DEPARTMENT OF STATISTICS
OF THE LUDWIG–MAXIMILIANS–UNIVERSITY MUNICH

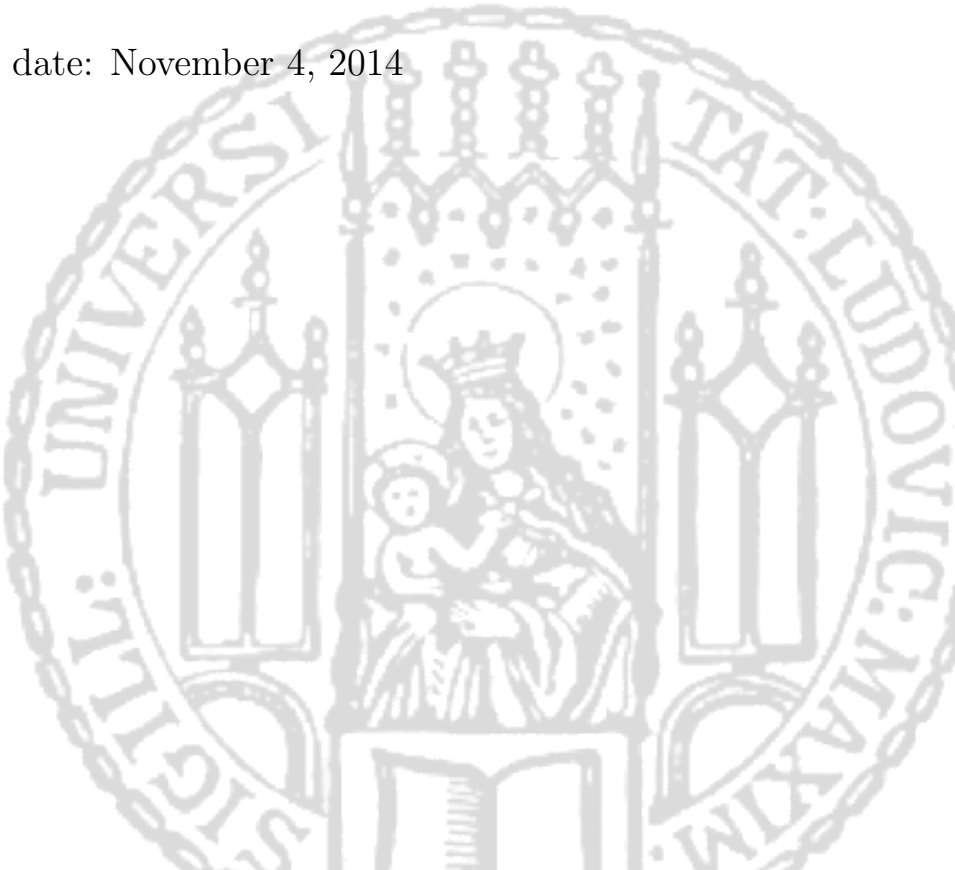
Master Thesis

RELEVANCE OF SOCIAL MEDIA ANALYSIS FOR OPERATIONAL MANAGEMENT IN HEALTH INSURANCE

Author: Maria Schmelewa

Supervisor: Prof. Dr. Christian Heumann
Dr. Fabian Winter

Submission date: November 4, 2014



Abstract

This master's thesis illustrates the future trend of insurance companies using social media data for improvement of business management in various fields of operation. Using the example of Facebook data for product development, the research question, how external data can be structured and combined with internal data and how the resulting database can be used in order to improve the product development process in health insurance, is analyzed. Here for, two comprehensive approaches exist, whereby this master's thesis will focus on the first proceeding. On the one hand, new products can be determined beforehand and it is considered, that the suggested products do not yet exist in this compilation and they also represent a market and client need. Thereby, the task is to find potential customers whom those products can be offered to. On the other hand, however, it can also be of interest to find new, not yet existing products according to demands of the market and of the existing clients derived from social media data.

To begin with, the term Big Data as the global source of external data, as well as its challenges are introduced. As social media is a substantial and very promising part of Big Data, it is illustrated what different sources and types of social media data there are and how such data can easily be extracted. Only a small amount of social media data is structured or semi-structured data, from which information content is easily extractable. Unstructured data represents the largest share of social media data, which has to be structured first in order to be able to obtain information on an automated basis. In the next step, therefore, important text analytics tools and techniques for structuring of unstructured data are introduced and described step by step. The combination of the then structured external data and the existing internal data warehouse, afterwards, creates a Big Data warehouse, which allows the evaluation and intelligence gathering from versatile, innovative and information-rich data. Here for, suitable matching techniques are illustrated. To conclude this master's thesis, such a combined dataset is simulated and analyzed with regard to product development by finding potential clients for new products. One possible approach here for is to score existing clients according to their externally matched information regarding their affinity towards the offered product. The alternative is to assign existing clients into so called lifestyle clusters according to their externally matched information and intuitively choose clusters, which contain potential new clients for the offered products.

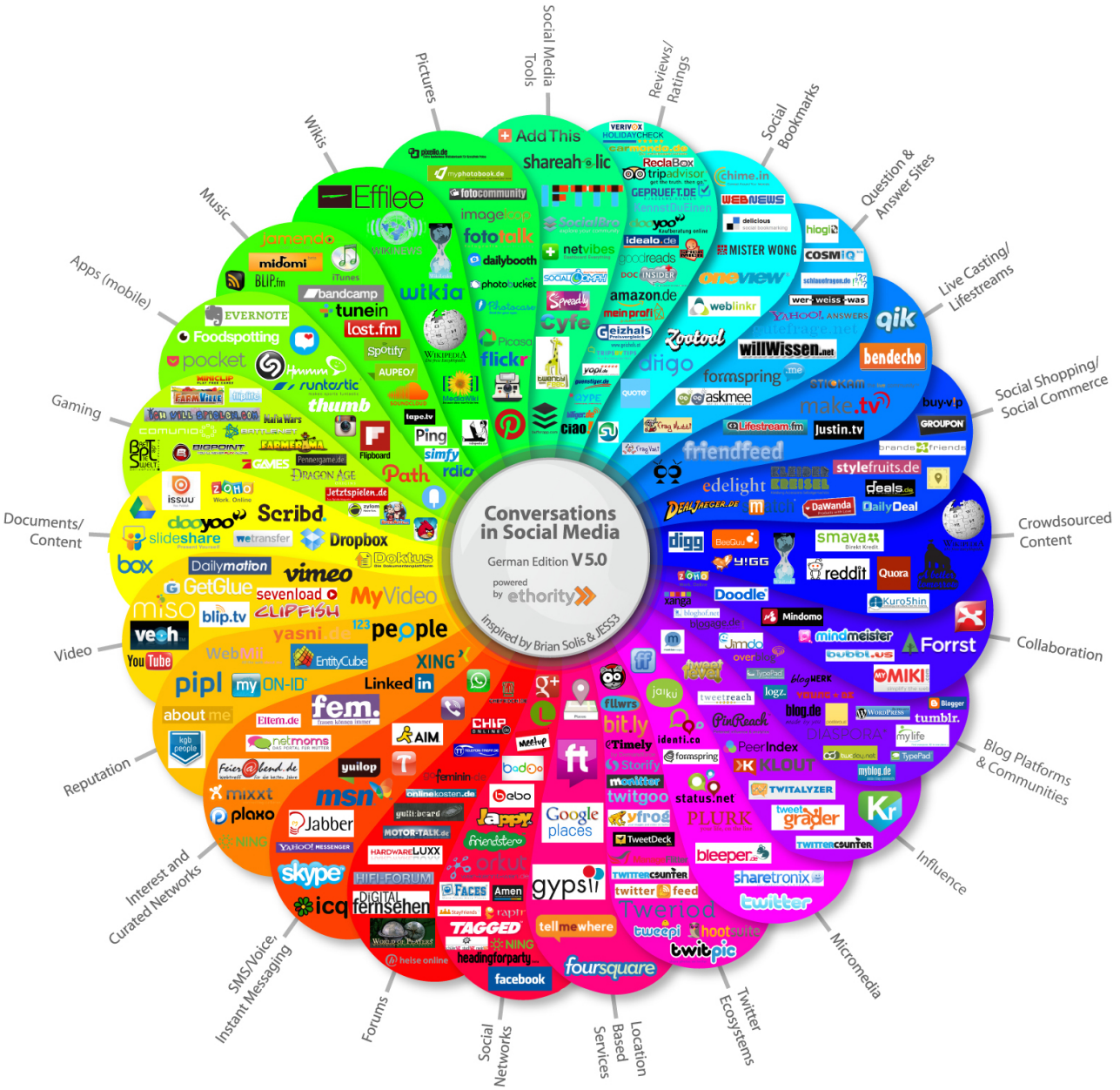
Contents

List of Figures

List of Tables

1 Social media data: a game changer for insurance

"Social media is a variety of digital media and technologies, which allow users to communicate with one another and to create media contents individually or jointly. This interaction includes mutual exchange of information, opinions, impressions and experiences, as well as contribution to creation of contents. Users actively make reference to those contents through comments, evaluations and recommendations and establish a social connection among one another that way."(Bundesverband Digitale Wirtschaft (BVDW) e.V., 2009) A more general definition of social media was written by Andreas Kaplan and Michael Haenlein. They define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content." (Kaplan and Haenlein, 2010) Hereby, "Web 2.0 describes World Wide Web sites that use technology beyond the static pages of earlier Web sites.[...] A Web 2.0 site may allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated content in a virtual community, in contrast to Web sites where people are limited to the passive viewing of content. Examples of Web 2.0 include social networking sites, blogs, wikis, folksonomies, video sharing sites, hosted services, Web applications, and mashups." (Wikipedia, 2014c)
Social media is a term that refers to a wide range of areas in today's society and, as shown in figure 1.1, it can be found in different fields of the media- and cyber-world. Not only intuitive branches, as social networks or blog platforms are referred to as social media, but also, for example video, gaming, music and pictures sites, as they also belong to "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."(Kaplan and Haenlein, 2010) Social bookmarks, Wikis, review and rating sites, question and answer sites, lifestreams, social commerce, crowdsourced content sites, collaboration and influence sites, location based service and document sites, apps and social media tools sites, and twitter, of course, make up the remaining branches of social media.
Due to fast growing technological developments, the ability to save larger amounts of data, "increasing proliferation and affordability of Internet enabled devices, such as personal computers, mobile devices and other more recent hardware innovations such as Internet tablets" (Aggarwal, 2011), analysis of social media data has become more and more important and of interest in various areas of application. It allows to not only reach out to others but to also analyze and understand them. Social media "may contain various kinds of contents such as text, images, blogs or web pages. The ability to mine these rich sources of information [...] provides an unprecedented challenge and also an opportunity to determine useful and actionable information in a wide variety of fields such as [sales and] marketing, social sciences, [the medical sector, public relations, human resources, crime prevention, as well as fraud detection] and defense." (Aggarwal, 2011) Even the "Bundesnachrichtendienst", the German Federal Intelligence Service, wants to start using social media in real time in order to detect fraud.

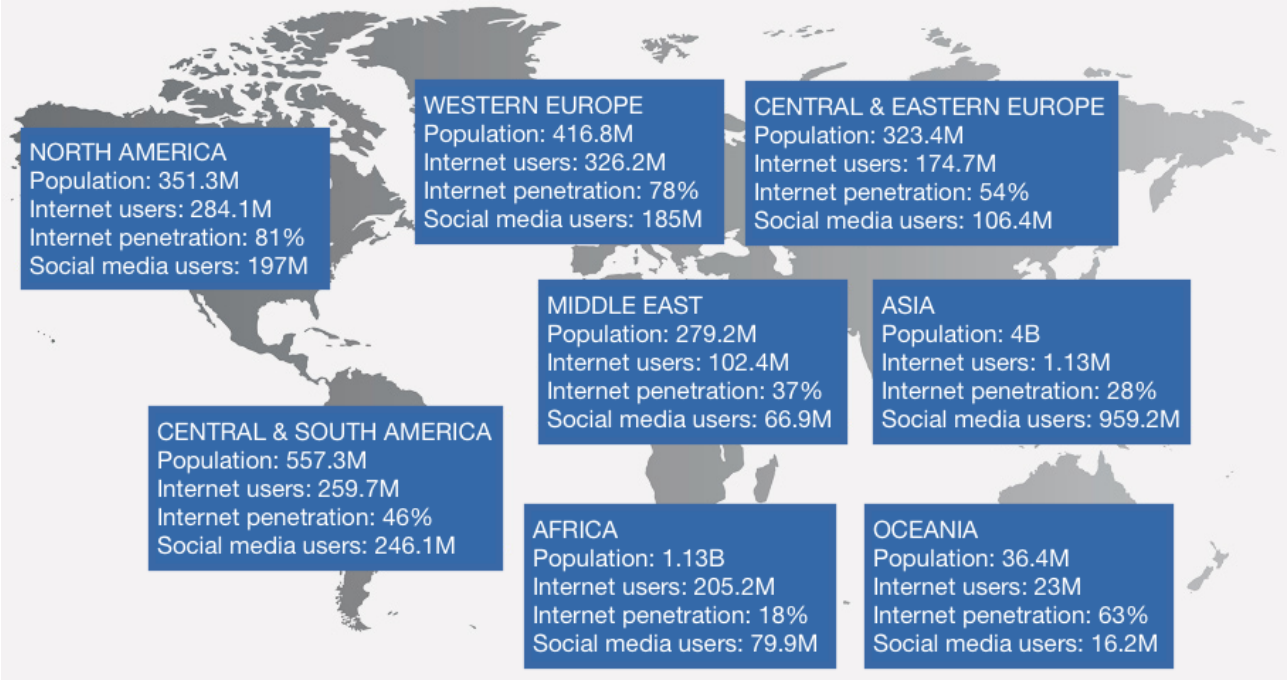


Source: Published by Ethority *http://social-media-prisma.ethority.de*

Figure 1.1: Social Media Prisma Version 5.0 - Branches of social media

In order to understand, why social media data has so much potential, some user numbers will be looked at, first. Looking at the estimated number of social media users worldwide in January 2014, as well as the share of Internet users on the overall population in that region (figure 1.2), one can see that especially in North America and Western Europe around 80% and in Central and Eastern Europe over 50% of the population had Internet access, from which over 50% used social media of some kind, by that time. In Central and South America, as well as in Asia around 90% of the Internet using population accessed social media. 70% of persons from Middle East and Oceania, who had Internet access also used social media. In total, there were around 1.9 billion people using social media around

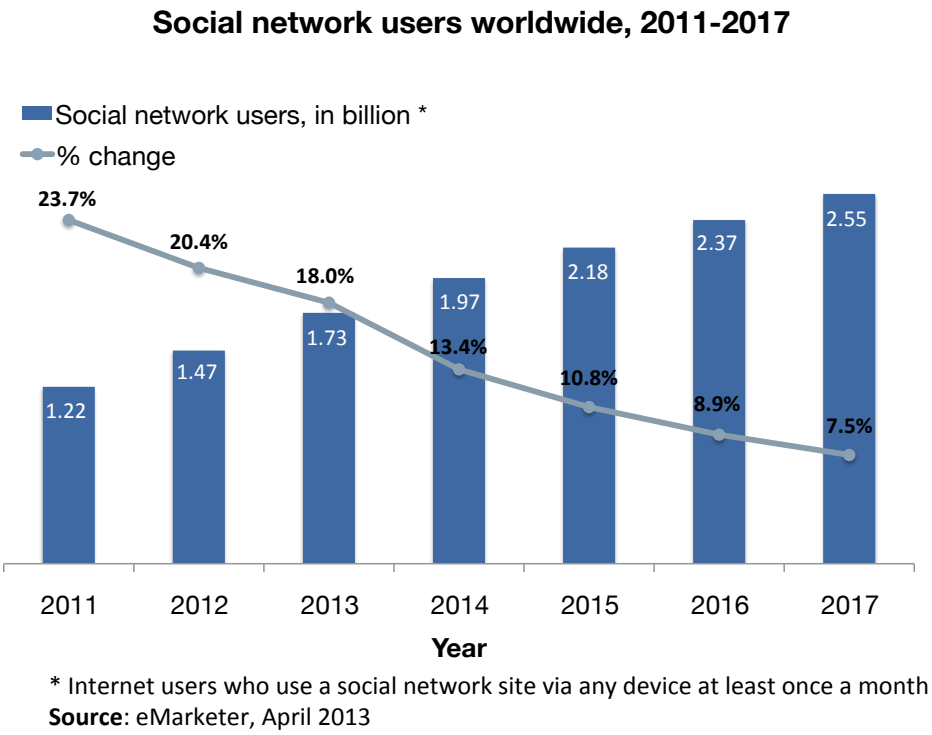
the world.



Source: Published by eMarketer *http://wearesocial.net/blog/2014/01/social-digital-mobile-worldwide-2014/*

Figure 1.2: Total Population, number of Internet users, their share on total population and number of social media users by region worldwide in January 2014

Social networks, as Facebook, Twitter and LinkedIn, make up a big part of social media. Looking in chronological sequence (figure 1.3), one can observe, that the number of users of social networks worldwide rose from 2011 until 2013, around 200 million each year. In year 2011, only 1.22 billion people used social networks, which increased to 1.73 billion active users worldwide in 2013. In April this year, Facebook alone had 1.28 billion users. 255 Million were registered at Twitter. This trend and continuing technological developments suggest an even further increase of user numbers, or more generally speaking number of people engaging in social media in the near future, which is estimated in figure 1.3 for the years 2014-2017. This also means that the amount of data, which will be created by users, will increase almost exponentially.



Source: Adapted from *http://www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976*

Figure 1.3: Number of social network users worldwide and the percentage increase compared to the previous year starting in the year 2011 with estimated numbers for years 2014-2017

Looking at user numbers of social networks in Germany, the same trend can be seen (figure 1.4).

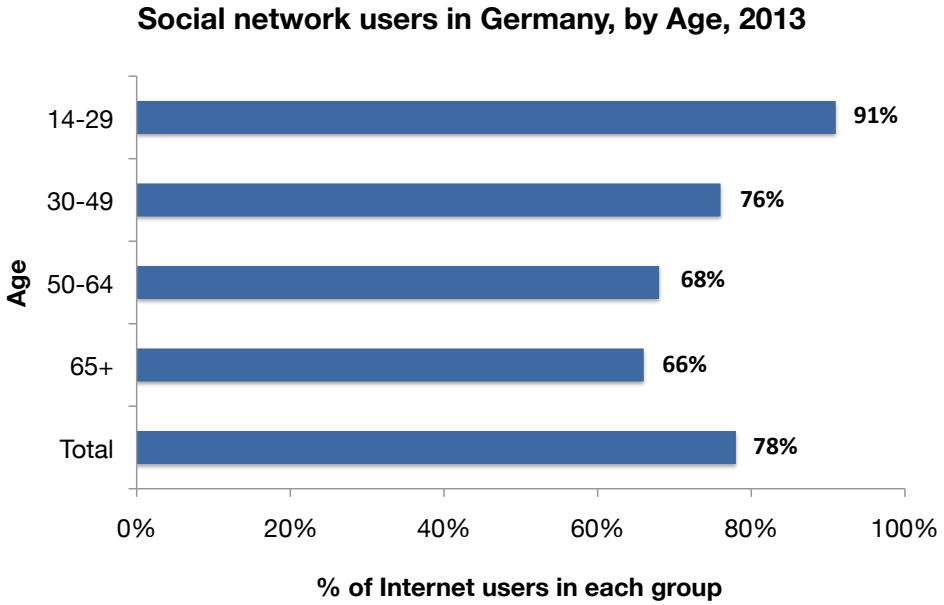
Social network users and penetration in Germany, 2011-2017							
	2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Social network users (millions) *	25.7	29.2	32.4	34.7	36.5	38.1	39.4
% change	20.8%	13.7%	11.1%	7.0%	5.3%	4.5%	3.4%
% of Internet users	46.1%	51.1%	55.6%	58.8%	61.5%	64.0%	66.0%
% of population	31.5%	35.9%	39.9%	42.8%	45.1%	47.2%	48.9%
* Internet users who use a social network site via any device at least once a month **Source:** eMarkter, April 2013							

Source: Adapted from *http://www.emarketer.com/Article/Social-Networks-Make-Strides-Germany/1010143*

Figure 1.4: Number of social network users in Germany and the percentage increase compared to the previous year starting in the year 2011 with estimated numbers for years 2014-2017

More and more people in Germany start using social media. In 2011, around 30% of the population were engaged in social networks at least once a month, whereas the number increased to almost 40% in

2013. The share of social network users on the Internet using population grew, as well. The prediction for years 2014-2017 also shows a clear upward trend.
But not only the young generation is involved in social networks, as can be seen in figure 1.5.



Source: Adapted from *http://www.emarketer.com/Article/Social-Networks-Make-Strides-Germany/1010143*

Figure 1.5: Percentage of social network users in each age group in Germany in 2013

In year 2013, 91% of the 14-29 year old German population 76% of adults between 30 and 49 years old and even over 65% of the population over 50 were engaged in social networks.
Despite the ever growing user numbers and thus additional information content, and all the possibilities, as well as opportunities social media brings with it, there is so much data and information left unconsidered. Nowadays, companies still underestimate the potential of social media as they either do not use this additional information at all or do not use it to full extent, amongst all due to lack of knowledge. The use of social media analysis in the customer service department, marketing, partially human resources and fraud detection has already spread, especially in the USA. In other areas of application, business enterprises are not known to use social media effectively or to its full potential, yet.
Considering insurance companies, most of which solely use their existing internal databases, as the claims or policy databases, are or will be in future confronted with the problem, that the internal data volume is limited and will only slowly increase over time. Thus, the information content will also stay limited and will not contribute much new knowledge to the enterprise. In some fields of application, this is not a problem of sufficient gravity. But in others, as for example sales and marketing or product development, internal data information is simply not enough, due to increasing competition, innovations and progress of the competitors. In order to succeed and be better than the others, additional complementary knowledge is needed.
External data and thus the information content volume is known to increase almost exponentially over

time. Figure 1.6 shows a purely hypothetical, fictitious image with internal data volume that exists in insurance companies and available external data from social media.

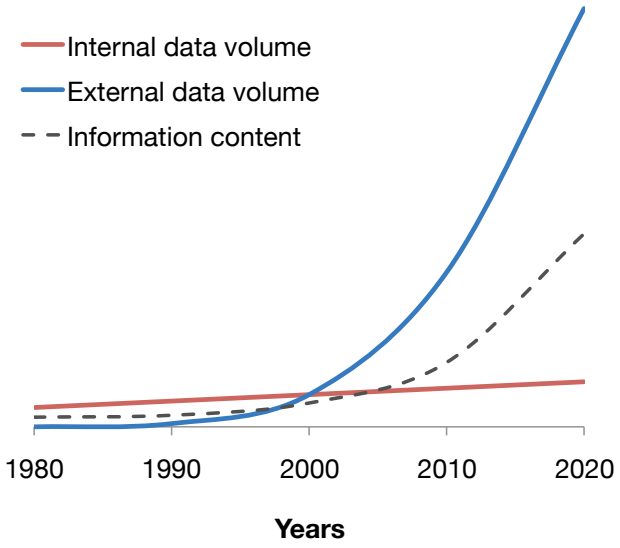


Figure 1.6: Hypothetical volume of internal and external data in insurance companies 1980 - 2020

The dashed line represents the purely hypothetical information content, which can be interesting for the area of application, if those two data sources are combined. Of course, not all external and internal data contains valuable information, which is the reason, why the line is always below the external data volume line. As can be seen, the information content will also increase with the combination of internal and complementary external data. This is due to the fact, that external data from social media containing real time information comprises knowledge and new insights that cannot be drawn from existing internal data from policies or claims. Such information are for example the activities, daily routines and interests of the insured. Thus, a combination of internal data with external data from social media will create a diversified, information-rich big data warehouse and will enable insurance companies to effectively use this additional knowledge to improve business management in various fields of application. Hence, the main question can be formulated as follows:

> **How can internal and external data be combined and how can the resulting database be used in order to improve business management of an insurance company?**

In order to keep this piece of work more application-oriented, this question will be looked at with regard to one business management area, the product development, as this approach has not been effectively worked out in this field of application, yet. In this context, not only the combination of different data sources, but also the structuring of information and the analysis on basis of this structured knowledge is of interest. Thus, the research question of this master's thesis reads as follows:

How can external data be structured and combined with internal data and how can the resulting database be used in order to improve the product development process in health insurance?

Here for, two comprehensive approaches exist, whereby this master's thesis will focus on the first proceeding. On the one hand, new products can be determined beforehand and it is considered, that the suggested products do not yet exist in this compilation and they also represent a market and client need. In this master's thesis the two products 'supplementary dental cover' and 'supplementary inpatient cover' are considered. Thereby, the task is to find potential customers whom those products can be offered to. On the other hand, however, it can also be of interest to find new, not yet existing products according to demands of the market and of the existing clients derived from social media data.

In order to answer this question, in the second chapter the term Big Data as the global source of external data, as well as its challenges will be introduced. As social media is a substantial and very promising part of Big Data, it will be illustrated what different sources and types of social media data there are and how such data can be extracted. Chapter 3 will look at the analytics and structuring of external unstructured data. In Chapter 4 the combination of the processed external and existing internal data will be dealt with. In each of these two chapters, the step-by-step analytics processes will be described first. Afterwards, statistical and analytical methods behind the process steps will be presented in-depth. Chapter 5 will then illustrate the evaluation of the combined database with regard to the product development process based on a simulated dataset, as this thesis is meant to be application-oriented. Thereby, the task is to find potential clients for a supplementary dental and for a supplementary inpatient cover. Conclusions and an outlook will round this Master's thesis off.

2 Definition, types and access of social media data

2.1 Big data and its challenges

"Big data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze. [But,] the right use of Big Data allows analysts to spot trends and gives niche insights that help create value and innovation much faster than conventional methods." (Wipro, 2014) Gartner defines Big Data as "[...] high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." (Gartner, 2014) Thereby, *volume* describes the "exponentially growing data volumes [...] due to the ever increasing use of the Internet around the world and the digitisation of everyday life." (Winter, 2014) Over the last decades, the amount of data that has been collected from a variety of sources, as social media or sensors and communicating devices, has increased exponentially. According to the study „Digital Universe" by IDC in year 2013, the amount of data will grow from 4400 exabytes equivalent to 4400000000 terabytes in 2013 to 44000 exabytes in 2020 (figure 2.1). (IDC, 2014)

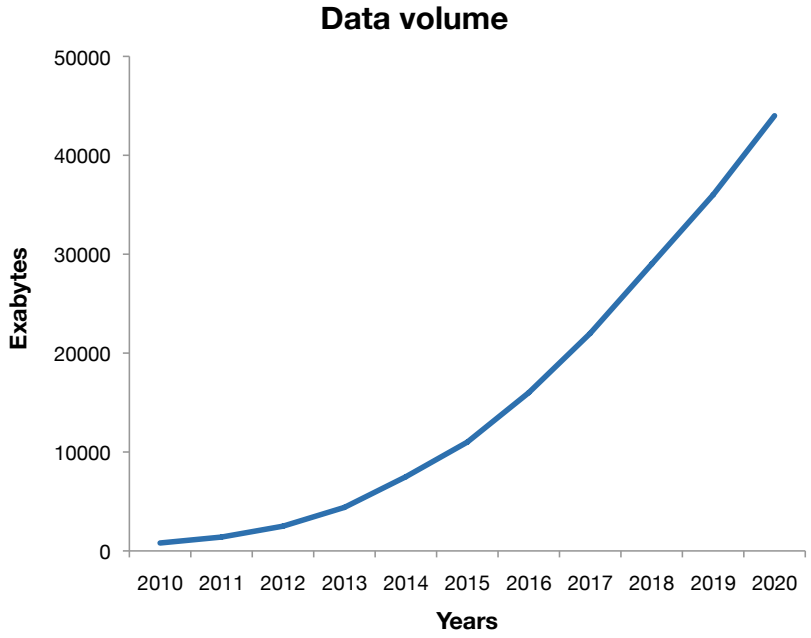


Figure 2.1: Volume of digital data 2010-2020

Only a tiny fraction of the digital universe has been explored for analytic value. "IDC estimates that by 2020, as much as 33% of the digital universe will contain information that might be valuable if analyzed." (see figure 2.2) (Gantz and Reinsel, 2012)

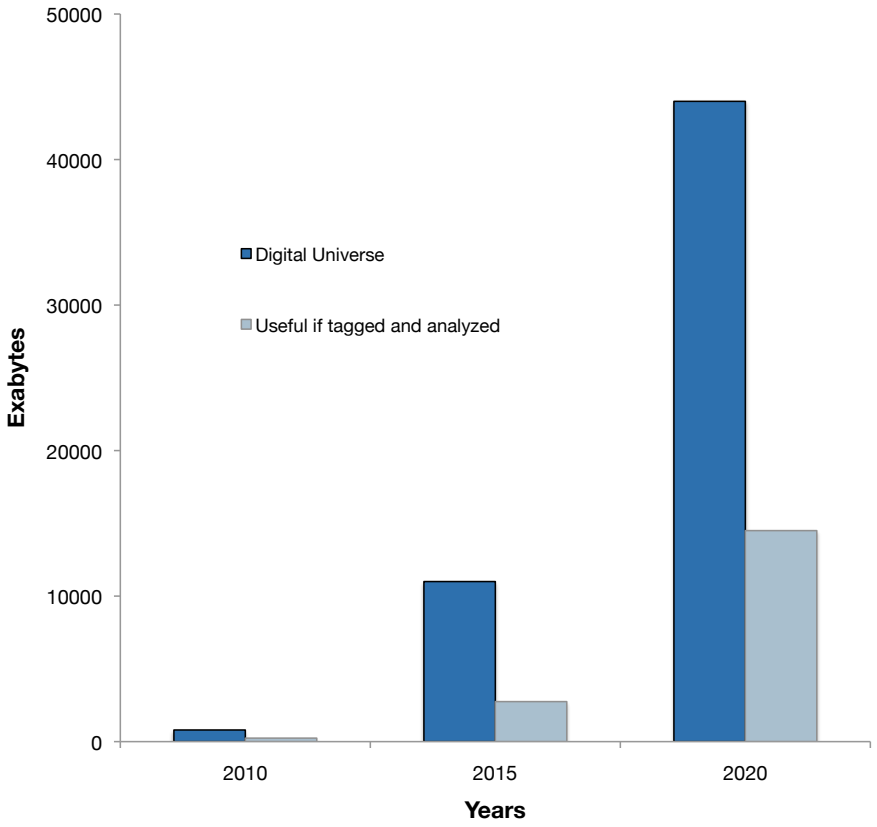


Figure 2.2: Opportunity for Big Data 2010-2020

The second term, *velocity* refers to the "real-time data processing[, as m]ore and more data have to be updated and available at an even faster pace. [The third 'v' for *variety*, says, that n]ot only the volume of data available, but also their quality is on the rise." (Winter, 2014) Thanks to newly-developed technologies and methods, as well as the progressing development process, this amount of data can be recorded and stored more effective.

However, in spite all the positive aspects Big Data has, it also entails challenges. As already described, Big Data is a large amount of data. Despite all technological innovations, it is still not possible to filter and store all that information, not to mention to analyze and use it correctly with the right statistical methods. Not only the size, but also the large variety of sources where it is being produced, makes it very complicated not to say impossible to manage such huge amounts of data. Regarding data quality, it is also difficult to distinguish between valuable, as well as true information and useless data. A critical point is data protection. It is necessary to stick to the law and regulations for data protection and also not all information is accessible or can be used from an ethical perspective.

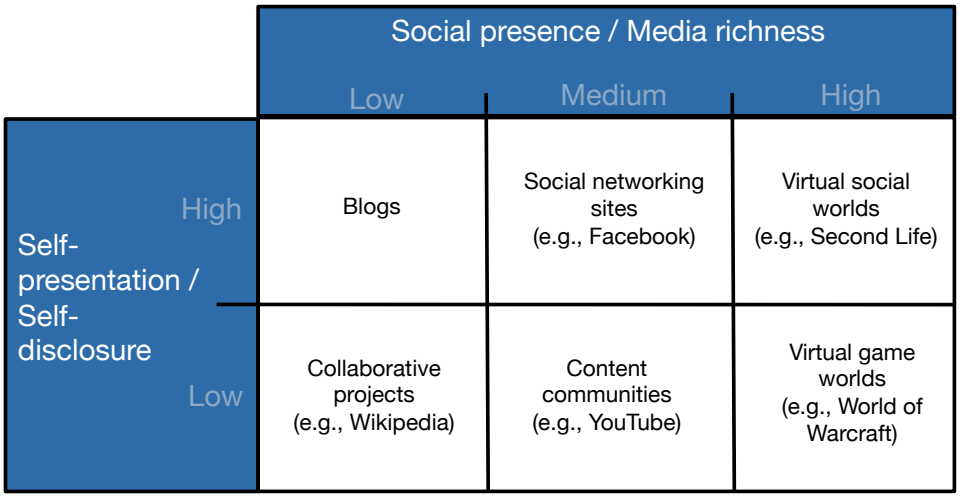
But, regardless of these points, Big Data provides great opportunities for our world. Big Data is produced in different sources and is stored in various forms, as links, texts, databases and so on. One important source of Big Data are sensors, the "Internet of Things" and communicating devices, for example GPS systems that send location and other information of the driver using it. In general, all electronic devices produce huge amount of usable data, as for example the Internet or surveillance systems. Credit card information also provides a huge amount of utilizable and valuable information.

At present, a very interesting, many opportunities offering and comprehensive source of Big Data for companies is social media data, which will be looked at closely in the following sections.

2.1.1 Sources and types of social media data

Social media data offers many new opportunities not only for individuals, but also especially for companies. Not only for marketing research, human resources, sales promotions and discounts, as well as relationship development, service, support and communication can social media be used, but also for E-Commerce, fraud detection, product development and many more. (Wikipedia, 2014b)
Beforehand, it is important to understand the difference between industrial or traditional media and social media. Social media differ from other media, because they are inexpensive and accessible in comparison. Everyone is able to publish and access any kind of information, whereas the process of publishing in industrial media undergoes many steps and stones. This fact consequently affects the quality of data. Another difference lies in the up-to-dateness of data. Whereas other media data can take month to be published, social media data is always up to date. (Wikipedia, 2014b)
As illustrated in figure 1.1, social media can be divided into 25 branches. According to the article "Social Media Kompass" published by Bundesverband Digitale Wirtschaft (Bundesverband Digitale Wirtschaft (BVDW) e.V., 2009), social media can be classified into four groups: communication, collaboration, multimedia and entertainment. The section *communication* includes blogs, micro-blogs, social networks, podcasts, newsgroups and Internet forums, as well as instant messengers. Wikis, social news pages and social bookmarking services are attributed to *collaborations*. *Multimedia* consists of photo, video and music sharing. The last group includes virtual worlds and online games. Kaplan and Haenlein selected a finer classification of six categories for social media according to the richness of the medium and the degree of social presence, as well as self-presentation and self-disclosure in their Business Horizons article (see figure 2.3). (Kaplan and Haenlein, 2010) *Virtual social worlds* are classified into the highest level, whereas *collaborative projects*, as Wikipedia score the lowest. All the other categories are rated in-between. "*Collaborative projects* enable the joint and simultaneous creation of content by many end-users[...]. With *collaborative projects*, one differentiates between wikis - that is, websites which allow users to add, remove, and change text-based content - and social bookmarking applications - which enable the group-based collection and rating of Internet links or media content.[...] *Blogs*, which represent the earliest form of Social Media, are special types of websites that usually display date-stamped entries in reversed chronological order. They are the Social Media equivalent of personal web pages and can come in a multitude of different variations, from personal diaries describing the author's life to summaries of all relevant information in one specific content area. *Blogs* are usually managed by one person only, but provide the possibility of interaction with others through the addition of comments.[...] *Content communities* exist for a wide range of different media types, including text (e.g., BookCrossing, via which 750,000+ people share books), photos (e.g., Flickr), videos (e.g., YouTube), and PowerPoint presentations (e.g., Slideshare). Users on *content communities* are not required to create a personal profile page; if they do, these pages usually only contain basic information, such as the date they joined the community and the number of videos shared.[...] *Social networking sites* are applications that enable users to connect by

creating personal information profiles, inviting friends and colleagues to have access to those profiles, and sending e-mails and instant messages between each other. These personal profiles can include any type of information, including photos, video, audio files, and blogs. *Virtual worlds* are platforms that replicate a three-dimensional environment in which users can appear in the form of personalized avatars and interact with each other as they would in real life.[...] *Virtual worlds* come in two forms. The first, *virtual game worlds*, require their users to behave according to strict rules in the context of a massively multiplayer online role-playing game.[...] The second group of *virtual worlds*, often referred to as *virtual social worlds*, allows inhabitants to choose their behavior more freely and essentially live a virtual life similar to their real life. As in *virtual game worlds*, *virtual social world* users appear in the form of avatars and interact in a three-dimensional virtual environment." (Kaplan and Haenlein, 2010)

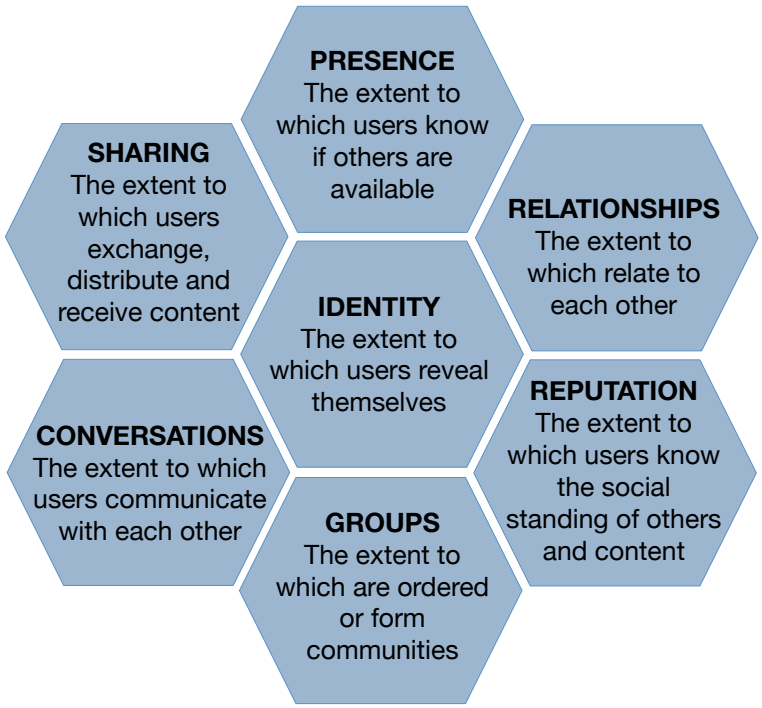


Source: Published by Kaplan and Haenlein, 2010 (Kaplan and Haenlein, 2010)

Figure 2.3: Classification of social media by social presence/media richness and self-presentation/self-disclosure

Nowadays, still not many companies realize the huge opportunities and the power that social media has. One problem is the lack of knowledge and expert know-how, another the insecurity regarding the correctness of the data, not taking social media seriously or sometimes even indifference. Therefore, Kietzmann, Hermkens, McCarthy and Silvestre presented a honeycomb framework that divides social media into seven functional building blocks: identity, conversations, sharing, presence, relationships, reputation, and groups. Looking at the seven building blocks (figure 2.4), one can get an idea of different social media functionality classes. "The *identity* functional block represents the extent to which users reveal their identities in a social media setting. This can include disclosing information such as name, age, gender, profession, location, and also information that portrays users in certain ways. [This information does not always have to be accurate and truthful, but, especially in social networking sites, mostly is.] [...] The *conversations* block of the framework represents the extent to which users communicate with other users in a social media setting. Many social media sites are designed primarily to facilitate conversations among individuals and groups. These conversations happen for all sorts of

reasons. People tweet, blog, et cetera to [...] be on the cutting edge of new ideas or trending topics[, to complain about issues, to get information or make their opinion heard.] *Sharing* represents the extent to which users exchange, distribute, and receive content.[...] The framework building block *presence* represents the extent to which users can know if other users are accessible. It includes where others are, in the virtual world and/or in the real worlds, and whether they are available. The *relationships* block represents the extent to which users can be related to other users, [by which is meant] that two or more users have some form of association that leads them to converse, share objects of sociality, meet up, or simply just list each other as a friend or fan. *Reputation* is the extent to which users can identify the standing of others, including themselves, in a social media setting. *Reputation* can have different meanings on social media platforms. In most cases, *reputation* is a matter of trust, but since information technologies are not yet good at determining such highly qualitative criteria, social media sites rely on 'mechanical Turks': Tools that automatically aggregate user-generated information to determine trustworthiness. The *groups* functional block represents the extent to which users can form communities and sub-communities. The more 'social' a networks becomes, the bigger the group of friends, followers, and contacts." (Kietzmann et al., 2011) Based on this framework, companies can decide what parts and to what extent they want to use social media for their field of interest.

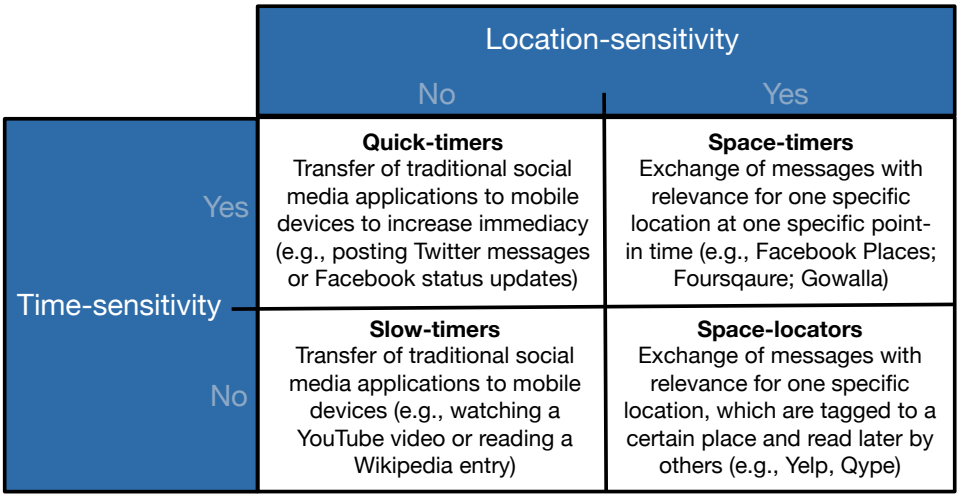


Source: Published by Kietzmann, Hermkens, McCarthy and Silvestre, 2011 (Kietzmann et al., 2011)

Figure 2.4: Honeycomb framework of social media

Another distinction has to be considered regarding how social media is accessed. In the early 1990s, cellular phones with access to the Internet were developed. At that time, the connection was very slow and expensive. Over the years many technological developments were made in this area. "The mobile marketing revolution only truly came to pass with the June 2007 launch of the iPhone. Since then, over

100 million iPhones have been sold worldwide [, not counting smartphones from other developers.]" (Kaplan, 2012) In the past decade, mobile devices have gained so much popularity, as they became more and more affordable and today, life cannot be imagined without them anymore. So nowadays, social media cannot only be accessed via PC, Mac and any other stand-alone computers, but also through most modern mobile devices, as smartphones, blackberries and tablets. That type of social media is called mobile social media and is defined "[...] as a group of mobile marketing applications that allow the creation and exchange of user-generated content.[...] Mobile social media differ from traditional social media applications in important ways. [In contrary to common social media, mobile social media is accessed] through a ubiquitous network to which consumers are constantly connected using a personal mobile device.[...] [Nowadays, many common social media websites, especially social networks, are also available as a mobile version. A distinction between] four types of mobile social media applications [can be made], depending on whether the message takes account of the specific location of the user (location-sensitivity) and whether it is received and processed by the user instantaneously or with a time delay (time-sensitivity). (see figure 2.5) [The incorporation of location- and time-sensitivity is the biggest distinction between common and mobile social media.] [...] Applications that are neither location- nor time-sensitive [are referred to] as *slow-timers* and [...] applications that take account of time and place simultaneously as *space-timers*. Applications that only reflect one of these two dimensions are referred to as either *space-locators* (location-sensitive but not time-sensitive) or *quick-timer* (time-sensitive but not location-sensitive)." (Kaplan, 2012)



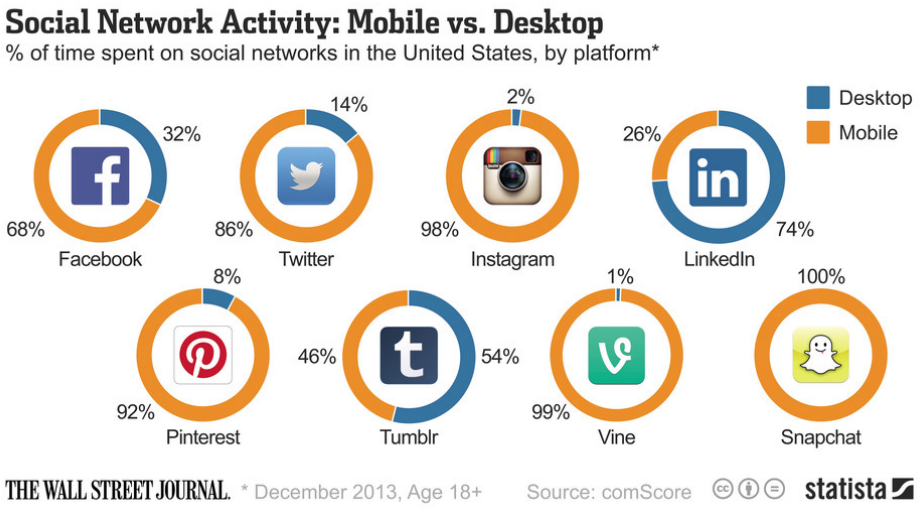
Source: Published by Kaplan, 2012 (Kaplan, 2012)

Figure 2.5: Classification of mobile social media applications

For example, Facebook Places and Foursquare can be classified as *space-timers*, whereas Yelp and Qype are assigned to *space-locators*. Twitter messages and Facebook status updates will be accounted as *quick-timers*. And last but not least, for example Youtube videos or Wikipedia entries, are referred to as *slow-timers*.

Looking at the ratio of mobile versus desktop social network activity in the United States in December 2013, one can see, that most time spent in social networks, as for example Facebook, Twitter and

Instagram was via mobile devices. (figure 2.6)



Source: Published by comScore/WSJ-Statista

http://blogs.wsj.com/digits/2014/04/03/data-point-social-networking-is-moving-on-from-the-desktop/

Figure 2.6: Mobile vs. desktop social network activity in the United States in December 2013

As social networks are one of the most promising sources of information regarding product development in modern-days, because they contain personal information of the users, as well as additional external components as consumer behavior, this master's thesis will mostly concentrate on this type of social media. "The most classical definition of a social network is one which is based purely on human interactions [, as for example Facebook or Twitter].[...] A number of technological enablers such as telecommunications, electronic mail, and electronic chat messengers (such as Skype, Google Talk or MSN Messenger) can be considered an indirect form of social networks, because they are naturally modeled as communications between different actors.[...] In addition, sites which are used for sharing online media content, such as Flickr, Youtube or delicious, can also be considered indirect forms of social networks, because they allow an extensive level of user interaction.[...] In recent years, it has even become possible to integrate real-time sensor-based content into dynamic social networks. This is because of the development of sensors, accelerometers, mobile devices and other GPS-enabled devices, which can be used in a social setting for providing a dynamic and interactive experience. [...] In fact, any web-site or application which provides a social experience in the form of user-interactions can be considered to be a form of social network."(Aggarwal, 2011)

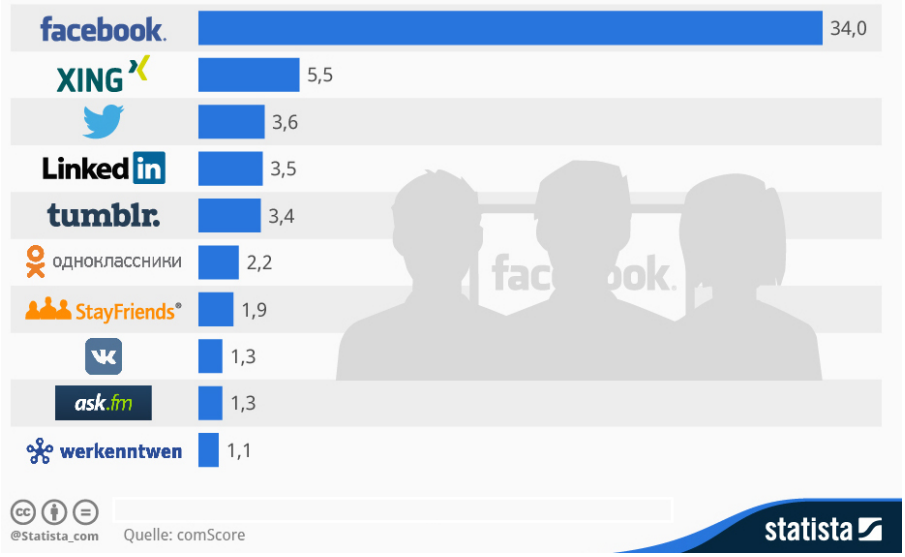
The focus of this thesis will lie on Facebook data, as this is the most used social network in almost every part of the world, as figure 2.7 shows. Although so many people use it, not everyone knows, that "Facebook [actually] started out as a niche private network for Harvard University students." (Kietzmann et al., 2011)

Looking at the user numbers of social networks in Germany (figure 2.8), one can see that Facebook dominated the social network market with a total of 34 million users in January 2014, by far.

Source: Published by Alexa.com and vincos.it

http://fashionbi.com/newspaper/social-media-engagement-and-preference-per-country

Figure 2.7: Top 3 social networks by membership by region worldwide in January 2014



Source: Published by comScore/WSJ-Statista

http://de.statista.com/infografik/907/top-10-der-sozialen-netzwerke-in-deutschland/

Figure 2.8: Top 10 social networks by membership in Germany in January 2014

Social network data and social media data in general can be divided into three different kinds of data, the structured and the unstructured data, as well as semi-structured data, which is considered as a form of structured data.

2.1.1.1 Structured and semi-structured data

In statistics, structured data, also referred to as quantitative data, describes data that is ordered in a fixed table structure. It is "organised in semantic chunks (entities)[, where] similar entities are grouped together [...] [.E]ntities in the same group have the same descriptions (attributes)[. The] descriptions for all entities in a group [...] have the same defined format, have a predefined length, are all present and follow the same order." (Wood, 2014)
"Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure." (Wikipedia, 2014a) As structured data, semi-structured data is also "organised in semantic entities [and] similar entities are grouped together [, as well. The difference is, that] entities in [the] same group may not have [the] same attributes [, that the] order of attributes [is] not necessarily important, [that] not all attributes may be required [and that] size [and type] of same attributes in a group may differ. " (Wood, 2014)
As we are interested in analyzing Facebook data, it will be used to illustrate examples of structured and semi-structured data. Every person creating a Facebook account has to fill out a registration form, which requires at least the first and last name, birthday, gender, spoken language and location of the user. Such information is considered structured data. Then, each Facebook user is assigned a user-specific id. In order to customize their own Facebook profile, further private information can be entered and is also considered as structured data. Such information includes the relationship status, hometown, attended school/university, work place, places lived at, religious and political views, hobbies/sports, favorite music/movies and "likes", as for example of different pages, interests, people, products or sports. Even contact information, as e-mail address or cellular phone number can be added. Of course, not all information will be available for every person, as providing it is optional. Another reason is the privacy. Each user can decide which information is open to public and which is only available to his or her friends.
A fictitious example for an extract of a dataset with information that is required from every Facebook user would look as follows:

id	name	birthday	gender	location
193852075	Peter Wood	14.06.1976	male	London, UK
19923742075	Mark Levene	NA	male	London, UK
139482075	Alex Poulovassilis	01.03.1965	male	London, UK

Table 2.1: Fictitious example for structured data with obligatory information

Further added information about the user for those fictitious three ids could look like the following:

id	hometown	relationship	school	work place
193852075	Philadelphia, USA	single	Philadelphia High	Birkbeck
19923742075	Vancouver, Canada	married	Hugh Boyd Secondary	Birkbeck
139482075	Athens, Greece	engaged	NA	Birkbeck

id	hobbies	music	movies	likes
193852075	Skiing, photography	Punk	Kill Bill	Marlboro
19923742075	Rafting, sailing	Classical Rock	Transformers	Boats
139482075	Climbing, boxing	House	Rocky I , II	Nutrition for life

Table 2.2: Fictitious example for further structured data

In addition, information about friends of those persons and thus their network structures can also be interpreted as structured data.
An example for semi-structured data for the same three invented persons from table 2.1 and 2.2 can be seen in the following table:

name:	Peter Wood
email:	ptw@dcs.bbk.ac.uk, p.wood@bbk.ac.uk
name:	
first name:	Mark
last name:	Levene
email:	mark@dcs.bbk.ac.uk
name:	Alex Poulovassilis
affiliation:	Birkbeck

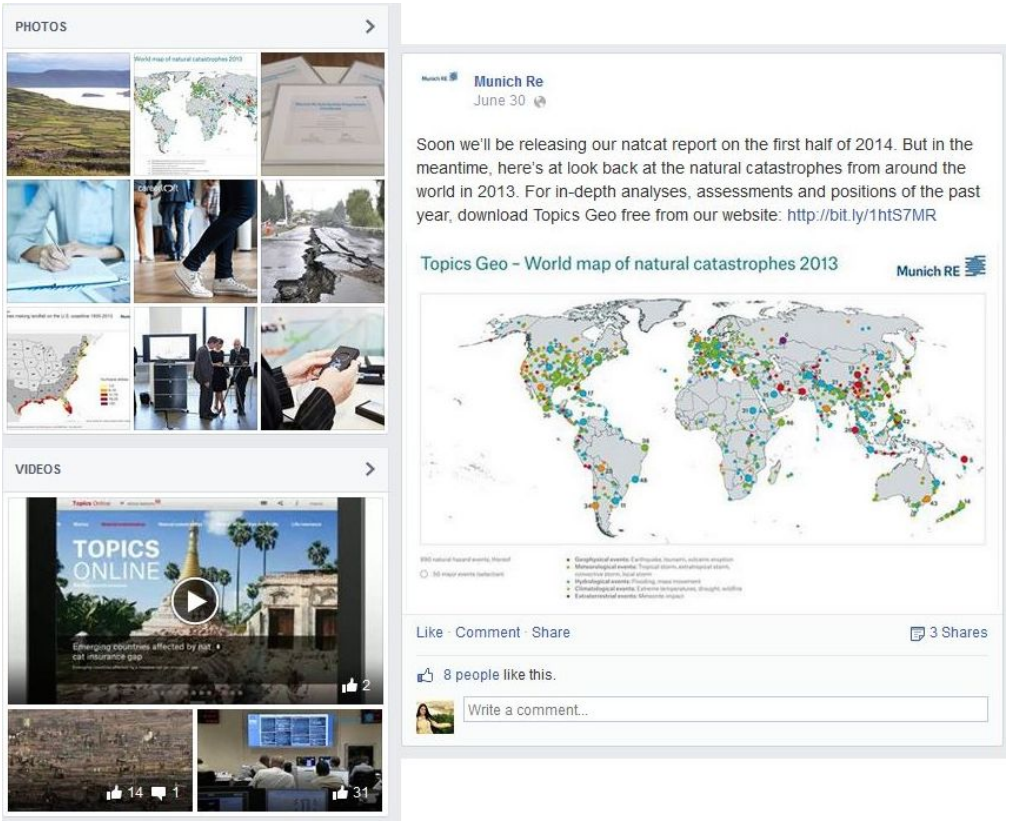
Table 2.3: Fictitious example for semi-structured data

The table still contains structured data but is not as identifying as fully structured data and thus has to be brought into a fully structured format for further analysis.
The second type of data is unstructured data.

2.1.1.2 Unstructured data

Unstructured data, often also referred to as qualitative data, "can be of any type. [It is] not necessarily following any format or sequence [or] any rules." (Wood, 2014) "Unstructured information is typically the direct product of human communications. [...] It is information that was not encoded for machines to understand but rather authored for humans to understand. We say it is 'unstructured' because it lacks the explicit semantics ('structure') needed by computer programs to interpret the information as intended by the humans author or required by the application." (Ferrucci et al., 2006) Thus, the unstructured format is not predictable and therefore has to be brought into a structured format

for further analysis. Not only text data is referred to as unstructured data, but also any type of multimedia, as videos, pictures and sounds. Users comment, share, post and follow various information and sites via Facebook. An example for three different kinds of unstructured data, namely text, photos and videos, can be found on the official Facebook page of Munich Re. (figure 2.9)



Source: *http://www.facebook.de/munichre*

Figure 2.9: Part of the Munich Re Facebook page

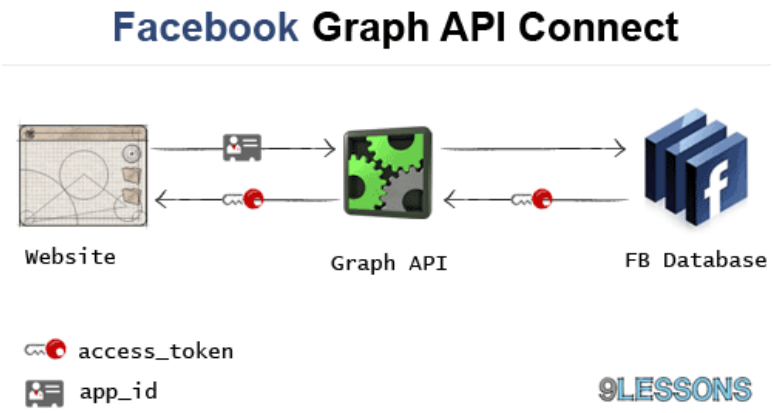
"Unstructured data [...] represents at least 85% of all data present in the enterprise and across the Web [...]." (Feldman et al., 2012) The remark has to be made, that this does not necessarily mean that the amount of information content of unstructured data is higher than of structured data. Every day, more and more new unstructured data is created, whereas the volume of structured data has its limits. In order to be able to analyze and evaluate such data, it has to be brought into a structured form first and can be then analyzed with statistical methods.

But, in order to be able to work with and analyze social media data, it has to be extracted from the corresponding websites, here network websites into datasets, first. There are three possibilities how to collect such data. The first one is the recruitment of participants in offline contexts. The second option is the recruitment of participants via Facebook applications. The third method, which is the focus of this thesis, is called 'data crawling'. 'Data crawling' means, that information from Facebook users is retrieved from the user profiles without them actively participating. (Wilson et al., 2012) A few statistical programs already exist, which enable the user to extract or 'crawl' and even evaluate social media data afterwards.

2.2 Collection of data

2.2.1 Software for data extraction

In the past few years, the demand for analysis of social media data has considerably increased. In order for statisticians to be able to evaluate this kind of data, information from the corresponding websites has to be extracted and brought into a dataset or rather tabular form, first. The first phase, meaning the extraction of data in accordance with the issue of interest is also called Information Retrieval (IR). Typical statistical software developers, as SAS, R or Statistica have already created packages for data extraction or IR from social media websites. Some even provide further analysis tools. Other software developers as IBM and small stand-alone software developers also dealt with this issue and created a few software packages also not only for data extraction, but for data analysis, especially of unstructured data, as well. In the following chapters, only extraction and analysis software and tools for Facebook data will be looked at. At this point, it has to be mentioned, that one has to be careful, which applications he or she uses for data extraction, as there are many softwares that are illegal. In the following sections, only legal and widespread software developers will be introduced.
Before continuing to look at different software packages more closely, one thing most available software has in common, will be introduced. Not only Facebook, but also other social media websites as Twitter or Youtube offer API's, which allow the access to data via any software in the first place. "An application-programming interface (API) is a set of programming instructions and standards for accessing a Web-based software application or Web tool. A software company releases its API to the public so that other software developers can design products that are powered by its service. [...] It is a software-to-software interface, not a user interface." (Roos, 2014) In easier terms, "an API is a programming language that allows two different application to communicate, or interface, with each other." (Gunelius, 2011) Facebook has it's own API, which is called Graph API. (see figure 2.10)

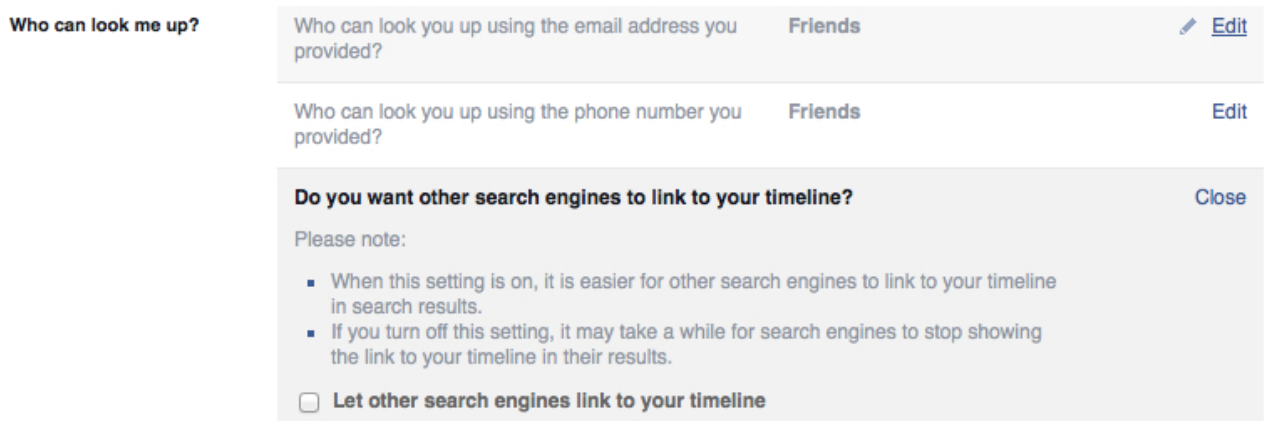


Source: *www.9lessons.info/2011/01/facebook-graph-api-connect_with_php_and.html*

Figure 2.10: Functioning of the Facebook Graph API Connect

"The Facebook social graph is a graphic representation of the Facebook community, showing Facebook users and connections. The Graph API allows developers access to a simplified view of this graph,

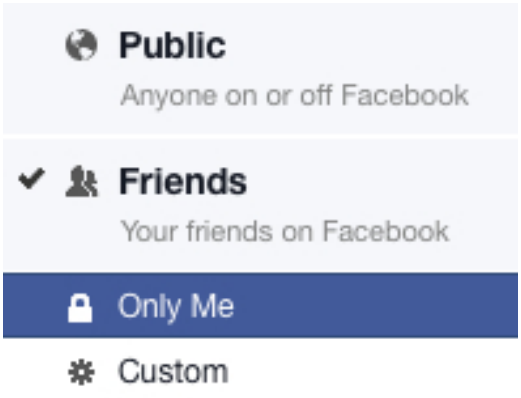
with representations of objects (people, photos, events and pages) and interconnectivity links." (API Portal by Anypoint Platform and Mulesoft, 2014) In order to access those Facebook objects and interconnectivity links, each user, website, app or software has to be registered as a Facebook developer, after which a unique app id is assigned. In exchange, an access token to Facebook data is created.
The only existing limitation regarding the access of such objects and interconnectivity links are the restrictions of users regarding the access of their data. Depending on their privacy settings, some information of the users is not accessible through the Graph API. First of all users can restrict their profile to be found using various search engines, as for example Google. (see figure 2.11)



Source: *https://www.facebook.com/settings/?tab=privacy&privacy_source=privacy_lite*

Figure 2.11: Restriction of finding Facebook profile using search engines

If the option to let other search engines link to the user's profile is switched off, the Facebook user can only be found by looking him or her up in Facebook itself.
In Facebook itself, there are 4 different levels of privacy of data that are displayed in the following figure.

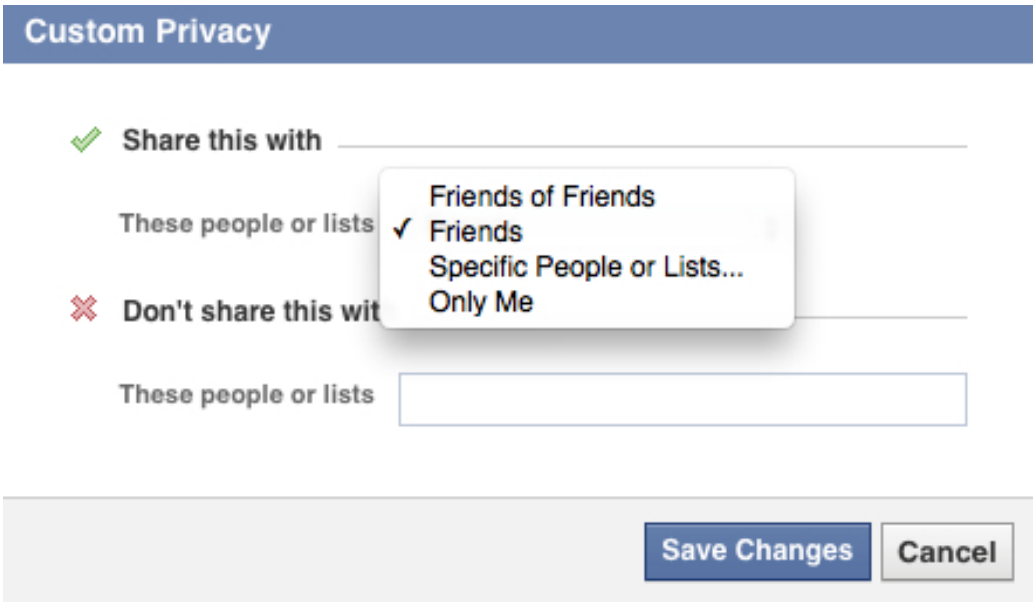


Source: *https://www.facebook.com/settings/?tab=privacy&privacy_source=privacy_lite*

Figure 2.12: Privacy levels in Facebook

Only me restricted information can only be accessed by the user him- or herself. This information is

neither visible to friends, nor to anyone else. All *Public* information from a user profile or any other page and data from friends' profiles can be crawled. The option *Custom* is classified into the following alternatives:



Source: *https://www.facebook.com/settings/?tab=privacy&privacy_source=privacy_lite*

Figure 2.13: Privacy levels in Facebook

In this case, the user can customize the limitation of audience to, for example *Friends of Friends*, specific people from his or her friends list or certain groups he or she is in. Then only those people have access to the user's information. In this context, reference is made to the data use policy of Facebook. (Facebook, 2014)
In order to get an idea, what data is contained in those Facebook objects and interconnectivity links and can be of interest for product development, an extract of what information is possible to extract from Facebook via Graph API is shown in figure 2.14. As can be seen, basic objects, connections and search queries of a user can be extracted from Facebook. Thereby, *Pages*, *Events* and *Groups* can also be categorized into the *Connections* section, as this information is external and only linked to a user's profile. (API Portal by Anypoint Platform and Mulesoft, 2014)

As already mentioned, a separate software is needed in order to access any information via Graph API, so a closer look at different Facebook data extraction softwares will be taken in the next step. As all software tools are building on accessing Facebook data through the same internal Facebook Graph API, the resulting outcomes from various extracting software environments are very similar. That and the fact that it is open source, is why only one software, namely R, will be presented in more detail in the following section. A few other developers will be introduced in the subsequent sections.

BASIC OBJECTS	CONNECTIONS	SEARCH
Users (Information on user e.g. ID, full name, gender, birth date, location, occupation) **Profile pictures** (User's profile pictures with comments, likes, etc.) **Status messages** (Status messages from users profile page) **Pages** (Various pages on Facebook or only pages user likes e.g. Marlboro fan page with all containing information as likes, members and comments) **Events** (Various events on Facebook or only events user attended e.g. boxing event with date, time, attending members, etc.) **Groups** (Various groups on Facebook or only groups user is member in e.g. 'I Love Chocolate' group with all members and all containing information) **Photos** (Photos from user profiles, pages, etc. with additional information as comments and likes) **Videos** (Videos from user profiles, pages, etc. with additional information as comments and likes) **Checkins** (Checkins of user e.g. user attended cigarette convention in Munich)	**Friends** (User's friends profiles with all available information) **News feed** (Activity from people, pages and groups user follows on Facebook) **Profile feed (Wall)** (All activity on user's profile wall as posts and comments) **Likes** (Posts, pages, groups, etc. user likes, e.g. Extreme Air Sports page) **Movies** (Movies user watched, wants to watch or likes, e.g. Fight club) **Music** (Music user likes or wants to listen to later, e.g. Heavy metal) **Books** (Books user read, wants to read or likes, e.g. Martial Arts) **Sports** (Kind of sport and sport teams user likes, e.g. boxing) **Photo Tags** (Photos user was tagged in) **Video Tags** (Videos user was tagged in) **Objects with location** (Information about objects that have location information attached. Objects can be those in which user or his/her friend has been tagged or were created by user or his/her friends)	**All public posts** **People** **Pages** **Events** **Groups** **Places** **Checkins** (Things user searched for on Facebook)

Figure 2.14: Possible application of Graph API with regard to available information for extraction

2.2.1.1 R

R is an open source programming language and software environment for statistical computing and graphics, where anyone can develop new packages. Pablo Barbera developed a package called *Rfacebook* enabling users to extract all above mentioned Facebook data via R (Barbera, 2014)
In order to be able to access data from Facebook, the user has to authorize R to connect through his or her own Facebook account. Therefore, the user has to be registered as a Facebook developer first. To make clear, a user does not necessarily have to be a natural person. A website, company or app can be referred to as a user, as well. After loading the required package *Rfacebook* and creating the authorization token to an Facebook R session, the user can extract any information not only about his or her friends, but also from any public profile and page via Facebook Graph API through his or her own developer account into a data matrix, as well as save it as a .txt or .csv file, for example.
It is possible, for example, to extract the whole list of friends or followers together with information as name, gender, spoken language, the link to the profile picture, birthday, location, hometown and relationship status, if available. An extract of the corresponding output would look as follows.

```
> friends <- getFriends(token = fb_auth)
```

id	name	gender	locale	birthday	location	relationship_status
758..	Marthin XY	male	en_US	03/26	Tomohon	NA
510..	Brittany XY	female	en_US	11/19	Vancouver, BC	Married
100..	Andreas XY	male	ru_RU	10/03/1989	Kyiv, Ukraine	Single
100..	Lydia XY	female	de_DE	04/30/1981	Ortisei, Italy	In a relationship

Table 2.4: Example for *Rfacebook* output with the list of friends

As already mentioned, not only information about the user's friends, but also of any profile which is available to the public can be accessed, as long as his or her Facebook id is known, as each person has his or her own id by which his or her profile can be found and accessed. Of course, only the information, that is publicly available, can be extracted. If the user would like to see any persons list of likes, he or she simply has to know the person's id and can then extract a matrix with the list of likes, which contains the id, names or terms of the liked page and its website. (see table 2.5)

```
> likes <- getLikes(id=100.., n=100, token = fb_auth)
```

id	names	website
129..	Mottolino Fun Mountain	www.mottolino.com
126..	Alpinestars	http://www.alpinestars.com
364..	Intense Cycles Inc	www.intensecycles.com
294..	Bikepark Hopfgarten Tirol	www.bikepark-hopfgarten.at

Table 2.5: Example for *Rfacebook* output with the list of likes of a specific user

An additional feature is the ability to display the network of friends or followers graphically. Typical for such networks are clusters, which indicate distinct circles of friends or followers with similar interests, hobbies, backgrounds or other characteristics. (see figure 2.15)

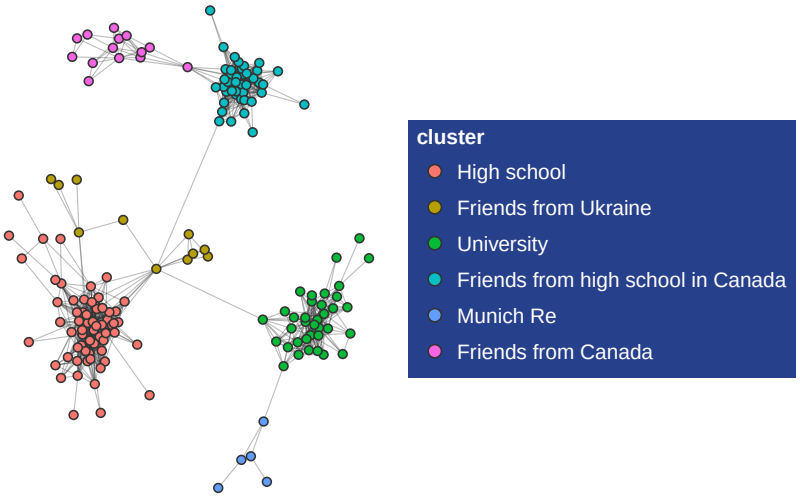


Figure 2.15: Example for *Rfacebook* output with the network structure of friends

This package does not only enable to extract information about persons, but also from different pages and to search for keywords throughout Facebook. If one is interested in "health insurance", for example, he or she can search Facebook for it. (see table 2.6)

```
> fb.posts_health <- searchFacebook(string="healthinsurance", token=fb_auth, n=100)
```

from_id	from_name	message	link
181..	D'Ebrar XY	Rejoice because ...	http://www.facebook.com/...
817..	Viewed XY	Medical Billing Errors Learn to Spot and Fix Them?and Save-Money.com	http://www.facebook.com/...
211..	High XY	In lawsuit, Connecticut man seeks health insurance without fee for coverage of abortions	http://www.facebook.com/...

id	likes_count	comments_count	shares_count
181.._102..	0	0	0
817.._893..	1	1	0
211.._827..	1	1	0

Table 2.6: Example for *Rfacebook* output with the list of users writing about "health insurance"

The output shows user ids who created these posts, by means of which all information about those users can be accessed. In addition, the name, comment or like, time of post and the number of likes, comments and shares of it is output.
One can also search individual pages, as for example Philippine Health Insurance Corporation, a health insurance company from the Philippines. Afterwards one can look at all posts made by one specific id that commented on this page. Information as to what and when this person posted something and the corresponding link will be extracted. In this case, Phil Health itself posted something on its page.

```
> page_ph <- getPage(page="PhilHealth", token=fb_auth)
> post <- getPost(post=page_ph$id[1], token=fb_auth)
```

from_id	from_name	type
211..	Philippine Health Insurance	photo
211..	Philippine Health Insurance	status
211..	Philippine Health Insurance	photo

message	link
Sa P6.60 kada araw, segurado na ang kalusugan mo at ng buong pamilya. Magpamiyembro na!	http://www.facebook.com/...
PhilHealth New Payment Schedule for Premium Contributions	NA
NA	http://www.facebook.com/...

id	likes_count	comments_count	shares_count
211.._717..	6277	359	644
211.._709..	824	107	146
211.._710..	2617	161	366

Table 2.7: Example for *Rfacebook* output with the post of a specific user on the Philippine Health Insurance Facebook page

Additionally, one can extract all information regarding who liked and who commented their posts.

```
> ph.likes <- getUsers(post$likes$from_id, token=fb_auth)
```

from_name	from_id
Happy Fajardo XY	456..
Gelouh XY	234..
Rosemarie XY	175..
Lucy XY	101..

Table 2.8: Example for *Rfacebook* output with the list of users who liked of post of a specific user

```
> ph.comm <- getUsers(post$comments$from_id, token=fb_auth)
```

from_id	from_name	message	likes_count	id
154..	Edgar XY	last year, P3.30 lng...	1	154.._235..
154..	Edgar XY	ngayanm dinoble	0	154.._568..
581..	Stephen XY	Safe ba talaga?	0	581.._875..
485..	Elaisa XY	paano	0	485.._196..
635..	Ronamyn XY	may i.d nb yan?	0	635.._811..
741..	Michelle XY	Paano b mgpamember?	1	741.._566..

Table 2.9: Example for *Rfacebook* output with the users commenting on the post of a specific user

Of course, basic information on those users can be extracted, as well, if available and open to the public.
Those outputs are just examples of what is possible with *Rfacebook*. In principal, all information from searches, as shown in figure 2.14 can be accessed and extracted into a structured dataset in table format for further analysis via *Rfacebook* through the Facebook Graph API. Information on the time of creation of all data is also available and can be extracted.
But not only open source software enables the access of Facebook data. There are a few other software developers that created such tools and will be introduced in the following section.

2.2.1.2 Further software

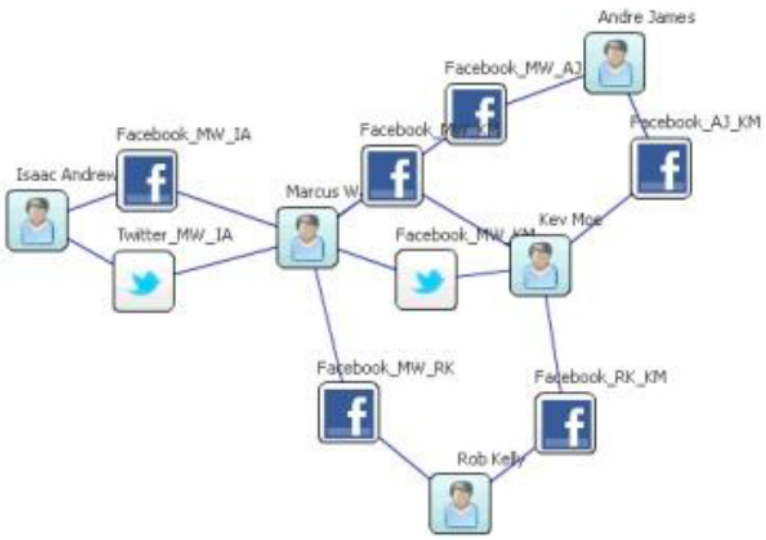
Not only open source program developers or small companies, as NetVizz, have created software for network data extraction. Two big software-developing companies SAS and IBM have been engaging in this subject for a few years now. In contrast to R, SAS and IBM have developed software not solely for data extraction, but as a bundle of tools combined into one package, which also enables further analysis of the extracted data. The main focus of those software products lies in the extraction and analysis of unstructured text data. Unlike with R, those software packages can also be individually adapted to the needs and the issue of interest of the user and are able to access data from many social media sources simultaneously. In the subsequent sections, those two software programs will be shortly introduced.

SAS

"SAS Social Media Analytics [(SMA)] is an on-demand offering that integrates, archives, analyzes and reports on the effects of online conversations occurring across professional, consumer-generated and social network media sites. [...] By enabling [one] to link sentiment to specific business concerns at an enterprise level, it allows for an in-depth monitoring and accelerated response time to shifts in the marketplace. [...] By merging market data (from blogs and other social media sources) and customer data (surveys or Web forms), market research professionals can validate then act upon a consumer need or sentiment shared across a customer base or market." (SAS Institute Inc, 2014)
The SAS SMA analytics process can be divided into three big steps. In the first step, data extraction

via API's and data management is performed. Afterwards, data and text mining via natural language processing, content categorization, sentiment analysis and a few other analytical methods are implemented. Thereby the un- or semi-structured information is brought into a structured form. As a final step, statistical analysis of the extracted structured information, such as the overall sentiment, real time tweets-sentiment and velocity of live events, as well as phrase clouds, and monitoring of relevant results, can be carried out. (Acharya, 2013) The second and third steps will not be elaborated upon, at this point, as the whole text and data analytics process will be the subject of the subsequent sections. As already mentioned, SAS SMA not only accesses data from Facebook, but also from other popular social networking sites like MySpace and Twitter, as well as from various review sites, blogs and many more social media sources. Further features of this software are the multilingual support, its mobile interfaces and the ability to communicate with consumers in social media conversations. (SAS Institute Inc, 2014)

SAS also developed a tool that allows the user to extract and display social networks. This tool is called Social Network Analysis (SNA). Thereby, one has to enter the user ids of interest manually. Afterwards, data is collected for those users in order to determine, if they have been interacting with each other in any way. Those persons are then visualized in a network diagram. Afterwards, additional data and information can be collected for the ids of interest via API's. This tool is mostly being used in fraud detection, but offers high potential for other fields of application. (Blomberg, 2012) The advantages over R is that information from Facebook and Twitter can be collected, simultaneously. A SAS SNA network diagram output for a fictitious user's network would look as follows.



Source: *(Blomberg, 2012)*

Figure 2.16: Example for *SAS SNA* output with the network structure of fictitious users

The second very popular software environment for social media data analysis is IBM's Watson Content Analytics.

IBM

IBM Watson Content Analytics is also a multifunctional software environment, which offers the user a variety of tools that can be customized according to ones needs, as well. (IBM, 2014b) IBM Watson Content Analytics does not only extract data. Its main function is the analysis of the extracted, mainly unstructured data.
As can be read in the White Paper on IBM Watson Content Analytics in more detail (Feldman et al., 2012), the analytics process can be divided into four big steps. In the first step, data extraction via data crawlers is performed. Depending on the crawler, not only social media data, but also data from various sources, as the world wide web in general, can be extracted. IBM WCA has 30 crawlers for 300 different types of data or documents. The Boardreader Crawler is responsible for social media data extraction through API's. The interface of the Boarder Crawler looks as shown in figure 2.17.

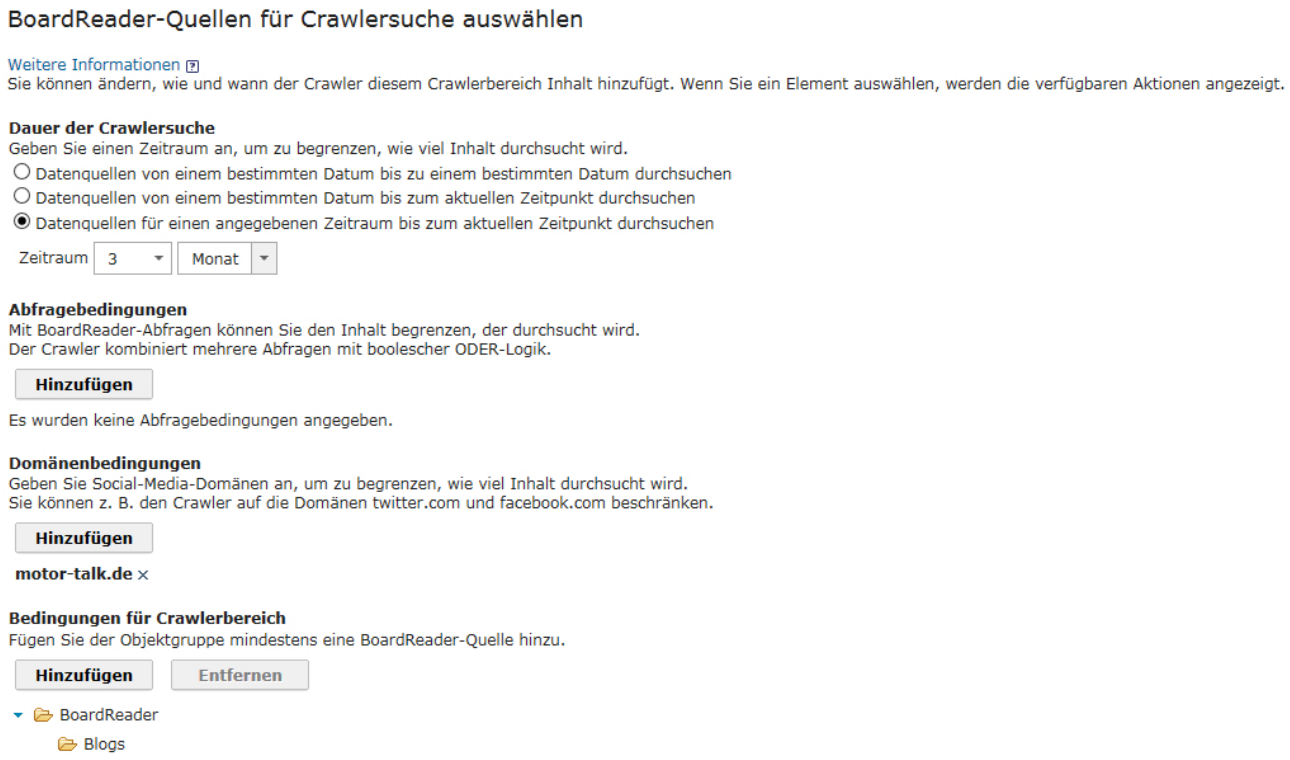
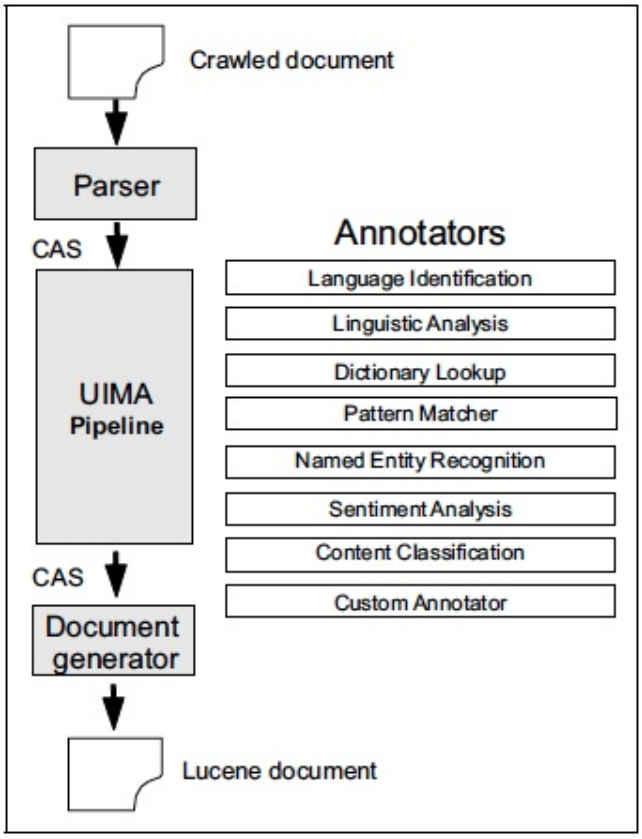


Figure 2.17: Example for *IBM Boardreadercrawler* interface

As one can see, the user can specify the sources, which he or she is interested in, and conditions, which have to be fulfilled for the search. Afterwards, one can specify the desired information (see table 2.14) that has to be extracted.
In the second step, document processing or rather data pre-processing is performed. Thereby, the unstructured data is "sent through the UIMA pipeline", where a text mining process is performed, which includes steps as language identification, text praising and filtering and sentiment analysis. (UIMA, 2013) Those document processing results are saved as indices, which are xml based annotations of the processed data. This means, that additional information is being added to the extracted and

pre-processed data. This metadata is used for the final step, the text, or rather statistical analysis. The following figure illustrates the whole process graphically.



Source: *(Feldman et al., 2012)*

Figure 2.18: Analytics process of *IBM Watson Content Analytics*

Now that data extraction tools and software were introduced in more detail, the most important and challenging part is the data preparation and analysis of such data. Structured information can be used in tabular form for analysis right away. Unstructured data, however, has to be brought into structured form first. As already presented, a few mostly not open source software environments offer such analysis tools. But what is actually the process behind those programs? A few terms and process steps were already mentioned. But how does the analytics of unstructured data actually work step by step and how can structured information, which has to be derived from the unstructured data, be combined with the existing internal data in order to be used for product development afterwards? These questions will be answered in the subsequent two sections in detail. Next section will look at the analytics and structuring of unstructured data. Chapter 4 will then deal with the matching process of the processed external and existing internal data. In each chapter, the step-by-step analytics processes together with a small example will be described, first. Afterwards, statistical and analytical methods behind the process steps will be presented in-depth.

3 Structuring and analysis of unstructured data

3.1 Unstructured data analytics processes

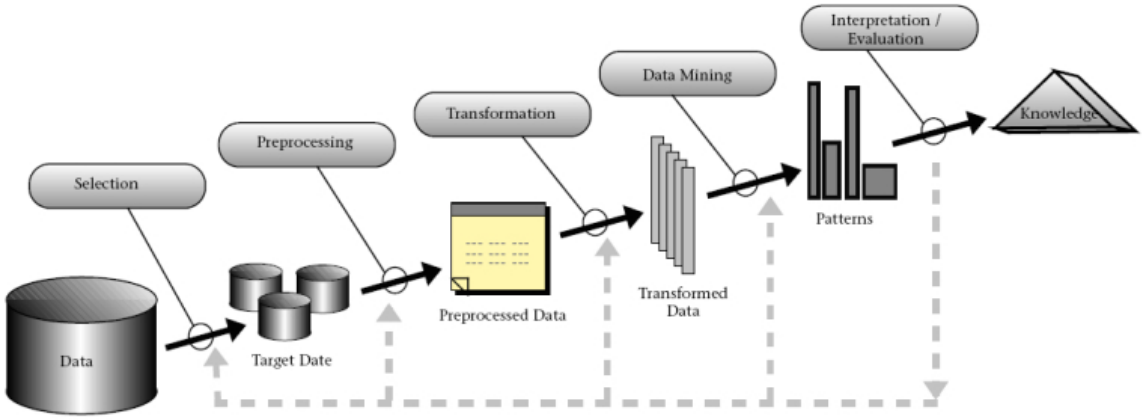
The content-based mining of social networks, as described in Aggarwal's *Social Network Data Analytics*, can be referred to the following four fields of application:

1. General data mining with arbitrary kinds of data
2. Text mining in social networks
3. Multimedia mining in social networks
4. Sensor and stream mining in social networks (Aggarwal, 2011)

In this thesis, the focus will lie on 'text analytics'. This term will be used for the denomination of the whole process of the analysis of text data instead of 'text mining', due to the fact that 'text mining' is usually considered as only one part of the entire process, although these two terms are both applied in literature as an umbrella term for the processing of text data to the same extent. As photos are also an important information source from Facebook, multimedia mining will be introduced, as well.

3.1.1 Text analytics processes

"Text Analytics is technology and process both, a mechanism for knowledge discovery applied to documents, a means of finding value in text. Solutions mine documents and other forms of 'unstructured' data. They analyze linguistic structure and apply statistical and machine-learning techniques to discern entities (names, dates, places, terms) and their attributes as well as relationships, concepts, and even sentiments [, as well as opinions]. They extract these 'features' to databases for further analysis and automate classification and processing of source documents." (Grimes, 2007) Thus, the aim of text analytics is to represent "textual documents that human beings can easily understand [...] in a form that can be mined by software. The raw documents need processing before the patterns and relationships that they contain can be discovered. Although the human mind comprehends chapters, paragraphs, and sentences, computers require structured (quantitative or qualitative) data. As a result, an unstructured document must be converted into a structured form before it can be mined." (SAS Institute Inc, 2012) This process from beginning to end is displayed in figure 3.1 graphically. Thereby, the first step 'Selection' has already been described in chapter 2.2. The 'Interpretation/Evaluation' process will be dealt with in chapter 5. Those two stages do not belong to the actual text analytics processes per se. That is why only the steps 'Preprocessing', 'Transformation' and 'Data Mining' will be discussed in this chapter.



Source: *(Miner et al., 2012)*

Figure 3.1: Entire process of text analytics

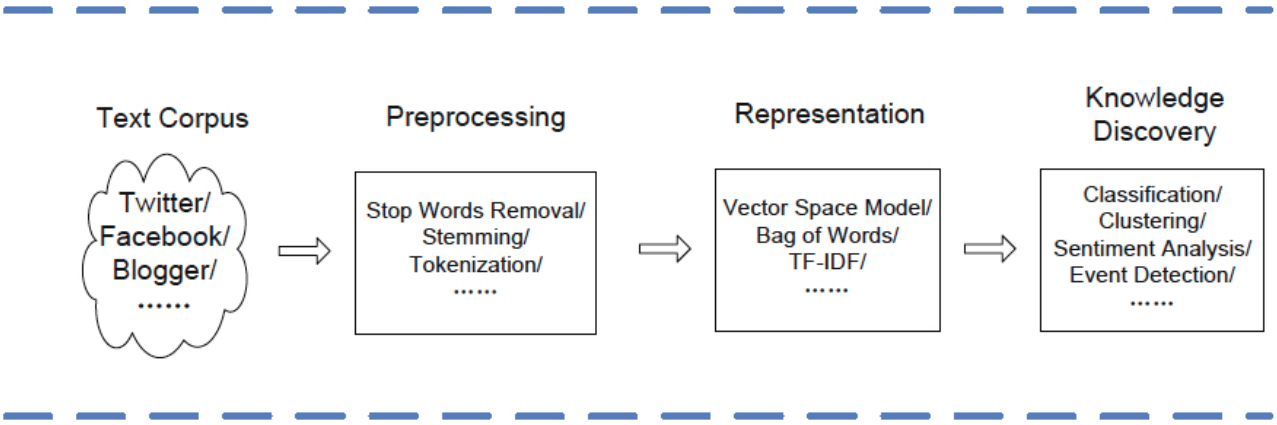
"Over time, the term 'text analytics' has evolved to encompass a loosely integrated framework by borrowing techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR) [, information extraction], [...] knowledge management [etc.]." (Chakraborty et al., 2013) The complexity of text analytics is displayed in the following graphic.

Source: *(Miner et al., 2012)*

Figure 3.2: Components of text analytics

Although every software has its own techniques and methods, as text analytics is quite a new research area and one optimal concept or approach does not exist, "fundamental methods for [modern and more complex] text [...][analytics] are natural language processing (NLP) and information extraction (IE) techniques. [...] NLP is a technology that concerns with natural language generation (NLG) and natural language understanding (NLU). IE involves directly with text mining process by extracting useful information from the texts. [...] IE can be described as the creation of a structured representation of selected information drawn from [natural language] texts. [...] IE produces structured data ready for post-processing [and further data mining.]" (Jusoh and Alfawareh, 2012)
As looking at all existing text mining techniques and methods will go beyond the scope of a master's thesis, only text analytics approaches which can be used for social network data will be displayed and explained in the following section step by step.
Before continuing to look at the individual text analytics process steps, it is important to understand the characteristics and arising challenges of text data. "The fundamental unit of a text is a *word*. Words are comprised of characters, and are the basic units from which meaning is constructed. By combining a *word* with grammatical structure, a *sentence* is made. Sentences are the basic unit of action in text, containing information about the action of some subject. *Paragraphs* are the fundamental unit of composition and contain a related series of ideas or actions. As the length of text increases, additional structural forms become relevant, often including sections, chapters, entire documents, and finally, a *corpus* of documents. In text mining study a *document* is generally used as the basic unit of analysis because a single writer commonly writes the entirety of a document and the document discusses a single topic." (Lee et al., 2010) Text data exists in different languages and formats. Its content can be very complex, as words and phrases create a context for each other. Sentences can also be nested, which makes the comprehension even more complex. Another problem arising from natural language not only while performing text analytics, but also in conversations between human beings, is the ambiguity of words and sentences. Ambiguity is thereby the "capability of being understood in two or more possible senses or ways." (Jusoh and Alfawareh, 2012) Same words can have different meanings, different words can have the same meaning depending on the context. Phrases or sentences can be interpreted and understood in different ways. Sarcasm can also be a big challenge, not only for software. A further challenge is noisy data, meaning that words can be misspelled, erroneous or abbreviated and sentences can be grammatically incorrect, as for example in chat conversations data. "The process of mining text in social media [presents additional challenges as it] requires the special ability to mine dynamic data which often contains poor and non-standard vocabulary." (Aggarwal and Zhai, 2012)
"A traditional text analytics framework consists of three consecutive phases: Text Preprocessing, Text Representation [or 'Transformation' as it is called in figure 3.1] and Knowledge Discovery [which is referred to the 'Data Mining' step in figure 3.1]." (Aggarwal and Zhai, 2012) Those phases, are divided up into different process steps (figure 3.3). In this thesis, the text analytics process will be looked at with the aid of an example, which is associated with potential reasons for the requirement of a dental cover, to illustrate each process step for better understanding. The following sentence will be used.

"A cig and coffe for breakfast is just wat makes me happy." (3.1)



Source: *(Aggarwal and Zhai, 2012)*

Figure 3.3: Traditional text analytics framework

3.1.1.1 Text preprocessing

"Understanding text information is fundamental to text mining. While the current approaches mostly rely on bag of words representation [where a sentence is viewed as a simple set of words without considering the syntactic structure of text (Lee et al., 2010)], it is clearly desirable to go beyond such a simple representation. Information Extraction (IE) techniques provide one step forward towards semantic representation. [...] [In order to apply IE techniques, the unstructured text has to be brought into a structured format through NLP first. This step is called text preprocessing.] Text preprocessing aims to make the input documents more consistent to facilitate text representation, which is necessary for most text analytics tasks." (Aggarwal and Zhai, 2012) "The task of natural language [...] [processing] is to take a sentence as input and return a syntactic representation that corresponds to the likely semantic interpretation of the sentence." (Ratnaparkhi, 1999)
This first phase is probably the most crucial, as all further analysis depends on the accuracy of the text preprocessing results. Depending on the specific text analytics technique of interest, preprocessing methods may vary. A common text preprocessing framework contains the following processing steps:

1. Tokenization
2. Language identification
3. Text normalization
4. Part of Speech (POS) Tagging + Lemmatization
5. Shallow Parsing
6. Syntactic Parsing
7. Filtering

1. Tokenization: Before being able to work with the given text, distinct sentences and words in them have to be determined. The aim is to split the string of characters into a set of tokens. The main cue for tokenization of sentences is hereby a period, exclamation or question marks at the end of the sentence, for the tokenization of words the white space between them. Additional features, as for example information regarding punctuation, special characters or types of capitalization can also be included. (Feldmann and Sanger, 2007) This method works for most languages, except for Chinese and some other Asian languages.
As to the example 3.1, a sentence will be detected based on the period, which will then be tokenized according to white space between words as follows.

A	**cig**	**and**	**coffe**	**for**	**breakfast**	**is**	**just**	**wat**	**makes**	**me**	**happy**	.

2. Language identification: After the text has been tokenized, it is crucial to identify the written language, as all further processing steps depend on that. Some software environments only offer text analytics for one specific language, so the text language has to be determined beforehand. Others as the Unstructured Information Management Architecture (UIMA), which provides a common framework for text preprocessing and IE, is able to first identify and then process 26 languages.
Different approaches can be used to identify the correct language out of many. One possible approach is called Small Words, where highly frequent words from different text data collections in each language are saved. The text of interest is then compared to those lists and the language with the most co-occurrences is chosen. (Grefenstette, 1995) Other, more advanced ones, compare the whole written text to dictionaries as for example Webopedia (Webopedia, 2014) in order to determine the correct language by means of co-occurences, as well. The third, most frequently used method is the N-gram approach. Thereby, unique sequences of n consecutive characters, also called n-grams, from the text are compared to the n-grams most common in text corpora in different languages. This approach will be closely looked at in chapter 3.2.
Applying this step on the exemplary sentence, will lead to the result, that this text is written in English, as all n-grams have the highest similarity with the n-grams in the English language.

3. Text normalization: In this next step, the individual word tokens are checked for typing errors and misspellings and if existent corrected afterwards, so that all words are spelled correctly.
There are several different methods for the correction of words, depending on the nature of the misspelling. A distinction has to be made between non-word errors, where the misspelled word is not equivalent to any word in a dictionary, for example *coffe*, and real-world errors, where the misspelled word corresponds to a word in the dictionary with a different meaning, for example *piece* and *peace*. The detection of real-world errors is way more complex, as the whole context of the word has to be considered. In this thesis, the focus will be made on non-word errors. Another distinction has to be made whether there is a single error or multiple errors made in a word. Two possible approaches for the correction of non-world errors are the Edit-Distance and the Noisy Channel Model approaches. Especially for multiple errors, the Noisy Channel approach, which will be closely looked at in chapter 3.2, is more accurate, as it also considers probabilities for the corrected word. (Tavast et al., 2012)

In the case of colloquial language, special dictionaries or lists containing known slang words are needed to find the equivalent correct terms. Thereby, the same approaches, as used for the correction of misspelled words, can be applied.
In the previous example, the words 'coffe', 'cig' and 'wat' will be corrected into standard English.

A cigarette and coffee for breakfast is just what makes me happy

Now, that the written text has been tokenized and normalized, morphological and lexical analysis can be performed.

4. Part of Speech (POS) Tagging: "POS tagging is the annotation of words with the appropriate POS tags based on the context in which they appear. POS tags divide words into categories based on the role they play in the sentence in which they appear. POS tags provide information about the semantic content of a word." (Feldmann and Sanger, 2007) The basic POS categories include nouns, verbs, adjectives, adverbs and prepositions. Depending on the issue of interest as well as the complexity of the text, further POS can be included. "Taggers can [thereby] be rule-based [meaning that they depend on grammatical rules, as for example on the morphological analyzer of English ENGTWOL (Voutilainen, 1995)], stochastic, or a combination [...].[of both]." (Abbott, 2010)
The stochastic, or as it also called statistical or corpus-based POS tagging, is nowadays the most frequently used approach, as it provides very accurate results with minimal amount of human effort or linguistic knowledge. Corpus-based POS taggers "[...] automatically learn to approximate syntactic and semantic knowledge for parsing from a large corpus of text, called a treebank, that has been manually annotated with syntactic information. " (Ratnaparkhi, 1999) Depending on the training data and tag set used, different POS tags for same words are possible. For the POS tagging of English texts, the Penn Treebank II tag set (*http://www.cis.upenn.edu/ treebank/*) is usually applied.
"Among recent top performing [machine learning] methods are Hidden Markov Models, maximum entropy approaches, and transformation-based learning. An overview of these and other approaches can be found in Manning and Schütze (1999, ch.10) [(Manning and Schütze, 1999)]. [The accuracy of these approaches lies between 95% and 98%. (Megyesi, 2002)] All these methods use largely the same information sources for tagging, and often almost the same features as well, and as a consequence they also offer very similar levels of performance." (Toutanova and Manning, 2000) That is why only one approach, namely the Maximum Entropy POS tagging will be introduced in chapter 3.2.
Applying the POS tagging on basis of the Penn Treebank II tag set to the exemplary sentence will lead to the following result. Wh-pronoun is thereby a special subclass of pronouns including a set of words beginning with wh-. verb(VBZ) refers to a 3rd personal singular present tense verbs.

A	**cigarette**	**and**	**coffee**	**for**	**breakfast**
determiner	noun(singular)	conjunction	noun(singular)	preposition	noun(singular)
is	**just**	**what**	**makes**	**me**	**happy**
verb(VBZ)	adverb	Wh-pronoun	verb(VBZ)	pronoun(personal)	adjective

"Usually, POS taggers at some stage of their processing perform morphological analysis of words. Thus, an additional output of a POS tagger is a sequence of stems (also known as "lemmas") of the input words." (Feldmann and Sanger, 2007)

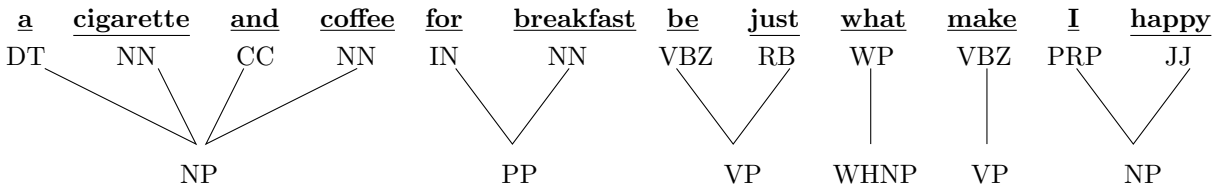
Lemmatization: Written text can contain different forms of a word, as well as families of derivationally related words with similar meanings. In some situations, it may be useful to search for one of these words to return information that contains another word in the set, as for example 'cigarette' and 'cigarettes'. The goal of both 'Stemming' and 'Lemmatization' is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form, a 'stem' and a 'lemma', respectively. (Manning et al., 2009)
"Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. [It does not consider the POS or the context in which the word occurs. There are three different classes of stemming algorithms, the truncating methods, where the suffixes and prefixes of a word are removed based on linguistic rules, statistical methods, where the words are stemmed based on statistical analysis and techniques, and mixed methods. (Jivani, 2011)]
Lemmatization [, the more accurate linguistic approach and of interest for this master's thesis,] usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma." (Manning et al., 2009) That means, that verbs are converted into their infinitive form and affixes as well as plurals, conjugations and declinations are removed in order to maintain a reasonable basic form of a real word. There is no simple rule or approach for lemmatization. Most algorithms are part rule-based, where POS tags are taken into consideration using for example ENGTWOL, which also contains morphology for lemmatization (Voutilainen, 1995), part dictionary-based, comparing words to a stemming dictionary as WordNet (*http://wordnet.princeton.edu*), and part machine learning results-based.
Applying lemmatization to the exemplary sentence, the word 'A' will become a lower case, the words 'is' and 'makes' will be converted into their infinitives and 'me' will be transferred into its basic form.

a	**cigarette**	**and**	**coffee**	**for**	**breakfast**
determiner	noun(singular)	conjunction	noun(singular)	preposition	noun(singular)
be	**just**	**what**	**make**	**I**	**happy**
verb(VBZ)	adverb	Wh-pronoun	verb(VBZ)	pronoun(personal)	adjective

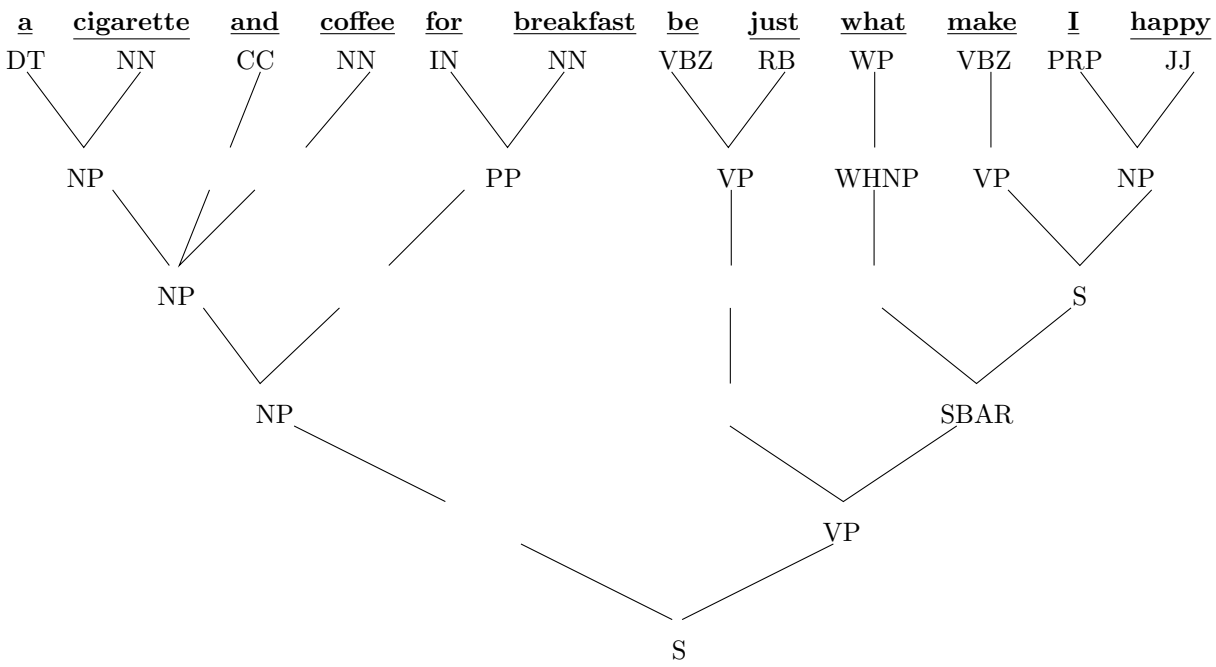
The next steps have the task to perform the actual syntactic analysis. "Linguistic analysis is the core component of any content analytics system. This component analyzes language at the syntactic or grammatical level, looking for the role that the words play in a sentence." (Feldman et al., 2012)

5. Shallow Parsing: Now, token sequences of interest can be identified. "Instead of providing a complete analysis (a parse) of a whole sentence, shallow parsers produce only parts that are easy and unambiguous. Typically, small and simple noun[, prepositional] and verb phrases are generated, whereas more complex clauses are not formed. [This procedure is also called 'chunking'.] Similarly, most prominent dependencies might be formed, but unclear and ambiguous ones are left out unresolved." (Feldmann and Sanger, 2007)
There are various approaches for chunking, some of which are purely linguistic and rule-based, others data-driven using different machine learning methods. As POS tagging has already been performed in the previous step, those results can be used for shallow parsing. There are various approaches for shallow parsing on the

basis of POS taggers introduced in the article of B. Megyesi, one of which will be introduced in chapter 3.2. (Megyesi, 2002) This approach treats chunking as a maximum entropy based POS tagging, introduced in the previous step.
Once again, looking example 3.1, the subsequent result after shallow parsing with POS taggers could be output, depending on the training set used. For overview purposes, only abbreviated POS and chunk tags as used in the Penn Treebank II tag set (*http://www.clips.ua.ac.be/pages/mbsp-tags*) are displayed. NP, thereby, stands for noun phrase, PP for prepositional phrase, VP for verb phrase and WHNP for Wh. noun phrase.



6. Syntactic Parsing: Shallow parsing does not lead to very informative results. Thus, if accurate parsing of whole sentences is of interest or necessary, the syntactic dependency between tokens has to be analyzed.
"Syntactic parsing components perform a full syntactic analysis of sentences according to a certain grammar theory. The basic division is between the *constituency* and *dependency* grammars. Constituency grammars describe the syntactic structure of sentences in terms of recursively built *phrases* - sequences of syntactically grouped elements. Most constituency grammars distinguish between noun phrases, verb phrases, prepositional phrases, adjective phrases, and clauses.[...] Additionally, the syntactic structure of sentences includes the *roles* of different phrases.[...] Dependency grammars, on the other hand, do not recognize the constituents as separate linguistic units but focus instead on the direct relations between words." (Feldmann and Sanger, 2007)
This process step enables further analysis, as the extraction of relationships between entities and events, facts or attributes. Syntactic parsing is very content specific and complex. There are various rule-based and corpus-based approaches. One possibility is to use the above mentioned maximum entropy approach combined with manually implemented grammars and rules. (Ratnaparkhi, 1999) (de Marneffe et al., 2006)
As to example 3.1, syntactic parsing according to *constituency* grammars, could lead to the following outcome.



While looking at the *constituency* grammars for this sentence, also the information regarding the role of the different phrases in a sentence will be output, which could produce one of the following outcomes, depending on the implementation used.

1. **a cigarette and coffee** (sentence subject) **for breakfast** (temporal) **be just** (predicate) **what make I happy** (purpose)

2. **a cigarette and coffee for breakfast be just what** (sentence subject) **make happy** (predicate) **I** (sentence object)

The resulting outcome can vary depending on content and implementation. In reality, the outcome after syntactic parsing is much more complex and contains a lot more information. The combination of both grammars together with manually implemented grammars can provide a relatively accurate analysis of the syntactic structure of written text, but is usually complicated to implement, very content specific and time-consuming.

7. Filtering: Due to reasons of resource intensity, in this last preprocessing step, all irrelevant and filler words have to be removed, as they do not contain any useful information and thus are not needed for further analysis. Such words, also called stop words, are depending on the context and issue of interest. In the case, that just clustering or keywords extraction is of interest, stop word lists, which include terms as 'is', 'and', 'for', 'if' et cetera, can be generated using the maximum entropy approach. Thereby all words, that are not associated with the subject, can be detected. All words contained in those stop word lists can then be removed from the output. (Sinka and Corne, 2003)
As to further process of this thesis subject, such a list has to be created manually, due to the fact, that words, as 'and' and 'for', which are contained in most stopword lists, can contain valuable information regarding the connection between entities and thus should not be removed.
In the exemplary sentence 3.1, only the words 'a' and 'just' will be removed, as those do not contain valuable information and are therefore irrelevant for further analysis.

As already stated, this process chain represents a basic framework for text preprocessing. Here, only a very simple exemplary sentence was used for demonstration purposes. In reality, written text is much more complex and difficult to analyze. All these process steps can be and usually have to be adapted to the individual requirements of the user with own training data, dictionaries and ontologies and also have to be constantly used for supervised or iterative learning by implementing machine learning algorithms. This learning process is very important for improvement of the accuracy of results, due to the fact that text analytics is exposed to various challenges as mentioned before. Process steps required for some applications are missing in this representation. However, some process steps displayed here may not be relevant for other applications or have to be changed in their order of execution.

3.1.1.2 Text representation

Now, that the unstructured text has been split into its components and the information of interest has been extracted, this information has to be transformed and saved into a structured format in order to be able to perform data mining or rather IE. The most common way to model documents is to transform them into sparse numeric vectors and then deal with them with linear algebraic operations. This representation is called "Bag Of Words" (BOW) or "Vector Space Model" (VSM). In these basic text representation models, the linguistic structure within the text is ignored and thus leads to "structural curse". (Aggarwal and Zhai, 2012)

Other transformation techniques include latent semantic analysis (LSA) and latent semantic indexing (LSI). Alternative methods are the counting of keywords in a text (TF) and counting the number of documents the term occurs in (DF), as well as the combination of TF and the inverse of DF. All these applications lead to the creation of a term-by-document matrix. More advanced approaches are n-grams and reduced dimensionality features (PCA).
Nevertheless, all these types of text representation are very simple and do not fully represent the information content contained in the unstructured text, as they do not consider the syntactic structure of the text. The most advanced text representation method up-to-date deals with this problem. This method is called indexing which is a framework, where the preprocessed data is saved in an XML format enriched with metadata, as for example POS tags and parsing results together with the location of tokens or phrases in text. At the same time, the data can be saved in a relational database before the indexing process.
An extract of the data from the previous example enriched with metadata in XML format, which will be output using the UIMA pipeline is shown below. The sentence and its length is identified in the first line.

```
xmi:id="60" name="filesize" value="65" decimal="l;65"/><uimatypes:UppercaseAlphabetic
xmi:id="532" sofa="15" begin="0" end="65"/><tt2:ParagraphAnnotation xmi:id="536" sofa="15"
begin="0" end="65"/><annotation_type:ContiguousContext
xmi:id="178" sofa="15" begin="0" end="1" tokenProperties="1" lemma="425"
dictionaryMatch="true" lemmaEntries="425" frost_TokenType="0" posTag="DT"
takmiPOS="determiner" tokenNumber="1"/><uimatypes:LowercaseAlphabetic
xmi:id="197" sofa="15" begin="2" end="11" tokenProperties="4" lemma="431"
dictionaryMatch="true" lemmaEntries="431" frost_TokenType="0" posTag="NN"
takmiPOS="noun"tokenNumber="2"><ftrs>singular</ftrs>
</uimatypes:LowercaseAlphabetic><uimatypes:LowercaseAlphabetic
xmi:id="543" sofa="15" begin="2" end="11" category="$._word.noun.general"
representation="cigarette"/><annotation_type:ContiguousContext
xmi:id="216" sofa="15" begin="12" end="15" tokenProperties="4" lemma="440"
dictionaryMatch="true" lemmaEntries="440" frost_TokenType="0" posTag="CC"
takmiPOS="conjunction"tokenNumber="3"><ftrs>coordinating</ftrs>
</uimatypes:LowercaseAlphabetic><uimatypes:LowercaseAlphabetic
xmi:id="549" sofa="15" begin="12" end="15" category="$._word.conj"
representation="and"/><annotation_type:ContiguousContext
xmi:id="235" sofa="15" begin="16" end="22" tokenProperties="4" lemma="449"
dictionaryMatch="true" lemmaEntries="449" frost_TokenType="0" posTag="NN"
takmiPOS="noun" tokenNumber="4"><ftrs>singular</ftrs>
</uimatypes:LowercaseAlphabetic><uimatypes:LowercaseAlphabetic
xmi:id="555" sofa="15" begin="16" end="22" category="$._word.noun.general"
representation="coffee"/><annotation_type:ContiguousContext
```

`posTag` and `takmiPOS` contain information regarding the part of speech of each word. `tokenNumber` refers to the position of the token in the sentence. `begin` and `end` encodes the position of the beginning and ending character of each word in the sentence, which itself is saved under `representation`. As all lemma forms are stored in a specific dictionary, the lemma position of each word in that dictionary is also saved under `lemma`.

3.1.1.3 Knowledge discovery

"When we successfully transformed the text corpus into numeric vectors [or other structured form], we can apply the existing machine learning or data mining methods [for knowledge discovery] [...]." (Aggarwal and

Zhai, 2012) As already stated, only Information extraction (IE) techniques that are relevant for social network analytics with regard to product development will be looked at, namely:

1. Keyword search
2. Entity, attribute, fact and event extraction
3. Sentiment and opinion analysis
4. Visualization of results

"IE [...] aims at pinpointing the relevant information and presenting it in a structured format - typically in a tabular format." (Feldmann and Sanger, 2007) In this process the syntactic structure is translated "into a semantic representation that is [a] precise and unambiguous representation of the meaning expressed by the sentence." (Jusoh and Alfawareh, 2012) In this case, the research question has to be decided beforehand, as all further approaches depend on the issue of interest.

1. Keyword search: It is often of interest to get the idea, what information is contained in the given text or in parts of the text. Therefore, most relevant sequences of one or more words have to be extracted. There are several different approaches for the extraction of keywords, as for example Rapid Automatic Keyword Extraction (RAKE), a graph-based ranking algorithm (TextRank) or a chi-square measure, which are introduced in the 'Text Mining: Applications and Theory' book. (Berry and Kogan, 2010) These approaches are only useful, if the user does not know, what he or she is looking for in the text.
In this case, as the business management area of application and thereby also the new products of interest have already been determined beforehand, one element of the RAKE method, namely the word co-occurrence graph, can be used for the distinct word tokens from the preprocessing step to find out, whether the content of the given text is of interest for the research subject. Therefore, the look is taken at, whether the words in the given text have a co-occurrence with a manually created set of terms connected with the need of potential new products. The more co-occurrences between the tokens of the given text and the manually created set of terms connected with the research question there are, the higher the probability, that the given text is of interest for further information extraction. If there is no co-occurrence what so ever, that information can be removed. All information connected to the research subject is then looked at in more detail in the next processing step.
In case of exemplary sentence 3.1 and a list of words and phrases associated with objects, interests and habits possibly requiring dental cover, the words 'cigarette' and 'coffee' will have a co-occurrence with the issue of interest. Thus, the text is of interest, as is may contain valuable information regarding the research subject, and will now be analyzed further in the next steps.

2. Entity, attribute, fact and event extraction: If, after the keyword search, information content, that is connected with the subject of interest, has been found, this information has to be examined and the actual information content extracted, more precisely the entities together with their attributes, facts and events. All this information can then be saved in a structured relational form.
In this step, also called "*domain analysis* [...] the system combines all the information collected from the [...] [text preprocessing] and creates complete frames that describe relationships between entities. Advanced domain analysis modules also possess an anaphora resolution component. Anaphora resolution concerns itself with resolving indirect (and usually protomic) references for entities that may appear in sentences other than the one containing the primary direct reference to an entity." (Feldmann and Sanger, 2007)
Hence, first entities, which refer to people, companies, objects, products, interests et cetera, have to be extracted by means of manually created lists for identification of entities. A distinction thereby has to be made between the writer of the text or other named entities and other entities as objects or products. As social network data

is the subject of attention, the writer is already known. All relational information regarding attributes, facts or events connected with those entities were detected during the syntactic parsing process and can now be used together with diverse dictionaries and user specific lexicons. This purely linguistic process is described in the Text Mining Handbook in detail and will thus not be elaborated on at this point. (Feldmann and Sanger, 2007) As to the exemplary sentence 3.1, no events will be found. The entities 'cigarette', 'coffee' and 'breakfast', as well as an entity referring to a person 'I', will be detected. If syntactic parsing was performed correctly, the attribute 'happy' will be assigned to the entity 'I'. Afterwards, the fact, that 'cigarette' and 'coffee' belong together and are connected to 'I' and thus to 'happy' can be determined.

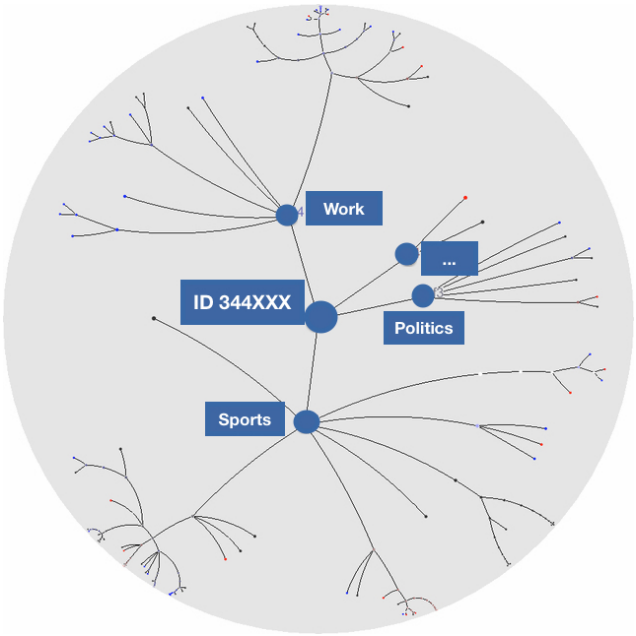
3. Sentiment analysis and opinion mining: In addition, it is often of interest to extract sentiments, attitudes or opinions of the persons of interest towards certain entities or aspects. "The field of sentiment analysis deals with categorization (or classification) of opinions expressed in textual documents. [...] Using sentiment analysis, an organization can identify and extract a customer's attitude, sentiment, or emotions toward a product or service. This is a more advanced application of text analytics that uses NLP to capture the polarity of the text: positive, negative, neutral, or mixed. With the advent of social networking sites, organizations can capture enormous amounts of customers' responses instantly." (Chakraborty et al., 2013)
Sentiment analysis can be performed on a document, a sentence or an entity and aspect level. Here for, currently four different types of approaches exist, namely "keyword spotting, lexical affinity, statistical methods, and concept-based techniques. Keyword spotting [...] [, which is] the most naive approach, [...] classifies text by affect categories based on the presence of unambiguous affect words [collected from different lexicons] such as happy, sad, afraid, and bored. [...] Keyword spotting is weak in two areas: it can't reliably recognize affect negated words, and it relies on surface features. [...] Lexical affinity [...] not only detects obvious affect words, it also assigns arbitrary words a probable 'affinity' to particular emotions. For example, lexical affinity might assign the word 'accident' a 75-percent probability of indicating a negative affect, as in 'car accident' or 'hurt by accident.' This approach usually trains probability from linguistic corpora. Although it often outperforms pure keyword spotting, there are two main problems with this approach. First, negated sentences (I avoided an accident) and sentences with other meanings (I met my girlfriend by accident) trick lexical affinity, because they operate solely on the word level. Second, lexical affinity probabilities are often biased toward text of a particular genre, dictated by the linguistic corpora's source. This makes it difficult to develop a reusable, domain-independent model. Statistical methods [...], which include Bayesian inference and support vector machines, are popular for affect text classification. [...] By feeding a machine-learning algorithm a large training corpus of affectively annotated texts, the system might not only learn the affective valence of affect keywords (as in the keyword-spotting approach), but also take into account the valence of other arbitrary keywords (similar to lexical affinity), punctuation, and word co-occurrence frequencies. [Nevertheless,] [...] statistical methods are semantically weak, which means that individually - with the exception of obvious affect keywords - a statistical model's other lexical or co-occurrence elements have little predictive value. As a result, statistical text classifiers only work well when they receive sufficiently large text input. So, while these methods might be able to affectively classify a user's text on the page level or paragraph level, they don't work well on smaller text units such as sentences or clauses. Concept-based approaches [...] use Web ontologies or semantic networks to accomplish semantic text analysis. This helps the system grasp the conceptual and affective information associated with natural language opinions. By relying on large semantic knowledge bases, such approaches step away from blindly using keywords and word co-occurrence counts, and instead rely on the implicit meaning/features associated with natural language concepts." (Cambria et al., 2013)
This process step is the most challenging, as not only ambiguity in words, but also of sentences, as well as sarcasm and other difficulties of understanding play an important role in the analysis. (Liu, 2012) For the research question of this master's thesis, however, this IE technique plays an important role, as the information whether

the user, for example wrote, that he or she likes smoking or hates smoking, makes a big difference. Although the facts 'cigarette', 'coffee', 'make', 'I' and 'happy' have been extracted in the previous step, the program has not determined that 'happy' propositions an affection of the entity 'I' towards 'cigarette' and 'coffee'. This is where sentiment analysis comes into play. As the focus lies on finding out the sentiment of potential clients for supplementary insurance towards a certain aspect, the opinion target, sentiment analysis on an entity and aspect level is of interest.
In the preceding example, the following sentiment can be detected.

Positive sentiment: Opinion holder 'I' towards entity/aspect 'cigarette' and 'coffee' without time information

"[As] industries, organizational structures, professional roles, and even specific task contexts can require highly differentiated, domain-specific term and entity definitions, structures, [sentiments] and relationships, [...] [all three IE extraction approaches can be adapted to the individual requirements by] specify[ing] their own lists of terms [and dictionaries] and define their significance in the context of the application." (Feldman et al., 2012)

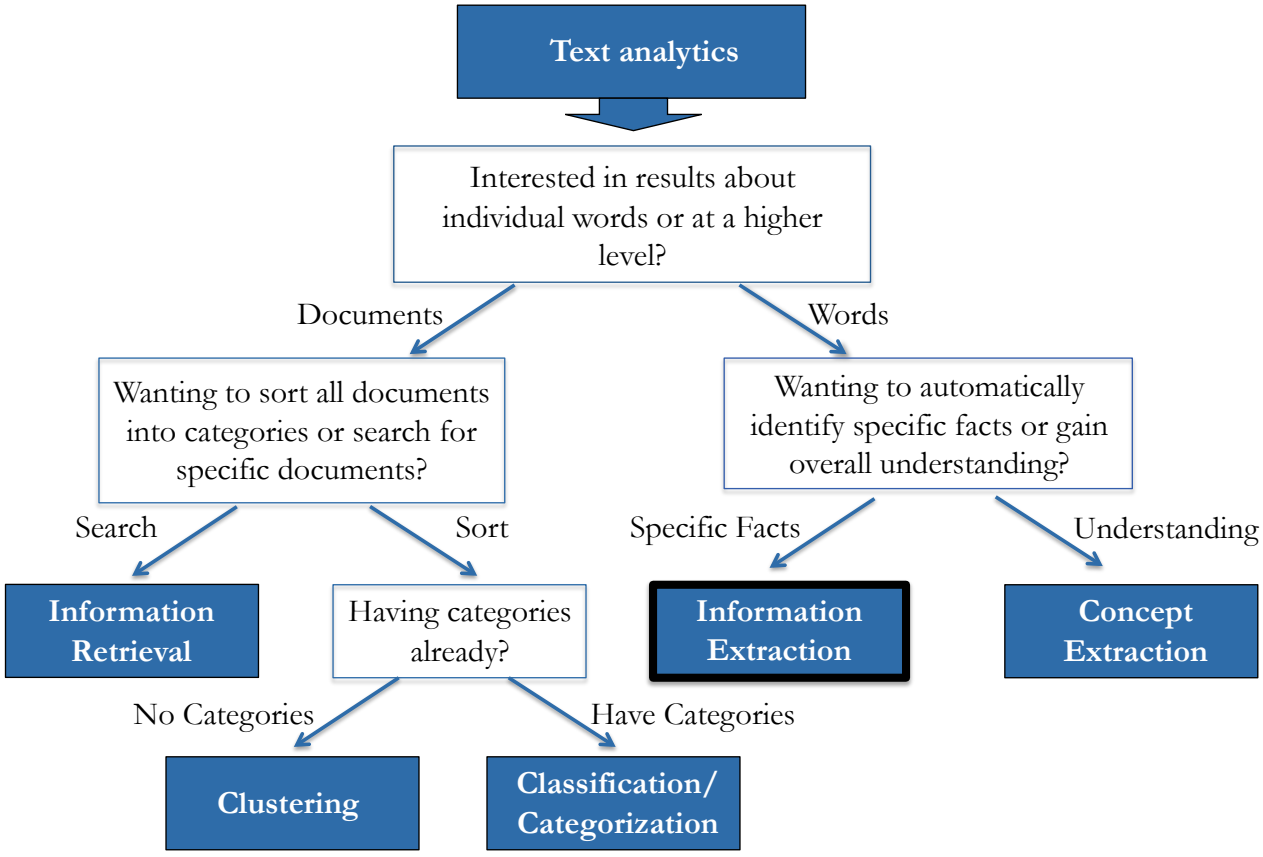
4. Visualization of results: In some cases, after IE, it is also useful to look at the extracted results graphically in order to get a better impression and understanding about the underlying information content and identify patterns in data. Especially the look at illustrated relationships or correlations between entities and their attributes can be of interest. As looking at external data separately from internal data is not the main goal of this thesis, the visualization of results will just be shortly introduced.
There are various visualization techniques. Simple ones include timelines, bar charts, pie charts, circle graphs, simple concept set and simple concept association graphs. More complex graphics illustrate self-organizing maps, hyperbolic trees or even 3-D effects. A combination of those methods is also possible. All these methods are presented in Feldmann's and Sanger's 'Text Mining Handbook' in more detail (Feldmann and Sanger, 2007)
Hyperbolic trees are the best-suited visualization method for the analysis of social network data, as they can illustrate important information content with regard to product development.



Source: Adapted from *http://en.wikipedia.org/wiki/Hyperbolic_tree*

Figure 3.4: Exemplary basic hyperbolic tree

"[This] [...] approach gives more display area to a part of a hierarchy (the focus area) while still situating it within the entire [...] context of the hierarchy. [...] A hyperbolic tree visualization allows an analyst always to keep perspective on the many attached relationships of a highlighted feature." (Feldmann and Sanger, 2007) Nodes in focus are placed in the center and given more room, while out-of-focus nodes are compressed near the boundaries (see figure 3.4). Most of these approaches have in common, that a larger amounts of text are required, as no useful information can be visualized, otherwise.
Further frequently used text mining applications, which will not be handled in this thesis, are displayed in figure 3.5. Thereby, choosing the right method depends on what kind of information is of interest. If the interest lies in simply finding relevant documents concerning one specific topic, 'Information Retrieval' is used. If the user wants to assign each document a few keywords in order to group them into topics he or she needs to use text 'Classification' or 'Categorization'. "In categorization problems [...] we are provided with a collection of preclassified training examples, and the task of the system is to learn the descriptions of classes in order to be able to classify a new unlabeled object. in the case of clustering, the problem is to group the given unlabeled collection into meaningful clusters without any prior information." (Feldmann and Sanger, 2007)



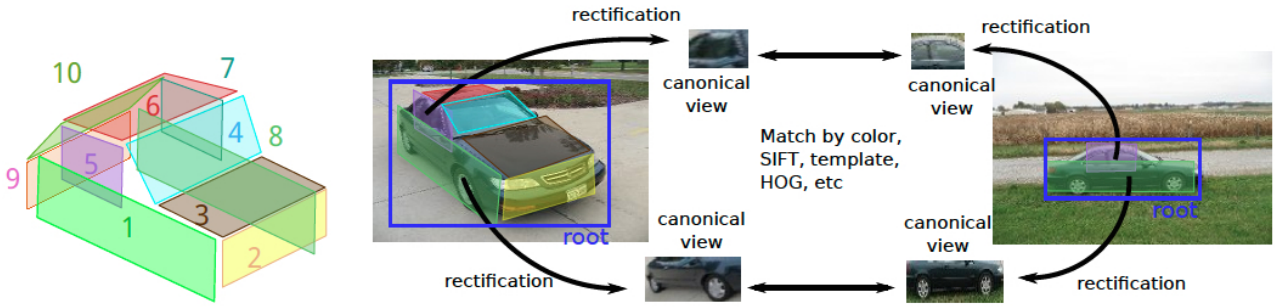
Source: *(Miner et al., 2012) Edited version*

Figure 3.5: Text analytics in different practice areas

Additional technologies available and not displayed in figure 3.5 include concept linkage, summarization, topic tracking and question answering. (Fan et al., 2006) Most of these applications are implemented in various software packages of different software providers. The most popular are SAS, R and IBM text mining tools.

3.1.2 Image mining processes

Another very important information source in social networks are images that can contain information regarding ones friends, hobbies and likings. Looking at a picture will reveal that information to the user, but how can this information be extracted by software?
"Image mining is more than just an extension of data mining to image domain." (Hsu et al., 2002) Up-to-date, image mining techniques include object detection, image indexing and retrieval, association rules mining, image classification and clustering, and neural networks. In this thesis, object detection will be looked at, as it is the most suitable procedure for mining social network images for information regarding product development. "The classical approach to object detection is to train object detectors from manually labeled bounding boxes in a set of training images and then apply the detectors on the individual test images. Despite previous success, this strategy only focuses on obtaining the best detection result within one image at a time and fails to leverage the consistent object appearance often existent when there are multiple related images. A promising alternative, called object co-detection, is to simultaneously identify 'similar' objects in a set of related images and use intra-set appearance consistency to mitigate the visual ambiguity." (Guo et al., 2013)
S. Y. Bao, Y. Xiang and S. Savarese (2012) suggest a very precise novel approach for object co-detection. They designed a co-detector that aims to detect objects in all images, to recognize if objects in various images are the same, which are then referred to as matching objects and to estimate the viewpoint transformation between the matching objects. The method they introduce "jointly detects and matches objects by their parts [by] leverag[ing] existing part-based object representation models. [...] Co-detection is related with and potentially useful to several other problems in computer vision [, namely,] object detection, [...] single instance 3D object detection, [...] image co-segmentation, [...] tracking by detection, [...] semantic structure from motion (SSFM), [...] single instance matching [and] [...] region matching [...]." (Bao et al., 2012) An example here for is shown in figure 3.6.



(a) A 3D part representation for a car.

(b) A 3D part representation allows to match objects across images by matching their parts after viewpoint rectification.

Source: *(Bao et al., 2012)*

Figure 3.6: Example for object representation

Appearance consistency between objects is measured by matching their parts, because part-based object representation is more robust to viewpoint changes and self-occlusions. Information from multiple images is combined by introducing an energy based formulation that models both the object's category-level appearance similarity in each image and the instance's appearances across images. As a detailed explanation of the actual process behind this method will go beyond the scope of this master's thesis, a reference is made to the article of S. Y. Bao, Y. Xiang and S. Savarese (2012). (Bao et al., 2012)
As an illustrative example with regard to the research question, photos with users holding cigarettes can be

considered in order to detect users who smoke and are thus potential clients for a dental cover. Thereby the difficulty lies in the fact, that not only the perspective can differ, but also the size of the cigarette as well as its shape and color, as can be seen in the subsequent figure.

Sources (From left to right, top to bottom):

www.locallyhealthy.co.uk./sites/default/files/imagecache/node_large/node_images/smokers_hand.jpg
static.guim.co.uk/sys-images/Money/Pix/pictures/2013/3/14/1363273382381/A-person-smoking-a-cigare-008.jpg
static.guim.co.uk/sys-images/Guardian/About/General/2010/7/27/1280247131464/Man-smoking-cigarette-006.jpg
www.dafttrunk.com/wp-content/uploads/2014/01/mk.jpg
i.telegraph.co.uk/multimedia/archive/02317/smoke_2317163b.jpg
i1.thejournal.co.uk/incoming/article5267266.ece/alternatives/s2197/407088JPG.jpg

Figure 3.7: Different perspectives and sizes of a cigarette

In order to be able to detect cigarettes in pictures, it necessary to create training images of cigarettes from different perspectives, in different sizes while holding it in different ways. To be able to determine the accuracy of the object co-detection not only for this particular, but for any object, first of all a certain number of images with and without those objects contained in them have to be chosen together with training images after which the hit ratio can be calculated. If necessary, the variety of object in the training images and thus the number of those has to be increased, in order to obtain higher hit ratios. Then, the co-detector can compare various images of the actual users with all the training images, which will allow more accurate and reliable matching results.

3.2 Statistical models behind the text analytics processes

Now, that all process steps have been described in the previous chapter, the statistical methods behind those text analytics processes will be displayed and explained.

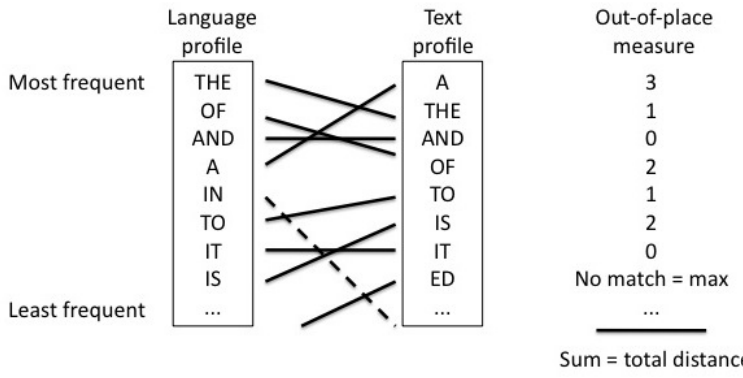
3.2.1 Text preprocessing

3.2.1.1 Language identification with the N-gram approach

The N-gram approach is the "most widely used method[...] to programmatically identify the language of a given text [...]. The basic idea is [...] [to] train a language identifier on a large corpus of text from a given language [, as every language uses certain n-grams more frequently than others]. 'Training' means gathering compression/frequency/information/data on n-gram occurrence." (Hanumathappa and Reddy, 2012) This approach is based on findings of William B. Cavnar and John M. Trenkle, 1994. (Cavnar and Trenkle, 1994)
For the example 3.1, all n-grams with $n = 1 - 4$ for the word 'cig' including the white space in the beginning and the end of the string, will look as follows. Thereby, uni-grams are just the letters of the word themselves.

uni-gram: _, C, I, G
bi-gram: _C, CI, IG, G_
tri-gram: _CI, CIG, IG_
quad-gram: _CIG, CIG_

Using the N-gram approach for language identification, first, sample texts, also called training sets, of sizes 20K to 120K bytes for each language have to be obtained and the n-gram frequency profiles have to be calculated. This means, that all overlapping 'n-character slices' of a longer string are used simultaneously to calculate the most frequent n-grams and rank them in descending order. Usually the most frequent 300-400 n-grams are considered, as they provide the best results. For better understanding, an ASCII character requires 1 byte, an utf-8 character 2 bytes of space. Afterwards, all n-grams from the given text have to be formed and the n-gram profile calculated. For both profiles, uni-grams, except for the words consisting of one character, as for example 'a' in the English language, are usually not considered, as they just reveal information about the alphabetical distribution of a language. In the third step, the overall distance between the language profiles and the text profile using the out-of-place measure have to be calculated. Out-of-place measure thereby means, that comparing one language profile with the text profile, the difference between ranks of same n-grams in those two profiles are calculated and summed up. If a n-gram from the text profile cannot be found in the language profile, the out-of-place measure is set to the number of n-grams used in the language profile.



Source: Fictitious language and text profiles for English language adopted from (Cavnar and Trenkle, 1994)

Figure 3.8: Illustration of the n-gram rank order approach from (Cavnar and Trenkle, 1994)

For better illustration, another fictitious text profile than from the exemplary sentence 3.1 is used, as this sentence is so short.
The language profile with the smallest distance to the text profile is then picked as the language of text. As to the example 3.1, English will still be chosen as the language of text, even if the total distance will be higher than with longer texts, because the distance to other language profiles is even larger.
Even for text profiles with less than 300 bytes using 300 to 400 n-grams in language profiles, an accuracy of over 98% can be achieved. Other advantages of this approach are, that it is robust, does not require linguistic knowledge and most importantly that spelling errors or foreign as well as slang words produce only a minimal offset, so that correct identification is not prevented.

3.2.1.2 Noisy channel model for text normalization

The task of spelling correction models is to detect errors in text and correct them afterwards. As only non-word errors are considered here, any word not in a dictionary is considered an error.
"The misspelling of a word is viewed as the result of corruption of the intended word as it passes through a noisy communications channel. The task of spelling correction is a task of finding, for a misspelling w, a correct word $r \in D$, where D is given dictionary and r is the most probable word to have been garbled into w. Equivalently, the problem is to find a word r for which $P(r|w)$ [...] is maximized." (Toutanova and Moore, 2002) Thus, the spelling correction based on noisy channel models consists of the following four steps[(Kernighan et al., 1990), (Brill and Moore, 2000)]:

1. Finding misspelled words $w \notin D$ and proposing candidate corrections $r \in R$ for them
2. Scoring each candidate correction r by applying Bayes' formula:

$$P(r|w) = \frac{P(w|r) \cdot P(r)}{\underbrace{P(w)}_{\text{cons.}}} \propto P(w|r) \cdot P(r), \tag{3.2}$$

   where the conditional probability $P(w|r)$ is the channel model or error model likelihood and $P(r)$ is the source model prior probability.
3. Choosing the candidate with the highest score according to

$$\hat{r} = argmax_r P(r|w) = argmax_r P(w|r) \cdot P(r). \tag{3.3}$$

So, in order to find the most probable correct word, $P(w|r)$ and $P(r)$ have to be calculated. Those probabilities are estimated by means of training data or prior knowledge. The error model comprises knowledge regarding the probability for misspelled word w given the candidate correction. The prior source model specifies the probability to find this candidate correction in a given language.
The source model prior probability $P(r)$ can be estimated through choosing a large amount of text data representing the specific language and calculating

$$\hat{P(r)} = \frac{freq(r) + 0.5}{N}, \tag{3.4}$$

where 0.5 is a correction due to possible zero counts, $freq(r)$ denotes the frequency of the candidate correction and N the total amount of words in this chosen text data corpus.
For the estimation of $P(w|r)$, different approaches exist, depending on the type of error. In the following, estimation of $P(w|r)$ for single and multiple error types will be looked at.

Single error

Kernighan et. al, 1990 introduced a model for the estimation of $P(w|r)$ for single errors in words, which is the most frequent type of spelling errors. They proposed, that the "[...] conditional probabilities are computed from four confusion matrices (see appendix [of their article]):

(1) del[x,y], the number of times that the characters xy (in the correct word) were typed as x in the training set,
(2) add[x,y], the number of times that x was typed as xy,
(3) sub[x,y], the number of times that y was typed as x, and
(4) rev[x,y], the number of times that xy was typed as yx. " (Kernighan et al., 1990)

As to the exemplary sentence from chapter 3.1, there are no other candidate corrections for the words 'coffe' and 'wat' than 'coffee' and 'what' in the English language. Thus, in this case, the conditional probabilities can be computed solely from the *deletion* matrix.
Probabilities $P(w|r)$ are then estimated from the matrices above by dividing by chars [x,y] in the case of del[x,y] and rev[x,y] or chars [x] in the case of add[x,y] and sub[x,y], the number of times that xy and x appeared in the training set, respectively.

$$P(w|r) \approx \begin{cases} \frac{del[x,y]}{chars[x,y]}, & \text{if deletion} \\ \frac{add[x,y]}{chars[x]}, & \text{if insertion} \\ \frac{sub[x,y]}{chars[x]}, & \text{if substitution} \\ \frac{rev[x,y]}{chars[x,y]}, & \text{if reversal.} \end{cases} \tag{3.5}$$

The five matrices *del, add, sub, rev,* and *chars* can then be computed with a bootstrapping procedure.
As to the exemplary sentence, estimating equation 3.3 for 'coffe' will lead to $\hat{r} =$ coffee, for 'wat' to $\hat{r} =$ what.

Multiple errors

Brill and Moore, 2000, proposed an improved error model for noisy channel spelling correction, which allows the correction of multiple errors, meaning all edit operations of the form $\alpha \to \beta$, where α and β are strings in an alphabet, as well as considering the position of misspelling in the word. (Brill and Moore, 2000) Therefore, $P(\alpha \to \beta|\text{PSN})$ denotes the probability of a user typing string α instead of β conditioned on the position of the error in the string.
In simplified terms for easier understanding, leaving out the position of error, the assumption of this error model is, that first a person picks a word to write and then this person chooses a partition of the characters of that word. Then each partition is typed, possibly with spelling mistakes. To illustrate this process, the word 'coffe' from the exemplary sentence is chosen. The person wanted to type the word $r = coffee$ and used the partitions $cof\ fee$. What he or she actually typed was $cof\ fe$. Thus, the probability $P(coffe|coffee)$ can be estimated through $P(cof|cof) \cdot P(fe|fee)$.
The general model for the estimation of the conditional probability $P(w|r)$ considering the position 'PSN', can then be formulated as follows:

$$P(w|r) = \sum_{S \in Part(r)} P(S|r) \sum_{\substack{T \in Part(w) \\ |T|=|S|}} \prod_{i=1}^{|S|} P(T_i|S_i), \tag{3.6}$$

where $Part(w)$ is the set of all possible ways of partitioning string w and $Part(r)$ the set of all possible ways of partitioning string r into adjacent substrings. T and S are particular partitions from $Part(w)$ and $Part(r)$, respectively, and both consist of j contiguous segments.
This equation can be approximated to

$$P(w|r) \approx max_{\substack{S \in Part(r) \\ T \in Part(w)}} P(S|r) \prod_{i=1}^{|S|} P(T_i|S_i) \tag{3.7}$$

by only considering the best partitioning of w and r.
The multiple error model proposed by Brill and Moore, 2000 provides a 94% correction quote compared to the single error model quote of 90% and is thus more precise. Naturally, this approach can also be applied to the correction of single errors.

3.2.1.3 Maximum Entropy approach for POS Tagging and Shallow Parsing

POS Tagging

The ulterior motive for this approach was to find a statistical model which solves the statistical classification problem, more precisely estimating the probability of a certain POS tag occurring in a given context of tags and words. The proposition to use a Maximum Entropy model for POS tagging was made by A. Ratnaparkhi in 1996, as a Maximum Entropy model combines forms of contextual information in a principled manner and does not impose any distributional assumptions on the training data. (Ratnaparkhi, 1996)
First of all, a large text corpus manually annotated with POS tags has to be chosen. The most popular and most often used annotated text corpus for learning probability distributions in English language nowadays is the Wall Street Journal corpus from the Penn Treebank project (Marcus et al., 1994) annotated with the Penn Treebank II tag set.
In the next step, the probability model is defined, which assigns a probability for every tag t in the set of all allowable tags T given a word and tag context h out of a set of possible word and tag contexts, or in other words, "histories" H.

$$p(h,t) = \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(h,t)} \tag{3.8}$$

$Z(h)$ is thereby a normalization constant ensuring a proper probability distribution, $\{\alpha_1, ..., \alpha_k\}$ are the positive parameters of the model and $f_j(h,t) \in \{0,1\}$ the corresponding features to the associated α_j. Those "features typically express a co-occurrence relation between something in the linguistic context and a particular prediction." (Ratnaparkhi, 1997) The specific word and tag context a feature can contain is given by

$$h_i = w_i, w_{i+1}, w_{i+2}, w_{i-1}, w_{i-2}, t_{i-1}, t_{i-2}. \tag{3.9}$$

"The model generates the space of features by scanning each pair (h_i, t_i) in the training data with the feature 'templates' given in [...] [figure 3.8]. Given h_i as the current history, a feature always asks some yes/no question about h_i, and furthermore constraints t_i to be a certain tag. The instantiations for the variables X, Y and T are obtained automatically from the training data." Thereby 'rare' words correspond to 'rare' and 'unknown' in the data of interest to tag.

Condition	Features	
w_i is not rare	$w_i = X$	& $t_i = T$
w_i is rare	X is prefix of w_i, $\|X\| \leq 4$	& $t_i = T$
	X is suffix of w_i, $\|X\| \leq 4$	& $t_i = T$
	w_i contains number	& $t_i = T$
	w_i contains uppercase character	& $t_i = T$
	w_i contains hyphen	& $t_i = T$
$\forall\ w_i$	$t_{i-1} = X$	& $t_i = T$
	$t_{i-2}t_{i-1} = XY$	& $t_i = T$
	$w_{i-1} = X$	& $t_i = T$
	$w_{i-2} = X$	& $t_i = T$
	$w_{i+1} = X$	& $t_i = T$
	$w_{i+2} = X$	& $t_i = T$

Source: (Ratnaparkhi, 1996)

Figure 3.9: Feature template for the generation of a feature space for the maximum entropy model

Considering the following sample of annotated training data

word	is	time	for	dinner	party
Tag	VBZ	NN	IN	NN	NN
Position	1	2	3	4	5

Table 3.1: Sample from fictitious annotated training data for POS tagging

these features can be generated from h_3:

$$\begin{aligned}
w_i &= \text{for} && \& \ t_i = \text{IN} \\
w_{i-1} &= \text{time} && \& \ t_i = \text{IN} \\
w_{i-2} &= \text{is} && \& \ t_i = \text{IN} \\
w_{i+1} &= \text{dinner} && \& \ t_i = \text{IN} \\
w_{i+2} &= \text{party} && \& \ t_i = \text{IN} \\
t_{i-1} &= \text{NN} && \& \ t_i = \text{IN} \\
t_{i-2}t_{i-1} &= \text{VBZ NN} && \& \ t_i = \text{IN}\ .
\end{aligned}$$

Assuming, that, for example the feature

$$f_j(h_i, t_i) = \begin{cases} 1 & w_i = \text{"}for\text{"} \text{ and } t_i = \text{ IN} \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

"is in the feature set of the model, its corresponding model parameter will contribute towards the joint probability $p(h_i, t_i)$ when w_i [...] [is the word 'for'] and when [...] [t_i = IN]. Thus a model parameter α_j effectively serves as a 'weight for a certain contextual predictor, in this case [...] [the word 'for' toward the probability of observing the tag 'IN']". (Ratnaparkhi, 1996)
Now, this model with the generated space of features "can be interpreted under the Maximum Entropy formalism, in which the goal is to maximize the entropy of a distribution subject to certain constraints." (Ratnaparkhi, 1996) "The idea of maximum entropy modeling is to choose the probability distribution p that has the highest entropy out of those distributions that satisfy a certain set of constraints. The constraints restrict the model

to behave in accordance with a set of statistics collected from the training data. The statistics are expressed as the expected values of appropriate functions defined on the contexts h and tags t. In particular, the constraints demand that the expectations of the features for the model match the empirical expectations of the features over the training data." (Toutanova and Manning, 2000) Only if those k constrains are all met, then p is consistent with the observed evidence. (Ratnaparkhi, 1997)
According to the general formula (Ratnaparkhi, 1997), the entropy of distribution p is defined as

$$H(p) = -\sum_{h \in H, t \in T} p(h,t) log p(h,t), \tag{3.11}$$

with the goal to find

$$p^* = argmax_p H(p). \tag{3.12}$$

The constraints that are imposed on the features' expectations are

$$Ef_j = \tilde{E}f_j, \qquad 1 \leq j \leq k, \tag{3.13}$$

with the models feature expectation

$$Ef_j = \sum_{h \in H, t \in T} p(h,t) f_j(h,t) \tag{3.14}$$

and the observed feature expectation

$$\tilde{E}f_j = \sum_{i=1}^{n} \tilde{p}(h_i, t_i) f_j(h_i, t_i). \tag{3.15}$$

Thereby, $\tilde{p}(h_i, t_i)$ denotes the observed probability in the training data. In order to estimate the parameters of this model, Generalized Iterative Scaling has to be used, which can be looked up in the introductory paper to maximum entropy by A. Ratnaparkhi. (Ratnaparkhi, 1997)

Shallow Parsing

This machine learning approach cannot only be used for POS tagging, but also expanded to shallow parsing of written text. M. Osborne proposed to use the maximum entropy approach for chunking as POS tagging without modifying the POS tagger's internal operation described above. This approach yields an accuracy of results of almost 95% which is comparable with other even more elaborate approaches. (Osborne, 2000)
The basic idea behind shallow parsing using the maximum entropy approach is to encode additional information regarding the surrounding lexical/POS syntactic environment in the training data and use this information to find and label chunks with the maximum entropy in the data of interest. Such additional information includes chunk-types labels and suffixes or prefixes of words.

word	w_1	w_2	w_3
POS Tag	t_1	t_2	t_3
Chunk	c_1	c_2	c_3

In their experiment, the best overall accuracy and performance for prediction of chunk labels was achieved when considering the current tag t_1, next tag t_2, current chunk label c_1, last two letters of the chunk label c_2, the first two and the last four letters of the current word w_1. Depending on the training set and text of interest, additional information can be included, as for example the previous and the third chunk label. Manual linguistic restrictions and also the combination with rule-based approaches is possible, as well.

3.2.2 Knowledge discovery

3.2.2.1 Keyword search using a word co-occurrence graph

The aim in this thesis is not to extract keywords, but search for certain ones. Here for, the word co-occurrence graph also applied in the RAKE method can be used. (Berry and Kogan, 2010)
First of all, a list L containing all words in lemma form associated with the research subject has to be assembled manually using various dictionaries. Afterwards, the given tokenized, normalized and lemmatized text is taken and a graph of word co-occurrences created. On the vertical axis, all distinct tokens from the given text, on the horizontal axis words from the manually created list L are placed. Then, it has to be coded, whether words on the vertical axis are contained in that manually created list and if so, calculated how often. Fields of non-matching words remain empty.
Such a word co-occurrence graph for example 3.1 with regard to dental supplementary cover could look as follows.

	box	chocolate	cigarette	coffee	fight	smoke	...
cigarette			1				
and							
coffee				1			
for							
breakfast							
be							
what							
make							
I							
happy							

Afterwards, for each given text the frequency of co-occurrences $freq(w)$ is summed up. In this example

$$\sum freq(w) = 2. \tag{3.16}$$

If the sum is equal to 0, this text is not considered for further analysis and can be removed due to utilization of resource capacity.

3.2.2.2 Sentiment analysis and opinion mining

"An *opinion* is a quintuple, $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The sentiment s_{ijkl} is positive, negative, or neutral, or expressed with different strength/intensity levels, e.g., 1 to 5 stars as used by most review sits on the Web. When an opinion is on the entity itself as a whole, the special aspect GENERAL is used to denote it. Here, e_i and a_{ij} together represent the opinion target." (Liu, 2012)
To determine an opinion on an aspect-based level, aspect-based sentiment analysis or opinion mining, as it is also called, has to be performed, which consists of the following six main tasks:

1. **Entity extraction and categorization:** Extract all entity expressions in document or text D, and categorize or group synonymous entity expressions into entity clusters (or categories). Each entity expression cluster indicates a unique entity e_i.
2. **Aspect extraction and categorization:** Extract all aspect expressions of the entities, and categorize these aspect expressions into clusters. Each aspect expression cluster of entity e_i represents a unique aspect a_{ij}.
3. **Opinion holder extraction and categorization:** Extract opinion holders for opinions from text or structured data and categorize them. The task is analogous to the above two tasks.
4. **Time extraction and standardization:** Extract the times when opinions are given and standardize different time formats. The task is also analogous to the above tasks.
5. **Aspect sentiment classification:** Determine whether an opinion on an aspect a_{ij} is positive, negative or neutral, or assign a numeric sentiment rating to the aspect.
6. **Opinion quintuple generation:** Produce all opinion quintuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) expressed in document or text D based on the results of the above tasks. (Liu, 2012)

Steps one and two can be considered as one step, due to the fact the assumption is made that all opinion targets as GENERAL aspects. Opinion targets can be identified using the word co-occurrence graph introduced in the previous chapter. All words from the text contained in list L can be considered GENERAL aspects and thus the opinion target. In order to categorize these aspects, for example referring to aspect 'Marlboro' and 'cigarette' as one category 'smoking', list L has to be enhanced with categories for each possible aspect.
The opinion holder, as already explained, is automatically extracted due to the fact that on social network sites written text is assignable to one specific user. Time extraction takes place automatically during the data extraction process.
The remaining task is the aspect sentiment classification. Using a purely lexical approach, which will be a suitable but merely simplified approach in this case, includes the following steps:

1. **Mark sentiment words and phrases:** For each sentence containing one or more aspects, all sentiment words and phrases in the sentence. The sentiment score of +1 is assigned to each positive and a sentiment score of -1 is assigned to each negative word according to a sentiment lexicon. Content specific, but not actual sentiment words, as "*long* battery life" can be determined through corpus based approaches. For the previous exemplary sentence, the output will be the following: "cigarette and coffee for breakfast be what make I happy [+1]".
2. **Apply sentiment shifters:** Sentiment shifters are words and phrases that can change sentiment orientations. There are several types of such shifters. Negation words like not, never, none, nobody, nowhere, neither, and cannot are the most common type. If such words are determined within the sentence, the sentiment score is changed to the apposite.
3. **Handle but-clauses:** Words or phrases that indicate contrary need special handling because they often change sentiment orientations, too. The most commonly used contrary word in English is "but". A sentence containing a contrary word or phrase is handled by applying the following rule: the sentiment orientations before the contrary word and after the contrary word are opposite to each other if the opinion on one side cannot be determined. Apart from "but", phrases such as "with the exception of", "except that", and "except for" also have the meaning of contrary and are handled in the same way.
4. **Aggregate opinions:** This step applies an opinion aggregation function to the resulting sentiment scores to determine the final orientation of the sentiment on each aspect in the sentence. Assuming the sentence

s with a set of aspect $\{a_1, \ldots, a_m\}$ and a set of sentiment words or phrases $\{sw_1, \ldots, sw_n\}$ with their sentiment scores obtained from steps 1-3, the sentiment orientation for each aspect a_i in s is determined by the following aggregation function:

$$score(a_i, s) = \sum_{ow_j \in s} \frac{sw_j.so}{dist(sw_j, a_i)}, \tag{3.17}$$

where sw_j is an sentiment word/phrase in s, $dist(sw_j, a_i)$ is the distance between aspect a_i and sentiment word sw_j in s. $sw_j.so$ is the sentiment score of sw_i. The multiplicative inverse is used to give lower weights to sentiment words that are far away from aspect a_i. If the final score is positive, then the opinion on aspect a_i in s is positive. If the final score is negative, then the sentiment on the aspect is negative. It is neutral otherwise. (Liu, 2012) For the exemplary sentence, the following score will be calculated:

$$score(cigarette, coffee) = \frac{+1}{8} + \frac{+1}{6} \approx 0.04, \tag{3.18}$$

which leads to a positive sentiment towards coffee and cigarette.

The following final opinion quintuple will be generated for the exemplary sentence:

(e_i/a_{ij} = cigarette, coffee, s_{ijkl} = positive, h_k = I, t_l = unknown here)

As already explained in the previous chapter, concept-based approaches are the most advanced and accurate approaches up-to-date and particularly suitable for social media data. In a sort, concept-based approaches represent a mixture between lexical approaches and semantic knowledge. M. Grassi, E. Cambria, A. Hussain and F. Piazza (2011) suggested a the Scentic Web approach "which exploits [Artificial Intelligence] AI and Semantic Web techniques to extract, encode, and represent opinions and sentiments over the Web. In particular, the computational layer consists in an intelligent engine for the inference of emotions from text, the representation layer is developed on the base of specific domain ontologies, and the application layer is based on the faceted browsing paradigm to make contents available as an interconnected knowledge base." (Grassi et al., 2011) As the introduction of this approach in further detail will go beyond the scope of this master's thesis, due to the fact that it is held purely on a statistical basis, reference is made to their paper for deeper insight into this method.

As a final remark, it has to be noted, that not much progress has been made in the text analytics field regarding new and more effective approaches and methods during the past decade. There is no ultimate approach or framework that provides optimal results. But, as demonstrated, a number of alterations and innovations was made in this research area which enables more accurate results than a few years ago.

4 Combination of external and internal data

Now, that all information of interest has been extracted from unstructured data, it can be combined again with all corresponding extracted, already structured personal information of users, as their title, first name, last name, birth date, gender, location et cetera. As a result, data in a tabular structured format containing all information on users with regard to the research question is obtained. This external dataset in a ready-to use form can now be combined with the existing internal data warehouse.

4.1 Matching processes

4.1.1 Exact matching using record linkage

First of all, external information related to existing clients has to be linked to the internal data warehouse. Therefore, record linkage can be used, as it combines data related to the same individual or entity from different datasets. Here for, usually, personally identifying and socio-demographic information on persons are used. Such information include the first and last names, birth date gender and location. In the following, the linkage of only two, the external and internal datasets is considered. Depending on the conformity of the information on a person in the two datasets, two different types of record linkage can be used. If all chosen information on one person in both datasets is identical, deterministic record linkage can be applied. As usually not all information on a person in social networks is available or correct, another approach called probabilistic record linkage has to be used in order to find the most probable match in the two datasets. To illustrate the processes of these two approaches, which will be described in this section, the following fictitious example will be used.

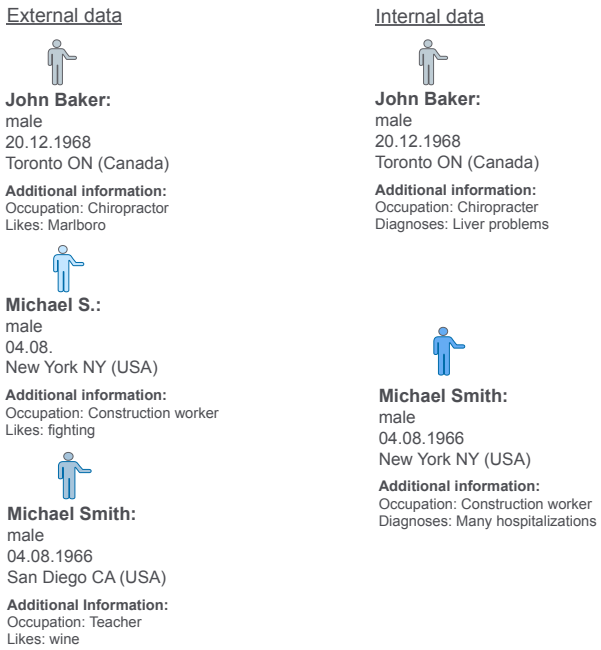


Figure 4.1: Fictitious example for the illustration of deterministic and probabilistic record linkage

Deterministic record linkage

Deterministic record linkage, also referred to as exact merging, is a very straightforward approach, as it simply links persons with identical identifiers as a match. Thereby, records are linked if they either agree exactly on all or on a predetermined set of matching fields or variables. But, matching variables will not necessarily uniquely identify matches, as the information from social network data used for deterministic linkage is limited, not always complete or correct, and does not provide unique identifiers such as the social security number. This approach is purely rule based and is only applicable if there is a complete conformity between the chosen matching fields. (Winkler, 1995)
Considering five personally identifying matching variables, namely the first and last names, the gender, birth date and city of residence in the example from figure 4.1 will lead to a complete match between John Baker in the external dataset and the existing client John Baker in the internal data. Then all information to this person from both datasets is combined as shown in the following figure.

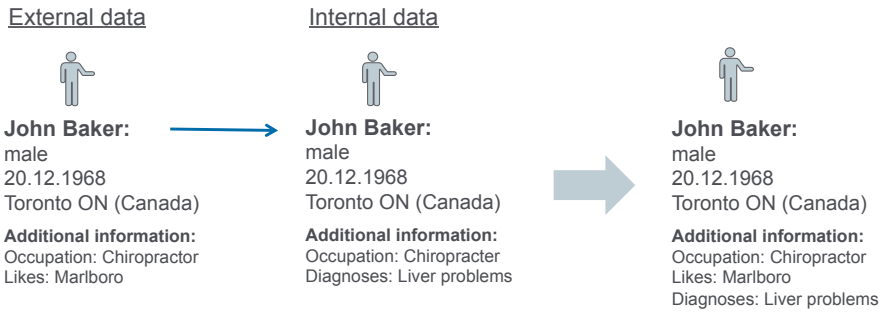


Figure 4.2: Result of deterministic record linkage for fictitious example

Any of the two Michael Smiths from the external data will not be linked to the Michael Smith in the internal data warehouse, as there is no exact match between all five personally identifying variables.
Due to the fact, that social network data is not always correct or available, as already mentioned, deterministic record linkage is not very practical. Instead, probabilistic record linkage, which is a statistical approach and is also applicable to records with complete conformity, should be used.

Probabilistic record linkage

This approach, also called probabilistic merging, uses, besides standard personally identifying information, as name and address, weighted additional information, as occupation and marital status to calculate the probability that two given records refer to the same entity. The weights for those additional potential identifiers are computed based on an estimated ability to correctly identify a match or non-match.
The most frequently used algorithms for this approach nowadays are still based on the model proposed by I. P. Fellegi and A. B. Sunter in 1969 after applying algorithms which remove typographical errors and standardize the matching fields' information, such as the first and last names. They developed this mathematical model in order to provide a theoretical framework for a computer-oriented solution. It is based on decision rules, where "a comparison is to be made between the recorded characteristics and values in two records (one from each file) and a decision made as to whether or not the members of the comparison-pair represent the same person or event [meaning whether they are a match or nonmatch in reality], or whether there is insufficient evidence to justify either of these decisions at stipulated levels of error. These three decisions are referred to as a link, a non-link, and a possible link. The first two decisions are called positive dispositions." (Fellegi and Sunter, 1969)
Record pairs with probabilities above a certain threshold are decided as links, record pairs with probabilities

below another threshold are decided as non-links and everything in between as possible links. Those cases usually have to be reviewed manually.
The parameters of the Fellegi-Sunter model can either be obtained directly from observed data or using maximum-likelihood-based methods, such as the Expectation-Maximization (EM) algorithm, in order to improve the overall matching performance and reduce false matches, as well as false nonmatches. False matches are thereby links made for entities that are nonmatching in reality and false nonmatches not linked truly matching records. Those are the two possible types of errors occurring in the Fellegi-Sunter model. (Winkler, 1995)
For the example from figure 4.1, the additional identifiers occupation and likes will be used in order to decide, that the first Michael Smith from external data is the Michael Smith in the internal dataset with the highest probability, as the occupation matches in both cases and the liking of fighting corresponds to frequent hospitalizations.

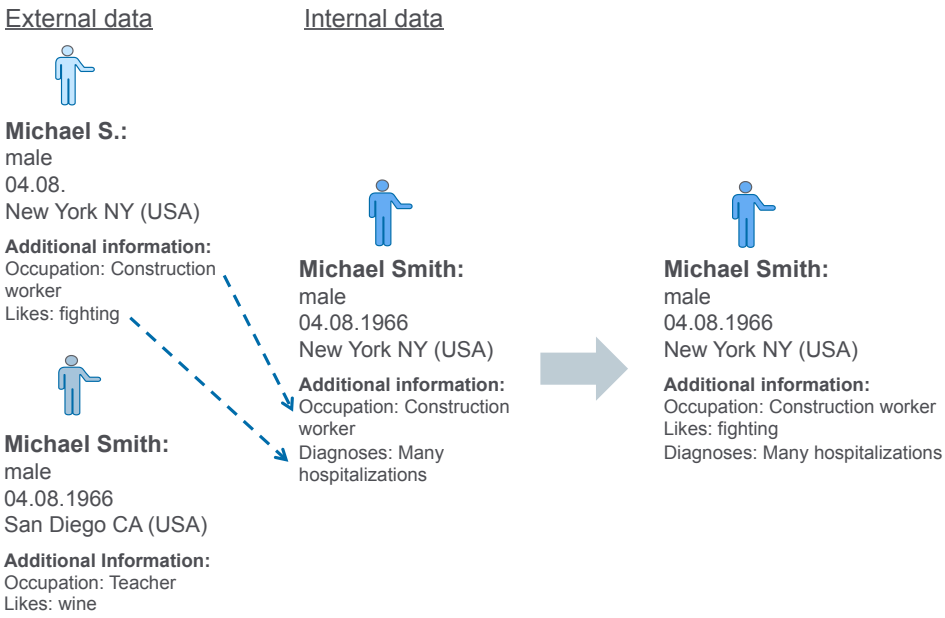


Figure 4.3: Result of probabilistic record linkage for fictitious example

It has to be considered, that in this case, as there is a lot of missing or false information in social network data, a manual review is still necessary in order to achieve accurate results.

4.1.2 Similarity matching

Now, that existing clients have been linked with additional information regarding the research question, if available, the obtained data warehouse can be used additionally to find potential new clients on Facebook similar to the existing clients. At this point, it has to be mentioned that, due to reason of data protection, it is not always possible to find exact matches. That is why this approach plays a very important role in the matching process. There are a few possible methods for the so-called similarity matching, as for example the principal component analysis. One other possible and very suitable approach in this case is the coarsened exact matching (CEM) algorithm, which is a statistical matching approach unlike the record linkage described in the previous section. Statistical matching methods bring together pairs of records with statistically similar characteristics, not necessarily representing the same entity. The CEM algorithm "[...] temporarily coarsens the data according to the researchers ideas [...] and then finds exact matches." (Iacus and King, 2011)

To illustrate this process, figure 4.3 will be considered. "Michael Smith" is an existing client who was linked with additional information from external data. After coarsening information such as the birth date into age group, the exact job title to occupational group, as well as city of location into country region, while keeping the likes fix, the following "statistical twins" will be assigned.

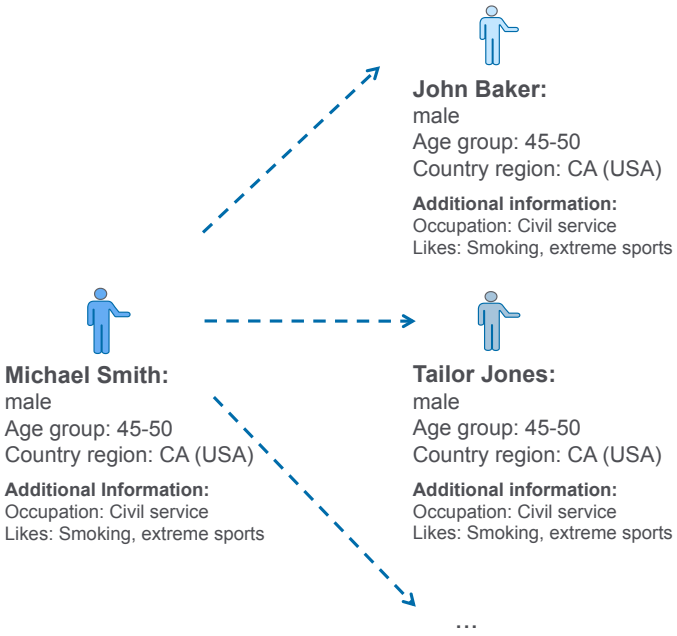


Figure 4.4: Result of coarsened exact matching for fictitious example

An alternative and very similar method is the matched pair design, which groups subjects into pairs based on blocking variables, whereby more than one pair can be formed. Blocking variables are thereby those variables according to which units are matched. The aim is to create homogeneous subgroups or blocks containing at least one existing client and one or more "statistical twins". Homogeneity is thereby determined by the distance between values of the blocking variables.

4.2 Statistical models behind the matching processes

Now, that approaches for the two different types of matching have been introduced, the statistical models behind those processes will be displayed. The following section is based on the original paper of Fellegi and Sunter, 1969, and Winkler, 1995. ((Fellegi and Sunter, 1969), (Winkler, 1995))

4.2.1 Probabilistic record linkage using the Fellegi-Sunter model

Two files A and B with contained entities a and b, respectively, are considered, where the set of ordered pairs from A and B are denoted as

$$A \times B = \{(a, b); a \in A, b \in B\}. \tag{4.1}$$

The record linkage process classifies pairs from this product set into two disjoint sets, namely the true matches

from files A and B

$$M = \{(a, b); a = b, a \in A, b \in B\} \tag{4.2}$$

and true nonmatches

$$U = \{(a, b); a \neq b, a \in A, b \in B\}. \tag{4.3}$$

The characteristics corresponding to the entities of A and B and used as matching variables, as for example name, birth date, gender, marital status, address, et cetera, are denoted as $\alpha(a)$ and $\beta(b)$, respectively. In the record linkage process, each characteristic or matching variable for two entities from files A and B is compared from which a comparison vector γ is generated, whose components are the coded agreements and disagreements on each characteristic, as for example "first names are the same", "last names are the same" and "birth dates are the same". The comparison vector γ is formally defined as follows:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], ..., \gamma^K[\alpha(a), \beta(b)]\}. \tag{4.4}$$

Additionally, each γ may account for the relative frequency of the occurrence of specific values of name components, such as the relative frequency of the occurrence of the last name "Smith". The comparison space Γ is then defined as the set of all possible realizations of γ.
A linkage rule assigns probabilities $P(A_1|\gamma)$ for a link, and $P(A_2|\gamma)$ for a possible link and $P(A_3|\gamma)$ for a non-link with $\sum_{i=1}^{3} P(A_i|\gamma) = 1$ to each possible realization of $\gamma \in \Gamma$ and is denoted as

$$d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\}, \gamma \in \Gamma, \tag{4.5}$$

where $d(\gamma)$ is a set of random decision functions.
There are two different types of errors associated with a linkage rule, due to diverse mistakes and incompleteness of records. Two members from files A and B can have identical records and will be linked, although they are unmatched. The probability for this type of error is given by

$$\mu = P(A_1|U) = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1|\gamma), \tag{4.6}$$

where $u(\gamma)$ denotes the conditional probability of γ given that $(a, b) \in U$.
The second type of error occurs in the case, if two members that are matching in reality have different records and will thus not be linked. The probability for this second type of error is given by

$$\lambda = P(A_3|M) = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3|\gamma), \tag{4.7}$$

where $m(\gamma)$ denotes the conditional probability of γ given that $(a, b) \in M$.
"An optimal linkage rule $L(\mu, \lambda, \Gamma)$ is defined for each value of (μ, λ) as the rule that minimizes $P(A_2)$ at those error levels. In other words, for fixed levels of error, the rule minimizes the probability of failing to make positive dispositions." (Fellegi and Sunter, 1969) In order to determine the optimal linkage rule the following steps have to be followed:

1. Exclude all γ if $m(\gamma), u(\gamma) = 0$
2. Define a unique ordering of the set of possible realizations of γ by positioning all γ in descending order according to their matching weight or score $R = m(\gamma)/u(\gamma)$. For equal value of R for more than one γ, those γ are ordered arbitrarily.

3. Apply the following decision rules:
   - If $R > T_\mu$, then designate record pair as a link
   - If $T_\lambda \leq R \leq T_\mu$, then designate record pair as a possible link
   - If $R < T_\lambda$, then designate record pair as a non-link

The cutoff thresholds T_μ and T_λ are determined by a priori error bounds on false matches and false nonmatches. If $\gamma \in \Gamma$ represents more than three variables, maximum-likelihood-based methods such as the Expectation - Maximization (EM) algorithm can be used in order to estimate the parameters needed for the decision rules (4.8). (Herzog et al., 2007)

4.2.2 Similarity machting using coarsened exact matching

"CEM is a monotonic imbalance bounding (MIB) matching method - which means both that the maximum imbalance between the [...] [existing clients and the potential new clients] may be chosen by the user ex ante, rather than discovered though the usual laborious process of ex post checking and repeatedly reestimating, and that adjusting the maximum imbalance on one variable has no effect on the maximum imbalance of any other." (Iacus et al., 2009) "CEM works in sample and requires no assumptions about the data generation process (beyond the usual ignorability assumptions)." (King et al., 2011) This very simple approach is usually used for finding "statistical twins" for units in a treatment group in order to estimate unbiased treatment effects and is described in S. M. Iacus, G. King and G. Porro, 2011 in full detail. (Iacus and King, 2011) Here, just the parts of the approach used for finding "statistical twins" from external data will be displayed.
"The CEM algorithm involves three steps:

1. Temporarily coarsen each control variable in [the set of matching variables] X as much as you are willing, for the purposes of matching. For example, years of education might be coarsened into grade school, middle school, high school, college, graduate school. [In general, it is left to the user to define the degree of coarsening depending on the measurement scale of each matching variable of interest.] [...]
2. Sort all units into strata [$s \in S$], each of which has the same values of the coarsened X.
3. Prune from the data set the units in any stratum that do not include at least one [existing client] [...] and one [...] [potential client from external data]. " (Iacus et al., 2009)

The existing clients assigned to stratum s are denoted as A^s and the number of A^s as m_A^s. All potential "statistical twins" from external sources assigned to stratum s are denoted as B^s and the number of B^s as m_B^s. The number of matched units are then denoted as $m_A = \cup_{s \in S} m_A^s$ and $m_B = \cup_{s \in S} m_B^s$ for existing and potential clients, respectively.
In order to enhance m_C^s, the maximum imbalance can be widened, meaning that the matching variables have to be coarsened even more. If less matched units are need and higher conformity between the matched units is of interest, matching variables can be coarsened less or even not at all.
As can be seen, this approach requires prior manual specification by the user in order to be able to find matching "statistical twins" and is thus very user specific and flexible.

4.2.3 Similarity machting using matched pair design

The matched pair design is another very popular approach for similarity matching, where subjects are grouped into pairs based on blocking variables according to the distance between values of these blocking variables.
Considering two disjoint sets of units $A = \{\alpha_1, ..., \alpha_K\}$, containing existing clients, and $B = \{\beta_1, ..., \beta_M\}$, containing potential new clients from Facebook, with $M \geq K$, units from A and B are matched, if a distance

$d(\cdot)$ calculated for the blocking variables of each pair is minimized. "The problem is to pick the best possible pairing, that is, to form the pairing to minimize the total distance within pairs. [...] [This issue] is called the optimal assignment problem or the optimal bipartite matching problem." (Lu and Rosenbaum, 2004).
Depending of the level of measurement of the blocking variables, different distance measures for the calculation of $d(\cdot)$ exist. As the data of interest contains both categorical and numerical variables, a distance measure for mixed-type data has to be considered. Here for, a General Distance Coefficient namely the Gower Distance can be used. "[...][T]he general distance coefficient between two data points $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$d_{gower}(\mathbf{x}, \mathbf{y}) = \left(\frac{1}{\sum_{k=1}^{d} w(x_k, y_k)} \sum_{k=1}^{d} w(x_k, y_k) d^2(x_k, y_k) \right)^{\frac{1}{2}}, \tag{4.8}$$

where $d^2(x_k, y_k)$ is a squared distance component for the kth attribute and $w(x_k, y_k)$ is [...] [either one or zero] depending on whether or not a comparison is valid for the kth attribute [...] [. I]f both data points $\mathbf{x}$ and $\mathbf{y}$ have observations at the kth attribute, then $w(x_k, y_k) = 1$; otherwise $w(x_k, y_k) = 0$. For different types of attributes, $d^2(x_k, y_k)$ is defined differently, as described below." (Gan et al., 2007)

- For numeric attributes, $d(x_k, y_k)$ is defined as

$$d(x_k, y_k) = \frac{|x_k - y_k|}{R_k}, \tag{4.9}$$

  where R_k is the range of the kth attribute.

- For ordinal attributes, values x_k and y_k have to be ranked and the corresponding ranks w_k and z_k of x_k and y_k, respectively, standardized according to

$$w_k^* = \frac{w_k - 1}{max(w_k - 1)} \qquad z_k^* = \frac{z_k - 1}{max(z_k - 1)}. \tag{4.10}$$

  Then, $d(x_k, y_k)$ is defined as

$$d(x_k, y_k) = \frac{|w_k^* - z_k^*|}{R_k}, \tag{4.11}$$

  where R_k is the range of the kth attribute.

- For binary and nominal attributes, $d(x_k, y_k)$ can be defined as

$$d(x_k, y_k) = \sum_{j=1}^{d} \frac{(n_{x_j} + n_{y_j})}{n_{x_j} n_{y_j}} \delta(x_j, y_j), \tag{4.12}$$

  where n_{x_j} and n_{y_j} are the numbers of objects in the data set that have categories x_j and y_j for attribute j, respectively, and

$$\delta(x, y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases} \tag{4.13}$$

Thereby, $d(x_k, y_k)$ has to fulfill the following assumptions:

1. Nonnegativity: $d(x_k, y_k) \geq 0$
2. Reflexivity: $d(x_k, x_k) = 0$
3. Commutativity: $d(x_k, y_k) = d(y_k, x_k)$
4. Triangle inequality: $d(x_k, y_k) \leq d(x_k, z_k) + d(z_k, y_k)$.

Matching pairs are then determined according to equation (4.8) choosing units from B with the minimal distance to the units from A:

$$min(d_{gower}(\mathbf{x}, \mathbf{y})), \tag{4.14}$$

More than one unit from B can be determined as a match for one unit from A by choosing units with the $2, ..., n$ minimum distances.

5 Data mining of combined database with regard to product development by means of a fictitious example

Now, that all information of interest has been extracted from Facebook, or any other social media source by that means, brought into a structured format and matched with the existing internal data warehouse, in this last step, the aim is to find existing clients who will buy a supplementary dental or inpatient cover. Hence, the target is cross-selling. Thereby, cross-selling refers to the practice of selling a complimentary or totally different product to an existing customer. As the number of existing clients of an insurance company is usually high, it is more efficient to target only those clients who have a high probability to buy the chosen new product.
In order to illustrate the process of finding potential clients for the new products, fictitious data containing existing clients and additional potential clients from Facebook with a few selected additional variables with information collected from Facebook is simulated and used. Afterwards, two approaches are introduced to find the target audience for the two new products.
The first suggested approach is scoring. Here for, it is assumed that a survey on the potential new clients from Facebook is performed regarding their interest in a supplementary dental and inpatient cover. By means of the simulated response information, a logistic regression is used to estimate weights for the influence of the external information on the probability for a person to buy a new product. Other approaches, such as the ridge regression (Malthouse, 1999), the Chi-squared Automatic Interaction Detection (CHAID) (Stroh, 2010) or other discriminant analysis (Knott et al., 2002) can be used for scoring, as well. The estimated scoring weights are then used to estimate the probability for existing internal clients to buy a supplementary dental or inpatient cover according to their external information.
The second approach introduced is clustering of units according to their lifestyle risk factors. Thereby, the whole dataset with the existing, as well as potential new clients from Facebook is used. Clustering, which is an unsupervised learning algorithm, is the process of grouping units together in such a way that units in the same group are more similar to each other than to those in other groups. There are many different clustering techniques, the most popular of which can be divided into three categories: Partitioning, hierarchical and model-based methods. (Rokach and Maimon, 2005) For clustering of lifestyle risk factors, where all units from the dataset are divided into a determined number of homogeneous clusters, each representing a certain lifestyle risk pattern according to their external information, the two-step cluster analysis is a very suitable approach. [(Busch et al., 2013), (Lv et al., 2011), (Landsberg et al., 2010)] The advantage of this algorithm is, that is can handle mixed type attributes in large datasets and also determines the number of clusters automatically. In the first step, units are divided into many small subclusters according to a Cluster Features (CF) Tree. In the second step, those subclusters are aggregated according to an agglomerative hierarchical clustering algorithm. In order to choose the target audience for the new products, clusters containing units with high lifestyle risks with regard to the new products are intuitively selected. (IBM, 2014a)
The flowchart in figure 5.1 shows the procedure in the following sections. Both approaches do not only focus on cross-selling, but also enable the acquisition of new clients from Facebook. Whereas, the first approach relies on actual data, the second approach is based on hypothetical assumptions.

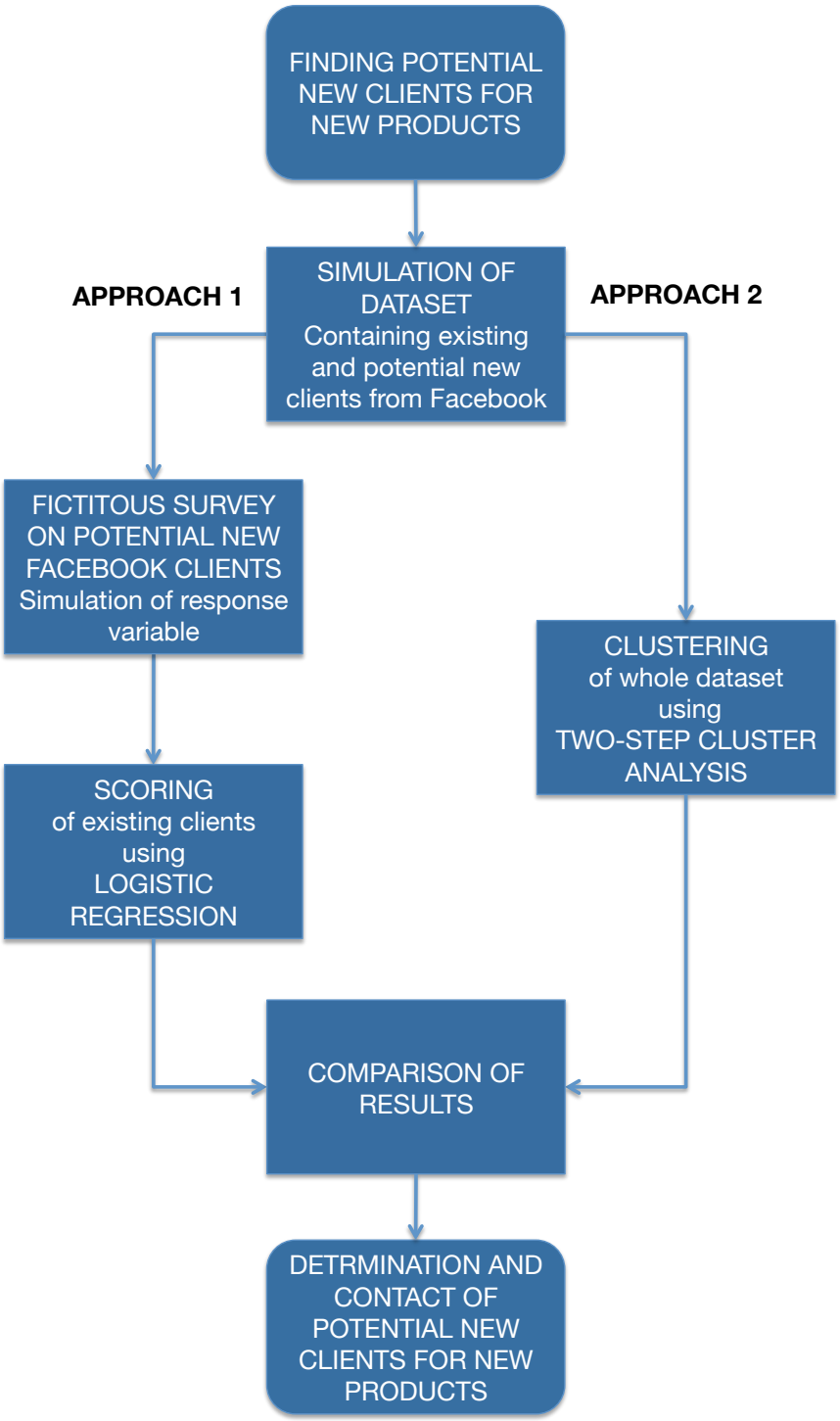


Figure 5.1: Flowchart of proceeding

For the selection of the target audience, results from both approaches shall be considered. For reasons of simplicity and clarity, only the relevant outputs will be displayed in the subsequent sections. All syntax and all results can be found on the attached CD.

5.1 Simulation of fictitious dataset with regard to supplementary dental and inpatient covers

The simulated dataset of 10000 observations contains the following personally identifying information, which is usually already available in internal data for existing clients.

- First name
- Last name
- Gender
- Birth date and age
- City and state of residence
- Occupation
- Marital status
- Internal risk profile score
- Customer value score

First and last names, as well as occupation are drawn from a list of most popular English names and professions. Birth date and age are simulated according to the assumption, that all persons are aged between 18 and 55 years at the time of 01.10.2014. Location of residence is drawn from a list of US cities with the largest population. Marital status, internal risk profile and customer value scores are simulated at random considering the fact, that younger persons are less likely to be married, have a high customer value and low internal risk profile score. Additionally, an indicator, whether the person is an existing or a potential new client from Facebook, is created. For those persons, the internal risk profile, as well as the customer value scores do not exist.
The following additional external information with regard to a dental supplementary cover

- Number of membership in groups, likes, attended events and pictures associated with sports harmful to dental: *nb_sport*
- Number of membership in groups, likes, attended events and pictures associated with smoking: *nb_smoke*
- Number of membership in groups, likes, attended events and pictures associated with harmful food and beverages: nb_food

and inpatient cover

- Number of membership in groups, likes, attended events and pictures associated with dangerous sports: *nb_danger_sport*
- Indicator, whether occupation is regarded as dangerous: *danger_occup*
- Number of posts/comments and searches regarding supplementary inpatient cover: *nb_interest_ip*

are simulated. All those variables, except the indicator variable, are assumed to follow a poisson distribution

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{5.1}$$

with overdispersion, as the observed variance in data is expected to be higher than the variance of normal poisson distribution, which equals to its mean. The overdispersion parameter is set as $\phi = 4$. Thus,

$$E(x) = \lambda, \qquad Var(x) = \phi E(x) = 4\lambda. \tag{5.2}$$

Regarding the number of membership in groups, likes, attended events and pictures associated with smoking, sports harmful to dental and dangerous sports, the assumption is made that younger persons have higher frequencies than older persons. Additionally, men are assumed to have higher frequencies than women, as far as sports harmful to dental and dangerous sports are concerned. Professions, such as policemen and firemen, fall under the category dangerous occupations. According to this, the indicator, whether occupation is regarded as dangerous, is assigned. Naturally, far more information from external data can be included. Here, only a few variables were chosen in order to illustrate the process in a simple and comprehensible manner.
In the next steps, this simulated dataset can be used to find potential clients to buy the new products.

5.2 Scoring using logistic regression

"Scoring models [...] use [...] information that a company has about a customer to predict whether or not the customer will be interested in a particular offer." (Malthouse, 1999) Hence, the aim is to calculate probabilities for existing customers to buy a new product using the external information provided in accordance with calculated scoring weights for each of the external predictor variables. As already mentioned, different scoring models exist depending on the type of data and number of predictor variables. Due to the fact, that there are only a few predictive variables available in the fictitious dataset, a simple logistic regression is used.

Determination of scoring weights

Since two totally new products are offered, there is no historical information on existing clients available regarding their interest in purchasing one of the supplementary covers. Thus, the scoring weights to calculate the probability of existing clients to purchase a new product have to be estimated using test data. In this case, the optimal way is to use the additional potential clients from Facebook determined through similarity matching and perform a survey, where these users are questioned, whether they have an interest in buying a supplementary dental and inpatient cover, separately.
In this thesis, their responses are simulated assuming, that users with high frequencies in the additional external information, respectively, are more willing to buy the new products. Those binary response variables

$$y_{response\_dental} = \begin{cases} 1, & \text{interested in buying a supplementary dental cover} \\ 0, & \text{not interested} \end{cases} \tag{5.3}$$

and

$$y_{response\_ip} = \begin{cases} 1, & \text{interested in buying a supplementary inpatient cover} \\ 0, & \text{not interested} \end{cases} \tag{5.4}$$

can then be used in a logistic regression

$$log\left(\frac{P(y_i = 1)}{P(y_i = 0)}\right) = log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \tag{5.5}$$

to estimate the influence of predictor variables x_{nb_sport}, x_{nb_smoke} and x_{nb_food} on $y_{response\_dental}$, as well as $x_{nb\_danger\_sport}$, x_{danger_occup} and $x_{nb\_interest\_ip}$ on $y_{response\_ip}$.

After performing logistic regressions on both response variables, the following outcomes are obtained (see table 5.1, table 5.2). Looking at the results one can see, that all predictor variables have a positive influence on the response variables and are highly significant. This is explained by the fact, that the response variables were simulated according to the assumption that the external information has a positive effect on the response, because the expectation is, that this interrelation does also exist in real data.

Coefficients	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.76077	0.04997	-35.24	<2e-16 ***
nb_sport	0.20115	0.01537	13.08	<2e-16 ***
nb_smoke	0.17834	0.01763	10.12	<2e-16 ***
nb_food	0.19292	0.01494	12.92	<2e-16 ***

Table 5.1: Outcome after logistic regression on response variable $y_{response\_dental}$ using fictitious data

Coefficients	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.62132	0.04621	-35.09	<2e-16 ***
danger_occup	1.81104	0.16957	10.68	<2e-16 ***
nb_danger_sport	0.20924	0.01470	14.24	<2e-16 ***
nb_interest_ip	0.37530	0.03162	11.87	<2e-16 ***

Table 5.2: Outcome after logistic regression on response variable $y_{response\_ip}$ using fictitious data

These estimates can now be used as scoring weights to calculate the probability for each existing client to buy one of the two new products.

Scoring of existing customers

In order to calculate probabilities, equation (5.5) has to be rearranged into

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}. \tag{5.6}$$

After weighting each predictor variable of the existing clients with the estimated scoring weights for each product, respectively, all clients have to be arranged using the estimated probabilities in descending order, for each product individually. An extract of the simulated data with estimated probabilities for existing clients to buy supplementary dental and inpatient covers, respectively, in descending order looks as follows.

first_name	last_name	gender	age	nb_sport	nb_smoke	nb_food	prob_dental
Calvin	Gordon	1	18	14	6	6	96,4 %
Howard	Stokes	1	37	10	3	1	72,7 %
Joel	Bridges	1	44	6	2	1	49,9 %
Tamara	Nielsen	2	49	2	2	0	26,9 %

Table 5.3: Extract of simulated data with estimated probabilities for existing clients to buy a supplementary dental cover

first_name	last_name	gender	age	danger_occup	nb_danger_sport	nb_interest_ip	prob_ip
Edwin	Barrett	1	18	18	1	6	99,8 %
Raymond	Dickson	1	21	14	0	0	78,7 %
Cassandra	Gardner	2	37	1	1	0	59,8 %
Nathan	Foley	1	50	0	0	1	22,3 %

Table 5.4: Extract of simulated data with estimated probabilities for existing clients to buy a supplementary inpatient cover

Customers in excess of a certain probability can now be targeted and contacted regarding the new products. Assuming a threshold of 50%, Calvin Gordon and Howard Stokes (see table 5.3) have to be contacted and offered a supplementary dental cover. Edwin Barrett, Raymond Dickson and Cassandra Gardner (see table 5.4) shall be targeted regarding a supplementary inpatient cover.
As a matter of course, it is advisable or even inevitable to consider further information, such as the location, customer value and the internal risk profile score before proceeding with the targeting.

5.3 Lifestyle clustering using two-step cluster analysis

"The [...] TwoStep Clustering Component is a scalable cluster analysis algorithm [, which is implemented in SPSS and] designed to handle very large datasets [unlike most traditional clustering algorithms]. Capable of handling both continuous and categorical variables or attributes, it requires only one data pass in the procedure. In the first step of the procedure, [...] the records [are pre-clustered] into many small subclusters [This is achieved by "[...] scanning the entire dataset and storing the dense regions of data records in summary statistics called cluster features which are stored in memory in a data structure called CF-tree [...]" (Chiu et al., 2001)]. Then, [...] the subclusters from the pre-cluster step [are clustered] into the desired number of clusters. If the desired number of clusters is unknown, the SPSS TwoStep Cluster Component [...] find[s] the proper number of clusters automatically." (SPSS Inc., 2001) The SPSS two-step cluster analysis, which will be described in more detail in the following according to Chin et al., 2001 and Teachwiki, 2008, is based on the framework of the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm with a few modifications, such as the ability to handle mixed type attributes and the automatic determination of the appropriate number of clusters. [(Zhang et al., 1996), (Chiu et al., 2001), (Teachwiki, 2008)]

Step1: Pre-clustering

In the pre-clustering step, a Cluster Features (CF) tree is constructed. A cluster feature CF is thereby the collection of statistics that summarizes the characteristics of a dense region. Given a cluster C_j with $j = 1, ..., J$ clusters, CF_j is defined as

$$CF_j = \{N_j, \bar{x}_{Aj}, s^2_{Aj}, N_{Bj}\} \tag{5.7}$$

with

- N_j - Number of K-dimensional data records in C_j
- $\bar{x}_{Aj}$ - Mean of continuous attributes of the N_j data records $x_i = (x_{Ai1}, ..., x_{AiK_A})$
- s^2_{Aj} - Variance of continuous attributes of the N_j data records $x_i = (x_{Ai1}, ..., x_{AiK_A})$

- $N_{Bj} = (N_{Bj1}, N_{Bj2}, ..., N_{BjK_B})$ - A $\sum_{k=1}^{K_B}(L_k - 1)$ - dimensional vector where the k-th sub-vector is of $(L_k - 1)$ dimensions, given by $N_{Bjk} = (N_{jk1}, ..., N_{jkL_k-1})$ in which N_{jkl} is the number of data records in C_j whose k-th categorical attribute takes the l-the category, $l = 1, ..., L_k - 1$

When merging two clusters C_j and C_s, the corresponding entries in CF_j and CF_s are simply added:

$$CF_{<j,s>} = \{N_j + N_s, \bar{x}_{Aj} + \bar{x}_{As}, s^2_{Aj} + s^2_{As}, N_{Bj} + N_{Bs}\}. \tag{5.8}$$

"A CF tree is a height-balanced tree with two parameters: branching factor B and threshold T. Each nonleaf node contains at most B entries of the form $[CF_i, child_i]$, where $i = 1, 2, ..., B$, "$child_i$" is a pointer to its i-th child node, and CF_i, is the CF of the subcluster represented by this child. So a nonleaf node represents a cluster made up of all the subclusters represented by its entries. A leaf node contains at most L entries, each of the form $[CF_i]$, where $i = 1, 2, ...L$. In addition, each leaf node has two pointers, "*prev*" and "*next*" which are used to chain all leaf nodes together for efficient scans. A leaf node also represents a cluster made up of all the subclusters represented by its entries. But all entries in a leaf node must satisfy a *threshold requirement*, with respect to a threshold value T: *the diameter (or radius) has to be less than* T." (Zhang et al., 1996) "[...] SPSS uses a CF-tree with $T = 0$, as well as a maximum of three levels of nodes and a maximum of eight entries per node at default. This combination may result in a maximum of 512 leaf entries, hence 512 sub-clusters [, as is shown in figure 5.2]." (SPSS Inc., 2001)

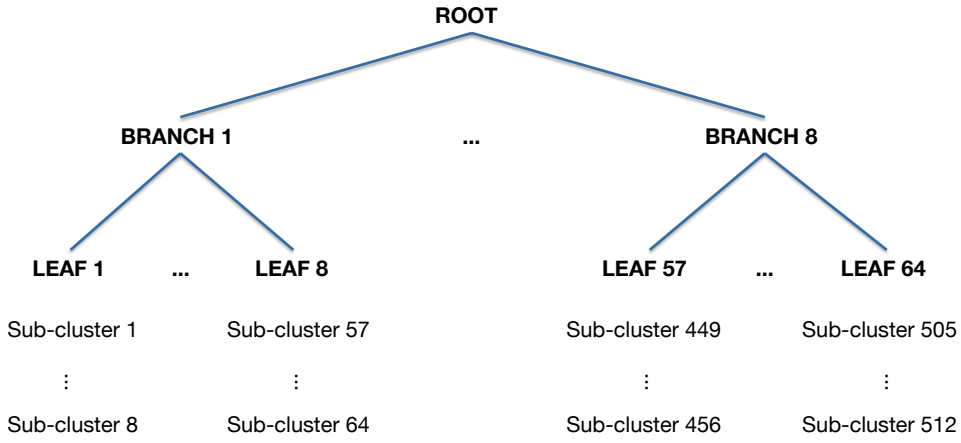


Figure 5.2: Structure of a CF under SPSS

In order to construct a CF tree, the first data record is placed starting at the root of the tree in a leaf node while saving variable information about that case. "[...] [All remaining] data records are [then] scanned sequentially from the dataset and the decision is made immediately whether the current record is to merge with any previously constructed dense region or to form a singleton by itself based on the distance criterion. During this data pass, a CF-tree is constructed to store the summary statistics of dense regions or singletons; it serves as a guidance structure to efficiently identify dense regions. [...] When a data record is passing through a non-leaf node, it finds the closest entry in the node and travels to the next child node. The process continues recursively and the data record descends along the CF- tree and reaches a leaf-node. Upon reaching a leaf node, the data record finds the closest entry. The record is absorbed to its closest entry if the distance of the record and the closest entry is within a threshold value; otherwise it starts as a new leaf entry in the leaf node. If the CF-tree grows beyond the maximum size allowed, it is rebuilt by a larger threshold criterion. The new CF- tree is smaller and hence has more room for incoming records. The process continues until a complete data pass is finished. [...] The distance measure used for two-step clustering is derived from a probabilistic model that the distance

between two clusters is equivalent to the decrease in log-likelihood function as a result of merging." (Chiu et al., 2001) The distance between two clusters C_j and C_s with the log-likelihood function before merging $\hat{l}$ and the log-likelihood function after merging those clusters $\hat{l}_{new}$ is defined as

$$d(j,s) = \hat{l} - \hat{l}_{new} = \xi_j + \xi_s - \xi_{<j,s>}, \tag{5.9}$$

where

$$\xi_\upsilon = -N_\upsilon \left(\sum_{k=1}^{K_A} \frac{1}{2} log(\hat{\sigma}_{\upsilon k}^2 + \hat{\sigma}_k^2) - \sum_{k=1}^{K^B} \sum_{l=1}^{L_k} \frac{N_{\upsilon kl}}{N_\upsilon} log \frac{N_{\upsilon kl}}{N_\upsilon} \right) \text{ with } \upsilon = (i,j,<i,j>) \tag{5.10}$$

and is calculated by using the information from the cluster features. Thereby, K_A represents the total number of continuous, K_B the total number of categorical variables. L_k denotes the number of categories for the k-th categorical variable and $N_{\upsilon kl}$ is the number of data records in cluster υ, whose categorical variable k takes the l category. $\hat{\sigma}_{\upsilon k}^2$ is the estimated variance of the continuous variable k for cluster υ.

Step2: Clustering

"After the CF-tree is built in step one, a collection of dense regions is identified and is stored in the leaf nodes of the tree. Since the number of dense regions is usually far less than the number of data records in the dataset and the summary statistics stored in the cluster features are sufficient for calculating the distance and related criterion, most clustering algorithms can be applied to cluster the dense regions very efficiently. [...] [In this] algorithm, a hierarchical clustering algorithm using the log-likelihood based distance measure is used. [...] [If the number of clusters is not pre-specified, a two-phase procedure is used to determine the appropriate number of clusters with a maximum of 15 clusters as specified in SPSS.]
The first phase is to detect a coarse estimate of the number of clusters in the data based on Bayesian Information Criterion. BIC is a likelihood criterion penalized by the model complexity, which is measured by the number of parameters in the model." (Chiu et al., 2001)

$$BIC(J) = -2 \sum_{j=1}^{J} \xi_j + J \left\{ 2K_A + \sum_{k=1}^{K_B} (L_k - 1) \right\} log(N). \tag{5.11}$$

"Models with small BIC are often considered good models. Usually as the number of clusters increases, BIC decreases first and then increases. [...] At each merge, decrease in BIC is calculated. A coarse estimate of the number of clusters is obtained as the number of clusters at which the decrease in BIC starts to diminish when the number of clusters increases." (Chiu et al., 2001) This is determined by looking at the ratio

$$R_1(J) = \frac{BIC(J) - BIC(J+1)}{BIC(1) - BIC(2)} \tag{5.12}$$

The maximum number of clusters is set equal to number of clusters where the ratio $R_1(J)$ is smaller than 0.04, as defined in SPSS.
"The second [...] [phase] refines the initial estimate by finding the greatest change in distance between the two closest clusters in each hierarchical clustering stage." (SPSS Inc., 2001) "Merging starts from the set of clusters resulted from phase one, and an estimate of the number of cluster[s] is obtained at the step where a big jump of the ratio change is observed. The rationale behind phase two is that a big jump in ratio change of the distance usually occurs when [...] two clusters [are merged] that should not be merged." (Chiu et al., 2001) Here for, J determined in phase one is used to calculate

$$R_2(m) = \frac{d_{min}(C_m)}{d_{min}(C_{m+1})}, \quad m = J, ..., 2 \tag{5.13}$$

with the minimum distance between two clusters $d_{min}(C_m)$ and the new number of clusters m. The optimal cluster number M is then determined by comparing the two biggest values $R_2(m_1)$ and $R_2(m_2)$, where $R_2(m_1) > R_2(m_2)$:

$$M = \begin{cases} m_1, & \text{if } R_2(m_1) > 1.15 \cdot R_2(m_2) \\ \underset{i=1,2}{max}(m_i), & \text{else.} \end{cases} \tag{5.14}$$

Lifestyle clustering of existing and potential new clients from Facebook

Now, the two-step clustering algorithms in SPSS can be applied to cluster not only the existing, but also the potential new clients from the simulated dataset. The attributes used for the lifestyle clustering are the same variables, as used for the scoring. Default settings from TwoStep SPSS are used and the number of clusters fixed at seven.
As an outcome, clustering of lifestyle factors with regard to the supplementary dental cover leads to a very fair cluster quality, as can be seen in figure 5.3.

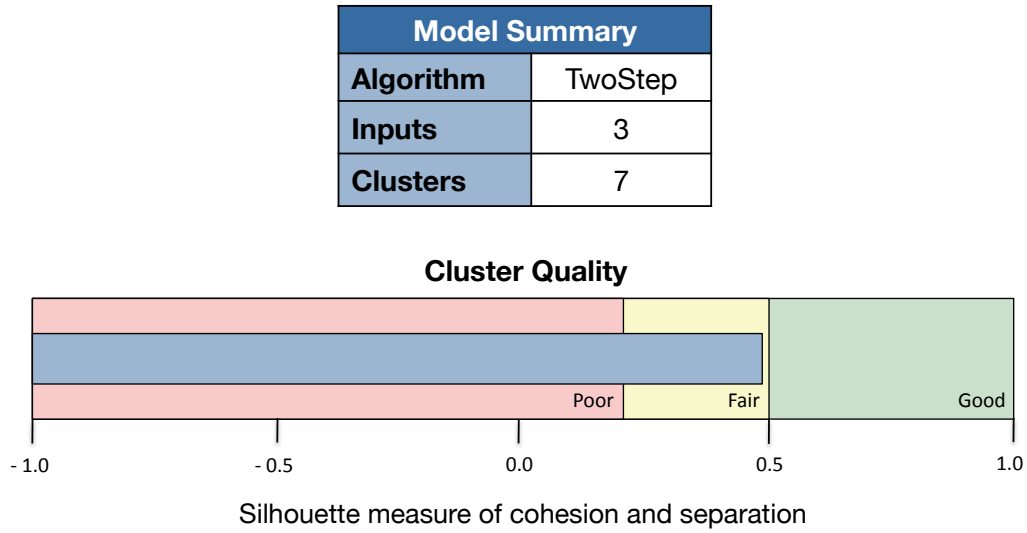


Figure 5.3: Model summary after two-step clustering regarding supplementary dental cover (modified)

Looking at the cluster classification shows that most persons were categorized into the first four clusters. (see figure 5.4) Clusters five, six and seven contain only 12.1% of all 10000 cases. Considering each created cluster individually, it can be seen that the first four clusters contain persons with small values in the lifestyle factors $nb\_sport, nb\_smoke, nb\_food$, as the mean values of each attribute in each cluster show. Clusters five, six and seven, however, each contain persons with either high values in nb_sport, nb_smoke or nb_food, thus showing lifestyle clusters for persons who do a lot of dangerous sports, who smoke a lot or consume a lot of harmful food and drinks. Making the hypothesis, that persons with high values in one of those variables are more willing to buy a supplementary dental cover, persons from clusters five, six and seven have to be targeted and offered a supplementary dental cover.

Cluster	1	2	3	4	5	6	7
Size	42.5% (4246)	18.5% (1846)	9.8% (976)	17.2% (1716)	1.6% (164)	4.3% (434)	6.2% (618)
Inputs	nb_sport 0.61	nb_sport 0.27	nb_sport 4.87	nb_sport 0.38	nb_sport 11.42	nb_sport 1.04	nb_sport 0.62
	nb_smoke 0.05	nb_smoke 0.13	nb_smoke 0.49	nb_smoke 1.98	nb_smoke 1.77	nb_smoke 7.13	nb_smoke 0.56
	nb_food 0.07	nb_food 2.10	nb_food 1.12	nb_food 0.55	nb_food 0.85	nb_food 1.07	nb_food 7.64

Figure 5.4: Cluster overview after two-step clustering regarding supplementary dental cover (modified)

Clustering of lifestyle factors with regard to the supplementary inpatient cover leads to an even good cluster quality, as can be seen in figure 5.5.

Model Summary	
Algorithm	TwoStep
Inputs	3
Clusters	7

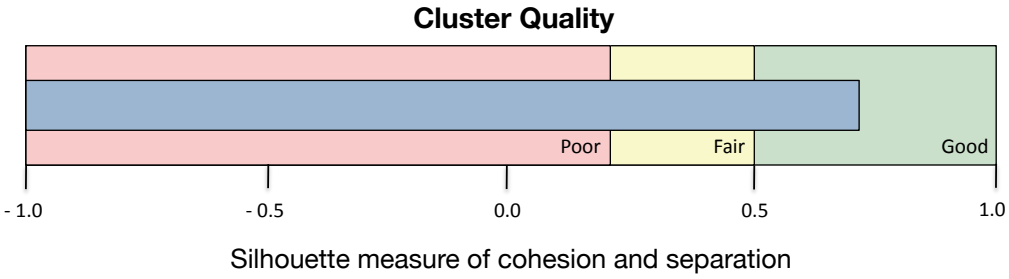


Figure 5.5: Model summary after two-step clustering regarding supplementary inpatient cover (modified)

Looking at the cluster classification for the three lifestyle factors $danger\_occup$, $nb\_danger\_sportandnb\_interest\_ip$ also shows that most persons were categorized into the first four clusters. (see figure 5.6) Clusters five, six and seven contain 12.6% of all 10000 cases. Again, considering each created cluster individually, it can be seen that the first four clusters contain persons with small values in the three lifestyle clusters, as the mean values of each attribute in each cluster show. Clusters five, six and seven, however, each contain persons with either high values in nb_danger_sport, $nb\_interest\_ip$ or person with a dangerous occupation $danger\_occup = 1$, thus showing lifestyle clusters for persons who do a lot of dangerous sports, who searched a lot for a supplementary inpatient cover or have a dangerous occupation. Making the hypothesis, that persons with high values in one of those variables are more willing to buy a supplementary inpatient cover, persons from clusters five, six and seven have to be targeted and offered a supplementary inpatient cover.

Cluster	1	2	3	4	5	6	7
Size	40.7% (4066)	24.4% (2437)	14.2% (1425)	8.0% (804)	7.4% (745)	2.0% (201)	3.2% (322)
Inputs	danger_occup 0	danger_occup 0	danger_occup 0	danger_occup 0	danger_occup 0	danger_occup 0	danger_occup 1
	nb_danger_ sport 0.00	nb_danger_ Sport 1.87	nb_danger_ sport 0.60	nb_danger_ Sport 0.73	nb_danger_ Sport 7.20	nb_danger_ Sport 1.02	nb_danger_ Sport 1.34
	nb_interest_ip 0.00	nb_interest_ip 0.00	nb_interest_ip 1.00	nb_interest_ip 2.32	nb_interest_ip 0.44	nb_interest_ip 5.06	nb_interest_ip 0.48

Figure 5.6: Cluster overview after two-step clustering regarding supplementary inpatient cover (modified)

5.4 Comparison of results from scoring and lifestyle clustering

After identifying potential new clients for the supplementary dental and inpatient covers, separately, it is now of interest, if the same persons have been chosen in those two approaches or, if not, how these results differ. In order to be able to compare the identified persons from both approaches, only existing internal clients will be considered. All internal clients having a probability of purchasing a supplementary dental or inpatient cover over 50% after scoring are chosen. Also, internal clients selected into cluster five, six and seven after two-step clustering are picked.
Comparing the identified persons from both approaches leads to the following results regarding the supplementary dental cover:

- Total number of data record with scored probability over 50% = 329
- Total number of data record in clusters five, six and seven = 592
- Number of same data records = 239 .

It can be seen, that around three quarters of persons identified in the scoring were also allocated into the most severe clusters. Vice versa, however only around half of the persons assigned to clusters with high valued attributes also have a high probability according to the scoring. These results can be explained by the fact, that scoring uses all attributes for the calculation of their impact, whereas clustering creates a separation of each attribute. Thus, units having average high values in each variable may receive a high score, while they will not appear in cluster five, six or seven. The other way round, a person having a high value only in one variable will be assigned to cluster five, six or seven, but not receive a high score.
Regarding the supplementary inpatient cover, the following results are obtained:

- Total number of data record with scored probability over 50% = 382
- Total number of data record in clusters five, six and seven = 621
- Number of same data records = 357 .

Here, almost all persons indentified through scoring were also allocated into clusters five, six or seven. Yet again, only around half of the internal clients assigned to clusters with high valued attributes also have a high probability according to the scoring.
It can be stated, that scoring leads to more conservative and accurate estimates as they are data based. Clusters

identified in the lifestyle clustering, however, contain information regarding the persons specific needs according to their assignment to a lifestyle clusters. This is the reason, why it should be considered to target identified persons from both approaches in order to address a wider range of potential clients knowing their exact needs. If the aim, however, is to achieve a better accuracy, persons identified through scoring should be targeted. Are the specific needs and interest of persons in the foreground, lifestyle clustering is the better approach.
In summary, it can be said that both methods are suitable approaches for the identification of potential new customers. Depending on the data and issue of interest, however, it is the task of the analyst to choose the appropriate approach.

6 Conclusions and outlook

This master's thesis displays the future trend in using social media data for product development using the example of Facebook. It was shown, how important social media data is nowadays and will be even more in future. The extraction of such data from various sources through API's is very simple and provides a lot of new and useful information, which is not only lucrative in the product development area but can also be of importance in various other business analytics areas, such as in human resources and fraud detection. The continuing technological progress simplifies the handling, as well as the analysis of such big amounts of data, which are often unstructured. For the structuring of unstructured data, important text analytics tools and techniques were introduced. The combination of external data and existing internal data warehouse, afterwards, creates a Big Data warehouse, which allows the evaluation and intelligence gathering from versatile, innovative and information-rich data.
In this master's thesis, the new products were determined beforehand and it was considered, that the suggested products do not yet exist in this compilation and they also represent a market and client need. The task was to find potential customers who those products can be offered to. Otherwise, however, it can also be of interest to find new, not yet existing products according to demands of the market and existing clients derived from social media data. It can be said, that the product development process represents a cycle between the identification of market as well as client needs, the so-called demand, and new product ideas, more precisely the offer.

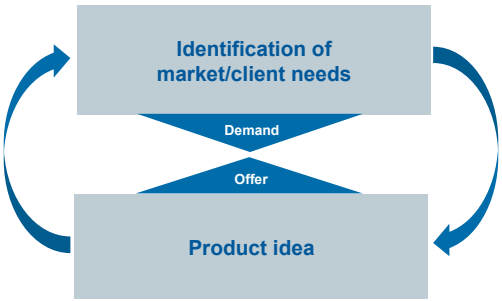


Figure 6.1: Two components of the product development process

For the identification of market as well as client needs not only specific information connected with one product, but also all possibly relevant information on persons from social network sites has to be extracted. Afterwards, all this information will have to be clustered according to similarity. Looking at the distinct clusters will then give the opportunity to detect new information regarding possible not yet covered areas of health insurance and thus ideas for new products and packages. Those can either be detected by looking at the clusters manually and deciding intuitively, or by conducting a survey.
This second component of the product development process was not discussed here, as the implementation is not yet realizable due to limitations in the still immature text analytics techniques, but is a really innovative approach and should be investigated for further development and research in the product development field of insurance companies.
In conclusion, it can be said, that social media data and Big Data in general are the future trend in data analytics and business management.

A. Contents of enclosed CD

The enclosed CD contains the following folders and subfolders

- The folder `RCode/` contains the following subfolders:
  - `RFacebook`: RCodes for extraction of Facebook data and display of friends' network + commented videos for data extraction in German and English
  - `Simulation`: RCodes and files for simulation of fictitious dataset
  - `Scoring`: RCodes and files for scoring of existing clients
  - `Comparison`: RCodes and files for comparison of scoring and clustering results
- The folder `SPSS/` contains the following subfolder:
  - `Clustering`: SPSS code and files for clustering of simulated dataset
- The folder `Master_thesis/` contains the following files:
  - `master_thesis_schmelewa_for_print.pdf`: Master's thesis print version
  - `master_thesis_schmelewa_online_version.pdf`: Master's thesis online version

Bibliography

Abbott, D. (2010). Introduction to Text Mining. www.vscse.org/summerschool/2013/Abbott.pdf.

Acharya, P. (2013). Six Social Media Analytics Technologies You Should Know About. http://vtalktech.wordpress.com/2013/01/13/social-media-analytics/.

Aggarwal, C. C. (2011). *Social Network Data Analytics*. Springer Science + Business Media, LLC.

Aggarwal, C. C. and Zhai, C. (2012). *Mining Text Data*. Springer Science+Business Media, LLC.

API Portal by Anypoint Platform and Mulesoft (accessed 2014). Facebook Graph API. http://api-portal.anypoint.mulesoft.com/facebook/api/facebook-graph-api.

Bao, S. Y., Xiang, Y., and Savarese, S. (2012). Object Co-Detection. *ECCV*.

Barbera, P. (2014). *Package Rfacebook*.

Berry, M. W. and Kogan, J. (2010). *Text Mining: Applications and Theory*. Wiley.

Blomberg, J. (2012). Twitter and Facebook Analysis: It's Not Just for Marketing Anymore. *SAS Global Forum*, 309.

Brill, E. and Moore, R. C. (2000). An Improved Error Model for Noisy Channel Spelling Correction. In *ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293.

Bundesverband Digitale Wirtschaft (BVDW) e.V. (2009). Social Media Kompass.

Busch, V., Stel, H. V., Schrijvers, A., and de Leeuw, J. (2013). Clustering of health-related behaviors, health outcomes and demographics in Dutch adolescents: a cross-sectional study. *BMC Public Health*, 13(1):1–11.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Cavnar, W. B. and Trenkle, J. M. (1994). N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*.

Chakraborty, G., Pagolu, M., and Garla, S. (2013). *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. SAS Institute.

Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, pages 263–268. ACM.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology, The Stanford Natural Language Processing Group*.

Facebook (2014). Data use policy. https://www.facebook.com/about/privacy/your-info.

Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping into the Power of Text Mining. *Communications of the ACM*, 49(9):76–82.

Feldman, S., Reynolds, H., and Schubmehl, D. (2012). Content Analytics and the High-Performing Enterprise. *IDC*.

Feldmann, R. and Sanger, J. (2007). *THE TEXT MINING HANDBOOK: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E. W., Hampp, T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L., and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. *Computer Science*, RC24122(W0611-188).

Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics.

Gantz, J. and Reinsel, D. (2012). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. www.emc.com/leadership/digital-universe/index.htm.

Gartner (2014). Big Data. http://www.gartner.com/it-glossary/big-data/.

Grassi, M., Cambria, E., Hussain, A., and Piazza, F. (2011). Sentic Web: A New Paradigm for Managing Social Media Affective Information. *Cognitive Computation*, 3(3):480–489.

Grefenstette, G. (1995). Comparing two language identification schemes. In *JADT 1995, 3rd International Conference on Statistical Analysis of Textual Data*.

Grimes, S. (2007). Defining Text Analytics. www.informationweek.com/software/information-management/defining-text-analytics/d/d-id/1051763?

Gunelius, S. (2011). What Is an API and Why Does It Matter? http://sproutsocial.com/insights/api-definition/.

Guo, X., Liu, D., Jou, B., Zhu, M., Cai, A., and Chang, S.-F. (2013). Robust Object Co-Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR.

Hanumathappa, M. and Reddy, M. V. (2012). Natural Language Identification and Translation Tool for Natural Language Processing. *International Journal of Science and Applied Information Technology*, 1(4):107–112.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer Publishing Company, Inc.

Hsu, W., Lee, M. L., and Zhang, J. (2002). Image Mining: Trends and Developments. *Journal of Intelligent Information Systems*, 19(1):7–23.

Iacus, S. M. and King, G. (2011). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*.

Iacus, S. M., King, G., and Porro, G. (2009). cem: Software for Coarsened Exact Matching. *Journal of Statistical Software*, 30(9).

IBM (2014a). TwoStep Cluster Analysis. http://pic.dhe.ibm.com/infocenter/spssstat/v22r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fspss%2Fbase%2Fidh_twostep_main.htm.

IBM (2014b). Watson Content Analytics. http://www-01.ibm.com/support/knowledgecenter/#!/SS5RWK/welcome.

IDC (2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

Jivani, A. G. (2011). A Comparative Study of Stemming Algorithms. *International Journal of Computer Technology and Applications*, 2(6):1930–1938.

Jusoh, S. and Alfawareh, H. M. (2012). Techniques, Applications and Challenging Issue in Text Mining. *International Journal of Computer Science Issues*, 9(2):431–436.

Kaplan, A. M. (2012). If you love something, let it go mobile: Mobile marketing and mobile social media 4x4. *Business Horizons*, 55(2):129–139.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1):59–68.

Kernighan, M. D., Church, K. W., and Gale, W. A. (1990). A Spelling Correction Program Based on a Noisy Channel Model. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, pages 205–210.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251.

King, G., Nielsen, R., Coberley, C., Pope, J. E., and Wells, A. (2011). Comparative Effectiveness of Matching Methods for Causal Inference.

Knott, A., Hayes, A., and Neslin, S. A. (2002). Next-product-to-buy models for cross-selling applications. *Journal of interactive marketing*, 16(3):59–75.

Landsberg, B., Plachta-Danielzik, S., Lange, D., Johannsen, M., Seiberl, J., and Müller, M. J. (2010). Clustering of lifestyle factors and association with overweight in adolescents of the Kiel Obesity Prevention Study. *Public Health Nutrition*, 13:1708–1715.

Lee, S., Song, J., and Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computer Information Systems*, Fall 2010.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Lu, B. and Rosenbaum, P. R. (2004). Optimal pair matching with two control groups. *Journal of computational and graphical statistics*, 13(2):422–434.

Lv, J., Liu, Q., Ren, Y., Gong, T., Wang, S., Li, L., and the Community Interventions for Health (CIH) collaboration (2011). Socio-demographic association of multiple modifiable lifestyle risk factors and their clustering in a representative urban population of adults: a cross-sectional study in Hangzhou, China. *International Journal of Behavioral Nutrition and Physical Activity*, 8(1):1–13.

Malthouse, E. C. (1999). Ridge regression and direct marketing scoring models. *Journal of interactive marketing*, 13(4):10–23.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press., Cambridge, MA.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Megyesi, B. (2002). Shallow Parsing with PoS Taggers and Linguistic Features. *Journal of Machine Learning Research*, 2(639-668).

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., Delen, D., and Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press.

Osborne, M. (2000). Shallow Parsing as Part-of-Speech Tagging. In *CoNLL-2000 and LLL-2000*, pages 145–147.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. University of Pennsylvania.

Ratnaparkhi, A. (1997). A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

Ratnaparkhi, A. (1999). Learning to Parse Natural Language with Maximum Entropy Models. *Machine Learning*, 34:151–175.

Rokach, L. and Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer US.

Roos, D. (accessed 2014). How to Leverage an API for Conferencing. http://money.howstuffworks.com/business-communications/how-to-leverage-an-api-for-conferencing1.htm.

SAS Institute Inc (2012). Getting Started with SAS Text Miner 12.1. *SAS Institute Inc.*

SAS Institute Inc (2014). SAS Social Media Analytics. Product documentation.

Sinka, M. P. and Corne, D. (2003). Evolving Better Stoplists for Document Clustering and Web Intelligence. In *HIS*, pages 1015–1023.

SPSS Inc. (2001). The spss twostep cluster component. a scalable component to segment your customers more effectively. White paper – technical report, Chicago.

Stroh, M. (2010). Scoring: Kaufwahrscheinlichkeiten ermitteln und gezielt Umsätze steigern. Whitepaper published by D&B Germany.

Tavast, A., Muischnek, K., and Koit, M. (2012). *Human Language Technologies: The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012*. IOS Press.

Teachwiki (2008). Die two-step clusteranalyse unter spss. http://mars.wiwi.hu-berlin.de/mediawiki/teachwiki/index.php/Die_Two-Step_Clusteranalyse_unter_SPSS.

Toutanova, K. and Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-2000)*.

Toutanova, K. and Moore, R. C. (2002). Pronunciation Modeling for Improved Spelling Correction. *Proceedings of the 40th Meeting of the Association for Computational Linguistics(ACL-2002)*, pages 141–151.

UIMA (2013). Apache UIMA project. http://uima.apache.org.

Voutilainen, A. (1995). NPtool, a detector of English noun phrases. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 157–164. Association for Computational Linguistics Dublin.

Webopedia (2014). Webopedia. www.webopedia.de.

Wikipedia (2014a). Semi-structured data. http://en.wikipedia.org/wiki/Semi-structured_data.

Wikipedia (2014b). Social Media. http://en.wikipedia.org/wiki/Social_media.

Wikipedia (2014c). Web 2.0. http://en.wikipedia.org/wiki/Web_2.0.

Wilson, R. E., Gosling, S. D., and Graham, L. T. (2012). A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science Perspectives on Psychological Science*, 7(3):203–220.

Winkler, W. E. (1995). *Matching and Record Linkage, in Business Survey Methods*, chapter 20, pages 355–384. John Wiley & Sons, Inc.

Winter, F. (2014). Business analytics provides structure when dealing with big data. Munich Re Web International.

Wipro (2014). Big Data. http://www.wipro.com/services/analytics-information-management/big-data.aspx.

Wood, P. (2014). Semi-Structured data. http://www.dcs.bbk.ac.uk/$\sim$ptw/teaching/ssd/notes.html.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: An efficient data clustering method for very large databases.

Declaration of Authorship

I declare that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other University.
Formulations and ideas taken from other sources are cited as such.

München, den November 4, 2014

(Maria Schmelewa)