parameter key="9" value="STOCK CHANGE.true.nominal.attribute"/>
        </list>
      </operator>
      <operator activated="true" class="generate_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Generate Attributes" width="90" x="179" y="120">
        <list key="function_descriptions">
          <parameter key="date" value="cut(metadata_file,0,10)"/>
          <parameter key="year" value="cut(date,6,4)"/>
        </list>
      </operator>
      <operator activated="true" class="filter_examples" compatibility="5.2.008" expanded="true"
height="76" name="Filter Examples" width="90" x="313" y="120">
        <parameter key="condition_class" value="no_missing_attributes"/>
      </operator>
      <operator activated="true" class="nominal_to_date" compatibility="5.2.008" expanded="true"
height="76" name="Nominal to Date" width="90" x="447" y="120">
        <parameter key="attribute_name" value="date"/>
        <parameter key="date_format" value="MM-dd-yyyy"/>
      </operator>
      <operator activated="true" class="nominal_to_date" compatibility="5.2.008" expanded="true"
height="76" name="Nominal to Date (2)" width="90" x="581" y="120">
        <parameter key="attribute_name" value="year"/>
        <parameter key="date_format" value="yyyy"/>
      </operator>
      <operator activated="true" class="date_to_nominal" compatibility="5.2.008" expanded="true"
height="76" name="Date to Nominal" width="90" x="715" y="120">
        <parameter key="attribute_name" value="year"/>
        <parameter key="date_format" value="yyyy"/>
      </operator>
      <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes" width="90" x="313" y="255">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="STOCK CHANGE|year"/>
      </operator>
      <operator activated="true" class="nominal_to_binominal" compatibility="5.2.008" expanded="true"
height="94" name="Nominal to Binominal" width="90" x="447" y="255"/>
      <operator activated="true" class="fp_growth" compatibility="5.2.008" expanded="true" height="76"
name="FP-Growth" width="90" x="581" y="255">
        <parameter key="min_support" value="0.05"/>
      </operator>
      <operator activated="true" class="create_association_rules" compatibility="5.2.008" expanded="true"
height="76" name="Create Association Rules" width="90" x="715" y="255">
```

```
      <parameter key="min_confidence" value="0.5" />
    </operator>
    <connect from_op="Read Excel" from_port="output" to_op="Generate Attributes" to_port="example set
input" />
    <connect from_op="Generate Attributes" from_port="example set output" to_op="Filter Examples"
to_port="example set input" />
    <connect from_op="Filter Examples" from_port="example set output" to_op="Nominal to Date"
to_port="example set input" />
    <connect from_op="Nominal to Date" from_port="example set output" to_op="Nominal to Date (2)"
to_port="example set input" />
    <connect from_op="Nominal to Date (2)" from_port="example set output" to_op="Date to Nominal"
to_port="example set input" />
    <connect from_op="Date to Nominal" from_port="example set output" to_op="Select Attributes"
to_port="example set input" />
    <connect from_op="Select Attributes" from_port="example set output" to_op="Nominal to Binominal"
to_port="example set input" />
    <connect from_op="Nominal to Binominal" from_port="example set output" to_op="FP-Growth"
to_port="example set" />
    <connect from_op="FP-Growth" from_port="frequent sets" to_op="Create Association Rules"
to_port="item sets" />
    <connect from_op="Create Association Rules" from_port="rules" to_port="result 1" />
    <connect from_op="Create Association Rules" from_port="item sets" to_port="result 2" />
    <portSpacing port="source_input 1" spacing="0" />
    <portSpacing port="sink_result 1" spacing="0" />
    <portSpacing port="sink_result 2" spacing="0" />
    <portSpacing port="sink_result 3" spacing="0" />
  </process>
 </operator>
</process>
```

Association Rule Mining (with stock prices, with delay)

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.2.008">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.2.008" expanded="true" name="Process">
    <process expanded="true" height="449" width="815">
      <operator activated="true" class="subprocess" compatibility="5.2.008" expanded="true" height="76"
name="Join_and_Delay" width="90" x="45" y="30">
        <process expanded="true" height="562" width="1020">
          <operator activated="true" class="read_excel" compatibility="5.2.008" expanded="true"
height="60" name="Read Excel" width="90" x="45" y="30">
            <parameter key="excel_file"
value="C:\Users\lennart.borst\Dropbox\Thesis\Crawler\Sample\STOCKPRICE_DATA\STOCK PRICE
DUMP\COMPLETE_DATA.xlsx"/>
            <parameter key="imported_cell_range" value="A1:H985"/>
            <parameter key="first_row_as_names" value="false"/>
            <list key="annotations">
              <parameter key="0" value="Name"/>
            </list>
            <list key="data_set_meta_data_information">
              <parameter key="0" value="id.true.integer.attribute"/>
              <parameter key="1" value="label.true.polynominal.attribute"/>
              <parameter key="2" value="metadata_file.true.polynominal.attribute"/>
              <parameter key="3" value="Date.true.date.attribute"/>
              <parameter key="4" value="Month-Year.false.nominal.attribute"/>
              <parameter key="5" value="Category.true.nominal.attribute"/>
              <parameter key="6" value="second tier category.true.nominal.attribute"/>
              <parameter key="7" value="Sentiment.true.nominal.attribute"/>
            </list>
          </operator>
          <operator activated="true" class="date_to_nominal" compatibility="5.2.008" expanded="true"
height="76" name="Date to Nominal" width="90" x="246" y="30">
            <parameter key="attribute_name" value="Date"/>
            <parameter key="date_format" value="MM-yyyy"/>
          </operator>
          <operator activated="true" class="generate_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Generate Attributes" width="90" x="380" y="30">
            <list key="function_descriptions">
              <parameter key="OrgDateID" value="label + &amp;quot; - &amp;quot; + Date"/>
            </list>
          </operator>
          <operator activated="true" class="set_role" compatibility="5.2.008" expanded="true" height="76"
name="Set Role" width="90" x="514" y="75">
            <parameter key="name" value="OrgDateID"/>
            <parameter key="target_role" value="id"/>
            <list key="set_additional_roles"/>
          </operator>
          <operator activated="true" class="read_excel" compatibility="5.2.008" expanded="true"
height="60" name="Read Excel (2)" width="90" x="45" y="390">
            <parameter key="excel_file"
value="C:\Users\lennart.borst\Dropbox\Thesis\Crawler\Sample\STOCKPRICE_DATA\STOCK PRICE
DUMP\monthly\STOCKPRICE_DUMP_MONTH.xlsx"/>
            <parameter key="imported_cell_range" value="A1:H510"/>
            <parameter key="first_row_as_names" value="false"/>
            <list key="annotations">
              <parameter key="0" value="Name"/>
            </list>
            <list key="data_set_meta_data_information">
              <parameter key="0" value="Company.true.polynominal.attribute"/>
              <parameter key="1" value="Date.true.date_time.attribute"/>
              <parameter key="2" value="Open.true.numeric.attribute"/>
              <parameter key="3" value="High.false.numeric.attribute"/>
              <parameter key="4" value="Low.false.numeric.attribute"/>
              <parameter key="5" value="Close.true.real.attribute"/>
              <parameter key="6" value="Volume.false.integer.attribute"/>
              <parameter key="7" value="Adj Close.false.numeric.attribute"/>
            </list>
```

```
        </operator>
        <operator activated="true" class="adjust_date" compatibility="5.2.008" expanded="true"
height="76" name="Adjust Date" width="90" x="179" y="390">
          <description>This opperator set the "delay"
CurrentIt +1, meaning that the relation between post of this month are related to the stock prices of last
month</description>
          <parameter key="attribute_name" value="Date" />
          <list key="adjustments">
            <parameter key="1" value="Month" />
          </list>
        </operator>
        <operator activated="true" class="date_to_nominal" compatibility="5.2.008" expanded="true"
height="76" name="Date to Nominal (2)" width="90" x="313" y="390">
          <parameter key="attribute_name" value="Date" />
          <parameter key="date_format" value="MM-yyyy" />
        </operator>
        <operator activated="true" class="generate_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Generate Attributes (2)" width="90" x="447" y="390">
          <list key="function_descriptions">
            <parameter key="OrgDateID" value="Company + &amp;quot; - &amp;quot; + Date" />
          </list>
        </operator>
        <operator activated="true" class="set_role" compatibility="5.2.008" expanded="true" height="76"
name="Set Role (2)" width="90" x="648" y="390">
          <parameter key="name" value="OrgDateID" />
          <parameter key="target_role" value="id" />
          <list key="set_additional_roles" />
        </operator>
        <operator activated="true" class="join" compatibility="5.2.008" expanded="true" height="76"
name="Join" width="90" x="782" y="210">
          <parameter key="join_type" value="left" />
          <list key="key_attributes" />
        </operator>
        <connect from_op="Read Excel" from_port="output" to_op="Date to Nominal" to_port="example set
input" />
        <connect from_op="Date to Nominal" from_port="example set output" to_op="Generate Attributes"
to_port="example set input" />
        <connect from_op="Generate Attributes" from_port="example set output" to_op="Set Role"
to_port="example set input" />
        <connect from_op="Set Role" from_port="example set output" to_op="Join" to_port="left" />
        <connect from_op="Read Excel (2)" from_port="output" to_op="Adjust Date" to_port="example set
input" />
        <connect from_op="Adjust Date" from_port="example set output" to_op="Date to Nominal (2)"
to_port="example set input" />
        <connect from_op="Date to Nominal (2)" from_port="example set output" to_op="Generate Attributes
(2)" to_port="example set input" />
        <connect from_op="Generate Attributes (2)" from_port="example set output" to_op="Set Role (2)"
to_port="example set input" />
        <connect from_op="Set Role (2)" from_port="example set output" to_op="Join" to_port="right" />
        <connect from_op="Join" from_port="join" to_port="out 1" />
        <portSpacing port="source_in 1" spacing="0" />
        <portSpacing port="sink_out 1" spacing="0" />
        <portSpacing port="sink_out 2" spacing="0" />
      </process>
    </operator>
    <operator activated="true" class="subprocess" compatibility="5.2.008" expanded="true" height="76"
name="aggregate_months" width="90" x="179" y="30">
      <process expanded="true" height="437" width="614">
        <operator activated="true" class="nominal_to_numerical" compatibility="5.2.008" expanded="true"
height="94" name="Nominal to Numerical" width="90" x="313" y="210">
          <parameter key="attribute_filter_type" value="subset" />
          <parameter key="attributes" value="Category|Sentiment|second tier category" />
          <list key="comparison_groups" />
        </operator>
        <operator activated="true" class="aggregate" compatibility="5.2.008" expanded="true" height="76"
name="Aggregate" width="90" x="447" y="210">
          <list key="aggregation_attributes">
            <parameter key="Sentiment" value="mode" />
            <parameter key="Category" value="mode" />
            <parameter key="second tier category" value="mode" />
            <parameter key="Open" value="average" />
```

```
          <parameter key="Close" value="average" />
        </list>
        <parameter key="group_by_attributes" value="Date" />
      </operator>
      <connect from_port="in 1" to_op="Nominal to Numerical" to_port="example set input" />
      <connect from_op="Nominal to Numerical" from_port="example set output" to_op="Aggregate"
to_port="example set input" />
      <connect from_op="Aggregate" from_port="example set output" to_port="out 1" />
      <portSpacing port="source_in 1" spacing="0" />
      <portSpacing port="source_in 2" spacing="0" />
      <portSpacing port="sink_out 1" spacing="0" />
      <portSpacing port="sink_out 2" spacing="0" />
    </process>
  </operator>
  <operator activated="true" class="subprocess" compatibility="5.2.008" expanded="true" height="76"
name="recoding_STOCK_SENTIMENT_AND_CATEGORIES" width="90" x="313" y="30">
    <process expanded="true" height="437" width="500">
      <operator activated="true" class="rename" compatibility="5.2.008" expanded="true" height="76"
name="Rename" width="90" x="246" y="30">
        <parameter key="old_name" value="average(Close)" />
        <parameter key="new_name" value="Close" />
        <list key="rename_additional_attributes">
          <parameter key="average(Open)" value="Open" />
          <parameter key="mode(Category)" value="tempCat" />
          <parameter key="mode(second tier category)" value="tempCat2" />
          <parameter key="mode(Sentiment)" value="tempSent" />
        </list>
      </operator>
      <operator activated="true" class="generate_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Generate Attributes (3)" width="90" x="380" y="30">
        <list key="function_descriptions">
          <parameter key="STOCK_CHANGE" value="if(Close - Open &gt; 0, &quot;increase&quot;, if(Close -
Open &lt; 0, &quot;decrease&quot;, &quot;level&quot;))" />
          <parameter key="Category" value="if(tempCat &gt;= 0 &amp;&amp; tempCat &lt; 0.5,
&quot;Flame&quot;, if(tempCat &gt;= 0.5 &amp;&amp; tempCat &lt; 1.5, &quot;social related&quot;,
if(tempCat &gt;= 1.5 &amp;&amp; tempCat &lt; 2.5, &quot;Task related&quot;, &quot;Off topic&quot;)))" />
          <parameter key="Sentiment" value="if(tempSent &gt;= 0 &amp;&amp; tempSent &lt; 0.5,
&quot;Negative&quot;, if(tempSent &gt;= 0.5 &amp;&amp; tempSent &lt; 1.5, &quot;Objective&quot;,
&quot;Positive&quot;))" />
          <parameter key="second tier category" value="if(tempCat2 &gt;= 0 &amp;&amp; tempCat2 &lt;
0.5, &quot;flame&quot;, if(tempCat2 &gt;= 0.5 &amp;&amp; tempCat2 &lt; 1.5, &quot;social: person&quot;,
if(tempCat2 &gt;= 1.5 &amp;&amp; tempCat2 &lt; 2.5, &quot;task: product&quot;, if(tempCat2 &gt;= 2.5
&amp;&amp; tempCat2 &lt; 3.5, &quot;social: performance&quot;, if(tempCat2 &gt;= 3.5 &amp;&amp;
tempCat2 &lt; 4.5, &quot;task: new product&quot;, if(tempCat2 &gt;= 4.5 &amp;&amp; tempCat2 &lt; 5.5,
&quot;social: organization&quot;, if(tempCat2 &gt;= 5.5 &amp;&amp; tempCat2 &lt; 6.5, &quot;off
topic&quot;, if(tempCat2 &gt;= 6.5 &amp;&amp; tempCat2 &lt; 7.5, &quot;social: personal effects&quot;,
if(tempCat2 &gt;=7.5 &amp;&amp; tempCat2 &lt; 8.5, &quot;task: sales technique&quot;, &quot;task:
market developments&quot;)))))))))" />
        </list>
      </operator>
      <connect from_port="in 1" to_op="Rename" to_port="example set input" />
      <connect from_op="Rename" from_port="example set output" to_op="Generate Attributes (3)"
to_port="example set input" />
      <connect from_op="Generate Attributes (3)" from_port="example set output" to_port="out 1" />
      <portSpacing port="source_in 1" spacing="0" />
      <portSpacing port="source_in 2" spacing="0" />
      <portSpacing port="sink_out 1" spacing="0" />
      <portSpacing port="sink_out 2" spacing="0" />
    </process>
  </operator>
  <operator activated="true" class="select_attributes" compatibility="5.2.008" expanded="true"
height="76" name="Select Attributes" width="90" x="447" y="30">
    <parameter key="attribute_filter_type" value="subset" />
    <parameter key="attributes" value="STOCK_CHANGE|Sentiment" />
  </operator>
  <operator activated="true" class="nominal_to_binominal" compatibility="5.2.008" expanded="true"
height="94" name="Nominal to Binominal" width="90" x="447" y="165">
    <parameter key="attributes" value="STOCK_CHANGE|Sentiment|" />
    <parameter key="include_special_attributes" value="true" />
    <parameter key="transform_binominal" value="true" />
    <parameter key="use_underscore_in_name" value="true" />
```

```
    </operator>
    <operator activated="true" class="fp_growth" compatibility="5.2.008" expanded="true" height="76"
name="FP-Growth" width="90" x="581" y="165">
      <parameter key="min_support" value="0.05" />
    </operator>
    <operator activated="true" class="create_association_rules" compatibility="5.2.008" expanded="true"
height="76" name="Create Association Rules" width="90" x="715" y="210">
      <parameter key="min_confidence" value="0.5" />
    </operator>
    <connect from_op="Join_and_Delay" from_port="out 1" to_op="aggregate_months" to_port="in 1" />
    <connect from_op="aggregate_months" from_port="out 1"
to_op="recoding_STOCK_SENTIMENT_AND_CATEGORIES" to_port="in 1" />
    <connect from_op="recoding_STOCK_SENTIMENT_AND_CATEGORIES" from_port="out 1" to_op="Select
Attributes" to_port="example set input" />
    <connect from_op="Select Attributes" from_port="example set output" to_op="Nominal to Binominal"
to_port="example set input" />
    <connect from_op="Nominal to Binominal" from_port="example set output" to_op="FP-Growth"
to_port="example set" />
    <connect from_op="FP-Growth" from_port="example set" to_port="result 1" />
    <connect from_op="FP-Growth" from_port="frequent sets" to_op="Create Association Rules"
to_port="item sets" />
    <connect from_op="Create Association Rules" from_port="rules" to_port="result 2" />
    <portSpacing port="source_input 1" spacing="0" />
    <portSpacing port="sink_result 1" spacing="0" />
    <portSpacing port="sink_result 2" spacing="0" />
    <portSpacing port="sink_result 3" spacing="0" />
  </process>
 </operator>
</process>
```

Appendix IV Model performance

In this section the model performance for each modeling technique is presented for the category, the second level category and the sentiment level. This is done with both a unbalanced data set (all the data) and a balanced data set.

Naïve Bayes

Results for first level categories

The following are the results from the first level categories (Flame, Off-topic, Social related and Task related), with both an unbalanced and a balanced sample.

Accuracy: 55,91% ± 4,08% (mikro: 55,89%)					
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	5	2	10	0,00%
pred. Social related	15	461	115	166	60,90%
pred. Task related	0	42	30	32	28,85%
pred. Off topic	1	36	10	59	55,66%
class recall	0,00%	84,74%	19,11%	22,10%	

Table 33 Naïve Bayes for first level category (unbalanced)

Accuracy: 42,76% ± 5,19% (mikro: 42,76%)					
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	4	1	9	0,00%
pred. Social related	10	115	57	92	41,97%
pred. Task related	4	55	90	59	43,27%
pred. Off topic	2	9	26	40	51,95%
class recall	0,00%	57,50%	57,32%	20,00%	

Table 34 Naïve Bayes for first level category (balanced)

Results for second level categories

The following are the results from the second level categories (Flame, Off-topic, Social: Performance, Social: Organization, Social: Personal effects, Social: Person, Task: Product, Task: New Product, Task: Sales techniques and Task: Market development), with both an unbalanced and a balanced sample.

Accuracy: 30,89% ± 4,15% (mikro: 30,89%)											
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	0	0	0	0	5	8	0	0	0	0,00%
pred. Social: Person	1	16	2	0	1	16	6	1	3	0	35,56%
pred. Task: product	0	3	2	0	1	1	9	0	0	1	11,76%
pred. Social: Performance	0	1	0	0	0	3	2	2	1	0	0,00%
pred. Task: New product	1	4	4	0	0	2	3	0	0	1	0,00%
pred. Social: Organization	9	72	31	23	16	176	103	64	32	24	32,00%
pred. Off topic	4	20	5	4	2	33	83	14	5	5	47,43%
pred. Social: Personal effects	1	16	4	2	2	25	41	27	9	3	20,77%
pred. Task: Sales Technique	0	3	1	0	0	2	10	2	0	0	0,00%
pred. Task: Market developments	0	2	3	0	0	4	2	0	0	0	0,00%
class recall	0,00%	11,59%	3,85%	0,00%	0,00%	65,92%	31,09%	24,55%	0,00%	0,00%	

Table 35 Naïve Bayes for second level category (unbalanced)

Accuracy: 22,91% ± 5,55% (mikro: 22,91%)											
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	1	0	0	0	2	2	0	0	0	0,00%
pred. Social: Person	0	8	2	0	0	4	2	2	1	1	40,00%
pred. Task: product	0	5	11	1	6	2	7	1	1	5	28,21%
pred. Social: Performance	1	4	3	2	1	1	3	1	4	0	10,00%
pred. Task: New product	0	1	6	0	0	4	2	0	0	1	0,00%
pred. Social: Organization	3	9	3	3	2	12	8	3	8	2	22,64%
pred. Off topic	2	5	3	0	1	1	9	2	1	1	36,00%
pred. Social: Personal effects	10	30	15	17	7	34	28	54	23	13	23,38%
pred. Task: Sales Technique	0	10	4	6	1	10	10	10	11	3	16,92%
pred. Task: Market developments	0	2	5	0	3	5	4	2	1	8	26,67%
class recall	0,00%	10,67%	21,15%	6,90%	0,00%	16,00%	12,00%	71,00%	22,00%	23,53%	

Table 36 Naïve Bayes for second level category (balanced)

Results for sentiment analysis

The following are the results from the sentiment (Objective, Negative and Positive), with both an unbalanced and a balanced sample.

Accuracy: 63,52% ± 2,00% (mikro: 63,52%)				
	true Negative	true Objective	true Positive	class precision
pred. Negative	581	205	76	67,40%
pred. Objective	23	42	8	57,53%
pred. Positive	8	39	2	4,08%
class recall	94,93%	14,69%	2,33%	

Table 37 Naïve Bayes for sentiment (unbalanced)

Accuracy: 51,03% ± 4,09% (mikro: 51,04%)				
	true Negative	true Objective	true Positive	class precision
pred. Negative	275	187	74	51,31%
pred. Objective	20	64	8	69,67%
pred. Positive	5	35	4	9,09%
class recall	91,67%	22,38%	4,65%	

Table 38 Naïve Bayes for sentiment (balanced)

K Nearest Neighbor

Results for first level categories

The following are the results from the first level categories (Flame, Off-topic, Social related and Task related), with both an unbalanced and a balanced sample.

Accuracy: 64,63% ± 3,58% (mikro: 64,63%)			**K = 15**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	12	508	82	190	64,14%
pred. Task related	2	15	63	12	68,48%
pred. Off topic	2	21	12	65	65,00%
class recall	0,00%	93,38%	40,13%	24,24%	

Table 39 k-NN for first level category, k=13 (unbalanced)

Accuracy: 57,41% ± 6,22% (mikro: 57,42%)			**K = 18**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	6	134	27	87	52,76%
pred. Task related	8	42	118	36	57,84%
pred. Off topic	2	24	12	77	66,96%
class recall	0,00%	67,00%	75,16%	38,50%	

Table 40 k-NN for first level category, k=18 (balanced)

Results for second level categories

The following are the results from the second level categories (Flame, Off-topic, Social: Performance, Social: Organization, Social: Personal effects, Social: Person, Task: Product, Task: New Product, Task: Sales techniques and Task: Market development), with both an unbalanced and a balanced sample.

Accuracy: 44,62% ± 4,25% (mikro: 44,61%)							**K = 15**				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	0	1	0	0	0	0	0	0	0	0,00%
pred. Social: Person	1	38	4	0	1	12	10	5	2	1	51,35%
pred. Task: product	2	3	16	1	7	2	4	0	2	3	40,00%
pred. Social: Performance	0	0	0	1	0	2	0	0	1	0	25,00%
pred. Task: New product	1	0	3	0	2	0	1	0	0	0	28,57%
pred. Social: Organization	4	66	18	18	8	199	87	57	24	15	40,12%
pred. Off topic	6	23	9	5	2	33	134	13	8	7	55,83%
pred. Social: Personal effects	2	7	0	1	0	16	20	32	3	1	39,02%
pred. Task: Sales Technique	0	1	0	3	1	2	8	2	10	0	37,04%
pred. Task: Market developments	0	0	1	0	0	1	3	1	0	7	53,85%
class recall	0,00%	27,54%	30,77%	3,45%	9,52%	74,53%	50,19%	29,09%	20,00%	20,59%	

Table 41 k-NN for second level category, k=13 (unbalanced)

Accuracy: 36,67% ± 6,55% (mikro: 36,65%)							**K = 18**				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	0	1	0	1	0	1	0	0	0	0,00%
pred. Social: Person	1	15	1	1	0	1	2	2	0	0	65,22%
pred. Task: product	3	5	23	1	7	2	5	1	3	10	38,33%
pred. Social: Performance	2	4	0	3	0	5	1	3	2	1	14,29%
pred. Task: New product	1	1	4	0	6	0	0	0	0	1	46,15%
pred. Social: Organization	0	9	6	2	1	18	7	7	5	1	32,14%
pred. Off topic	3	9	3	2	0	6	22	2	2	1	44,00%
pred. Social: Personal effects	4	21	5	15	2	29	23	53	7	5	32,32%
pred. Task: Sales Technique	2	10	4	5	1	12	10	6	31	2	37,35%
pred. Task: Market developments	0	1	5	0	3	2	4	1	0	13	44,83%
class recall	0,00%	20,00%	44,23%	10,34%	28,57%	24,00%	29,33%	70,67%	62,00%	38,24%	

Table 42 k-NN for second level category, k=18 (balanced)

Results for sentiment analysis

The following are the results from the sentiment (Objective, Negative and Positive), with both an unbalanced and a balanced sample.

Accuracy: 64,43% ± 1,77% (mikro: 64,43%)			**K = 11**	
	true Negative	true Objective	true Positive	class precision
pred. Negative	582	236	74	63,25%
pred. Objective	29	50	10	56,18%
pred. Positive	1	0	2	66,67%
class recall	95,10%	17,48%	2,33%	

Table 43 k-NN for sentiment, k=11 (unbalanced)

Accuracy: 54,48% ± 5,74% (mikro: 54,46%)			**K = 6**	
	true Negative	true Objective	true Positive	class precision
pred. Negative	226	139	55	53,81%
pred. Objective	70	134	25	58,52%
pred. Positive	4	13	6	26,09%
class recall	75,33%	56,85%	6,98%	

Table 44 k-NN for sentiment, k=6 (balanced)

SVM

Results for first level categories

The following are the results from the first level categories (Flame, Off-topic, Social related and Task related), with both an unbalanced and a balanced sample.

Accuracy: 55,59% ± 0,50% (mikro: 55,59%)			**γ = 6,0 & C = 3,0**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	16	543	155	265	55,46%
pred. Task related	0	0	2	0	100,00%
pred. Off topic	0	1	0	2	66,67%
class recall	0,00%	99,82%	1,27%	0,75%	

Table 45 SVM for first level category, γ=6,0 & C=3,0 (unbalanced)

Accuracy: 45,38% ± 3,19% (mikro: 45,38%)			**γ = 3,0 & C = 9,0**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	5	110	77	54	44,72%
pred. Task related	0	1	6	2	66,67%
pred. Off topic	11	89	74	144	45,28%
class recall	0,00%	55,00%	3,82%	72,00%	

Table 46 SVM for first level category, γ=3,0 & C=9,0 (balanced)

Results for second level categories

The following are the results from the second level categories (Flame, Off-topic, Social: Performance, Social: Organization, Social: Personal effects, Social: Person, Task: Product, Task: New Product, Task: Sales techniques and Task: Market development), with both an unbalanced and a balanced sample.

Accuracy: 39,53% ± 3,21% (mikro: 39,53%)							**γ = 3,0 & C = 6,0**				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	0	1	0	0	0	0	0	0	0	0,00%
pred. Social: Person	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: product	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Performance	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: New product	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Organization	3	52	24	21	9	173	53	67	28	20	38,44%
pred. Off topic	13	86	28	8	12	94	214	43	22	12	40,23%
pred. Social: Personal effects	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: Sales Technique	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: Market developments	0	0	0	0	0	0	0	0	0	2	100,00%
class recall	0,00%	0,00%	0,00%	0,00%	0,00%	64,79%	80,15%	0,00%	0,00%	5,88%	

Table 47 SVM for second level category, γ=3,0 & C=6,0 (unbalanced)

Accuracy: 29,29% ± 4,57% (mikro: 29,28%)							**γ = 3,0 & C = 12,0**				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Person	7	36	13	5	3	13	17	14	6	4	30,51%
pred. Task: product	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Performance	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: New product	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Social: Organization	2	18	17	8	9	30	16	10	18	14	21,13%
pred. Off topic	4	11	16	3	7	21	33	5	13	10	26,83%
pred. Social: Personal effects	3	10	6	13	2	11	9	46	13	4	39,32%
pred. Task: Sales Technique	0	0	0	0	0	0	0	0	0	0	0,00%
pred. Task: Market developments	0	0	0	0	0	0	0	0	0	2	100,00%
class recall	0,00%	48,00%	0,00%	0,00%	0,00%	40,00%	44,00%	61,33%	0,00%	5,88%	

Table 48 SVM for second level category, γ=3,0 & C=12,0 (balanced)

Results for sentiment analysis

The following are the results from the sentiment (Objective, Negative and Positive), with both an unbalanced and a balanced sample.

Accuracy: 62,60% ± 0,87% (mikro: 62,60%)			**γ = 6,0 & C = 6,0**	
	true Negative	true Objective	true Positive	class precision
pred. Negative	612	282	86	62,45%
pred. Objective	0	4	0	100,00%
pred. Positive	0	0	0	0,00%
class recall	100,00%	1,40%	0,00%	

Table 49 SVM for sentiment, γ=6,0 & C=6,0 (unbalanced)

Accuracy: 61,91% ± 5,76% (mikro: 61,90%)			**γ = 3,0 & C = 12,0**	
	true Negative	true Objective	true Positive	class precision
pred. Negative	243	113	55	59,12%
pred. Objective	57	173	31	66,28%
pred. Positive	0	0	0	0,00%
class recall	81,00%	60,49%	0,00%	

Table 50 SVM for sentiment, γ=6,0 & C=6,0 (balanced)

Dimensionality reduction (SVD)

Results for first level categories (K-NN)

The following are the results from the first level categories (Flame, Off-topic, Social related and Task related), with a balanced sample.

Accuracy: 60,37% ± 6,79% (mikro: 60,38%)			**K = 15, Dimensions=40**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	6	129	31	70	54,66%
pred. Task related	3	24	110	23	68,75%
pred. Off topic	7	47	16	107	60,45%
class recall	0,00%	64,50%	70,06%	53,50%	

Table 51 k-NN for first level category with dimensionality reduction through SVD, k=15, number of dimensions=40(balanced)

Results for second level categories (K-NN)

The following are the results from the second level categories (Flame, Off-topic, Social: Performance, Social: Organization, Social: Personal effects, Social: Person, Task: Product, Task: New Product, Task: Sales techniques and Task: Market development), with a balanced sample.

Accuracy: 38,67% ± 5,58% (mikro: 38,66%)							**K = 20, Dimensions=20**				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	0	2	0	1	0	0	1	0	0	0	0,00%
pred. Social: Person	7	35	8	2	0	9	19	10	5	2	36,08%
pred. Task: product	1	4	22	1	11	5	1	0	5	8	37,93%
pred. Social: Performance	0	2	1	2	0	6	0	4	2	1	11,11%
pred. Task: New product	0	0	1	0	0	1	0	0	0	0	0,00%
pred. Social: Organization	0	10	8	6	3	21	5	3	5	3	32,31%
pred. Off topic	4	6	5	2	3	11	29	4	2	4	41,43%
pred. Social: Personal effects	2	9	3	12	1	17	8	47	4	3	44,34%
pred. Task: Sales Technique	2	6	2	3	1	4	9	5	27	1	45,00%
pred. Task: Market developments	0	1	2	0	2	1	3	2	0	11	50,00%
class recall	0,00%	46,67%	42,31%	6,90%	0,00%	28,00%	38,67%	62,67%	54,00%	32,35%	

Table 52 k-NN for second level category with dimensionality reduction through SVD, k=20, number of dimensions=20 (balanced)

Results for sentiment analysis (SVM)

The following are the results from the sentiment (Objective, Negative and Positive), with a balanced sample.

Accuracy: 61,59% ± 4,95% (mikro: 61,61%)			**γ = 15, C = 8 and Dimensions=14**	
	true Negative	true Objective	true Positive	class precision
pred. Negative	193	69	40	63,91%
pred. Objective	104	211	36	60,11%
pred. Positive	3	6	10	52,63%
class recall	64,33%	73,78%	11,63%	

Table 53 SVM for sentiment with dimensionality reduction through SVD, γ=15,0, C=8,0 and number of dimensions=14 (balanced)

Results for first level categories (SVM)

The following are the results from the first level categories (Flame, Off-topic, Social related and Task related), with a balanced sample.

Accuracy: 61,43% ± 5,65% (mikro: 61,43%)			**γ = 1, C = 15 and Dimensions=40**		
	true Flame	true Social related	true Task related	true Off topic	class precision
pred. Flame	0	0	0	0	0,00%
pred. Social related	3	123	29	58	57,75%
pred. Task related	3	22	103	16	71,53%
pred. Off topic	10	55	25	126	58,33%
class recall	0,00%	61,50%	65,61%	63,00%	

Table 54 SVM for first level category with dimensionality reduction through SVD, γ=1, C=15 and number of dimensions=40 (balanced)

Results for second level categories (SVM)

The following are the results from the second level categories (Flame, Off-topic, Social: Performance, Social: Organization, Social: Personal effects, Social: Person, Task: Product, Task: New Product, Task: Sales techniques and Task: Market development), with a balanced sample.

Accuracy: 41,41% ± 7,48% (mikro: 41,43%)							**γ = 1, C = 15 and Dimensions=40**				
	true Flame	true Social: Person	true Task: product	true Social: Performance	true Task: New product	true Social: Organization	true Off topic	true Social: Personal effects	true Task: Sales Technique	true Task: Market developments	class precision
pred. Flame	1	2	0	0	0	0	0	0	1	0	20,00%
pred. Social: Person	5	41	6	1	1	11	14	6	2	0	47,13%
pred. Task: product	3	2	27	2	9	3	2	2	6	6	43,55%
pred. Social: Performance	0	3	0	4	0	5	0	2	4	0	22,22%
pred. Task: New product	0	0	3	0	1	0	1	0	0	2	14,29%
pred. Social: Organization	1	9	8	9	3	25	9	10	8	5	28,74%
pred. Off topic	2	13	2	3	1	11	34	8	4	3	41,98%
pred. Social: Personal effects	2	3	2	6	1	11	11	41	4	3	48,81%
pred. Task: Sales Technique	2	2	2	4	1	8	2	3	21	2	44,68%
pred. Task: Market developments	0	0	2	0	4	1	2	2	0	13	54,17%
class recall	6,25%	54,67%	51,92%	13,79%	4,76%	33,33%	45,33%	54,67%	42,00%	38,24%	

Table 55 SVM for second level category with dimensionality reduction through SVD, γ=1, C=15 and number of dimensions=40 (balanced)

Appendix V Association Rule Mining

No.	Premises	Conclusion	Support	Confidence	Lift
1	Sentiment = Negative	Category = social related	0,38	0,60	1,094
2	Sentiment = Positive	Category = social related	0,05	0,62	1,115
3	Category = Task related	Sentiment = Negative	0,10	0,63	1,014
4	Category = social related	Sentiment = Negative	0,38	0,68	1,094
5	Category = flame	Sentiment = Negative	0,02	1,00	1,608

Table 56 Association rules for first level categories and sentiment.

Premises	Conclusion	Support	Confidence	Lift
second tier category = Task: product	Sentiment = Negative	0,03	0,62	0,989
second tier category = Social: Person	Sentiment = Negative	0,09	0,62	0,990
second tier category = Social: Personal effects	Sentiment = Negative	0,07	0,64	1,023
second tier category = Task: Market developments	Sentiment = Negative	0,02	0,65	1,040
second tier category = Social: Organization	Sentiment = Negative	0,20	0,72	1,156
second tier category = Task: Sales Technique	Sentiment = Negative	0,04	0,72	1,158
second tier category = Social: Performance	Sentiment = Negative	0,02	0,79	1,275
second tier category = flame	Sentiment = Negative	0,20	1,00	1,608

Table 57 Association rules for second level categories and sentiment.

Premises	Conclusion	Support	Confidence	Lift
year = 2010	Category = social related	0,07	0,51	0,926
year = 2006	Category = social related	0,11	0,57	1,034
year = 2009	Category = social related	0,09	0,57	1,034
year = 2008	Category = social related	0,10	0,59	1,060
year = 2012	Category = social related	0,04	0,59	1,073
year = 2011	Category = social related	0,07	0,64	1,156

Table 58 Association rules for first level categories and the year.

Premises	Conclusion	Support	Confidence	Lift
year = 2006	Sentiment = Negative	0,09	0,50	0,804
year = 2009	Sentiment = Negative	0,09	0,59	0,941
year = 2007	Sentiment = Negative	0,11	0,60	0,963
year = 2008	Sentiment = Negative	0,11	0,65	1,044
year = 2010	Sentiment = Negative	0,09	0,68	1,093
year = 2011	Sentiment = Negative	0,08	0,73	1,176
year = 2012	Sentiment = Negative	0,04	0,75	1,199

Table 59 Association rules for sentiment and the year.

Premises	Conclusion	Support	Confidence	Lift
Category = Task related	STOCK CHANGE = increase	0,08	0,51	1,038
STOCK CHANGE = increase	Category = social related	0,28	0,57	0,988
Category = flame	STOCK CHANGE = increase	0,01	0,57	1,153
STOCK CHANGE = decrease	Category = social related	0,28	0,58	1,014

Table 60 Association rules for change in stock and the first level category.

Premises	Conclusion	Support	Confidence	Lift
second tier category = Social: Organization	STOCK CHANGE = increase	0,15	0,50	1,009
second tier category = Task: Market developments	STOCK CHANGE = increase	0,02	0,50	1,009
second tier category = Task: product	STOCK CHANGE = decrease	0,03	0,50	1,025
second tier category = Task: Market developments	STOCK CHANGE = decrease	0,02	0,50	1,025
second tier category = Social: Personal effects	STOCK CHANGE = increase	0,06	0,51	1,022
second tier category = Task: Sales Technique	STOCK CHANGE = increase	0,03	0,53	1,072
second tier category = flame	STOCK CHANGE = increase	0,01	0,57	1,153
second tier category = Social: Performance	STOCK CHANGE = decrease	0,02	0,60	1,229
second tier category = Task: New product	STOCK CHANGE = increase	0,01	0,62	1,242

Table 61 Association rules for change in stock and the second level category.

Premises	Conclusion	Support	Confidence	Lift
Sentiment = Objective	STOCK CHANGE = increase	0,16	0,51	1,024
Sentiment = Positive	STOCK CHANGE = increase	0,05	0,54	1,087
STOCK CHANGE = increase	Sentiment = Negative	0,29	0,58	0,974
STOCK CHANGE = decrease	Sentiment = Negative	0,30	0,61	1,019
STOCK CHANGE = level	Sentiment = Negative	0,01	0,73	1,221

Table 62 Association rules for change in stock and the sentiment.

Premises	Conclusion	Support	Confidence	Lift
Sentiment = Negative	STOCK_CHANGE = decrease	0,45	0,5	1,05
Sentiment = Objective	STOCK_CHANGE = increase	0,07	0,75	1,47
STOCK_CHANGE = increase	Sentiment = Negative	0,44	0,87	0,95
STOCK_CHANGE = decrease	Sentiment = Negative	0,45	0,95	1,05

Table 63 Association rules for change in stock and the sentiment (with a 1 month delay for sentiment)

Premises	Conclusion	Support	Confidence	Lift
Sentiment = Objective	STOCK_CHANGE = increase	0,05	0,5	0,96
Sentiment = Objective	STOCK_CHANGE = decrease	0,05	0,5	1,07
Sentiment = Negative	STOCK_CHANGE = increase	0,48	0,53	1,00
STOCK_CHANGE = decrease	Sentiment = Negative	0,42	0,90	0,99
STOCK_CHANGE = increase	Sentiment = Negative	0,48	0,91	1,00

Table 64 Association rules for change in stock and the sentiment.