University of Economics, Prague
Faculty of Informatics and Statistics
Department of Information Technologies

E-service Quality Measurement
From Customers' Point of View

Doctoral Dissertation

Filip Vencovský
filip.vencovsky@vse.cz

Supervised by
Prof. Ing. Jiří Voříšek, Csc.

Study program
Applied Informatics

Prague, 2018

Cite this document as:

Vencovský, F., 2018. E-service Quality Measurement From Customers' Point of View, PhD dissertation, University of Economics, Prague.

Acknowledgements

First, I would like to thank my supervisor Prof. Jiří Voříšek for patient guidance. My sincere thanks go to Dr. Tomáš Bruckner, my friend and colleague, for inspiring discussions, motivation and feedback. I thank Faculty of Informatics and Statistics for funding the student research grant, my fellow colleague Lucie Šperková and all other faculty members.

This thesis would not be possible without my Ph.D. visit to Service Science Factory (SSF) in Maastricht and following eight-month research stay at Business Intelligence & Smart Services (BISS) institute in Heerlen. I would like to thank Dr. Dominik Mahr, scientific director of SSF, Prof. Rudolf Müller, scientific director of BISS, Dr. Benjamin Lucas and all other colleagues from both institutes.

Foremost, I am grateful to Prof. Jos Lemmink, professor of marketing and service innovation at Maastricht University, for immense knowledge, insightful discussions and advice during my stay at SSF and BISS.

Last but not the least, I would like to thank my family and friends for moral support. My deepest appreciation goes to Aneta for all her love and support.

Declaration

I hereby declare that this doctoral dissertation is my own work, and that I have not used any sources and aids other than those stated in the dissertation.

Prague, 31st January 2018

_____

Signature

Abstract

This thesis focuses on the quality diagnosis of services delivered using IT artefacts over the internet that are also known as e-services, IT services or self-services. The thesis reflects new challenges and new possibilities that digital economy brings. This work examines the relationship between consumers' feedback and service quality. The research issue is split in two particular branches. One branch is related to feedback sentiment and the other to feedback emotionality. The first particular issue refers to classification from unstructured feedback, the second to service quality implication. In order to solve the issues, the thesis presents the literature review of current service quality research and the review of opinion mining methods that can help with a large customer-generated online service feedback. This work contains the first literature review of papers that use opinion mining technique to service quality diagnosis. In contrast to the reviewed work, the thesis conceptualises service quality theory and synthesises it with opinion mining taxonomy. The case study from e-banking service is used for demonstration of opinion mining methods for service quality diagnosis. The emotions are examined in the second case study. Finally, the qualitative study on the level of a review is conducted.

Keywords:

Digital service economy, Customer feedback, E-service quality, Quality diagnosis, Opinion mining

Abstract in Czech / Abstrakt

Tato práce se zabývá vyhodnocováním kvality služeb, které jsou dodávány prostřednictvím IT artefaktů a internetu, známých jako e-služby, samoobslužné nebo IT služby. Práce reflektuje výzvy a příležitosti, které s sebou přináší digitální ekonomika. Zkoumá vztah mezi zákaznickou zpětnou vazbou a kvalitou služby. Tento problém dělí do dvou různých částí. První se váže k sentimentu zpětné vazby a druhá k její emocionalitě. Jelikož je zkoumaný problém široký, byl rozdělen do několika dílčích problémů. První řešený problém spočívá v klasifikaci sentimentu, potažmo emocionality ze zpětné vazby. Druhý problém se vztahuje k tomu, jakým způsobem mohou takto klasifikovaná data implikovat kvalitu služby. Pro vyřešení těchto problémů je provedena rešerše současné literatury o kvalitě služeb, ve které je uvedena konceptualizace kvality, přístupy k jejímu měření a výzkum dimenzí kvality. Dále je provedena rešerše literatury z oblasti opinion mining, ve které jsou popsány metody extrakce sentimentu, vyjádřených emocí a aspektů služby. Tato práce také obsahuje první rešerši literatury, která využívá metod z oblasti opinion mining pro vyhodnocení kvality služeb. Oproti článkům z rešerše se tato práce věnuje jednak konceptualizaci kvality služby a jednak její syntéze s oblastí opinion mining. Případová studie z prostředí elektronického bankovnictví je použita pro ověření vybraných metod pro extrakci aspektů a klasifikaci sentimentu. Emoce vyjádřené ve zpětné vazbě a jejich extrakce jsou zkoumány ve druhé případové studii, kde je ověřen jejich vztah ke kvalitě služby. Nakonec je provedena kvalitativní studie, která ověřuje předpoklady z rešerše týkající se vztahu mezi nestrukturovaným textem a strukturovaným hodnocením kvality na jednotkové úrovni.

Keywords:

Digitální ekonomika, Zákaznická zpětná vazba, Kvalita e-služeb, Vyhodnocení kvality, Opinion mining

Content

Chapter 1: Introduction

1.1 Research motivation

The rising importance of digital economy brings new challenges and also new possibilities for service research. Customers consume services in virtual space where traditional approaches to consumer experience are no longer valid. Service delivery, which traditionally relies on service personnel, depends now also on the performance of a complex network of technological artefacts. Service personnel cease to be visible to consumers that see only a web page or mobile application interface. Although the new interfaces offer more detailed information about service parameters and content, both sides suffer from the lack of face to face communication where problems can be quickly solved, and additional questions asked.

The lack of communication has rapidly impacted the task of service quality diagnosis, although the goal of the task remains unchanged. Service managers still need to understand what consumers think about their service, what are the strengths and weaknesses of the service. What differs is a character of information available to service managers. Digital service interfaces offer accurate data about consumers behaviour, how they use a service, which brings an excellent opportunity for analysis but does not allow to understand how consumers really see a service. Only capturing and analysing of consumers' point of view could lead to the right service improvement and result in the level of service that majority of consumers perceives as expected or above expected.

Digital environment also offers an easy way for direct quality surveying. Leading authors of service research (Bitner, Zeithaml, Gremler 2010) see technology impact as a benefit because it enables listening to customers through online research. "… there need to be no interviewers--and therefore no interviewer errors or bias that occur when the interviewer is in a bad mood, tired, impatient, or not objective." On the other hand, online research of service consumers' voice has other limitations that service researchers have to explore.

Service science literature describes many techniques of service quality evaluation, but only a few papers use full-text sources to gather information about service quality. Especially for services that deliver a value online to a large number of consumers around the globe is crucial to collect and analyse full-text data beside traditional surveys. If a service has to deliver promised values in a quality that consumer expects, only the information that consumers feel weakly or strongly satisfied with the service is not enough. Traditional structured service quality surveys are not flexible and limit consumers in expressing themselves (Qu, Zhang, Li 2008). More profound investigation methods like interviews are time and source consuming. There have been many disadvantages associated with unstructured data for a long time, including the necessity of painstaking content analyses and coding procedures (Chowdhury, Reardon, Srivastava 1998; Song, Lee, Yoon, Park 2015). Proper computer-aided analysis of written consumer opinions enables service managers to understand what consumers think about a service in a real time and in

a flexible way. Moreover, it is necessary to put such findings regarding consumers' opinions into a context of service quality concepts.

The opinion mining field can be helpful on this issue. It analyses a natural language in textual form and results in structured information about opinions expressed in the text. This thesis uses sentiment classification, emotion classification and aspect mining methods from the area.

The main challenge is not only how to gather and collect information about a service quality using the opinion mining methods. Collected information from a consumer's writing is one view on a service quality. Information gathered from a survey is another view to the same service and its quality. Service managers usually work with rating data. Can they rely on it? Which view describes consumer's attitudes the best?

In conclusion, the research motivation lies in

(1) rising importance of digital services that lack traditional ways of personnel-consumer communication,

(2) possibility of gathering online service feedback containing textual data and its computer analysis,

(3) lack in theory of linking online textual feedback to service quality concept.

1.2 Research scope definition

This thesis focuses on services that are based heavily on a use of information technologies and are delivered online. These services are also called IT service, e-service, web-based service, in some cases also technology-based service, technology-assisted service or self-service. Focal service is defined more in depth in chapter 1.3.

Quality of services that a service provider delivers over information technologies to a large number of consumers is especially worth researching. In this case, quality diagnosis relies on technological service interfaces. Quality of the other services where personal communication is possible is not the object of this thesis.

Service quality is an abstract and elusive construct (Lepmets, Cater-Steel, Gacenga, Ras 2012). It is more than a result of technical attributes; it is a relationship with a service consumer (van Bon, Jong, Kolthof, Pieper, Rozemeijer, Tjassing, van der Veen, Verheijen 2007). Due to that fact, the thesis aims at service quality diagnosis from consumer's point of view. It approaches technical or production quality of service only through a consumer.

Regarding service quality management, the scope of the thesis stays in service quality diagnosis process and does not cover the other management processes such as quality improvement.

1.3 Focal service definition

Although research scope defined a focal service as a service that is based heavily on a use of information technologies and is delivered online, it is necessary to address service related terms that are in the literature.

Service

The most broad definition of service is "the action of helping or doing work for someone" (Oxford Dictionaries 2018). With rising importance of service economy, marketing research brought more detailed definition and service characteristics. Although the characteristics differ in literature, these four are commonly addressed: *intangibility* – services are to a large extent abstract and intangible; *heterogeneity* – services are non-standard and highly variable; *inseparability* – services are typically produced and consumed at the same time, with customer participation in the process; *perishability* – it is not possible to store services in inventory. (McDonald, Frow, Payne 2011)

A definition from service-dominant logic philosophy (Lusch, Vargo 2006) is more complex and emphases *competences*, "… the application of specialized competences (operant resources – knowledge and skills) through deeds, processes, and performances for the benefit of another entity or the entity itself."

Definition of IT service management (Taylor, Lloyd, Rudd 2007), on the other hand, and emphases *value*, "A service is a means of delivering value to customers by facilitating outcomes customers want to achieve without the ownership of specific costs and risks."

In this thesis, service is seen in accordance with these definitions. Moreover, a service is seen as a complex system that has a core that defines its purpose, service may be surrounded by sub-services that have supporting, facilitating or extending functions. A classification of services into these categories depends on a point of view.

Service system

"...service system is a systematic interaction of parts that functions to perform a service. The smallest system is a single person; the largest is the global economy." (Mele, Polese 2011) The authors also identified four key dimensions of service system: customers, people, information and technology.

In this thesis, service is also seen as a system – business system, that is supported by an information system in order to achieve service goals.

E-service

It depends where the main benefit for customers is created. (Fassnacht, Koese 2006) made distinction between stand-alone *e-service* (main benefit) and support services e-service (facilitation). They divide stand-alone e-service into pure service offerings (like e-banking) and content service offerings (e.g. news). The main point of e-service is to facilitate traditional services (Ladhari 2010). (Parasuraman, Zeithaml, Malhotra 2005) see e-service as the extent to which a web site facilitates efficient and effective shopping, purchasing and delivery.

E-service and the following terms are seen as synonyms in this thesis, because all have in common absence of personal communication and consumers rely on digital interface only.

Self-service and technology-assisted service

"Many services are not delivered in person by employees, but rather are delivered via technology in the form of *self-service* or *technology-assisted service*" (Bitner, Zeithaml, Gremler 2010). "Self-service technologies are technological interfaces that enable customers to produce a service independent of direct service employee involvement" (Meuter, Ostrom, Roundtree, Bitner, Encounters 2000).

High-tech services

A service can be seen as *high-tech* or *high-touch* (Grönroos 2000). Hi-touch services depend highly on service personnel, whereas high-tech services are based on extensive use of automated systems and information technologies. In case of high-tech services, consumers are not in touch with service personnel; only with a technological terminal. Nevertheless, high-touch services may need technologies just as high-tech services may need service personnel in case of an incident.

Web-based services and online-services

"Web-based technologies have been used to automate product distribution and customer services, including transaction and payment systems, call centres, customer relationship management systems, as well as the underlying analytics, reporting, and operations of these systems" (Yang, Fang 2004).

"*Web-based services* (hereinafter *online services* or e-services) offer customers a panoply of benefits such as enhanced control, ease of use, and reduced transaction charges" (Yang, Fang 2004; Zeithaml, Parasuraman, Malhotra, Central 2002).

IT service

IT service, in general, is a term that describes services that depend strongly on IT artefacts, which is in accordance with the previous terms.

However, IT service may also refer to services that are provided by IT provider. IT provider is a subject whose business model depends strongly on IT artefacts. The services that one can mark as IT services, that does not depends strongly on IT artefact, but are provided by IT provider, are typically help-desk services. An internal IT department can be seen as IT provider. The second meaning is not the scope of this thesis.

“An IT service is a service provided to one or more customers by an IT service provider. An IT service is based on the use of IT and supports the customer’s business processes. An IT service is made up from a combination of people, processes and technology and should be defined in a Service Level Agreement.” (van Bon, Jong, Kolthof, Pieper, Rozemeijer, Tjassing, van der Veen, Verheijen 2007).

Software as a service

Software as a service, among the other cloud based services is commonly used in IT industry. It refers to a specific situation when a service provider offers functionality of applications that are working on provider’s infrastructure. (Bruckner, Buchalcevová, Chlape, Řepa, Stanovská, Voříšek 2012). This act may be also seen as e-service, but the definition clearly emphasises technological and producers’ point of view.

1.4 Goal and method definition

This thesis examines a problem that can be formalised into a theoretical model. The core relationship represents implication of service quality *q* using consumers' feedback *D*, where $D = \{d_1, ..., d_n\}$ and *d* is feedback given by a service consumer. Consumers' feedback may consist of unstructured feedback *u* and structured feedback *v*. Structured feedback is typically a scale or another close-ended item by which service consumers express their attitudes towards a service. Unstructured feedback is a coherent set of characters that create sentences, groups and words; and have an intended meaning coded by a service consumer that relates to a service. Consumer feedback needs an interpretation to capture a consumers' opinion, or sentiment *s*, about a service. Then the sentiment needs further analysis to explore service quality implications. Alternatively to the sentiment, the emotionality *m* can be captured from the unstructured feedback. Its relationship with service quality *q* will enhance the implication based on sentiment expression.

The goal of the thesis is to find how can consumers' feedback imply service quality by answering following research questions.

The object of the research is e-service quality.

The subject of the research is online consumer feedback.

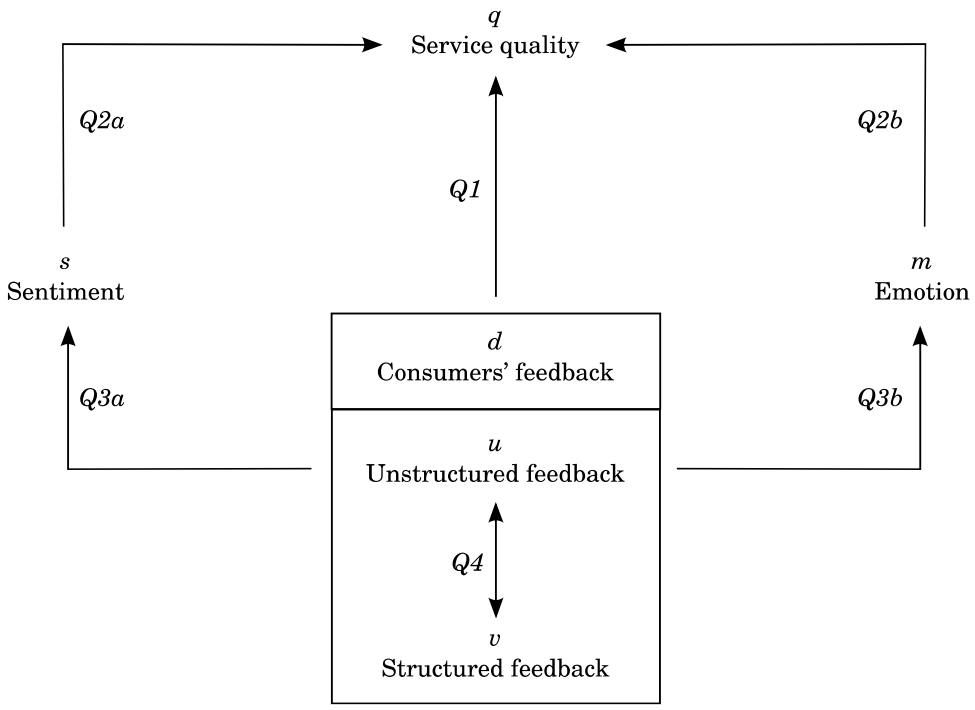

Figure 1.1: Scheme of the problem definition (author)

All the relationships from the model can be converted into following research questions. *Q1*, the core question, must be explained with use of questions *Q2*, *Q3* and *Q4*.

Q1 How can consumers' feedback *d* imply service quality *q*?

Q2a How can consumers' feedback sentiment *s* imply service quality *q*?

Q2b How can consumers' feedback emotionality *m* imply service quality *q*?

Q3a How can be consumers' feedback sentiment *s* extracted from *d*?

Q3b How can be consumers' feedback emotionality *m* extracted from *d*?

Q4 What is the interplay between structured consumers' feedback *u* and unstructured consumers' feedback *v*?

First of all, a service quality literature review will be conducted to explore *Q1* in general and to map the current state of knowledge about *Q4*. Although it is computer analysis what is the main considered innovation for service quality diagnosis, a critical factor for the use of such analysis is to understand what challenges service quality research faces nowadays and what are related research gaps.

The opinion mining literature will be explored as well as emotion research literature to answer *Q3*. This step involves mainly theoretical background mapping and identification of suitable methods that can help to fill the gaps. Special attention will be paid to studies that already use opinion mining methods in relation to service quality and try to address the same questions as this work, *Q2* and *Q3* in particular.

Two case studies will validate theoretical assumptions from previous review chapters. The first study works with online feedback from banking service and is focused on the research questions *Q2a* and *Q3a*. The relationship between the subtypes of consumers' feedback will be discussed as well to help answer *Q4*. The second study processes online feedback from call service. It focuses on the research questions *Q2b* and *Q3b*. A necessary part of *Q2* that deals with aspect mining is discussed in the comparison study.

To explore the relationship between consumers' feedback and service quality more in depth and to answer the research question *Q4*, a qualitative study will be undertaken. In contrast to statistical methods that will use previously mentioned studies, this study will aim at discovering hidden phenomena and discussing theoretical assumptions through detailed qualitative research on a small sample of quality reviews.

1.5 Content of the thesis

Chapter two, Service quality literature review, deals with the theoretical background of the thesis. It presents known models of service quality, approaches to quality measurement and observed phenomena. **Chapter three**, Opinion mining for service quality improvement, describes how can opinion mining contribute to service quality research. It presents sentiment classification, emotion classification and aspect mining methods and review studies that applied such methods for service quality diagnosis. In the next chapters, practical contribution is presented. **Chapter four** includes a case study from the online banking area. It explores content of online banking message board and provides method for continuous service quality monitoring. **Chapter five** describes a case study that focuses on emotion classification and relation of expressed model emotions to service quality. **Chapter six** examines deeper the relationship between structured and unstructured quality review data. **Chapter seven** discusses gaps in the service quality literature and contribution of presented studies to these gaps.

Chapter 2: Service quality literature review

2.1 Introduction

This chapter surveys service quality literature to reveal contemporary challenges and research gaps. The core aspect of the review is a concept of service quality in the digital society and its diagnosis. Understanding of service quality phenomena will help to answer the research question *Q1* and Q2.

Discussions how quality should be defined have last for years and differ according to the point of view. In general, (Garvin 1984) identified five categories, or approaches, to the concept of quality: transcendent approach, manufacturing-based approach, user-based approach, product-based approach and value-based approach. Due to the nature of service and importance of the role that consumers play in service production and delivery (Gronroos, Voima 2013; Bitner, Zeithaml, Gremler 2010), a quality of service is assessed mainly from consumers' point of view. Consumers also stand in the centre of following service quality models.

2.2 Service quality models

The research area of service quality is rich in definitions, models and measurement issues (Seth, Deshmukh, Vrat 2005). The majority of models and definitions support the theory that ***service quality*** **is a result of comparison between consumers' service quality expectation and their perceptions of service quality they have experienced** (Seth, Deshmukh, Vrat 2005). If consumers perceive that an actual level of an attribute associated with a service exceeds their expected level (positive disconfirmation of expectations), the service is likely to be evaluated favourably. Conversely, if a level of an attribute falls under the expected level, the service might be evaluated less favourably on that attribute (Parasuraman, Zeithaml, Berry 1988; Brown, Swartz 1989; Boulding, Kalra, Staelin, Zeithaml 1993; Gotlieb, Grewal, Brown 1994; Cronin Jr, Taylor 1994; Brown, Venkatesh, Goyal 2014; Brady, Cronin Jr 2001).

Although the difference between expectation and perception leads to perceived quality, these two drivers are hardly separable for an individual (Cronin Jr, Taylor 1992), because **the process is happening automatically while evaluating a service**. If a consumer expresses perceived quality on a scale, for example, it already contains the result of the process. Nevertheless, the discussion whether a performance-expectation scale or only a performance scale captures a service quality better persists (Mauri, Minazzi, Muccio 2013).

A finding of an extensive literature research (Seth, Deshmukh, Vrat 2005) shows that service quality models reflect the shift from traditional services to IT services. The reviewed literature still discern service quality as a very complex and depending on specific service conditions such service setting, situation, time, need and their influence on customer expectation and perception.

All reviewed quality models aspire for identification of factors that influence service quality. For instance, the first service quality model (Grönroos 1984) presents three quality factors: technical quality, functional quality and corporate image of the organisation. The technical quality relates to what was delivered whereas the functional quality relates to how the service was delivered. This division persists and applies to all service fields. In the healthcare services, for instance, the latest paper (James, Calderon, Cook 2017) distinguishes between the technical quality of service such as diagnosis or treatment and process quality such as waiting time or kindness. Another example, the definition of quality in IT service literature (Taylor, Iqbal, Nieves 2007, p. 259) refers more to internal quality. "The ability of a product, service, or process to provide the intended value."

The quality factors mentioned in the reviewed literature can be roughly divided into four groups. The first group refers to the the **internal service quality**. This considers all the factors that service provider can manage. These factors are remarkably positively correlated with the consumers' perceived quality (Caruana, Pitt 1997). The second group considers **service conditions** and refers to the service characteristics of inseparability, variability and perishability that directly influence service delivery process. These factors could be daytime, weather, speed of internet connection etc. The third group is linked with **social dynamics** and phenomenons like Word-of-Mouth. A corporate or service image could be an example of such factor. The last group refers to **consumers' state of mind**. It considers personal need which is the reason why an individual demands a service and past experience with a provider or similar services.

According to the method how researchers validate stability of proposed dimension, **a quality dimension is a general concept that consists of a set of attributes that correlate while influencing the perceived service quality**. The most of service quality studies (Seth, Deshmukh, Vrat 2005; Yarimoglu 2014; Duan, Cao, Yu, Levy 2013; Palese, Piccoli 2016) still use dimensionality from the SERVQUAL (Parasuraman, Zeithaml, Berry 1988):

(a) tangibles: physical facilities, equipment, and appearance of personnel;

(b) reliability: to perform the promised service dependably and accurately;

(c) responsiveness: to help customers and provide prompt service;

(d) assurance: courtesy knowledge, ability of employees to inspire trust and confidence;

(e) empathy: caring, individualized attention the firm provides its customers.

Conceptualisation of service quality does not only involve the multiple dimensions, but also multiple layers (Dabholkar, Thorpe, Rentz 1995). Some authors such as (Brady, Cronin Jr 2001) underwent systematic service quality dimensions refinements and proposed a different groups: (a) personal interaction quality, (b) physical service environment quality, (c) outcome quality; or extended service quality concept of a new

phenomenon. For instance, (Jia, Reich, Pearson 2008) proposed another dimension that corresponds with service climate.

Refinement of service quality dimensions in the modern economy, where regular services are transformed into so-called digital, electronic or IT services seems to be much more challenging. “New technologies and the increasing awareness of the dynamic nature of services underline the need for an updated analytical perspective which takes into consideration the crucial factors for the company evolution in uncertain and more competitive environments.” (Mauri, Minazzi, Muccio 2013) “New models and frameworks will be needed to accommodate, predict, and control these widespread technology changes.” (Bitner, Zeithaml, Gremler 2010)

2.3 E-service quality models

E-service quality can be defined as the extent to which a Web site facilitates efficient and effective shopping, purchasing and delivery of products and services (Zeithaml, Parasuraman, Malhotra, Central 2002). (Santos 2003) defines e-service quality more from the consumer point of view as an overall customer assessment and judgement of e-service delivery in the virtual marketplace. What customers expect from these new, innovative, technology-driven services does not necessarily fit the early models of service expectations (Parasuraman, Zeithaml, Malhotra 2005; Bitner, Zeithaml, Gremler 2010).

Six studies that explore the quality of services delivered through IT systems are worth taking into consideration (Zeithaml, Parasuraman, Malhotra, Central 2002; Wolfinbarger, Gilly 2003; Yang, Fang 2004; Parasuraman, Zeithaml, Malhotra 2005; Bitner, Zeithaml, Gremler 2010; Jiang, Jun, Yang 2016). One of the earliest works (Zeithaml, Parasuraman, Malhotra, Central 2002) on e-service quality pointed to an expectation issue. Customers were able to express their expectation only about the quality attributes that correspond with the traditional service approach. The attributes regarding website were not articulated. Similar thoughts about consumer **technology readiness** were presented by (Parasuraman, Zeithaml, Malhotra 2005). “Customer-specific attributes (e.g., technology readiness) might influence, for instance, the attributes that customers desire in an ideal web site and the performance levels that would signal superior e-SQ.”

(Gefen 2002) found that the five service quality dimensions collapse to three with online service quality: (a) tangibles; (b) a combined dimension of responsiveness, reliability, and assurance; and (c) empathy. (Wolfinbarger, Gilly 2003) defined factors that influence e-services of online shopping: (a) web site design, (b) fulfilment/reliability, (c) privacy/security and customer service. The authors of (Parasuraman, Zeithaml, Malhotra 2005) proposed these e-service quality dimensions: (a) efficiency, (b) system availability, (c) fulfilment, (d) privacy.

(Yang, Jun, Peterson 2004) used online content analysis to find e-service quality dimensions. The authors of the study combined the traditional quality model with information system quality, as the e-service is strongly based on IT components. Another content study by (Yang, Fang 2004) sees **technology adoption model** (Davis, Bagozzi 1989) **as an important part of e-service quality**. The model works with two factors: ease of use and usefulness. Perceived ease of use refers to the degree to which a person believes that using a particular system would be free of effort, and perceived usefulness refers to the degree to which a person believes that using a particular system would enhance his or her job performance. The content analysis (Yang, Fang 2004) resulted in following dimensions: (a) responsiveness, (b) reliability, (c) credibility, (d) competence, (e) access, (f) courtesy, (g) continuous improvement, (h) communication, (i) service portfolio, (j) content, (k) timeliness, (l) security, (m) aesthetic, (n) ease of use, (o) system reliability, (p) system flexibility. After extensive literature review (Ladhari 2010) stated six e-SQ dimensions that consistently recur: (a) reliability/fulfilment, (b) responsiveness, (c) ease of use/usability, (d) privacy/security, (e) web design, (f) information quality/benefit. The most recent e-service dimensionality research (Jiang, Jun, Yang 2016) identified five key e-service quality dimensions: (a) care, (b) reliability, (c) products portfolio, (d) ease of use and (e) security.

The idea to **combine information systems quality with service quality** was also presented by (Jingjun, Benbasat, Cenfetelli 2013), moreover, they took in consideration information quality to the consideration. They found that system quality has an indirect effect on service quality, indicating that **customers rely less on system quality and more on information quality** (that itself is influenced by system quality) **in forming a perception of service quality**.

There is also an opinion that **service quality can be approximated to a component quality if the service component plays a key role in the service** (Kritikos, Pernici, Plebani, Cappiello, Comuzzi, Benrernou, Brandic, Esz, Parkin, Carro 2013). “If services are considered as standalone software modules, then their quality can be determined by the attributes that traditionally characterize software quality and, thus, by the attributes defined in the ISO 9126 model.” It is also supported by (Lepmets, Cater-Steel, Gacenga, Ras 2012; Choi, Yoo 2009). “If any software application or module in IT service behaves incorrectly, service quality, and customer satisfaction will decrease significantly.”

The findings presented in the last two paragraphs seems to be contradictory. An explanation probably lies in a specific service composition. Perceived quality of services that strongly relies on the use of information technologies tends to be influenced more by perceived quality of information systems.

In conclusion, **the services that rely more on information technologies depend more, among the other factors, on technology acceptance and IT component quality**, which usually refers to internal or technical service quality. The technical quality of IT components is usually seen as IT service quality in IT

service management literature. For instance, (Voříšek, Pour 2012) refer to IT service quality dimensions such as availability, response time, security, reliability, flexibility and performance.

Therefore, **there is no service quality model that explains quality perception for all possible e-service implementations**. Literature review (Mauri, Minazzi, Muccio 2013) calls for a use of a less static approach in order to stick to the dynamic nature of the service sector and of the concept of quality. The **right model should be reasonably selected according to research and available data** (Duan, Cao, Yu, Levy 2013).

A possible solution for keeping a model of service quality is to **build a service aspect hierarchy**, according to entity modelling from opinion mining field (Hu, Liu 2004; Liu 2012, 2015). The aspect hierarchy can include all factors mentioned in this text, but should be tailored to a concrete service. A good lens for a top level would be a selected service dimensionality. The dimensions will then crumble to lower levels in hierarchy and will describe, for instance, hardware components quality and software quality in the tangible service dimension. The service model, according to opinion mining theory, in presented in chapter 3.2 Modelling service quality using opinion mining terminology, page 24.

However, a service aspect hierarchy in **expert understanding of service quality might not correspond with consumer point of view** (Song, Lee, Yoon, Park 2015) and should be developed rather in an iterative process.

Another issue relates to dimension weights. Empirical studies that explore service dimensionality (Gefen 2002; Wolfinbarger, Gilly 2003; Parasuraman, Zeithaml, Malhotra 2005; Yang, Fang 2004) resulted also in different dimension weights. In practice, **each consumer sees dimensions of a different importance**. The importance might even change during interpretation of quality reported by another consumer (Duan, Cao, Yu, Levy 2013). Different perceived weight of dimensions was solved, for example, by (Redchuk 2010). He proposed calculating the individual weights while measuring perceived quality.

2.4 Surveys

The surveys are the first possible research subject that contain consumer feedback, or generally said user generated feedback, and is used for service quality measurement (*Q1*).

The SERVQUAL instrument or its modification (Parasuraman, Zeithaml, Berry 1988; Parasuraman, L, Zeithaml 1991; El-Bayoumi 2012; Jiang, Klein, Carr 2002; VanDyke, Kappelman, Prybutok 1997; Vanparia, Tsoukatos 2013; Kettinger, Lee 1997; Pitt, Watson, Kavan 1997; Ladhari 2008; Pitt, Watson, Kavan 1995; Li, Suomi 2009; Choudhury 2014; Ruijin, Yunchang 2010; Alanezi, Kamil, Basri 2010; Kettinger, Lee 1994; Leong, Hew, Lee, Ooi 2015; Li, Tan, Xie 2002) stays a dominant quality measurement tool in the literature. A shift towards e-service quality represents survey instruments such (Parasuraman, Zeithaml, Malhotra 2005; Li, Tan, Xie 2002; Tsang, Lai, Law 2010; Han, Baek 2004).

However, there is an issue that relates to the measurement techniques based on close-ended questions. The surveys composed of the SERVQUAL or other's service dimensionality do not offer enough space to express consumers' attitudes to service quality and do not provide identification of concrete pain points of a service. Even if authors focus their papers on an alternative measurement techniques, the papers still deals with some sort of measurement scales (Ladhari 2008).

Literature reviews of e-service quality measurement were conducted by (Ladhari 2010) and (Yarimoglu 2015). The both studies show that **all reviewed measurement instruments are based on surveys**. With regard to the rising amount of online user generated content where quality judgements are already present, monopoly of SERVQUAL-like quality surveys opens a large research gap that can be filled by quality analyses with use of opinion mining methods.

2.5 Online consumer reviews

Whereas the quality surveys prevail in the literature, in practice, online services use reviews as a de facto standard. Despite that fact, only a minority of works about the online reviews refers to service quality. For instance, (Qu, Zhang, Li 2008) used reviews to map service quality dimensions. A further investigation of the relationship between consumer reviews and service quality is needed. The issue relates to the research question *Q1* of this thesis.

An online review is a type of user-generated content (UGC). It is a form of feedback where individuals express their opinions about products, services or organisations. **Online reviews are a common form of socialised data representing spontaneously shared opinions by customers on review platforms** (Mudambi, Schuff 2010). It can take a form of structured or unstructured data. Online reviews are typically short texts accompanied by a Likert-type scale. Online service reviews enable two primary functions: to assist the decision-making of service consumers and to assist service providers in service quality improvement (James, Calderon, Cook 2017).

As online reviews are a type of user-generated content, an issue of information data quality relates to the UGC in general. Casual users often lack domain expertise and cannot be held accountable for the quality of data they contribute (Lukyanenko, Parsons, Wiersma 2014). The consumers' data from online reviews describe directly what customers think about the service. However, an extraction and interpretation are still needed. Moreover, a reliability of measures based on consumer review is questionable.

The observation of examples from websites such as Google Play or Amazon showed that review guidelines differ from the quality surveys. For instance, description of review body on Google Play (Google Inc. 2017) says, *"Tell others what you think about this app. Would you recommend it and why?"*. The rating scale labels look as follows: *"Hated it"*, *"Didn't like it"*, *"Just OK"*, *"Liked it"*, *"Loved it"*. Amazon

reviews offer a similar labels: *"I hate it", "I don't like it", "It's okay", "I like it", "I love it"*. (Amazon.com Inc. 2017) notes in the help section of its website: *"Customer Ratings allow you to share information on the product attributes you consider important and rate those attributes on a 5-star scale."*

In accordance with the presented scales, (Liu 2015) refers rating scales to the sentiment polarity and intensity. Based on the observation and reviewed literature it is possible to claim that:

(a) the review reflects general opinions towards the products or services and ratings;

(b) the review ratings express a sentimental attitude towards the product or service.

2.5.1 Review rating characteristics

Literature research shows that online reviews do not have the uni-modal distribution, but rather bi-modal in a shape of the letter U or J (Hu, Pavlou, Zhang 2006; Zhu, Zhang 2012; Hu, Pavlou, Zhang 2014; Hu, Zhang, Pavlou 2009). (Hu, Zhang, Pavlou 2009) also found that a product with the bi-modal distribution on Amazon resulted in the uni-modal distribution in controlled experiment. They think that the different distribution is caused by a different motivation to write a review. "People tend to write reviews only when they are either extremely satisfied or extremely unsatisfied. People who feel the product is average might not be bothered to write a review." The assumption can be modified to fit service quality theory: **Individuals tend to review a service more likely if their perceived experience exceeded their expectation or did not meet their quality expectations**.

The bi-modality of online reviews causes that **an average score is an insufficient measure** (Hu, Pavlou, Zhang 2006). Average score of the online reviews does not reveal the 'true' product quality since the consumers' opinions do not converge to, or concentrate around the mean**,** as commonly suggested in the literature. It rather reflects the balance point of very different opinions.

There are also **significant cultural differences in reviewing behaviour**. Users of Chinese movie review site Douban.com, for instance, tend not to write extreme reviews in comparison to IMDB.com users (Koh, Hu, Clemons 2010), but the behaviour of the people from the same cultural background may change if they review anonymously (Zhang, Luo, Li 2012). The similar difference between Chinese and English reviews is shown in the study from hospitality sector by (Zhang, Yu, Li, Lin 2016).

Differences may also occur across service quality sectors. Study from healthcare sector (Gao, Greenwood, Agarwal, McCullough 2015) shows that patients rate less likely physicians that provide a low quality service.

Categories in star ratings are too broad to capture consumers' specific concerns. Content and correlation analysis (Qu, Zhang, Li 2008) showed that even the itemised, multidimensional, ratings are

over-aggregated. Ratings also do not cover all subcategories that naturally appear in the review text and may cause misunderstanding in interpretation.

2.5.2 Review dynamics and service quality

As a public information, each review may work as an input for a service quality evaluation. Although many studies try to find whether reviews have an influence on purchase behaviour (Xia, Bechwati 2010), only a few of them examined review dynamics in relation to service quality. This issue is being explored in the Word-of-Mouth (WoM) discipline.

Result of (Duan, Cao, Yu, Levy 2013) indicates that **a better total ranking of product or service actually may lead to lower user review ratings in the future**. It may be caused by raised expectation, which is already an observed characteristic of perceived quality (Parasuraman, Zeithaml, Berry 1988; Brown, Swartz 1989; Gotlieb, Grewal, Brown 1994; Cronin Jr, Taylor 1992; Brown, Venkatesh, Goyal 2014; Brady, Cronin Jr 2001). "When top-ranked [services] do not meet users' high expectations, they are more likely to share the negative (i.e., lower than expectation) experience by posting negative reviews." (Duan, Cao, Yu, Levy 2013).

The same authors also found that **a higher cumulative average rating from past reviews is correlated to a higher average rating from current reviews**. The results may be explained by a strategy of how individuals with limited information-processing capacity tend to choose a review rating. These individuals follow the previously observed model of service quality (Zhang, Yu, Li, Lin 2016).

(Duan, Cao, Yu, Levy 2013) also found that **perceived quality of attribute and its weights may change when an individual reads and interprets a review text**. Other authors examined social dynamics of reviews in general (Gao, Gu, Lin 2006) or how people are influenced by review (Robinson, Goh, Zhang 2012).

2.5.3 Relation between review rating and review text

Although an interest of researchers is attracted mainly by review ratings, unstructured content of review also matters. For instance, research of online reviews in the Word of Mouth context (Xia, Bechwati 2010) shows that individuals perceive factual reviews as more useful than experiential reviews.

Many works have been published about better information value that unstructured data bring (Chowdhury, Reardon, Srivastava 1998; Qu, Zhang, Li 2008). A study (Tsang, Prendergast 2009) validated, for instance, the assumption that text has a greater influence on interestingness of review and individual's purchase intention. A work (Mudambi, Schuff 2010) that examined perceived helpfulness of review found that deeper reviews (measured in text length) are perceived as more helpful.

Several works investigated the issue of interplay between structured and unstructured review data. (Qu, Zhang, Li 2008; Zhang, Yu, Li, Lin 2016) found the relationship as statistically significant. It was also proved by an older study (Chowdhury, Reardon, Srivastava 1998) where researchers stated **a high degree of correspondence between structured and unstructured review data**. Limitation of these studies is that they do not explain phenomenon of interplay more in depth.

One of the studies that dive more in the relationship interplay (Duan, Cao, Yu, Levy 2013) observed that **different quality attributes collected from a text have a different weight on an overall rating.** The question is how these assumptions can be validated and used for quality diagnosis.

The recent study about the relationship between structured and unstructured user generated content (Zhang, Yu, Li, Lin 2016) validates a theory that **structured-unstructured review interrelationship can be significantly affected by a variety of cognitive factors, such as extreme dissatisfaction and opinion heterogeneity.**

The first mentioned factor is related to the theory of emotional memory (Bradley, Greenwald, Petry, Lang 1992) and prospect theory (Kahneman, Tversky 1979). The corresponding negative emotion that is ignited, such as anger, hatred and disgust, enhances the memory, and helps to recall the experience. Therefore, **the extremely dissatisfied consumer,** who has experienced an unhappy cognitive process and encoded negative emotional memory, **is more willing to generate itemized rating and textual comments that bear negative sentiments, and correspondingly result in higher intensity of the structured-unstructured review interrelationship. The other individuals, without the extremely unsatisfactory experience, exert less effort to the review process, which causes inconsistencies in the structured-unstructured review interrelationship.** This theory may also support findings that negative ratings are significantly correlated with long reviews (Palese, Piccoli 2016).

Results of (Zhang, Yu, Li, Lin 2016) also validated the hypothesis that structured rating scales are consistent with a review text when subcategories are rated heterogeneously.

According to (Zhang, Yu, Li, Lin 2016), **individuals with limited information-processing capacity** (Malhotra 1982) **deal better with structured form of review than with a free text.** (Walters 1961; Chowdhury, Reardon, Srivastava 1998). They think that **these people will tend to follow the previously observed model of rating.** For instance, luxury hotels may receive higher ratings, or higher rated services may again receive higher ratings (Duan, Cao, Yu, Levy 2013).

The research of (Tsang, Prendergast 2009) showed that consistent reviews produce higher trustworthiness. In addition, trustworthiness is hampered if a review has a negative text with a positive rating.

2.6 Chapter summary and discussion

This chapter puts a first theoretical baselines of service quality conceptualisation and related theories. It also defined possible sources of consumers' feedback and theories of social behaviour connected with quality reviewing process. Following text outlines first theoretical assumptions significant for this thesis.

Perceived service quality is a phenomenon connected strongly with an individual and it is influenced by various factors. Quality is measured using quality dimensions. Quality does not have multiple dimensions only, but also multi levels. Hence, quality can be expressed in a hierarchical model of quality attributes, where quality dimensions are on the top level. Service quality model is not a stable construct, and it needs to be tailored to different contexts. Service consumers perceive a different importance of quality dimensions. Weighting quality dimensions and attributes could be crucial factor for service quality interpretation from text. A general model of service quality of hierarchical structure is described in the figure 2.1.

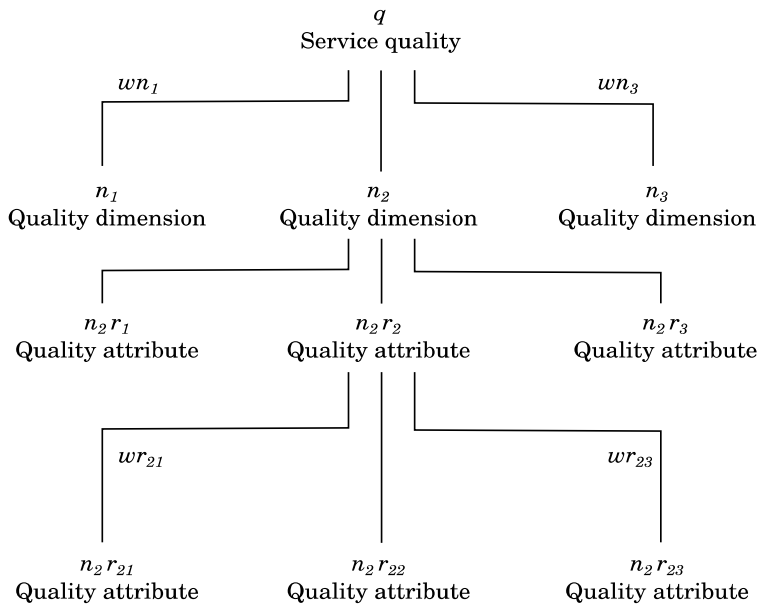

Figure 2.1: General service quality model (author)

The reviewed literature propose surveys as a tool for service quality measurement. Surveys usually contain multiple scale items that corresponds with service dimensions. The common issue is that consumers cannot freely express their opinions. Service managers can only see that one or more dimensions cause worse consumer experience, but they can only assume what the concrete source of the quality fall is. On the other hand, the close-ended quality surveys are easy to communicate as a general quality indicator.

Online consumer reviews are a type of user generated content (UGC) and a specific type of survey. Based on the academic literature and review examples, it can be said that online consumer reviews reflect general opinions towards the products or services and express a sentimental attitude towards the product or service. The online reviews are typically a short text accompanied by a Likert-type scale. Thus, it contains both structured and unstructured data. The online reviews are rather a poll than a representative research tool. It reflects the bi-modal distribution of consumers' rating. The reason of this phenomena is the willingness to review a service. Only consumers with strong dissatisfaction or strong satisfaction tend to write a review.

Because structured quality surveys or ratings are discussable as a full source of information about quality, other options such as unstructured data analysis need to be explored. The reviewed studies showed that interplay between these two types of sources is significant, nevertheless it is affected by a variety of cognitive factors, such as extreme dissatisfaction, opinion heterogeneity or information-processing capacity.

Theoretical assumption that is worth noting is that quality rating depends on cultural differences. These differences may be significant even in different service branches.

Although reviews are widely used, academic literature lacks proper explanation of diagnosing service quality from them. Four studies that explore on-line reviews in this context describe chapter 3.6.

Chapter 3: Opinion mining for service quality improvement

3.1 Introduction

The literature review in this chapter extends the review from the previous chapter. It goes deeper in unstructured data analysis and focuses on opinion mining research and its conceptualisation, methods and their potential for service quality diagnosis improvement.

> "One significant effort to respond to the need for automating the analysis of customer reviews has been using sentiment analysis (also known as opinion mining)" (Song, Lee, Yoon, Park 2015)

Opinion mining should help with service quality diagnosis and identification of strengths and weaknesses of a service in order to deliver a better service experience for each consumer. Opinion mining is a field of study that analyses peoples' opinions, expressed sentiment, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text (Liu 2015).

The first part of the chapter consists of service analysis and modelling using opinion mining concepts. The analysis is necessary for the application of opinion mining methods to service quality research. The second part answers the research question *Q3* by reviewing opinion mining techniques for consumers' opinion extraction and automated classification of sentiment and emotions.

The last part presents the first critical literature review in the area of service quality diagnosis using opinion mining methods. The review brings the light on the main research question *Q1* by covering questions *Q2a* and *Q3a*.

The analysis of consumers' feedback can be undertaken on different levels. The highest level is a whole message, review or email. In the opinion mining literature, the highest level is usually called *document*. The lower levels are sentences or particular service aspects. The lower levels provide higher fidelity of service image perceived by a consumer, but they are more difficult to extract.

3.2 Modelling service quality using opinion mining terminology

Prof. Liu (Liu 2015) put together the most complex set of conclusions and definitions related to opinion miming. These definitions create a baseline of this chapter. He uses the term *opinion* as a broad concept that covers sentiment, evaluation, appraisal, or attitude and associated information such as opinion target and the person who holds the opinion, whereas the term *sentiment* as the underlying positive or negative feeling implied by opinion.

According to (Liu 2015, p. 22), an opinion can be expressed as a quintuple (*e*, *a*, *s*, *h*, *t*), where *e* is a target entity, *a* is a target aspect of entity *e* on which an opinion has been given, *s* is a sentiment of the opinion on the aspect *a* of the entity *e*, *h* is an opinion holder, and *t* is an opinion posting time; *s* can be positive, negative, or neutral, or a rating. *e* and *a* together represent an opinion target, marked usually as *g*.

Prof. Liu also defines related terms. A *sentiment target*, also known as an opinion target, of an opinion is an entity or a part or attribute of the entity upon which the sentiment has been expressed.

An *entity e* is a product, service, topic, person, organisation, issue, or event. It is described as a pair, (*T*, *W*), where *T* is a hierarchy of parts, sub-parts, and so on, and *W* is a set of attributes of *e*. Each part or sub-part also has its own set of attributes.

Sentiment is an underlying feeling, attitude, evaluation, or emotion associated with an opinion. It is represented as a triple (*y*, *o*, *i*), where *y* is the type of the sentiment, *o* is the orientation of the sentiment, and *i* is the intensity of the sentiment.

Sentiment can be seen as linguistic-based, psychology-based, and consumer research–based. In opinion mining, consumer research–based classification is used; it can be divided broadly into two categories: rational sentiment and emotional sentiment (Chaudhuri 2006; Liu 2015). Rational sentiments are composed of rational reasoning, tangible beliefs, and utilitarian attitudes. They express no emotions. Emotional sentiments are composed of non-tangible and emotional responses to entities that go deep into people's psychological states of mind. Sentiment orientation can be positive, negative, or neutral. Neutral usually means the absence of sentiment. Sentiment orientation is also called polarity, semantic orientation, or valence in the research literature.

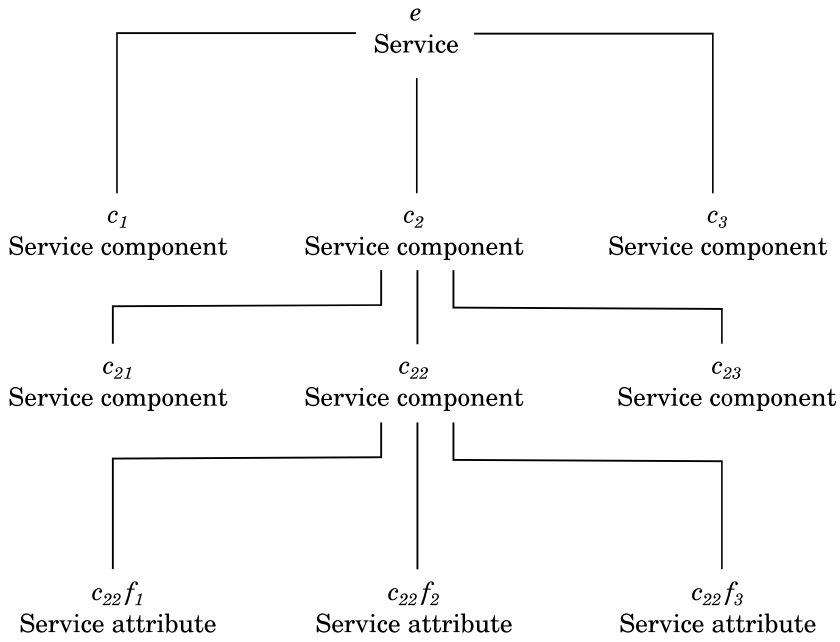

Figure 3.1: Service model of component hierarchy and attributes (author)

In service quality research, the opinion target entity e is a focal service. The service is composed of a hierarchical set of service components (or parts) T and a set of its attributes W. The components are sub-services, service artefacts, processes and personnel; $T = \{t_1, t_2,\ldots,t_m\}$ and $W = \{w_1, w_2,\ldots,w_n\}$. Each component t_i can be divided into more sub-components $\{t_{i1}, t_{i2},\ldots,t_{ip}\}$. Attributes[1] and components[1] are both parts of service aspects $A = \{a_1, a_2,\ldots,a_r\}$. Some authors (Song, Lee, Yoon, Park 2015) note that only the crucial aspects can be seen as service components or attributes. A scheme of hierarchically ordered service components and corresponding attributes is showed in figure 3.1.

Service consumers act as opinion holders h. They perceive service components t and service attributes w in a certain way, and they express their opinion about them.

Target service e is addressed in text by a finite set of entity expressions $\{ee_1, ee_2,\ldots, ee_s\}$. Each aspect $a \in A$ of service e can be expressed with one of a finite set of its aspect expressions $\{ae_1, ae_2,\ldots, ae_u\}$. An opinion document d contains opinions about a finite set of entities $\{e_1, e_2,\ldots, e_v\}$ and a subset of aspects of each entity. The opinions are expressed by a finite set of opinion holders $\{h_1, h_2,\ldots, h_w\}$ and are given at a particular time point t. In case of quality review, an opinion document d is a review with one opinion holder, who is a reviewer and a service consumer in the same time.

1 Name convention differs in the reviewed literature. The best arguments on conventions are provided by (Liu 2015). He propose to use *aspect* instead of *feature*, because feature can be also seen as attribute and is also used in machine learning. He uses *Component* instead of *object*, because object can be misunderstand for verb object from grammar.

3.3 Capturing sentiment and service quality

The question is, whether and when sentimental attitude towards service components or attributes corresponds with service quality concept. According to the service quality theory described in chapters 2.2 and 2.3, there are two main **differences between the discussed quality evaluation methods and simple measuring of sentimental attitude towards a service**.

The first difference is that the service quality research divides service quality into a certain number of dimensions. These **dimensions do not directly correspond to service components nor service attributes**. The quality dimensions rather represent states of selected attributes or components in general.

The second difference is that **quality surveys do not use the sentimental attitude for quality evaluation but the level of agreement with quality evaluation statements**. For example, (Li, Tan, Xie 2002) propose "Online ordering is simple" as one of the statements from *competence* quality dimension. It is clear from the example that the statements already contain a sentiment with a certain orientation and intensity. Respondents can only moderate the intensity of the expressed sentiment or turn its polarity by choosing a level of agreement with the statement.

The challenge of service quality research is to classify sentiment of service aspects and to assign these to service quality dimensions. Classification is usually done model-based using standard machine learning approaches or lexicon-based using pre-defined dictionary.

3.3.1 Model-based classification

According to (Liu 2015), document-level sentiment classification is considered the simplest sentiment analysis task because it treats **sentiment classification as a traditional text classification problem** with sentiment orientations and polarities as classes.

On the other hand, service quality researchers need to be careful when choosing training data. For example, one of the most cited papers (Pang, Lee, Vaithyanathan 2002) employed supervised learning to classify sentiment of movie reviews. They chose review ratings as data for sentiment training. As it was said in chapter 2.5, review ratings do not always correspond with sentimental expression from review content. The relationship is determined psycho-socially and may comprise other influences on perceived quality than those expressed in the text. Therefore, **training data for quality related sentiment should be more tailored and prepared by an expert supervisor**.

Besides training data preparation, **features need to be set up**. Liu (2015) summarises six types of features that are suitable for machine learning: (1) the most commonly used terms, their frequency and position; (2) part-of-speech of a word, especially with adjectives; (3) pre-defined sentiment words or phrases, which must be carefully selected to fit in a specific domain context; (4) additional rules that express an opinion;

(5) sentiment shifters, which are used to change sentiment polarity in language, for example an adverb *not*; (6) parsed syntactic dependency or dependency trees. There could be also found features such as emoticons (Wiebe, Wilson, Cardie 2005) in literature.

Machine learning input is then standard bag-of-features matrix, as per (Pang, Lee, Vaithyanathan 2002), for example, where $\{f_1, \ldots, f_m\}$ is a predefined set of m features that can appear in a document. $n_i(d)$ is the number of times f_i occurs in document d. Then, each document d is represented by the document vector $d := (n_1(d), n_2(d), \ldots, nm(d))$.

3.3.2 Lexicon-based classification

Sentiment lexicons (Liu 2015) offer more basic approach to sentiment classification. Lexicons contain coded or categorised list of words. For example, SO-CAL (Taboada, Brooke, Tofiloski, Voll, Stede 2011) method uses numerically expressed sentiment modifiers attached to each word. These modifiers are then used for overall sentiment calculation. One of the most popular lexicon is LIWC (Tausczik, Pennebaker 2010; Salas-Zarate, Lopez-Lopez, Valencia-Garcia, Aussenac-Gilles, Almela, Alor-Hernandez 2014; Lim, Yoon, Kim, Kim 2012; Pennebaker, Mehl, Niederhoffer 2003); commonly used in psychological related research of natural language.

In conclusion, sentiment classification methods rely on input quality, meaning training data for supervised learning or lexicon quality; moreover, input data should fit a specific domain. General, “off-the-shelf”, method comparison (Ribeiro, Araújo, Gonçalves, André Gonçalves, Benevenuto 2016) revealed that none of the tested methods exceeded sixty percent in F1 performance measure. Hence, there still exists space for improvements.

3.4 Emotionality within service quality diagnosis

It is a general opinion that emotions are expressed, besides the other ways, in a written text. The statement that emotions are therefore present in quality reviews is then very clear. Unfortunately, there are three issues related to the topic: (a) there is no clear consensus in the emotion research literature; (b) theories about the relationship between emotions and service quality seems to be immature; (c) how to classify emotions expressed in a text.

3.4.1 Emotionality measurement

Since text-mining techniques are well covered by the literature, the biggest challenge is to understand what an expressed emotion means regarding the service quality. The issue of emotion research is too broad and cannot be fully explored within this thesis. The chapter presents only the related findings and overlaps with service quality research.

> "The study of emotion is one of the most confused (and still open) chapters in the history of psychology … more than 90 definitions of 'emotion' were proposed over the course of the 20th century" (Plutchik 2001)

Since there is no consensus on the definition of emotion in the literature, only a rough overview based on the work of Plutchik and Scherer will be discussed.

"Emotions are not simply linear events, but rather are feedback processes. The function of emotion is to restore the individual to a state of equilibrium when unexpected or unusual events create disequilibrium." (Plutchik 2001). Plutchik also notes that feeling states tend to be followed by impulses to action. However, it is questionable, if in case of perceived quality, any impulses that lead to writing a specific quality review arise.

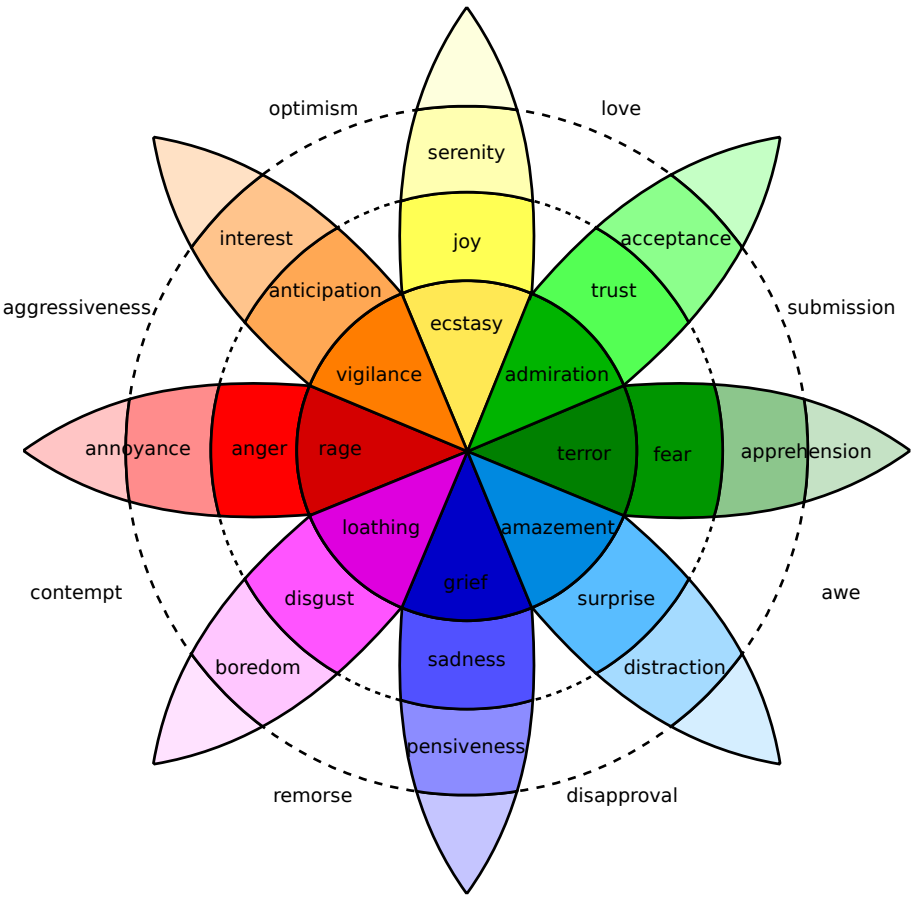

Figure 3.2: Plutchik's wheel of emotions (Wikimedia Commons 2011; Plutchik 1980)

Philosophers and psychologists have proposed sets of primary, or basic, emotions that count from 3 to 11 (Plutchik 2001). All the sets include *fear*, *anger* and *sadness*; the most include *joy*, *love* and *surprise*. The

widely used psycho-evolutionary theory (Plutchik 1980, 2001) assumes there are eight basic[2] emotion dimensions arranged in four pairs: joy versus sadness, anger versus fear, trust versus disgust and surprise versus anticipation.

(Plutchik 2001) defines emotion as a kind of homeostatic process in which behaviour mediates progress toward equilibrium. Primary emotions can be conceptualised in a way analogous to a colour wheel – by placing similar emotions close together and opposite emotions on the other side, like complementary colours. The primary emotions, just as colours, can be mixed together in order to create new emotions. For example, *joy* and *acceptance* lead to a new mixed emotion – *love*; *disgust* with *anger* lead to *hatred* or *hostility*. Such mixtures are called primary dyads in the theory. Wheel of emotions is illustrated in figure 3.2.

(Scherer 2005) explored emotions in a broader complex of affective phenomena and distinguished between preferences, attitudes, moods, affect dispositions, interpersonal stances, aesthetic emotions and utilitarian emotions.

Proper emotion measurement explored was explored by (Scherer 2005)[3], unfortunately, there is no other access than to ask the individual to report on the nature of his experience. For that task (Scherer 2005) proposed The Geneva Emotion Wheel[4], which is, similar to Plutchik's, a circumplex. The related method is based on semantic descriptions of affect states. The description is a text label, one word, on the circumference. During a self-evaluation, an individual selects one or more points on a graphical representation of the wheel. Another circumplex model was proposed by Russell (Posner, Russell, Peterson 2005), the valances are *pleasant – unpleasant* and *activation – deactivation*.

2 (Scherer 2005) proposes using the term "modal emotions" rather than "basic emotions" because there is no consensus on what basic means in this context.

3 Ideal measure of emotions according to (Scherer 2005) should consists of (1) the continuous changes in appraisal processes at all levels of central nervous system processing (i.e. the results of all of the appraisal checks, including their neural substrata), (2) the response patterns generated in the neuroendocrine, autonomic, and somatic nervous systems, (3) the motivational changes produced by the appraisal results, in particular action tendencies (including the neural signatures in the respective motor command circuits), (4) the patterns of facial and vocal expression as well as body movements, and (5) the nature of the subjectively experienced feeling state that reflects all of these component changes, which is nowadays still impossible.

4 The Geneva Emotion Wheel is a product of The Geneva Emotion Research Group http://www.unige.ch/cisa/gerg.html

3.4.2 Classification of emotions from text

Emotions, or affective states in general, can be classified using similar techniques as sentiment. The only difference is, that affective states are multidimensional and constitute a more complex issue.

Lexicon based approach is the most used approach of classification in the reviewed literature. One example is the Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Mehl, Niederhoffer 2003; Tausczik, Pennebaker 2010; Lim, Yoon, Kim, Kim 2012; Salas-Zarate, Lopez-Lopez, Valencia-Garcia, Aussenac-Gilles, Almela, Alor-Hernandez 2014). It contains multiple categories of words that refer to different affective states or human activities. The LIWC method count frequencies of words in the pre-defined categories.

Another example of lexicon based emotion classification is NRC Emotion Lexicon (Mohammad, Turney 2013). The lexicon follows Plutchik's emotion model. An advantage of this lexicon is that it does not contain only emotion words, but emotion associations with more than fourteen thousand English words. The associations are expressed with a degree of intensity with a particular modal emotion. Then, the feedback can be converted using the word associations into an emotion vector.

Similar to sentiment classification, but more reliable is to train a model using supervised learning. The only difference is that the emotion is more a dimensional problem, therefore multiple class learner is a good choice. A learning of emotions and comparison with NRC lexicon performance is described in a case study in chapter 5.

3.5 Aspect mining

As modelled in chapter 3.2, an aspect and sentiment are the main parts that build an opinion. When sentiment is classified, it is necessary to link it with corresponding aspects of a service. The goal of this chapter is to discuss techniques for automatic service aspects extracting, and do discuss their ability to provide clear managerial insights and to contribute to the research question *Q3*. According to (Liu 2015), there are four main approaches to extracting explicit aspects:

(a) Extraction by finding frequent nouns and noun phrases.

(b) Extraction by exploiting syntactic relations. There are two main types of relations:

  (i) Syntactic dependencies depicting opinion and target relations.

  (ii) Lexico-syntactic patterns encoding entity and part/attribute relations.

(c) Extraction using supervised learning.

(d) Extraction using topic models.

General stages of aspect mining (Vencovsky, Lucas, Mahr, Lemmink 2017) are following: In the beginning, data need to be gathered, cleaned and pre-processed. Additional data enrichment like part-of-speech classification or name recognition can be undertaken. Then, raw output like terms, n-grams, co-occurrences or syntactic relations can be retrieved. These data contain possible service aspects but need to be organised in a certain way and filtered first. After that, processed aspects have to be linked with a service context and discussed with service experts to build a final list. An organisation of aspects into categories and discussion with experts is crucial because service aspects are often implicit (Liu 2015) and cannot be revealed easily.

The paper (Vencovsky, Lucas, Mahr, Lemmink 2017) focuses on the stage of pre-processing and retrieving data from a dataset. It is only one paper that provides a comparison of service quality aspect extraction techniques. It examines four different techniques for mining service aspects; three of them are frequency-based; the fourth is based on syntactic relations; the techniques are following: (a) weighting terms by TFC-ICF, (b) n-grams, (c) word co-occurrences and (d) grammatical dependencies.

3.5.1 Frequency-based aspect extraction

All the three frequency-based techniques are built on the assumption that expressed aspects are linked with an expressed sentiment (Liu 2015). In contrast to fundamental frequency measures (Hu, Liu 2004), the paper uses frequency of aspects that are more likely in high and low rated reviews. The same approach was already used for a similar purpose in the past (Ma, Xue, Zhang 2016).

Weighting terms by TFC-ICF could be a good example of the frequency-based extraction techniques. Unlike (Ma, Xue, Zhang 2016) that also defines TFC-ICF, the formula, similar to original TF-IDF, was used in (Vencovsky, Lucas, Mahr, Lemmink 2017).

Let f_{ij} be the raw frequency of term t_i in category c_j and f_{kj} a frequency of each term t_k in category c_j. Let $|C|$ be the total count of categories in the dataset and c_{fi} the number of categories in which term t_i appears at least once. Then, the relative frequency of term t_i in category c_j is given by

$$tfc_{ij} = \frac{f_{ij}}{\sum_k f_{kj}}, \quad (1)$$

and inverted category frequency icf_i for term t_i is given by

$$icf_i = \log\left(1 + \frac{|C|}{cf_i}\right). \quad (2)$$

Finally, the criterion for weighting service terms w_{ij} is given by

$$w_{ij} = tfc_{ij} \times icf_i. \quad (3)$$

After weighting terms in categories according to a rating, more abstract groups were set up. Two joint groups were proposed: low and high rating.

In order to get better results, the study used part-of-speech (POS) filtering and filtered out all parts-of-speech except for nouns in pre-processing stage. The POS tags are predicted based on English left3words model from Stanford NLP library (Manning, Bauer, Finkel, Bethard, Surdeanu, McClosky 2014).

The frequency-based techniques, trialled in the comparison, demonstrated a certain ability to satisfactorily isolate and extract service aspects, but also produced a lot of noise output in contrast to grammatical dependency technique.

3.5.2 Extraction using syntactic relations

The grammatical dependencies technique (the fourth trialled in the comparison) **offered much better results**. It works with the grammatical structure of the text and makes complex queries possible. Dependency networks are widely used in computational linguistics for machine translation tasks. One of the essential papers on the subject is written by (Quirk, Menezes, Cherry 2005). Strong method applications can be found in bioinformatics, e.g. (Erkan, Ozgur 2007), and in the other UGC contexts, e.g. (Hassan 2010; Joshi, Das, Gimpel, Smith 2010). No previous research had explored this approach in the service management context before (Vencovsky, Lucas, Mahr, Lemmink 2017).

As discussed in the general method statement, data can be enriched after pre-processing stage. In the comparison (Vencovsky, Lucas, Mahr, Lemmink 2017), the enrichment was done, based on a comparison in

(Stent, Choi, St, St, York 2015), by Neural Network Dependency Parser from Stanford NLP group (Chen, Manning 2014). The parser employed pre-trained dependency model and provided Universal Dependencies annotations. Annotated dependencies with meta-data were then loaded into a graph database. The database offers a relative freedom of choosing query language. Neo4J database with Cypher Query language was used in that case. Figure 3.3 shows an example sentence tree. Nodes are labelled with POS tags and edges with dependency types.

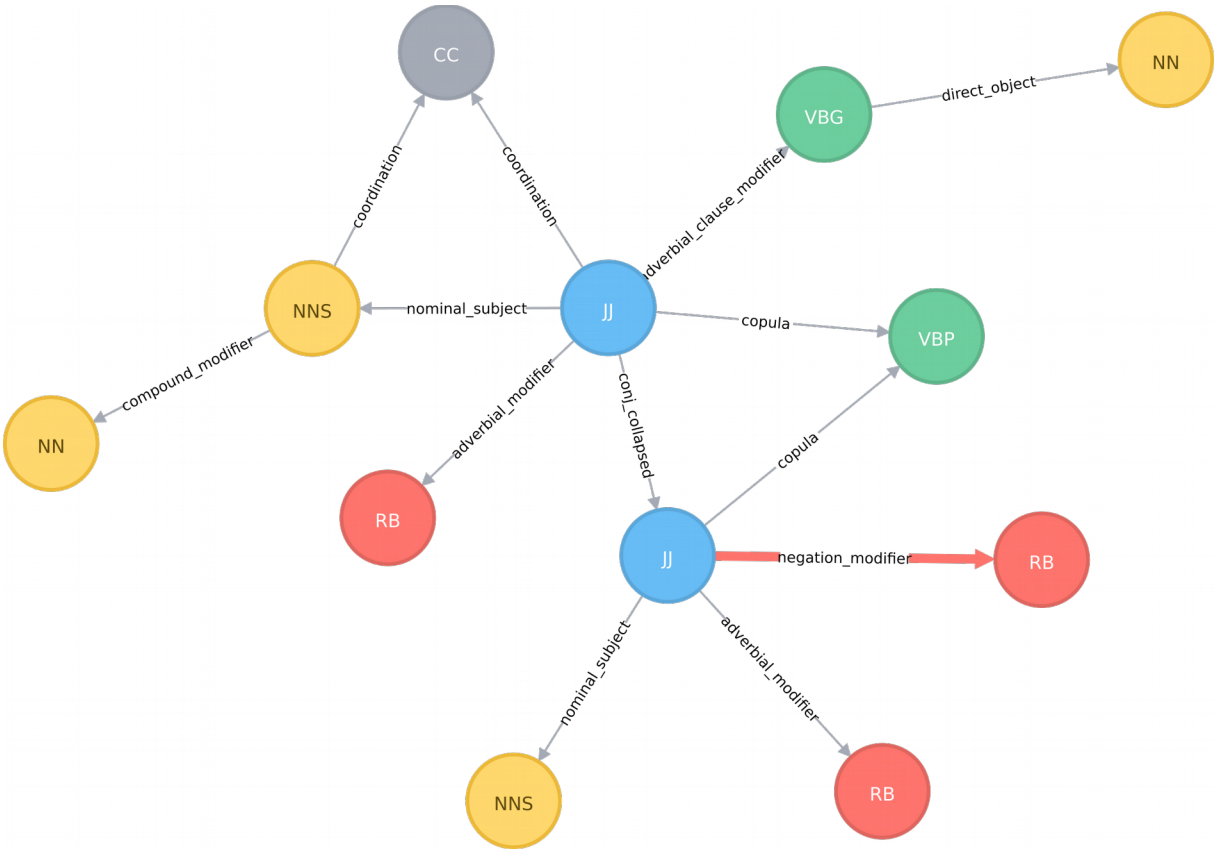

Figure 3.3: Example of a Sentence Tree with Universal Dependencies Annotations

There are several proposed patterns recommended for aspect extraction. In the paper (Vencovsky, Lucas, Mahr, Lemmink 2017), six different language patterns were identified based on sample sentences. Patterns are combinations of grammatical dependencies and pre-defined words.

No	Pattern[5]	Example
1	JJ_1—*nsub*→NN_1; (JJ_1—*amod*→RB_1); (JJ_1—*neg*→RB_2)	*"very friendly staff"*; *"very helpful staff"*; *"very good service"*
2	JJ_1—*nsub*→NN_1; JJ_1—*cop*→VBD_1; (JJ_1—*neg*→RB_1)	*"food was great"*; *"room was nice"*
3	NN_1—*amod*→RB_1; NN_1—*adjmod*→JJ_1; (JJ_1—*neg*→RB_2)	*"front desk"*; *"French restaurant"*
4	VB_1—*dobj*→NN_1; (VB_1—*aux*→VB_2); (NN_1—*adjmod*→JJ_1); (VB_1—*amod*→RB_1); (VB_1—*neg*→RB_2); where VB_1 match "expect"	*"do not expect fast service"*; *"expect drinks"*
5	VB_1—*dobj*→NN_1; (VB_1—*nsub*→NN_2); (VB_1—*amod*→RB_1); (VB_1—*neg*→RB_2); where NN_2 match "expectations"	*"food far exceeded expectations"*; *"resort met expectations"*
6	VB_1—*nmod*_prep→NN_1; NN_1—*cmod*→IN_1; (VB_1—*neg*→RB_1) where VB_1 match regular expression "complain.*"	*"complaints about resort"*; *"no complaints at all"*

Table 3.1: Language patterns for aspect extraction (Vencovsky, Lucas, Mahr, Lemmink 2017)

The first two patterns work with nominal subject and offer the clearest results. The third pattern outputs similar results as the first, non-grammatical, techniques. The rest of patterns enhances the coverage of possible service aspects with a minimum amount of noise. Moreover, negation in phrases is a part of the extracted patterns and thus enables us to work with customer feedback that does not include rating points.

In comparison to (Qiu, Liu, Bu, Chen 2009, 2011), researchers were also observing grammatical patterns, but they worked with sentiment words which were not needed in the previous case. The method the authors employed is called *double propagation*. They use set of dependencies called *MR* consisting of relations such as modifications, subjects and objects. Relations cannot be satisfactorily mapped to the previously presented patterns because they do not use common names from Universal dependencies project. The main principle of the paper is to find as much as possible connections between sentiment words and aspects. What remarkable is the idea of repetition of the propagation process with newly found sentiment words and aspects until no new sentiment words or aspects are found. More advanced method for aspect mining using grammatical dependencies is proposed by (Liu, Liu, Zhang, Kim, Gao 2016); the proposed approach is *recommendation-based* and, according to their validation, it resulted in higher performance than the double propagation offers.

The grammatical dependencies technique represents a good alternative to topic analysis for service aspect mining. The dependencies technique could even help with sentiment analysis. **The extracted phrases are independent of the word order and can easily capture negation of expressions. The phrases are extracted for all used aspects even when they are in conjunction.** For instance, a sentence, *"I like the website and app"*, contains two aspects: a website and an application. A query based on dependencies will

5 Abbreviation meanings: adverb (RB), adjective (JJ), noun (NN), verb (VB), past tense verb (VBD), preposition (IN), nominal subject (*nsub*), adverbial modification (*amod*), negation modification (*neg*), copula (*cop*), adjectival modification (*adjmod*), direct object (*dob*), case modification (*cmod*); Patterns in parenthesis are optional parts of a pattern.

extract *"I like website"* and *"I like app"* as well. These extracted phrases will be a good source of data for supervised learning.

Unlike topic analysis, **the dependencies technique does not offer a grouping of related terms**. The non-grammatical relationship between used words needs to be investigated. A couple of methods can help in this task. For instance, by using the same graph database it is possible to connect parsed words with WordNet database (Miller 1995; Finlayson 2014). Figure 3.4 illustrates connection of parsed dependencies to WordNet. By using the hypernym-hyponym relationship it is possible to query and extract[6] even abstract terms. Moreover, a similarity between words can be measured using the same database. The limitation of connection to WordNet is, that the right meaning of the word should be identified. The other possible approach of enhancing grammatical dependencies technique is to use word similarity based on the word usage. An example method is Deep learning skip-gram based on Word2Vec algorithm (Mikolov, Corrado, Chen, Dean 2013).

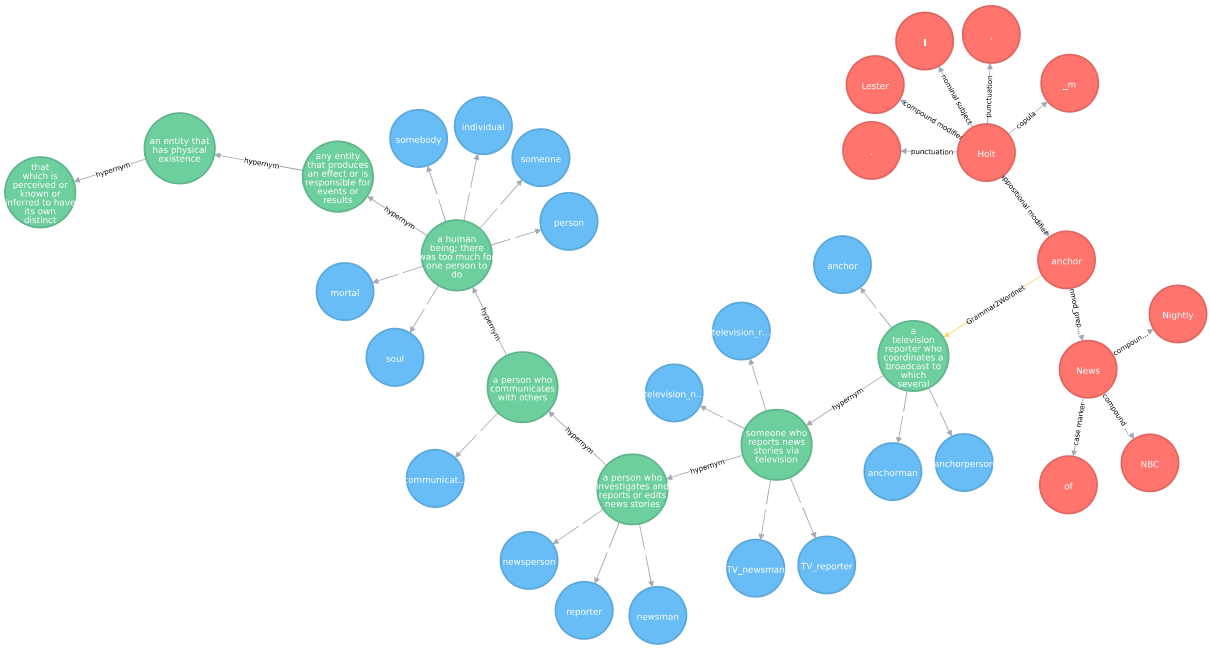

Figure 3.4: Example of connection of parsed dependencies to WordNet (author)

6 Example code in Cypher for recursive extracting all hyponyms of the term *somebody* from consumer feedback using WordNet and Stanford dependency parser in Neo4J database:

```
match (n:Word)-[:Wordnet]->(s1:Synset),
(s1:Synset)-[r:Wordnet*1..20]->(s2:Synset),
(s2:Synset)-[:Wordnet]->(w:WnWord)
where w.Lemma = "somebody" AND all(x in r where x.RelType = "hypernym")
return distinct n.Token, n.POS, count(*)
order by count(*) desc
```

3.5.3 Extraction using topic models

Liu (2015) noted that statistical topic modelling performs two tasks of aspect mining at once: extraction and organisation into groups. Topic models need a large corpus of text documents. **A set of word clusters and topic distribution for each document are the usual outputs of a topic model.** A word cluster is called *topic* and is a probability distribution over words in the corpus; topics are aspects or, more precisely, aspect categories. The topic distribution of a document gives the proportion of each topic in the document. The probabilistic Latent Semantic Analysis (pLSA) (Hofmann 2001) and Latent Dirichlet Allocation (LDA) (Blei, Ng, Jordan 2003) are the most common methods of topic modelling; both methods are unsupervised.

According to a comparison (Lee, Baker, Song, Wetherbe 2010) of topic mining methods, pLSA handles polysemy, but does not provide a probabilistic model at the level of documents and is inappropriate for long documents. The pLSA model overall performs better for documents with a single topic.

On the other hand, LDA provides full generative model with multinomial distribution for words in topics and Dirichlet distribution over topics; it handles long documents or documents with mixed length, but performs better with shorter documents. The advantage of topic models is that **they output adjective together with nouns in a topic and thus results are easy to label**. LDA is incapable to model relations among topics, but offers a good performance for documents with multiple topics.

A different comparison study (Lu, Mei, Zhai 2011) shows that, in general, LDA with the optimal settings works better in text categorisation than PLSA. Despite its better overall performance, LDA has one more parameter to tune than PLSA. When a suboptimal *alpha* hyper-parameter is used, the performance of LDA could become worse than PLSA and even worse than the uni-gram baseline.

As mentioned, LDA is tuned by alpha and *beta* hyper-parameters. For simplification, a high *alpha* value leads to a higher number of significant topic per document, whereas high *beta* leads to higher number of significant words per topic.

In tasks where the latent semantic representation of topics is used alone, such as classification and clustering, optimal performance is achieved when *alpha* is set to be small, e.g. between 0.1 and 0.5. Comparison experiment (Lu, Mei, Zhai 2011) also showed that there is no stable single optimal value of *alpha*. The choice of this parameter also depends on the collection.

As relatively easy to use while offering great performance, topic modelling was used as the aspect mining method of case study in the chapter 4. According to the comparison results, LDA was the main candidate. In the end, its newer modified version called Parallel LDA (Wang, Bai, Stanton, Chen, Chang 2009) was used instead of the standard LDA.

3.6 Literature review of opinion mining in service quality field

Despite the fact that most of consumers' feedback do not correspond with the richness of previously presented quality models, only six studies examine it from the service quality perspective (Palese, Piccoli 2016; Song, Lee, Yoon, Park 2015; Ashton, Evangelopoulos, Prybutok 2015; Duan, Cao, Yu, Levy 2013; Lo 2008; James, Calderon, Cook 2017). Since there have not been any critical literature review on this topic, this thesis presents the only one.

The first work (Lo 2008) does not operate with service quality models. It examines web service quality through user content classification into pre-defined categories. The authors classified a user content and developed a *p-chart control* based on a complaint rate. Service managers can use the chart for service quality real-time diagnosis. Although the manufacturing industry is the typical user of the p-chart, the authors transfered it to the service field. It would not be an issue if the analysed data were from service logs where normality can be assumed. As the research (Hu, Pavlou, Zhang 2006; Hu, Zhang, Pavlou 2009; Hu, Pavlou, Zhang 2014; Zhu, Zhang 2012) proved, the opinion dispersion is rather bi-modal in the case of online reviews which is caused by review motivation drivers. Nevertheless, the proposed chart still provides certain information about a service quality. The correctness of classification varied from 83% to 94% among categories for thousand labelled messages. The other issue relates to the selected level of the text analysis. Even if a content of a message relates to more than one class, the algorithm classifies it to one class only.

Study	(Lo 2008)
Object	Web service quality
Subject	Messages for a web manager on a discussion forum
Quality measurement	Complain rate / p-chart
Dimensionality/Classes	Technical problems, Complaints, Other
Level	Document / one class
Methods	Supervised learning / SVM classification, Keyword extraction / TF-IDF

(Ashton, Evangelopoulos, Prybutok 2015) follows Lo's study and also proposes the p-chart for service quality measurement. The authors extracted factor loadings for quality related factors, discretised them, and counted a number of high-loading customer comments per time frame of interest. They used Latent Semantic Analysis (LSA) for classification of topics. The discretising solved Lo's issue with classifying a document in multiple classes. The representativeness of service feedback is higher in this case because all consumers have to reply to the question before they cancel the service subscription. However, the motivation to leave a meaningful response might not be high. The main issue is related to service quality measurement. Compared to the traditional itemised scales, consumer can freely express his opinion. Although selected service attributes do not reflect quality models from the literature, they can be regarded

as quality dimensions, because they are significant from consumers' point of view. Unfortunately, quality is surveyed only by cancelling consumers which might not be the best sample for generalisation of a service quality.

Study	(Ashton, Evangelopoulos, Prybutok 2015)
Object	Online retail service quality
Subject	Open-ended question from survey after the service cancellation
Quality measurement	Quality attributes / p-chart
Dimensionality/Classes	Labelled clusters (The service, Wait time too long for product to arrive (mail), Delays in shipping and receiving, Product availability waiting time is too long, The customer wants a more extensive selection, Received damaged products)
Level	Document / multiple classes
Methods	Unsupervised learning / LSA classification

The next study comes from the hospitality sector. (Duan, Cao, Yu, Levy 2013) mapped sentences from a review to the SERVQUAL quality dimensions and evaluated their sentiment. They trained a model to classify sentences into quality dimensions. The model uses for training a set of keywords which were manually selected by reading a sample of 500 reviews from the corpus (N = 64 806). As keywords, they used only nouns that should represent service dimension categories. They employed Cornell movie-review dataset for sentiment polarity training. It was the first study that focused on the quality dimensions from the service research literature. Unfortunately, it is not reliable to classify the whole documents into quality dimensions, because they may contain more than one mention of service aspects.

Study	(Duan, Cao, Yu, Levy 2013)
Object	Hotel service quality
Subject	Online reviews
Quality measurement	Service performance
Dimensionality/Classes	SERVQUAL (tangibles, reliability, assurance, responsiveness, empathy)
Level	Sentence / one class
Methods	Supervised learning / Naïve Bayes classification for sentiment and SERVQUAL dimensions

The study by (Song, Lee, Yoon, Park 2015) is an excellent example of work that incorporates service quality models in text analysis. The authors tried to capture service quality as a difference between service performance and expectations. Performance was measured by weighted service aspect's sentiment polarity from a review, expectation as a total aspect frequency across all reviews. The authors see service quality as a set of hierarchical relationships between service aspects**.** In case study they performed, the service aspects' hierarchy is identified by domain experts. Hence, they captured quality from producers' point of view. The expectation side of the quality equitation is the main issue because quality expectations are a

strongly individual concept. “Consumer expectations and perceptions change constantly and the quality assessment, expectations and perceptions vary according to the meaning that consumers give to their experience.” (Mauri, Minazzi, Muccio 2013; Schembri, Sandberg 2011) Aspect mentions from previously read reviews act only as one of many factors that influence individual’s expectation for that aspect. The measure of total frequency across all reviews is also discussable.

Nevertheless, the equitation for weighting service aspects mentioned in a review text based on an overall review is the main contribution of the study. The weighting technique corresponds with the assumption of (Hu, Zhang, Pavlou 2009) and (Zhang, Yu, Li, Lin 2016) that consumers mention only the aspects which did not meet or exceed their expectations. The aspects with positive or negative sentiment are weighted according to the overall rating score. The authors assume that the importance of aspects with negative sentiment is higher for overall negative ratings and vice versa. The aspects with positive sentiment have higher importance for overall positive ratings. Sentiments were classified only in the sense of polarity, as positive or negative. The opinion mining methods also offer intensity classification which might make the quality diagnosis from this study more precise.

Study	(Song, Lee, Yoon, Park 2015)
Object	Mobile navigation service quality
Subject	Online reviews
Quality measurement	Expectation confirmation / P (weighted aspect sentiment polarity) - E (total relative aspect frequency)
Dimensionality/Classes	Service aspect hierarchy identified by domain experts
Level	Word
Methods	Aspect extraction based on dictionary and part of speech, sentiment classification based on dictionary

The study from Louisiana State University (Palese, Piccoli 2016), as well as (Duan, Cao, Yu, Levy 2013; Song, Lee, Yoon, Park 2015), also uses SERVQUAL dimensions to capture service quality. The authors employed topic modelling with a weak supervision by adding seed words for each service quality dimension. Sentences from online reviews were the subject of classification. The developed model was successfully validated by graduate students with 93.3% accuracy. The work was framed by (Lu, Ott, Cardie, Tsou 2011; Grün, Hornik 2009) approach of multi-aspect sentiment analysis with topic models. The authors use sentiment expressed on a scale. The scale is, unfortunately, on the review level, so the dimensions cannot be assessed appropriately. The authors used Gibbs sampling method (Griffiths, Steyvers 2004) for topic mining, which can be perceived as one of the most advanced methods for topic mining. Unlike (Song, Lee, Yoon, Park 2015), this study does not deal with weighting service dimensions on the general nor individual level.

Study	(Palese, Piccoli 2016)
Object	Online retail service quality
Subject	Online reviews
Quality measurement	Service performance
Dimensionality/Classes	SERVQUAL (tangibles, reliability, assurance, responsiveness, empathy)
Level	Sentence (aspect), Document (sentiment)
Methods	Multi-aspect sentiment analysis Topic analysis / Gibbs sampling and Latent Dirichlet Allocation with seed words

The recent study (James, Calderon, Cook 2017) that examines service quality and employs opinion mining methods comes from the healthcare sector. The authors used Latent Dirichlet Allocation (LDA) technique to extract topics from online reviews. Extracted topics verified previously proposed dimensionality of healthcare service quality (López, Detz, Ratanawongsa, Sarkar 2012) and corresponded to the well-established quality model that distinguishes between technical and functional quality (Grönroos 1984). The authors tried to explore the relationship between overall perceived quality on a rating scale and quality dimensions gathered from a text. Unfortunately, they composed overall perceived quality from four different quality items without proper weighting or validating their stability, so the observed relationship is discussable. The used methods, especially the level of analysis, do not reflect the challenges that service quality measurement from textual feedback is currently facing. The sentiment of reviews is calculated with dictionary approach, which is not the most advanced method as discussed in previous chapters.

Study	(James, Calderon, Cook 2017)
Object	Healthcare service quality
Subject	Online reviews
Quality measurement	Quality attributes
Dimensionality/Classes	systems, interpersonal, and technical
Level	Document
Methods	Topic analysis / Latent Dirichlet Allocation aspect mining, Dictionary base sentiment classification

3.7 Chapter summary and discussion

The literature review in this chapter focused on opinion mining concepts, methods and service quality improvement potential. It mapped opinion mining concepts on service and proposed a joint service model in part 3.2. Service is represented as an opinion target. Service components and attributes are opinion target aspects. Service consumers, as opinion holders, express sentimental attitude towards these aspects. Service quality is a more complex construct and does not directly imply from service attributes or components sentiment. It means that even if the research question *Q3* is answered, it is necessary to explore relationship between aspects sentiment and service quality, which refers to the question *Q2*.

The part 3.3 answers the research question *Q3a* by reviewing opinion mining techniques for extraction of consumers' opinions and classification of sentiment and emotions. The analysis of consumers' feedback can be undertaken on different levels. On the highest level, it is a whole message, review or email. On the lower levels, it is a sentence or a single service aspect. The lower levels provide higher fidelity of service image perceived by a consumer, but are more difficult to extract. Service components or attributes have to be matched with sentiment target aspects from consumer's feedback. Sentiment is usually classified based on the intensity and orientation of words derived from a lexicon that have a certain relationship to the sentiment target aspect or according to supervised learned model predictions.

Emotion measuring and classification is the content of part 3.4. Emotion research is not consensual in emotion definition and taxonomy. An individual can pass through many different affective states, not all the states can be marked as emotion. Emotions are hard to measure because of the intrinsic nature of feeling. Self-reporting is the most frequent method of emotion measurement. Several models of emotion have been issued, the most influential ones are from Plutchik and Scherer; these models are arranged in a circumplex. This thesis adopted eight modal emotion model for problem simplification. Self-reported emotions can be captured from a text using the similar methods as sentiment. The literature review of emotion measurement and its capturing from text answers the research question *Q3b*.

Service aspects extraction is described in part 3.5. Aspects could be extracted by finding frequent nouns and noun phrases, exploiting syntactic relations, using supervised learning or using topic models. Topic analysis and syntactic relation exploiting showed notably great results. Latent Dirichlet Allocation (LDA) and Latent Semantic Allocation (LSA) are the most frequent methods for topic mining. The advantage of topic models is that they not only extract significant aspects, but also form them into groups. Extraction techniques based on syntactic relations struggle to perform meaningful aspect groups, but they offer detailed analysis of language and they are able to capture nuances in expressions such as negation or aspect conjunction.

In the part 3.6, the first critical literature review of the papers that use opinion mining techniques for service quality diagnosis was conducted. Four studies used online reviews as a source of feedback (Duan, Cao, Yu, Levy 2013; Song, Lee, Yoon, Park 2015; Palese, Piccoli 2016; James, Calderon, Cook 2017). The two others used messages for service managers and cancellation surveys (Lo 2008; Ashton, Evangelopoulos, Prybutok 2015). Two studies use charting of communication service quality to service managers (Lo 2008; Ashton, Evangelopoulos, Prybutok 2015). For answering the research question *Q2a*, it is necessary to evaluate their approach to service quality.

Three studies conducted the feedback analyses on a document level (Lo 2008; Ashton, Evangelopoulos, Prybutok 2015; James, Calderon, Cook 2017), two on a sentence level (Duan, Cao, Yu, Levy 2013; Palese, Piccoli 2016) and one on the word level (Song, Lee, Yoon, Park 2015). The studies do not show that higher level leads to an information loss. The level of analysis might not be an issue regarding aspect extraction, it could be undergone with multiple classes, the problem is caused mainly by sentiment analysis which on the higher level might result in one sentiment for more than one service aspect.

The half of the studies used topic mining for aspect extraction. Two of them used LDA (Palese, Piccoli 2016; James, Calderon, Cook 2017), one LSA (Ashton, Evangelopoulos, Prybutok 2015). The two other studies (Lo 2008; Song, Lee, Yoon, Park 2015) extracted sentiment using on the term frequency (TF) combined with part of speech (POS) and TF-IDF. The last study (Duan, Cao, Yu, Levy 2013) did not extract aspects at all. The most advanced approach presented (Palese, Piccoli 2016) by using Gibbs sampling method (Griffiths, Steyvers 2004; Casella, George 1992). However, even weak supervised method needs to be discussed with the results of unsupervised topic modelling after a certain time or in a different context. According to (Liu 2015), life long learning could be a good approach for sustainable aspect extraction. Aspect extraction based on POS by (Song, Lee, Yoon, Park 2015) is also interesting, but there is a more advanced, complex and precise approach to syntactical relation.

Sentiment classification used three of the studies. (Duan, Cao, Yu, Levy 2013) used Cornell movie-review dataset for sentiment polarity training with Naïve Bayes learner. (Song, Lee, Yoon, Park 2015; James, Calderon, Cook 2017) used dictionary approach. The other studies (Lo 2008; Ashton, Evangelopoulos, Prybutok 2015) did not classify the sentiment, they employed only frequency measurement of document categories. (Palese, Piccoli 2016) used rating scale instead of sentiment classification of text.

Generally speaking, several approaches are available for sentiment classification (*Q3a*). It cannot be claimed that one approach is better than the other. The analysis on the **higher level causes certain information loss but it is more reliable**. For instance, a multi-aspect analysis outputs two aspects in one document and a sentiment analysis of this document returns neutral sentiment. Does it mean that the both aspects are perceived neutrally, or that one aspect is perceived as strongly positive and the other one as strongly negative? **Lower level of analysis**, on the other hand, **may suffer from sentiment or class**

prediction error caused by a low number of document features or an error caused by sentiment classification based on a few words that can be found in the sentiment or emotion lexicon.

The reviewed papers do not prove that lexicon based approach performs worse than machine learning approach in this particular task. Nevertheless, the lexicon approach needs language rules of a good quality. These two approaches will be examined more in depth in the following case studies. Especially the chapter 5 contains comparison of supervised learning and lexicon approach in an emotion classification task.

Regarding the research question *Q2a*, three of the studies tried to **map user content to SERVQUAL dimensions** (Palese, Piccoli 2016; Song, Lee, Yoon, Park 2015; Duan, Cao, Yu, Levy 2013). The SERVQUAL dimensions have been validated in many studies, but they might not fit in all contexts, especially in e-service area where the authors of SERVQUAL observed a new dimensionality. The dimensionality can be different in other service branches such as e-retailing or video-streaming services. In contrast to the previously mentioned studies, (Ashton, Evangelopoulos, Prybutok 2015; James, Calderon, Cook 2017) used **aspect groups that appear naturally in the feedback.** Although this approach did not explain how extracted aspect groups explain service quality, it is not necessarily wrong. According to (Yang, Jun, Peterson 2004) that used topic analysis on service reviews as a method for service quality dimension determination, the identified topics needs to be examined according to the service literature about whether can be seen as service dimensions in the particular context or not.

The reviewed studies, as it also implies the service quality literature, showed controversies about expectation measurement. (Song, Lee, Yoon, Park 2015) captured the expectation separately from perceived service performance. According to their work, e**xpectations are expressed as an overall frequency of aspect mentions**. This **approach is inappropriate** because: (a) The expectations are an individual concept and cannot be calculated among all consumers. (b) The frequency of aspect mentions relates only to the importance of aspect and frequent disconfirmation of expectations. (c) Expressed attitude towards service quality is already the result of the function of expectation and perceived service performance.

The idea of weighting service aspects according to the overall rating from (Song, Lee, Yoon, Park 2015) is notable, even thought its measuring of expectation is discussable.

This literature review examined only e-service quality studies what can be considered as its limitation (Lo 2008; Ashton, Evangelopoulos, Prybutok 2015; Song, Lee, Yoon, Park 2015; Palese, Piccoli 2016). The other two studies explored hospitality and healthcare service sectors, but they faced the same challenges of service quality diagnosis from consumers' feedback as the studies from e-service area.

The review brought the light on the main research question *Q1* by covering both sub-questions *Q2a* and *Q3a* in the sentiment analysis branch. Appendix I summarizes the six reviewed papers in one table.

Technology nowadays influences the way service providers listen to consumers' expectations when planning appropriate service delivery. Customer feedback is already available in a big volume, especially for services delivered online. The questions of how consumers' perceived quality and expectations can be collected were partially answered, but more studies need to be undertaken to explore the issue more in depth. The reviewed studies that emphasise presentation of service quality findings lack proper service quality research background and vice versa. A study that combines both approaches and validates methods from this chapter in the new e-service branch is presented in the chapter 4.

Especially the research questions *Q2b* and *Q3b* that explore relationship of service quality with emotions remain unclear. The study chapter 5 brings the light on these research questions and fills the gap from this literature review.

Chapter 4: Case study of online banking service

4.1 Introduction

This chapter presents a practical application of opinion mining on an online banking service. The chapter is based on study (Vencovsky, Bruckner, Sperkova 2016) that was presented on the 3th European Conference on Social Media. Unlike the paper, this chapter contains more supporting arguments and extended discussion. Regarding the research question *Q3a*, the study uses topic mining for aspect extraction and sentiment classification on the document level. Extracted aspect groups are discussed in relation to service model from the chapter 3.2, service quality dimensions from the chapter 2.3 and research question *Q2a*.

The study, similarly as the papers (James, Calderon, Cook 2017; Palese, Piccoli 2016) reviewed in chapter 3.6, uses Latent Dirichlet Allocation (LDA) and Gibbs sampling for topic analysis as methods. It explores consumers' feedback on the service in the way it naturally appears in the data. The study employs Cornell movie review database and Naïve Bayes learner for sentiment analysis.

The goal of the study is to enable continual service quality monitoring using managerial dashboard which is similar to the charting in (Ashton, Evangelopoulos, Prybutok 2015; Lo 2008).

Contributions of the study are to:

(a) explore dimensionality of online banking service in relation to research question *Q2a*;

(b) validate of methods related to research question *Q3a*:

  (i) Parallel LDA and Gibbs sampling for extracting service quality dimensions;

  (ii) SMO algorithm for classification of service feedback;

  (iii) Naïve Bayes learner for service dimension sentiment classification;

(c) validate a dashboard tool for service quality monitoring as the managerial implication of how service quality diagnosis can be employed for service quality improvement.

4.2 Data description

Data for the analysis consist of a set of 2430 full text contributions of 1564 users in 635 locations of 13 states from an official internet forum of a certain undisclosed bank. Every full text contribution item encompasses following data attributes: *title*, *full text contribution*, *author nickname*, *period of registration* of the user in the forum, *reply title*, *reply full text*, *time of the post*, *province*, *location*.

4.3 Method

The method of the contribution dataset analysis consists of following steps:

- Sentiment analysis of contributions and removal of sentiment related terms;
- Topic extraction from the dataset;
- Categorisation of contributions by topic;
- Sentiment analysis of contributions within topics;
- Depiction of the topic sentiment on dashboard.

The initial step was to consider the right level of analysis. Due to the intention of a use of analysis results in a managerial dashboard, the level of document (review) was selected. A dashboard would allow to browse results of analysis on the lower levels and to put review as the focal unit in the same time was not found on the market.

The next goal was to categorise topics of reviews according to their full text content. Among the reviewed methods of topic discovery, LDA method (Latent Dirichlet Allocation) (Blei, Ng, Jordan 2003) or its parallel version (Wang, Bai, Stanton, Chen, Chang 2009) was considered. This method takes the corpus (set of all quality reviews) as an input and results in a generative probabilistic model that contains latent topics characterized by specific distribution of words in documents (individual reviews).

Since the topics sentiment needed to by analysed, sentiment expressions were not be used for topic identification. Thus sentiment words needed to be purged from review full texts before a topic extraction. Without sentiment words removal, topics within a model would tend to keep sentiment from training data. Identification of sentiment related words was done by sentiment coding on this particular dataset.

Regarding this dataset, it was assumed that all contributions are related to e-banking, due to the purpose of the bank's official discussion forum where public reviews and communication take place. If there were other topics present in the data, analysis would identify them.

After an initial topic analysis on the dataset, a known topic assignment was used for future classification. For this purpose, supervised learning method using SMO (sequential minimal optimization) algorithm (Keerthi, Shevade, Bhattacharyya, Murthy 2001; Platt 1998) was employed.

Then, sentiment for each topic was determined using Cornell movie review database and Naïve Bayes learner. Although overall sentiment score per topic is not sufficient metric due to the theory of U shape distribution, topic sentiment score was calculated as both arithmetic mean and median.

To conduct deeper analysis of posts, their sentiment and relation to service quality, all results of analysis were loaded on a dashboard. As a dashboard, Kibana, an extension of the Elasticsearch, was used. An advantage of the selected dashboard solution is that Elasticsearch allows to filter and browse analysed reviews and to aggregate review sentiment and rating from the search results.

4.4 Results

4.4.1 Topic analysis

The Parallel Latent Dirichlet Analysis (PLDA) (Wang, Bai, Stanton, Chen, Chang 2009; Newman, Welling 2009) was employed to analyse the review topics. The preprocessing step consisted of punctuation erasing and purging of terms with less than 3 characters, and it also consisted of filtering numbers and stopwords, sentiment and frequency filtering. Terms were converted to lowercase and to their root forms by applying Porter algorithm from Snowball stemmer library. Sentiment related words were removed from review texts in the same stage, it affected thirty-four unique terms in total. In order to filter terms by their frequency, relative term frequency (R-TF), absolute term frequency (A-TF) and inverse document frequencies (IDF) were counted first. Terms, that occurred less than three times in the whole corpus, were filtered out if average R-TF value was below the threshold of 0.03 points and IDF value was 0.6 points.

Because PLDA algorithm needs number of topics as an input, the same approach as described in (Antons, Kleer, Salge 2015) was used. The approach consists of finding the optimum log-likelihood value. Log-likelihood parameter describes cross validation of extracted topics. As PLDA returns the parameter after analysis, optimization loop was employed to find the optimal count of topics. The loop was set to find the optimum between one to fifty topics. Unfortunately, the results showed the best log-likelihood value outputted for one topic. Even different optimisation of hyper-parameters did not bring any meaningful result.

Since the results of optimisation wre unacceptable, graphical representation of PLDA outputs was considered to explore the issue more in depth. Figure 4.2 shows the distribution of thirty different variations of the topic count. The presented graph is a result provided by the OpenOrd algorithm implemented in the Gephi graph tool. Each topic is represented by one node and is connected through edges with ten different keywords that were identified by PLDA algorithm. An edge is thick according to a weight of a keyword in a certain topic. The more topics are connected to keywords, the more density the graph shows. The keywords that are outputted often together create a network of nodes with larger density – a topic cloud. As the result of the graph analysis, five topic clouds wre identified. According to presented keywords, clouds wre labelled as (1) *Cards and Access*, (2) *Account*, (3) *Support*, (4) *Website and Features* and (5) *Mobile Application*.

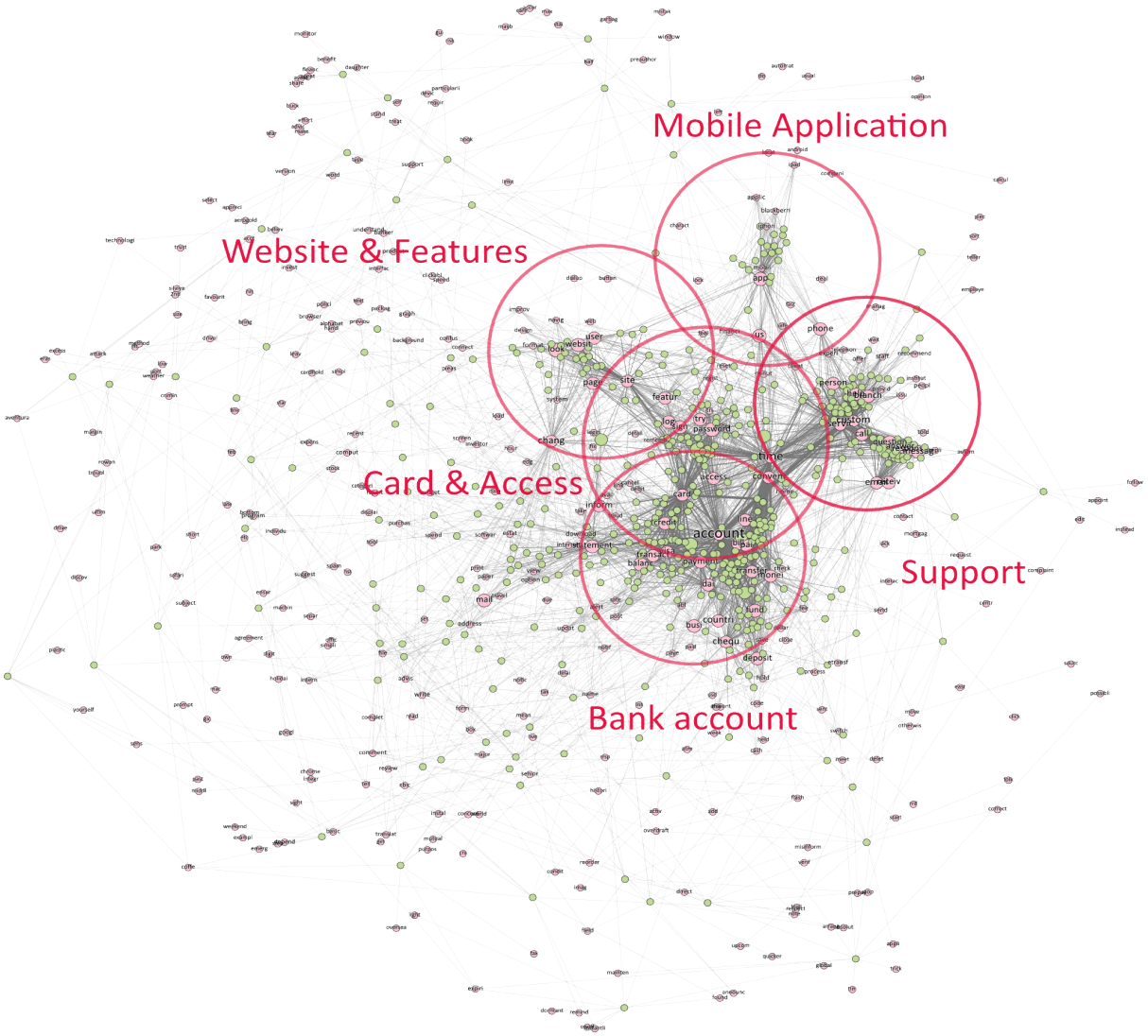

Figure 4.1: Topic-Keyword graph (Vencovsky, Bruckner, Sperkova 2016)

After the study (Vencovsky, Bruckner, Sperkova 2016) had been published, another examination of topic count was undertaken - this time using the Gibbs sampling method (Casella, George 1992) from the R package called ldatuning. The result of the topic count analysis returned similar results as the graph analysis. According to the indexes *CaoJuan2009* and *Deveaud2014*, the optimal number of topics was also five. The results are presented in figure 4.2.

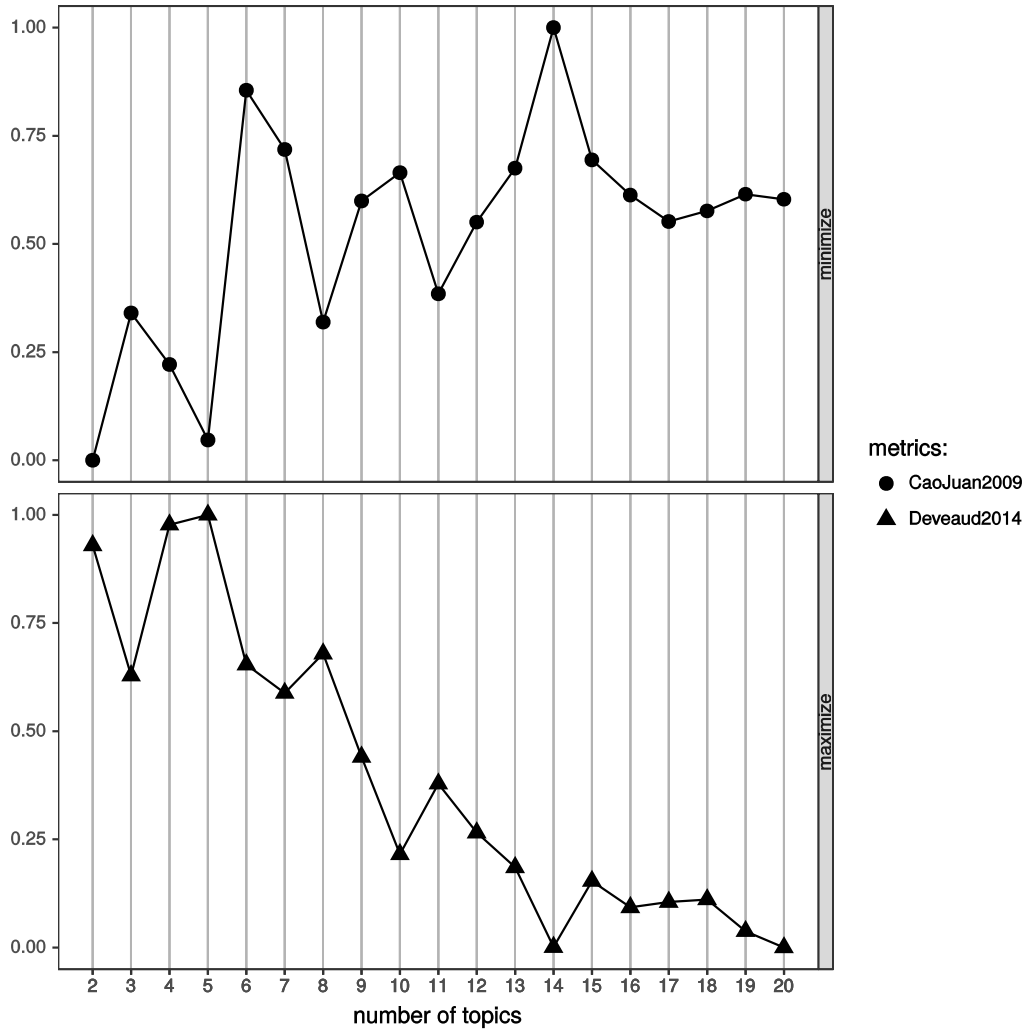

Figure 4.2: Number of topic optimization using Gibbs sampling method (author)

To see whether any other set of topics offers meaningful results, keyword sets of near topic count was investigated. The most meaningful set was identified with seven topics. Figure 4.3 presents the topic-keyword graph based on the seven topics. Unlike previously labelled topics, in the new set, the *Card and Access* group was separated in two new groups. *Topic 0* from the figure represents *credit card* as a significant group of aspects that relate to credit card and services such as paying the bills. *Topic 2* was also strongly connected to the *card* keyword, but represents rather an *access* to the banking service via card details. Furthermore, another division of the group was taken in case of the *Support. Topic 3* represents service support that is based more on a *personal basis* and voice communication, it is identified by keywords such as *call*, *person* and *branch*. *Topic 4*, on the other hand, represents *written communication* mediated by the online support system and emails. The other labels are *website and features* for *topic 1*, which are group of aspects related to an website and its attributes, *account* for *topic 5* and *banking application* for *topic 6*.

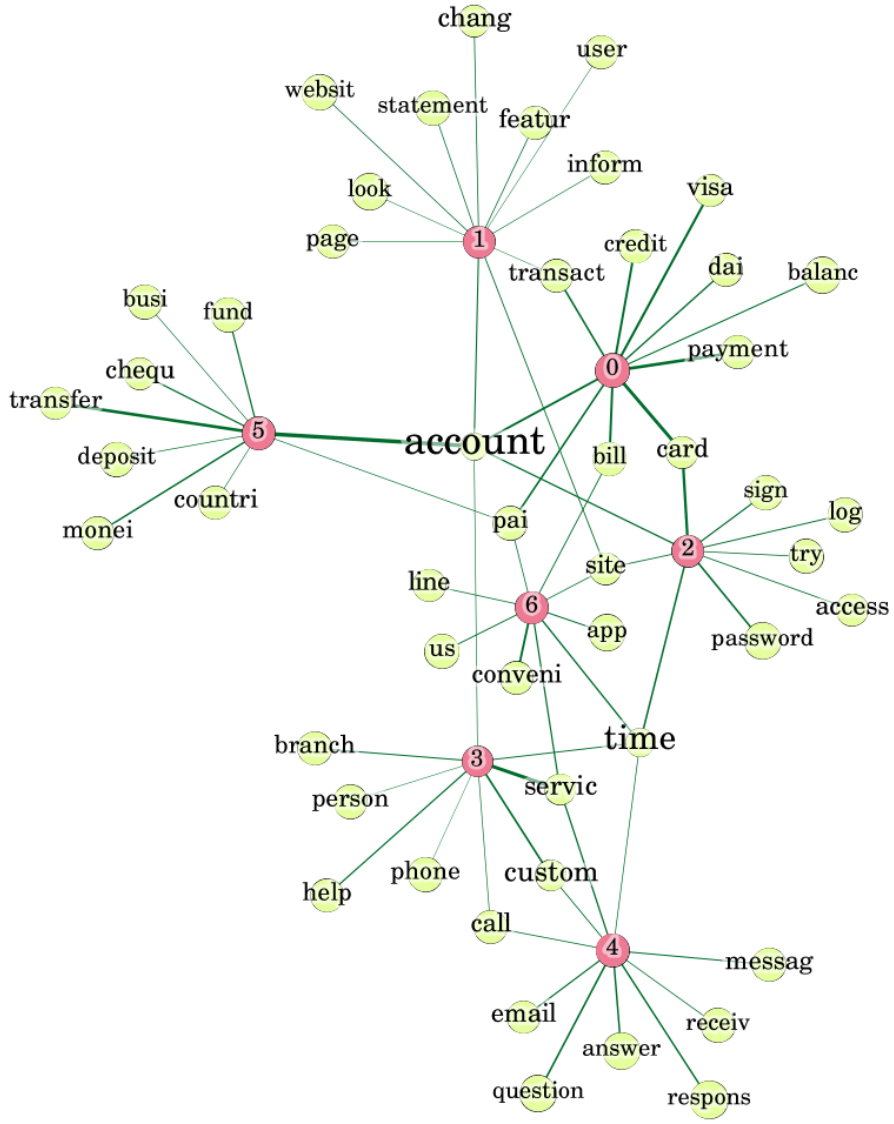

Figure 4.3: Topic-Keyword graph: 7 topics (Vencovsky, Bruckner, Sperkova 2016)

Beside topic keywords, outputs of PLDA are also document topic values. A document topic value is a value that represents how a document relates to a particular topic on a scale of 0 to 1. Higher value means that the document is more connected with a topic. Figure 4.4 shows distributions of topic value for each combination of topics in a chart. The chart implies that topics are rather distinctive. It indicates that, globally, one review is likely about one topic. Topic frequencies are presented in figure 4.5.

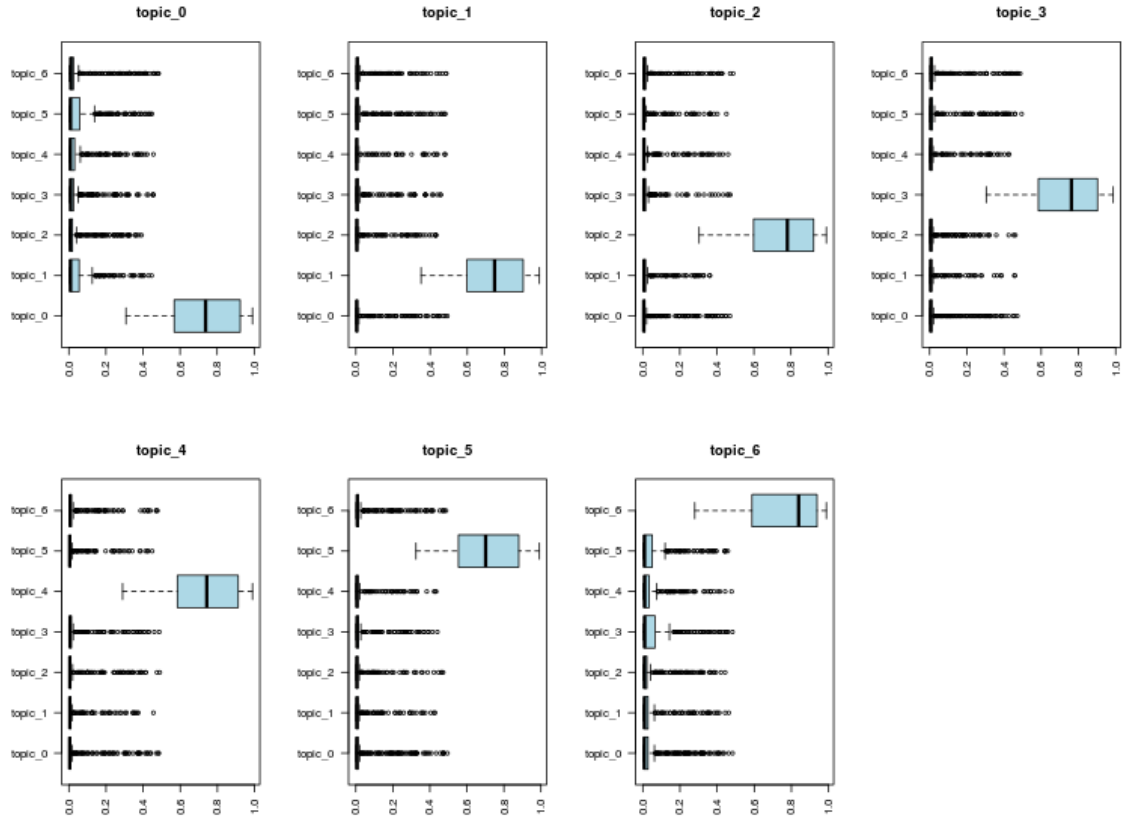

Figure 4.4: Topic value distribution (Vencovsky, Bruckner, Sperkova 2016)

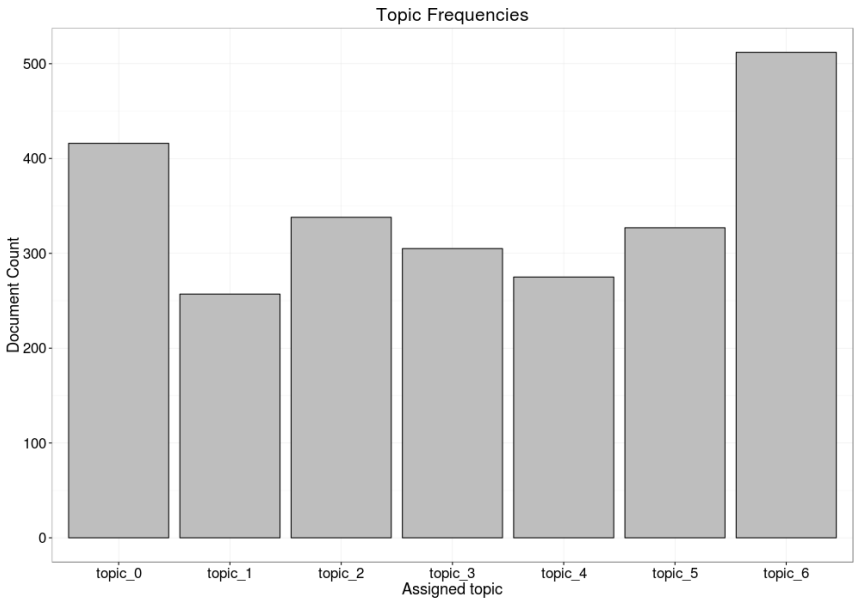

Figure 4.5: Topics distribution (Vencovsky, Bruckner, Sperkova 2016)

Topic code	Topic label
topic_0	credit card
topic_1	website and features
topic_2	access (to the banking website)
topic_3	service support - personal basis
topic_4	service support - written communication
topic_5	bank account
topic_6	banking application

Table 4.1: Topic labels and numbers map

4.4.2 Sentiment analysis

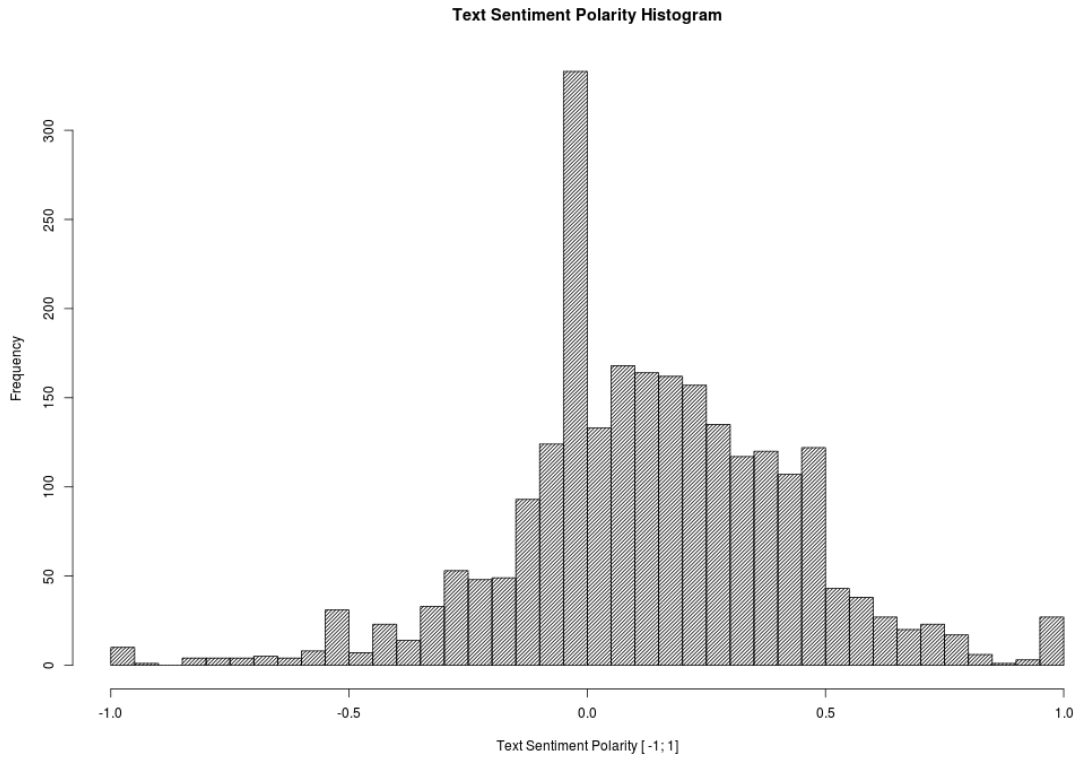

Figure 4.6: Histogram of review sentiment (Vencovsky, Bruckner, Sperkova 2016)

In order to count review text sentiment, NLTK Python library (Bird, Klein, Loper 2009) and its corpus *Sentiment Polarity Dataset Version 2.0* (Pang, Lee 2004) was employed. Model was trained using NLTK NaiveBayes. Sentiment polarity was counted for each review body. Classifier outputs sentiment as a combination of sentiment polarity and intensity with values from *-1* to *1*. Figure 4.6 shows the sentiment polarity histogram. The plot indicates that most of the contributions are neutral. The positive reviews prevail over the negative ones in this case.

Classified sentiment of reviews was applied to identified topics. Each review is represented by one topic and a sentiment value. As per the method chapter, generalisation of sentiment or rating is not reliable, but works as a good indicator. Table 4.2 shows average values of topic sentiment polarity and overall rating. *Access* was considered as the most problematic part of service. It had the lowest average overall rating and also the most negative polarity. Average values also implied that the most positive opinions wre linked to *banking application*.

Topic	**Topic frequency**	**Average sentiment polarity**	**Average overall rating**
credit card	416	0.094	3.154
website and features	257	0.088	2.669
access (to the banking website)	338	-0.027	2.080
service support - personal basis	305	0.226	3.662
service support - written communication	275	0.092	2.476
bank account	327	0.116	3.254
banking application	512	0.300	4.547

Table 4.2: Average sentiment statistics per topic

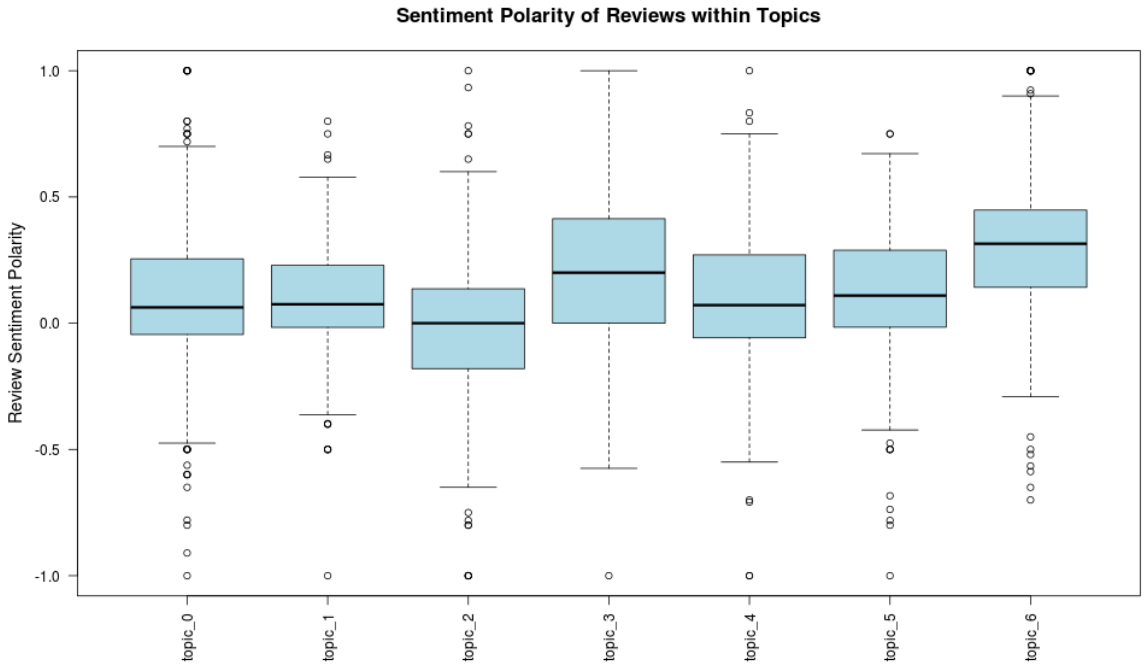

Figure 4.7: Sentiment polarity of reviews within topics (Vencovsky, Bruckner, Sperkova 2016)

As it was presented in box plot with sentiment values 4.7, almost a half of contributions that were connected to a group of aspects labelled as service *access* had a negative sentiment polarity. Previously mentioned aspects regarding *banking application* were seen much more positively than the rest of groups. Reviews that contained aspects related to *personal* and *written support* indicated the highest variance in the sense of the sentiment polarity. The other topics had similar sentiment value distribution. User rating presented in table 4.3 supports these statements.

Overall rating	credit card	website and features	access to the banking service	service support - personal basis	service support - written communication	bank account	banking application
5.0	118	59	31	172	67	105	391
4.0	80	34	36	25	24	62	65
3.0	70	27	39	14	21	50	23
2.0	44	37	55	21	24	31	11
1.0	104	100	177	73	139	79	22

Table 4.3: Overall user rating of service

4.4.3 Future categorisation of reviews

Each review that service customers will create on a bank site in the future should be linked to a corresponding topic. For that purpose, SVM classifier (Keerthi, Shevade, Bhattacharyya, Murthy 2001; Platt 1998) was employed. The classifier was already used in service quality context by (Lo 2008). Polynomial kernel was used to train a model on 1944 reviews, that made 80 percent of the data set. Remaining 20 percent were used to evaluate model accuracy. Resulting average F-measure was 0.708 points with Cohen's kappa of 0.655. F-measure is a measure based on precision and recall (McCallum, Nigam 1998). Table 4.4 shows more detailed information about accuracy results. According to the results, it is possible to claim that the model is suitable for operative service quality diagnosis tasks. For keeping the model valid, it was proposed to repeat topic analysis and classification learning process in a half-year period, because new topics may emerge.

Topic	True Positives	False Positives	True Negatives	False Negatives	Precision	Sensitivity	F-measure
credit card	50	17	386	33	0.746	0.602	0.667
website and features	34	15	420	17	0.694	0.667	0.680
access (banking website)	55	16	401	14	0.775	0.797	0.786
service support - personal basis	39	25	400	22	0.609	0.639	0.624
service support - written communication	42	11	420	13	0.792	0.764	0.778
bank account	44	19	402	21	0.698	0.677	0.688
banking application	80	39	345	22	0.672	0.784	0.724

Table 4.4: SVM accuracy statistics

4.4.4 Depiction of the topic sentiment in a dashboard

To provide interactive dashboard that is usable on daily basis, the enriched review data were indexed and loaded into Elasticsearch server first. Predefined English analyser that includes English stop word filter and English stemmer were used for indexing. Then, *Customer review dashboard* was created in Kibana.

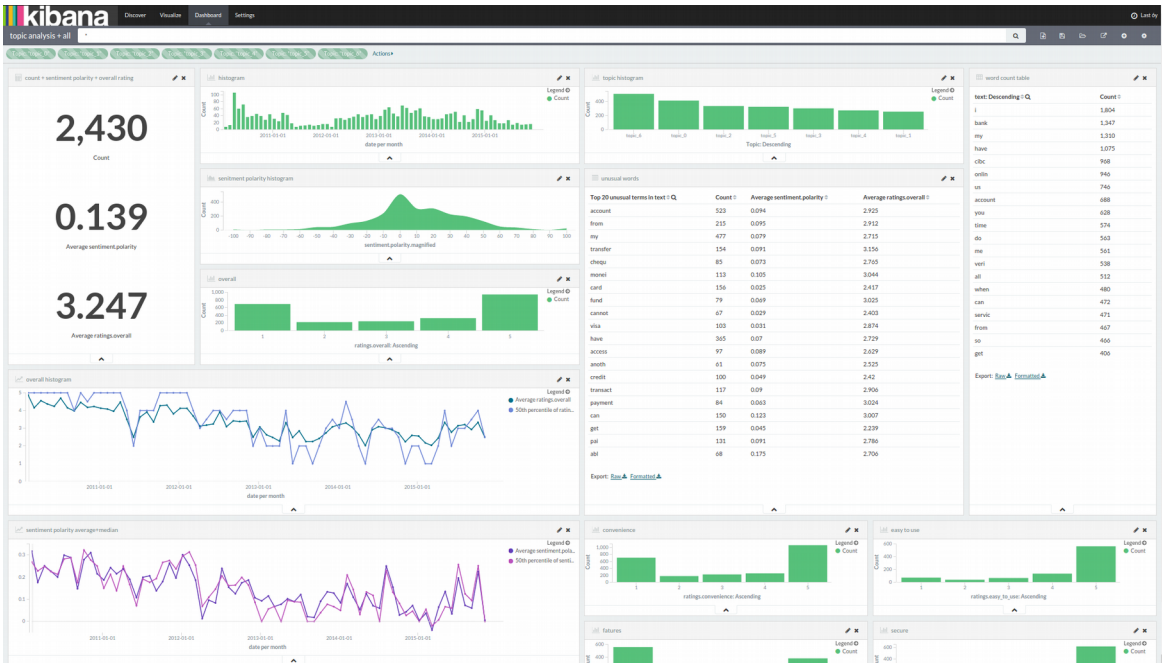

Figure 4.8: Customer review dashboard overview (Vencovsky, Bruckner, Sperkova 2016)

Customer review dashboard is presented in figure 4.8. It consisted of following containers. The container on the top left corner shows overall metrics such as number of contributions, average sentiment polarity

and average rating. The line chart on the top represents number of posts per month on a timeline. Below the chart, there is distribution of sentiment and distribution of rating on the whole dataset. In the bottom on left site of the dashboard, two line charts show average rating of contributions per month on a timeline and average sentiment polarity of contributions per month on a timeline. The right side of the dashboard shows distribution between topics in the bar chart on the top. Below, there is a list of unusual terms which are determined by Elasticsearch "uncommonly common" algorithm (Gormley, Tong 2015) and the most frequent words among the reviews together with rating and sentiment aggregation data.

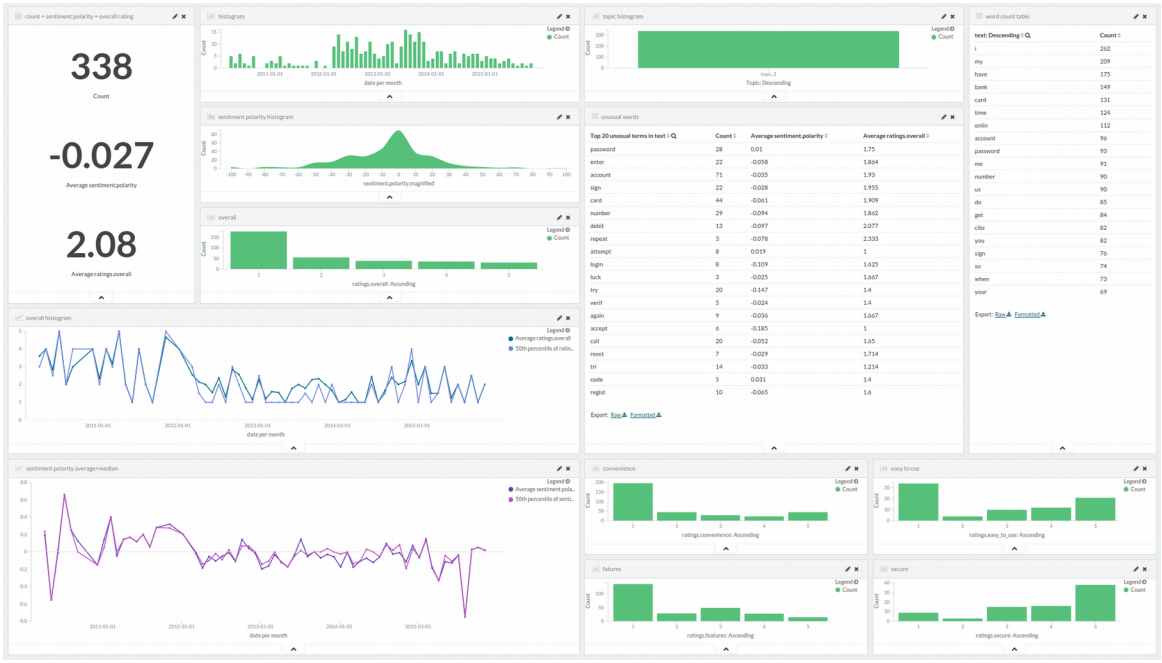

Figure 4.9: Review topic filter on dashboard (Vencovsky, Bruckner, Sperkova 2016)

User of the dashboard can change the data upon which the aggregation is counted by choosing different filter options. Dashboard in figure 4.9 shows the same components but with different filters applied. In the case of figure 4.9, it is *topic_2* (*access*) only. Figure 4.10 displays data filtered to low rated reviews only. The last example, figure 4.11, limit data sub-set to a specific time period.

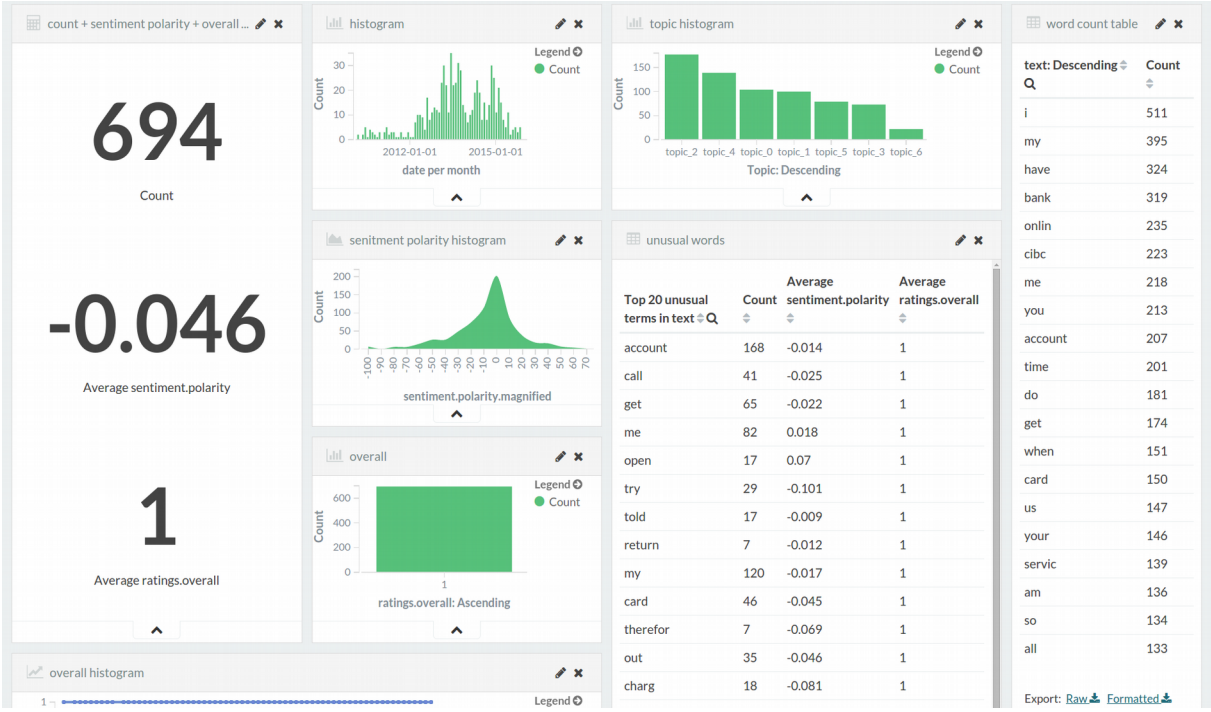

Figure 4.10: Low rating dashboard filter (Vencovsky, Bruckner, Sperkova 2016)

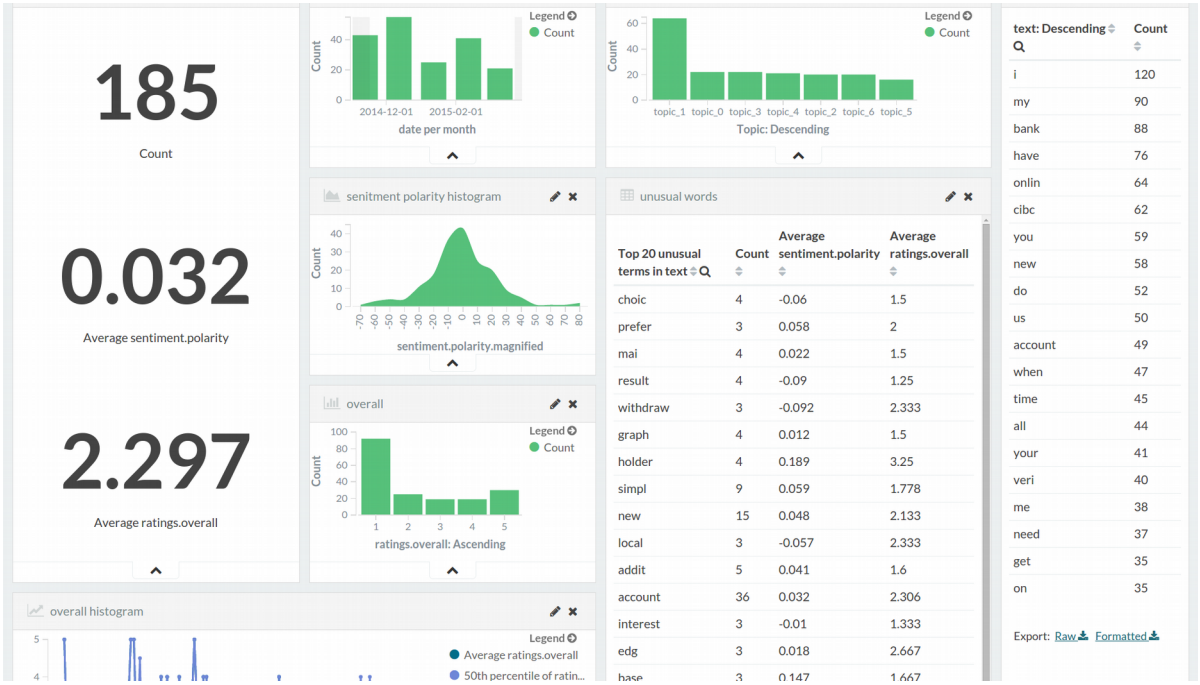

Figure 4.11: Time period filter (Vencovsky, Bruckner, Sperkova 2016)

As the most interesting chart for service quality diagnosis, the average sentiment polarity in time (in figure 4.12) was selected. Dashboard chart allows user to select a specific time period. Selection, for example a

period with decreasing sentiment, will change the data subset to a new one and render all graphs and tables according to the new subsets. And thus, user is able to identify group of aspects that are linked with negative reviews. The real cause of consumers' dissatisfaction is then more likely to be identified.

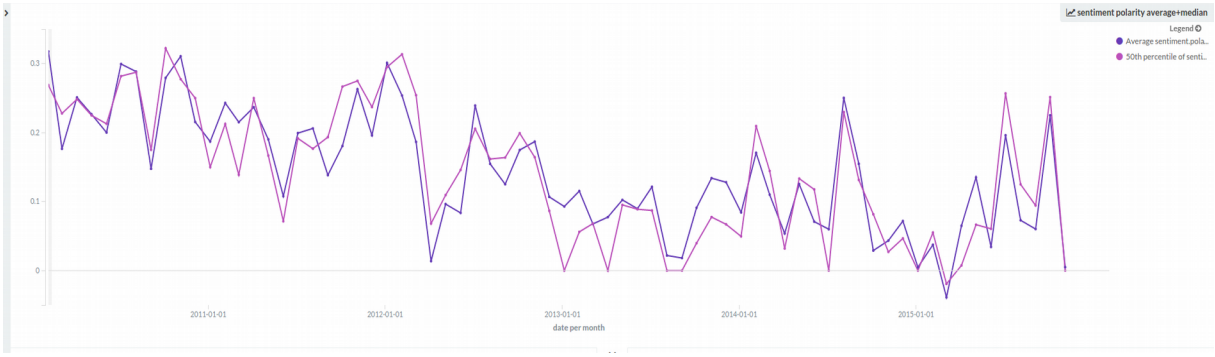

Figure 4.12: Aspect sentiment on timeline (Vencovsky, Bruckner, Sperkova 2016)

Sentiment distribution chart can be used for service quality analysis. For example, chart in figure 4.13 indicates unusually larger volume of reviews with sentiment higher than 0.05. The group of aspects that are linked with *topic_6* concerns *banking application*.

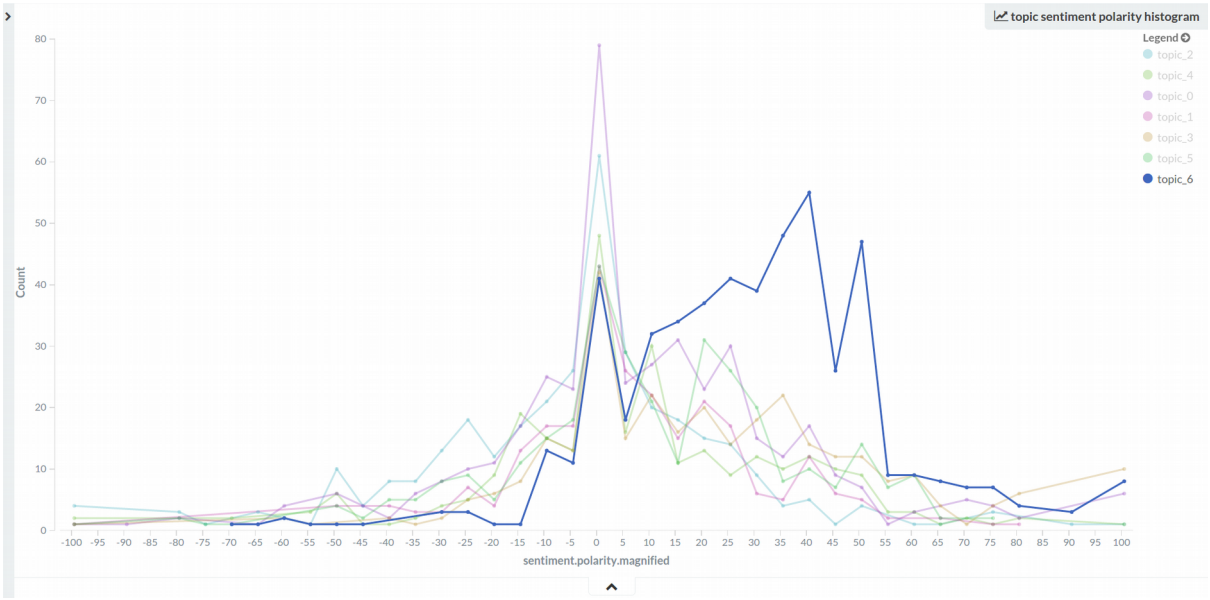

Figure 4.13: Sentiment histogram by topic (Vencovsky, Bruckner, Sperkova 2016)

4.5 Case study discussion and summary

The purpose of the study was to enable continual service quality monitoring and diagnosis. Regarding the thesis goals, research questions *Q3a, How can be consumers' feedback sentiment extracted from a document,* and *Q2a, How can consumers' feedback sentiment imply service quality*, were addressed.

As discussed in chapter 3, before sentiment analysis, it is necessary to extract significant service aspects that can be later linked with expressed sentiment. Several methods were presented for this purpose. In this study, significant aspect groups were extracted from the content using topic analysis. The analysis included the use of Parallel LDA algorithm, Gibbs sampling method, network analysis and manual topic labelling.

After identification of the aspect groups from the content, sentiment of the reviews was extracted. As a method for sentiment classification, supervised learning was chosen, and was also assessed as a reliable and stable method for sentiment extraction on a document level. Naïve Bayes algorithm was used for learning and Cornell Review database as training data. Results of sentiment analysis showed normal distribution-like shape, usually observed for quality reviews, as for example in (Song, Lee, Yoon, Park 2015).

Literature review in chapter 3 built assumptions for the research question *Q3a, How can be consumers' feedback sentiment extracted from a document.* This study validated these assumptions through sentiment extraction for significant aspect groups of a specific e-service.

Following research question, addresses by this study, is *Q2a, How can consumers' feedback sentiment imply service quality.* The study contributes to this question through exploration of dimensionality of gathered feedback. The dimensionality of online banking service was explored in the way it naturally appears in the feedback. Given the fact that consumers' feedback was collected voluntarily, quality dimensions cannot be generalised neither to e-banking nor this particular e-banking service. On the other hand, service quality dimensions are representative for available sample and likely contain the most controversial or negatively perceived service aspects. This is supported by motivation of consumers to write a review in the situations when they are strongly dissatisfied or satisfied. The assumption is based on the literature review from chapter 2 and work the of (Hu, Zhang, Pavlou 2009) and (Zhang, Yu, Li, Lin 2016).

According to the broadest e-service quality model (Yang, Fang 2004), the only service quality dimensions that clearly correspond with the model are ***access*** and both service *support* topics that can be seen as Yang's ***communication***. What are the other aspect groups extracted through topic analysis like? They refer rather to different sub e-services or service artefacts and can be divided into two groups: (1) *Credit card* and *bank account* are **sub-services of a core banking service** which online banking facilitates. (2) *Website banking*

application and *mobile banking application* are pure **IT sub-services** or **IT artefacts** (from different point of view) that support the banking e-service.

Because the same method for quality dimension exploration was used as in (Yang, Fang 2004), the only difference is the data sample which is limited to one particular e-service in this study. Quality dimension research cannot be generalised to all e-banking services, but for this focal service, it **resulted in two service quality attributes, two core sub-services and two IT sub-services**. As consumers see IT sub-services as significant service aspect, it indicates the importance of information technology for service operation and confirms the nature of e-service.

To demonstrate practical implication of service quality diagnosis using unstructured data feedback, the sentiment together with mapping to service aspects and original data were then loaded into a managerial dashboard. The managerial dashboard allowed service mangers to monitor service quality over time and offered them the possibility to create hypotheses about service quality and potential improvements. For future analysis of reviews, the study proposed a classification based on SMO algorithm. The final accuracy of the trained classification model was 70%.

This study represents one part of continual service improvement process. The goal of this process is to identify strong and weak points of the service based on the voice of customers. This could be reached through continuous monitoring of consumer feedback in a real time and diagnosis of service quality based on quality rating and sentiment of service aspects. This study does not explore processes of service improvement that follow quality diagnosis and monitoring.

For instance, the group of aspects identified as *topic_2* refers to *access* to e-banking service. This group of aspects is more likely to be connected with a negative sentiment than the others. In other words, the aspect of the service that is responsible for access to the core service is frequently mentioned by consumers in reviews classified as negative. The aspect caused that the whole service package did not meet expectations of large number of service consumers.

The managers could filter data to see only the reviews of this particular category and to read a couple of review samples. They could also connect data from different CRM-like systems and see whether negative sentiment is linked to a client information such as client location or account type.

Although it was not goal of this study, it can be said that overall service quality of this particular e-banking in observed time period could be probably increased. Proposed activities would be to (a) lower consumers' expectations, (b) enhance quality of access aspect and (c) focus more on support services.

The analysis of structured and unstructured review data validates a theory of bi-modal distribution of online review ratings (Hu, Zhang, Pavlou 2009). The dashboard solved the issue of average measure inadequacy

through enabling data exploration more in depth. However, the study does not explain what an actual relationship between the two types of feedback is. The issue is related to the research question *Q4* and was explored in chapter 6.

The presented method of service quality diagnosis has certain limitations. The document level that the study uses may cause an information loss. Theoretically, consumers can mention more than one group of service aspects in one review. The box plot in the figure 4.4 implied that the aspect groups do not overlap and therefore the information loss should be minimal. As discussed in the chapter 3.7, analysis on document level is enabled by learning algorithms such as Naïve Bayes and SMO to achieve higher accuracy score and to return better results. Although the review was linked with single aspect group in the dashboard, the other eventual aspects were also extracted. For the reason of managerial dashboard simplification, the other groups were leaved unused.

Chapter 5: Case study of call service online feedback

5.1 Introduction

The goal of this study[7] is to enrich the actual feedback of call service and to enable the diagnose of service quality from a textual content. This study focused on different approaches to content analysis in comparison to the previous study. The aspiration of this study, in particular, is to capture the sentimental attitude in a more complex way and to contribute to the research question *Q3b, how can be consumers' feedback emotionality extracted from a document*. The second aspiration is to classify feedback into reasonable categories, similarly to the previous case. In the second part of the study, relationship between modal emotions and service quality is explored in order to bring the light on the research question *Q2b, how can consumers' feedback emotionality imply service quality?*.

The focal service is a call service which acts as a supporting service in a package that delivers the main value for customer through a core service, an administration of a saving fund.

The study uses feedback data from call service consumers. After customers finish a phone call with the company, the company asks them to fill in a survey. The survey represents a current tool for analysis of the call service quality. It contains both close-ended and open-ended questions. The open-ended questions are accompanied by three close-ended questions, together they create a pair of rough and detailed information given by consumers on one subject. The subjects with the question pairs are: *q1* perceived level of personal service, *q2* perceived level of effort to get their question answered and *q3* feeling about a core service. The core service is a pension product in this case. The two first questions represent call service quality dimensions. A service of a good quality means that the consumers agree with the statement that they perceive the service to be on a high level for *q1* and agree with the statement that it costs them low effort to get their question answered for *q2*.

Contributions of the study are:

(a) exploration of the relationship between perceived service quality and expressed emotions to address the research question *Q2b*;

(b) comparison of multi-class and numeric learners for learning topic and emotion classification of service feedback to contribute to the research question *Q3b*;

(c) comparison of lexicon and supervised learning approach for emotion classification to contribute to the research question *Q3b*;

7 The study has not been published. It was conducted by the following team: Filip Vencovský (researcher), Jos Lemmink and Benjamin Lucas (supervisors), Loes Moritz (manual categorisation, coding), Guy Simons (coding, statistic methods), Roel Lubberink and Maurice Kusters (marketing department).

(d) development of categories of online service feedback for support service to enable quality diagnosis and monitoring for specific feedback types as a practical implication;

(e) validation of dashboard tool for service quality diagnosis as a practical implication.

5.2 Data description

The data this study used consisted of 10 000 survey responses from 8 170 different service consumers. The company had been gathering the data for more than one year period from September 2016 to October 2017. The responses are written in Dutch language.

5.3 Method

The first task was to classify the answers for each of the three questions into reasonable categories. It was mostly for practical implications of the study of quality diagnosis of this particular service. The method for categories development consisted of unsupervised classification task and expert categories development based on preliminary categories from reading of one hundred sample responses. The findings from these two approaches were discussed and final categories were consensually developed. The proposed unsupervised techniques were Parallel Latent Dirichlet Allocation (LDA) (Wang, Bai, Stanton, Chen, Chang 2009), cascade K-Means (Calinski, Harabasz 1974) and term vector network analysis. A team of experts consisted of a person responsible for the survey and a higher marketing department manager.

After the categories development, a model for future classification was trained to enable continual service monitoring. Multiple learning algorithms and feature preparation were tested to receive the best training results.

In order to receive more information from feedback than the basic sentiment analysis offers, modal emotional dimensions of text were classified. This activity was necessary for validation of theoretical assumption from chapter 3 and for addressing the research question *Q3b*. Because there was not any existing models of emotion classification described in the literature, a model in this study had to be trained on manually coded data. The learning algorithm was selected according to the best measured accuracy.

Emotion was classified according to Plutchik's emotion model (Plutchik 1980): *joy, sadness*, *fear, anger, disgust*, *anticipation, trust* and *surprise*. Each answer was assessed by a level of association with every modal emotion. The level of association was expressed on a four-items scale, where zero means *no association,* one means *weak association,* two means *moderate association* and three means *strong association* with an emotion.

The same approach to emotion intensity and emotion dimension was employed in the work of (Mohammad, Turney 2013). Apart from the proposed emotion association method, the work resulted in an emotion lexicon. The lexicon was used in this study for comparison of lexicon approach and supervised learning approach to emotion classification. Emotionality of document can be expressed through an emotion vector. The vector for numeric values look as follows:

$$Response_i\ Question_j\ (Anger\langle 0;3\rangle,\ Anticipation\langle 0;3\rangle,\ Disgust\langle 0;3\rangle,\ Fear\langle 0;3\rangle,\ Joy\langle 0;3\rangle,\ Sadness\langle 0;3\rangle,\ Surprise\langle 0;3\rangle,\ Trust\langle 0;3\rangle)$$

Models for the comparison with the lexicon were trained by multi-class learners as well as by numeric learners. The comparison was based on F-measure values. F-measure works with precision and recall (McCallum, Nigam 1998). Methods for classification learning are: Support Vector Machine (SVM) with Plynomial, Hyper Tangent and RBF kernels, PNN Learner (DAA), Gradient Boosted Trees Learner, Random Forrest Learner, Fuzzy Rule Learner, RProp MLP Learner, Naive Bayes Learner and Decision Tree Learner.

For numeric predictions, twelve different learners are used: Random Forrest Learner, Fuzzy Rule Learner, Gradient Boosted Regression Learner, PNN Learner (DAA), Tree Ensamble Learner, Pace Regression Learner, RBF Regression Learner, Linear Regression Learner, Isotonic Regression Learner, RProp MLP Learner, Polynomial Regression Learner and Simple Tree Learner.

Three different feature sets for each question were prepared for learning. The first set consisted of word uni-grams, the second set of word bi-grams and third set of their mutual combination. For emotion classification, where function words might play an important role, features were prepared twice. Once with and once without stop words.

The research question *Q2b* was approached by measuring a relationship between modal emotions and service quality. The service quality metric was based on close-ended scales and emotion associations from feedback full text. The close-ended scale of *q1* refers to overall service quality. The close-ended scale of *q2* will be taken as a significant service aspect. Answers to both questions are expressed on a five-point Likert scale. Two items on the scale represent agreement of different intensity, two represent disagreement of different intensity and one neutral attitude. For the purpose of simplification, the items were grouped in two sets. One set contained two agreement items, the other one contained two disagreement items. Neutral items were omitted. Bi-polarisation of the variables enabled to use of binary logistic to explore the relationship between expressed emotions and service quality. The sub-question that relates to the research question *Q2b* was posed: *Can agreement with high level of service quality be explained by intensity of modal emotions?*

5.4 Results

5.4.1 Feedback content categorisation

The unsupervised clustering methods did not result in reasonable outputs. Parallel LDA algorithm needs to have concrete number of topics to extract, but in this case, number of categories was not known. The only possible solution was to run an optimization loop, similar to (Antons, Kleer, Salge 2015), that identifies the optimum according to log-likelihood measure. The loop resulted in the optimum equal to number one which is not an acceptable solution for the category development task. Even hyper-parameter optimisation was not successful.

Another approach was tried using Calinski and Harabasz technique (Calinski, Harabasz 1974) to find the optimal cluster count according to the CH index. The technique is one of the most efficient techniques for the number of clusters determination (Milligan, Cooper 1985; Everitt, Landau, Leese 2001; Calinski, Harabasz 1974; Kozak 2012). For *q1*, the optimum returned four clusters with volumes 67%, 23%, 5% and 4%. The returned clusters were then labelled. The labelling was based on keywords extracted according to TCF-ICF weights (see the chapter 3 for more details). Unfortunately, the keywords did not lead to any reasonable interpretation of proposed clusters. The visualisation of the network of keywords and documents also supports the finding that data do not have any clear clusters.

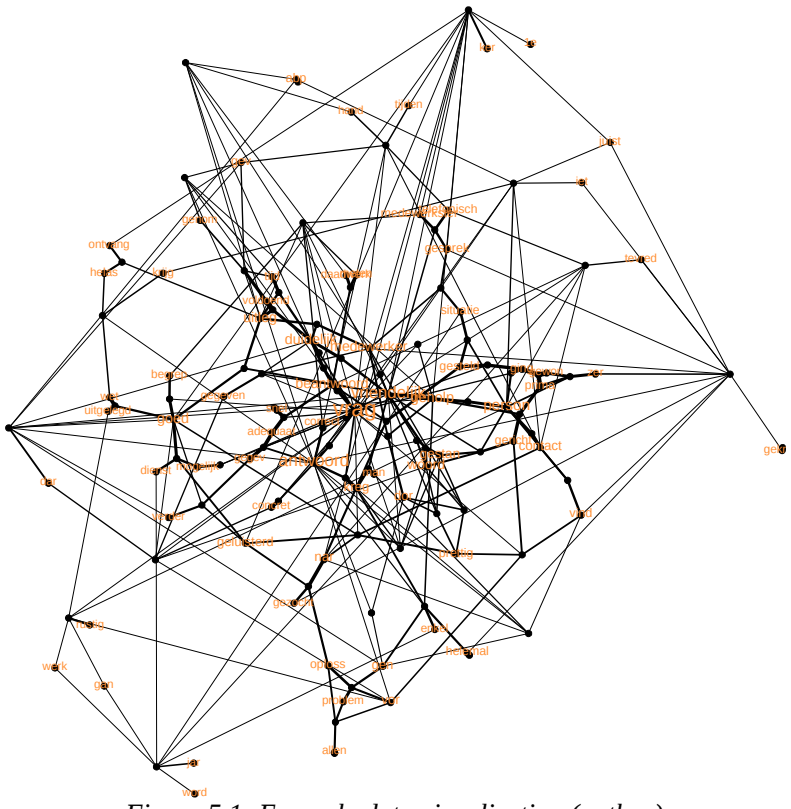

Figure 5.1: Example data visualisation (author)

Because unsupervised methods were not successful, team of domain experts proposed a possible content categories for each question. Subsequently, the team of experts modified categories after reading of one hundred random text samples. The categorisation process finally resulted in 14 categories of *q1*, 10 categories of *q2* and 14 categories of *q3*.

5.4.2 Response category and expressed emotion classification

Learning of categories was based on two thousand survey responses that were manually coded with the previously developed categories. Eighty percent of data were used for learning and twenty for testing. The best performance was presented by SVM learner with polynomial kernel, Gradient Boosted Tree learner and Random Forest learner. The best accuracy was 67% for *q1* with SVM learner, 61% for *q2* with Fuzzy Rule learner and 52% for *q3* with SVM learner. The differences in the accuracy can be explained by the fact that respondents did not always answer all three questions.

Emotion classification learning was based on the coded sample of six thousand answers. Eighty percent of data were used for learning and twenty for testing. The best overall results were performed by Gradient Boosted Tree learner, Random Forest learner, Fuzzy Rule learner, SVM RBF learner and PNN learner (DAA). The accuracy details are presented in the table 4.4. The overall accuracy includes performance across different data pre-processing and feature preparation. The overall maximal accuracy score, presented in the table 4.4, works only with the maximal algorithm accuracy across all of the eight emotions. The best performance were returned by learners for surprise, fear and anticipation (99.6% - 97.7%). Worse results were returned by learners for anger, disgust and sadness (92% - 88%). The worst results were returned for trust and joy (78% - 54%). Details are in the table 4.1.

Emotion	**Average accuracy**	**Best accuracy**
Surprise	78.78%	99.60%
Fear	76.96%	98.30%
Anticipation	76.64%	97.70%
Anger	68.24%	92.00%
Disgust	66.89%	91.00%
Sadness	65.33%	88.00%
Trust	57.76%	78.00%
Joy	37.25%	54.00%

Table 5.1: Emotion classification accuracy

The results reflect distribution of expressed emotions intensity. The emotions with the best accuracy results were not often present in the text. Conversely, the emotion with the worst accuracy results were present more often.

Learner	Overall accuracy	Overall maximal accuracy
Fuzzy Rule Learner	82.93%	86.51%
Gradient Boosted Trees Learner	84.01%	85.69%
Random Forest	83.71%	85.65%
PNN Learner (DAA)	81.87%	84.68%
SVM RBF	81.89%	83.24%
RProp MLP Learner	79.76%	83.09%
Decision Tree Learner	81.19%	82.71%
Naive Bayes Learner	75.07%	80.26%
SVM Polynomial	4.64%	5.58%
SVM Hyper Tangent	4.69%	5.41%

Table 5.2: Learning algorithms accuracy comparison

From the overall accuracy result, it is clear that using stop words did not generally influence the results. The overall accuracy showed only a small difference. Overall, uni-grams performed the best. Uni-grams and bi-grams together resulted in a slightly worse performance. The worst performance with three percent points under the overall maximum accuracy was presented by bi-grams.

Stop-words	Overall accuracy
No	66.00%
Yes	65.96%

Table 5.3: Overall accuracy of preprocessing with and without stop-words

Feature preparation	Overall accuracy
Uni-grams	67.41%
Uni-grams and bi-grams	66.08%
Bi-grams	64.45%

Table 5.4: Feature preparation accuracy comparison

Learning of numeric intensity performance was measured by the mean square error. Numeric learners used the same data as an input as category learners. Better results were performed by learners for surprise, fear and anticipation (0.004 – 0.016). Worse results were returned by learners for anger, disgust and sadness (0.042 – 0.055). The worst results were for trust and joy (0.68 – 0.69). The best performance was achieved by Random Forest learner, Fuzzy Rule learner, Gradient Boosted Regression learner, Tree Ensemble learner.

In lexicon-based approach,there was one emotion vector calculated for each question, based on a frequency and intensity of used words. The calculated values were then normalized with a full-text word count and multiplied by three to make them comparable with numeric prediction and manual codes.

Fear, surprise and anticipation showed the best results according to mean square error rate (0.05 – 0.31). Anger, disgust and sadness performed worse results (0.52 – 0.73) and the worst results were found in trust with 1.13 and joy with 3.16.

From the numeric results in table 5.5, it is obvious that the lexicon-based approach performed worse. It also had a strong limitation in words that can be recognised in text. The NRC emotion lexicon consists of 14 182 English words. To make lexicon work with Dutch language, a machine translation was needed. It resulted into only 7 850 Dutch words that s were not validated. It calls for lexicon approach improvement, especially for multi-language environments.

Emotion	Class learning	Numeric learning	Lexicon approach
Surprise	99.6 %	0.0045	0.19
Fear	98.3 %	0.0155	0.05
Anticipation	97.7 %	0.0165	0.31
Anger	92 %	0.0420	0.52
Disgust	91 %	0.0553	0.68
Sadness	88 %	0.0500	0.73
Trust	78 %	0.0693	1.13
Joy	54 %	0.0689	3.16

Table 5.5: Emotion classification performance comparison

From the result table 5.5, it is clear that a consistency in performance between all classification approaches exists. The possible prediction performance of modal emotion can be discussed only in a very limited way, because emotions expressed in different languages are hardly similar. For example, the work (Burget, Karásek, Smékal 2011) explored emotions expressed in Czech language. Contrarily to this study, they classified only six emotions (trust and anticipation were missing). The subject of classification was news headlines. Anger and disgust, followed by fear and sadness resulted in the best accuracy. Joy and surprise, on the other hand, were considered as the worst predictable. Similarity can be seen in good predictability of fear and bad predictability of joy.

Emotion	Best results (this study)	Best results of (Burget, Karásek, Smékal 2011)
Surprise	99.6 %	71%
Fear	98.3 %	81%
Anticipation	97.7 %	-

Emotion	Best results (this study)	Best results of (Burget, Karásek, Smékal 2011)
Anger	92 %	87%
Disgust	91 %	95%
Sadness	88 %	75%
Trust	78 %	-
Joy	54 %	72%

Table 5.6: Rough comparison of emotion predictability

The categories and emotionality were predicted for the rest of the data based on the setting of the best results. The enriched data were then loaded into a dashboard that was similar to the one described in the chapter 4.

5.4.3 Relationship between service quality and expressed emotions

Unlike the studies reviewed in the chapter 3.6 and the first presented case study from the chapter 4, this study explored relationship between service quality and emotions expressed by service consumers. The following part brings the light on the research question *Q2b*.

In the survey, service quality is directly examined using close-scaled items *q1* and *q2*. The first question refers to the overall perceived quality. The second question relates to the consumer effort and refers to a significant service aspect. Based on that, it is possible to observe which emotions are related to perceived service quality.

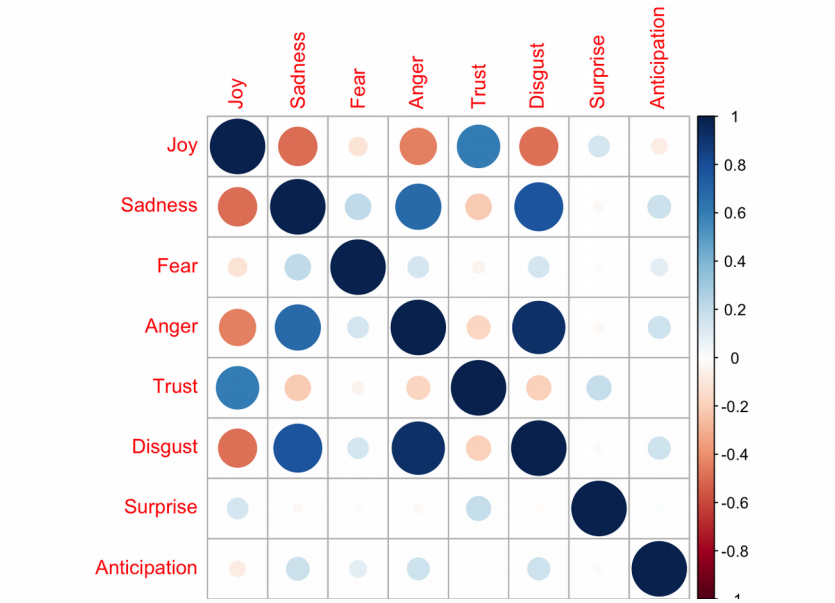

Figure 5.2: Service quality and modal emotions (author)

First of all, correlations of the modal emotions were calculated. Results for both questions indicated higher mutual correlations only for *joy*, *sadness*, *anger* and *disgust*. Figure 5.2 charts correlations for the *q1*, the level of service quality. It implies that *joy* correlated negatively with *sadness*, *anger and disgust* and positively correlated with *trust*. *Sadness*, on the other hand, positively correlated with *anger*, *disgust* and negatively with *joy*. And a*nger* correlated strongly positively with *disgust*, moderately with *sadness* and negatively with *joy*. The model of correlations for *q2*, as a significant quality aspect, also supported this findings (figure 5.3)

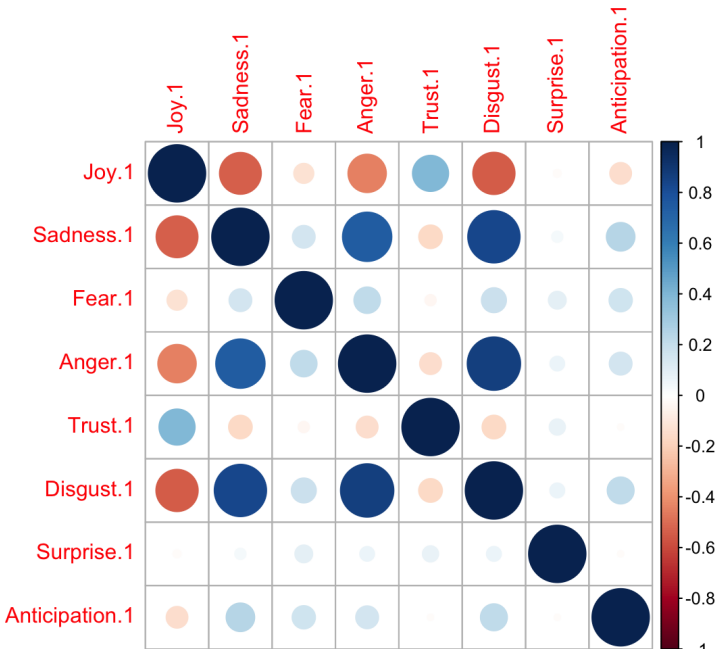

Figure 5.3: Consumer effort and modal emotions (author)

Regarding the theory of emotion circumplex (Plutchik 1980), discussed in the chapter 3.4.1, emotions are arranged in the pairs of bipolar items. The correlation data from this study validated these assumptions only for joy and sadness and only weakly for trust and disgust. It could be explained by the fact that this study is limited to one particular service; nature of emotions observed by Plutchik reflects rather emotions felt in everyday life.

Relationship between modal emotions and service quality had not been described before this study. According to the result, the model explains relationship with a high significance. **For overall perceived service quality, there were two strongly significant and two significant relationships observed.** (a) Joy expressed in a consumer feedback strongly positively implies the perceived service quality, (b) sadness moderately negatively implies perceived service quality, (c) fear weakly positively implies perceived service quality and (d) disgust weakly negatively implies perceived service quality. The results are

presented in table 5.7. Based on these results it is possible to claim that **the agreement with high level of service quality can be explained by intensity of modal emotions**.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.6466	0.3740	-7.077	1.47e-12 ***
Joy	-1.7281	0.3192	-5.413	6.20e-08 ***
Sadness	1.1648	0.1498	7.775	7.55e-15 ***
Fear	-0.7386	0.3490	-2.117	0.0343 *
Anger	0.4450	0.2980	1.493	0.1354
Trust	0.5982	0.3846	1.555	0.1199
Disgust	0.5736	0.2758	2.079	0.0376 *
Surprise	0.8161	0.6899	1.183	0.2369
Anticipation	-0.1441	0.2352	-0.613	0.5400

Table 5.7: Service quality and modal emotions, binary logistic results

For customer effort as a significant service quality aspect, there were three strongly significant relationships observed. (a) Joy expressed in a consumer feedback strongly positively implies the service quality aspect of consumer effort, (b) sadness moderately negatively implies the perceived service quality aspect and (c) disgust strongly negatively implies the perceived service quality aspect. The results are presented in table 5.8.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.2961	0.3308	-6.942	3.87e-12 ***
Joy	-1.9297	0.4028	-4.791	1.66e-06 ***
Sadness	1.0369	0.2252	4.605	4.13e-06 ***
Fear	11.2072	1244.6136	0.009	0.993
Anger	-0.2548	0.3909	-0.652	0.514
Trust	0.3376	0.7848	0.430	0.667
Disgust	1.6991	0.3172	5.357	8.46e-08 ***
Surprise	1.0360	1.3667	0.758	0.448
Anticipation	0.2096	0.3804	0.551	0.582

Table 5.8: Customer effort as a service aspect and modal emotions, binary logistic results

5.4.4 Dashboard

The content enriched in content categorisation and expressed emotions classification was loaded in a dashboard to validate dashboarding of consumer feedback as a source for service quality continual monitoring and diagnosis. The dashboard used the same technological background as the dashboard from

the previous study: Elasticsearch was used as a database and search engine; Kibana as a dashboard tool. The dashboard was validated as a source for service quality continual monitoring and diagnosis by the team of marketing experts.

5.5 Case study summary and discussion

The purpose of the study was to enable continual service quality monitoring and diagnosis. Regarding the thesis goals, research questions *Q3b, how can be emotion measured from consumers' feedback*, and *Q2b, how can emotions imply service quality* was achieved with help of service feedback enrichment by categorisation of textual content and classification of expressed emotions.

First part of the study focused on capturing expressed emotions in relation to the question *Q3b*. Based on theoretical assumptions from chapter 3, different learning methods and different ways of data preprocessing were employed. It resulted in the finding that **the accuracy of prediction differs across different modal emotions**. It was also observed that **the accuracy is influenced by frequency and intensity of expressions of modal emotions in a text**. The most frequent emotions, *joy* and *trust,* had also the worst accuracy across all classification methods. In contrast, the lowest frequency and the highest accuracy were measured for *surprise*, *fear* and *anticipation*.

The literature review in chapter 3.4 explored and discussed the possible ways of expressed emotion measuring as assumptions for the research question *Q3b*. **This study validated these assumptions and it can be claimed that it is possible to capture expressed emotions from consumers' service feedback using supervised learning and emotion lexicon.** But, the capturing of expressed emotions from text has certain limitations. The limitations lie in frequency and intensity of emotions expressed in a text and in specificity of emotional expressions for different languages.

Second part of the study focused on the research question *Q2b*. The question, *how can emotions imply service quality*, is answered by exploring relationship between overall service quality and emotions expressed in consumer open-ended answers. Based on the results, it is possible to claim that **an agreement with high level of service quality can be explained by intensity of modal emotions**. The modal emotion that strongly implies service quality in a positive way is *joy*. *Sadness* and *disgust* imply service quality strongly, but in a negative way.

The monitoring of consumers' feedback emotionality, especially in quality survey, can improve the feedback information value. The data presented together with the original data in a managerial dashboard allows service managers to create hypotheses about service quality. For instance, they can monitor consumer trust over time and see connection with business objects and core services that were discussed during the call. On the other hand, a knowledge that a certain emotion occurred does not mean anything. It

needs to be investigated more in depth in relation to other available data. A possible managerial implication could be to set a level of emotionality as a goal.

The limitation of this study is that it focused only on one language. The intensity of emotion expressed in a text can differ between languages and cultures. The second limitation is that this study explored the textual consumers' feedback, but emotions towards the service can be expressed in more ways. Beside basic self-evaluation, measuring of emotions can be undertaken using voice analysis, facial recognition or even more directly by EEG headsets[8].

This study was the first study that explored emotionality of service feedback. Nevertheless, additional research on this topic is needed. Especially, a larger sample of services from more service branches needs to be gathered.

8 For instance, EMOTIV can measure emotions such as *focus*, *stress*, *excitement*, *relaxation*, *interest* and *engagement*. (https://www.emotiv.com/)

Chapter 6: Qualitative research of online reviews

6.1 Introduction

One of the main identified issues of service quality diagnosis from online reviews lies in the relation between the review rating and the review body. The issue is referenced by the research question *Q4*, *What is the interplay between structured consumers' feedback and unstructured consumers' feedback,* but it also influences the answer to *Q1*, *How can consumers' feedback imply service quality.* Based on theoretical assumptions from the literature review in chapters 2 and 3, the studies in chapters 4 and 5 showed how certain aspects or aspect groups could be extracted from reviews. The studies even found that it is possible to link them with corresponding sentiments or emotions.

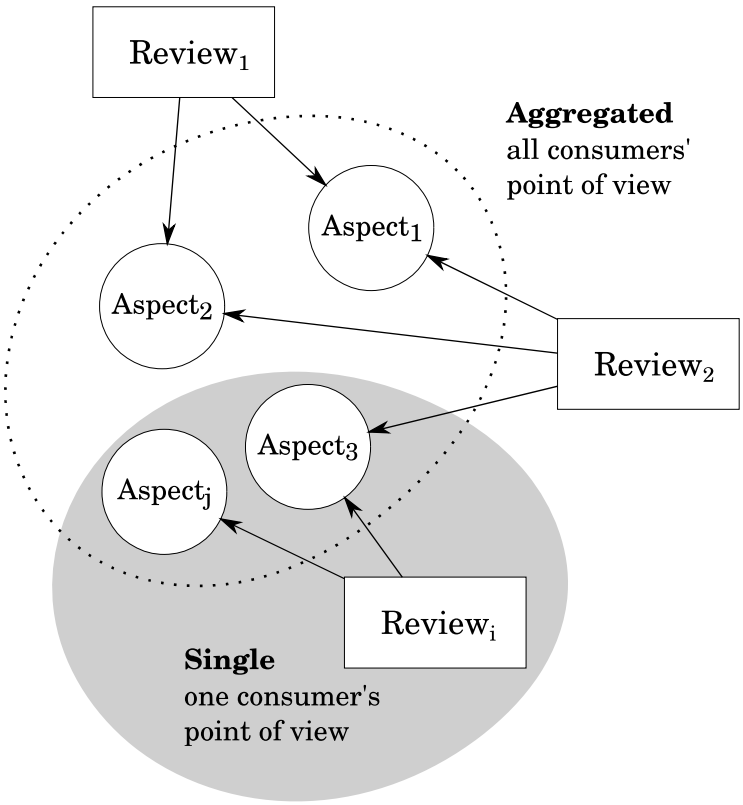

Figure 6.1: Levels of consumer point of view (author)

Figure 6.1 illustrates levels of consumer point of view, the levels on which service quality can be evaluated. The previous studies explored service quality on the aggregated level where all particular opinions were put together and categorised according to service quality aspects.

Aggregated level serves as an indicator of aspects' performance. For example, a service provider needs to know whether communication with a consumer is perceived positively. However, this aggregated information is hard to join to a service quality model. As the literature review in chapter 2 shows, the missing part is a weighting of service aspects' sentiment or emotionality.

Setting the right aspects and dimensions weights is challenging. Empirical studies (Gefen 2002; Wolfinbarger, Gilly 2003; Parasuraman, Zeithaml, Malhotra 2005; Yang, Fang 2004) that explored service dimensionality resulted in different quality dimension weights, but, which weights should be used and how to map them to service if every service has different dimensions? Moreover, if the nature of service is perishable and the provision of service is an individual act, it is always different, dependent on circumstances. Even the same service artefacts may be perceived differently by the same consumer over time.

Because service is an individual act, it is necessary to explore quality on the level of a single quality review. Only this could lead to the right assumptions about quality aggregation. Reviewed literature reflects this effort as well. (Duan, Cao, Yu, Levy 2013) observed that different quality attributes collected from a text have different weights on the overall rating. On the other hand, (Song, Lee, Yoon, Park 2015) used overall rating for setting individual weights of aspects for calculation of service quality.

Hence, it is essential to understand how quality is expressed in a review (*Q1*), what is the relationship between quality rating and sentiment of addressed service aspects, which directly refers to the research question *Q4*, and whether it is possible to identify the aspect weights (*Q2*).

Because quality on the level of a single review needs to be explored, the analysis in this study is not based on big data and statistical approaches, as the previously presented studies. It focuses on discovery and description of related phenomena, rather than its quantification, and thus qualitative research was chosen.

6.2 Method

Before answering the the research questions, suitable data have to be gathered and analysed. University information system was selected as a focal service in this case. Participants were also consumers of focal service. Data were gathered through a questionnaire in the common form of online quality review.

Online service' quality of university information system was surveyed using two questions. The survey was captioned as *"UIS (University Information System) User Review"*. The first question was a five-point rating scale labelled *"Overall perceived quality"*. The first point was labelled *"Low"*; the fifth point was labelled *"High"*. No other description was provided because an implicit interpretation of online review with the same structure was suitable for the research. The scale was followed by an open-ended question labelled *"Full-text review"*. A description was included in the question, *"Please, write at least five sentences long*

review that reflects your subjective experience with university information system." Research data were then analysed using principles of grounded theory (Strauss, Corbin 1990) research. In specific terms, it means that:

(a) no hypothesis was set;

(b) a theory was built on data only;

(c) open coding of research data was used;

(d) research memos were written.

Five different categories of code were collected reflecting these five research questions:

(1) What type of statements do consumers use to build a review?

(2) What kind of language patterns do reviewers use?

(3) How do reviewers address service aspects?

(4) How do reviewers express sentiment?

(5) How does overall review rating relate to review body?

Because there are few perspectives that need to be considered, more than one open coding process was undertaken. Analyses considered seven particular coding steps: (1) sentence coding, (2) language pattern coding, (3) parsed dependencies coding, (4) aspect coding, (5) sentiment pattern coding, (6) sentiment coding, (7) comparison with an overall quality rating.

To enable the coding, each review needed to be divided into single clauses and classified by parser into word dependencies and parts-of-speech. Sentiment of clauses was also classified for comparison purpose.

6.3 Results

The survey resulted in thirty responses collected from students of University of Economics, Prague. They had been asked to follow the description cited in the method section of this study. The students filled the survey online, but also were present in a classroom during the data collection phase. Because the quality survey was optional, not all of attending students submitted it.

The results are limited given the fact that the survey was in English and the respondents were not native English speakers. Notably, the question on language patterns is difficult to generalise. On the other hand,

service customers are not always native speakers, and natural language analysis must adapt to a specific vocabulary and language patterns used by customers.

In total, seventy-eight clauses were coded. Each clause was evaluated according to eight different perspectives. For example, the sentence *"I can really fast find some informations about my courses."* was processed as follows:

The sentence was coded (1) as a *quality statement*.

Language pattern of the statement was coded (2) as an *ability to do + activity aspect pattern + sentiment + intensity modifier*.

Moreover, every single parsed dependency was coded (3) separately in order to identify patterns of quality judgements[9]:

- VB_1-nsubj → PRP_1 (*actor*);
- VB_1-aux → MD (*ability*);
- VB_1-dobj → NN_1 (*core/aspect*);
- NN_1-nmod → NNS (*aspect*);
- VB_1-advmod → RB_1 (*sentiment/aspect*);
- RB_1-advmod → RB_2 (*sentiment intensity modifier*).

There were three aspects expressed in the single example statement coded (4) as follows:

- "find information", *2-word VB/NN Activity aspect*;
- "courses", *NNS Component aspect*;
- "fast", *RB Attribute aspect*.

The sentiment pattern was coded (5) as *RB over adverb modifier dependency plus RB as sentiment intensity modifier*.

The sentiment of the statement was coded (6) as *positive*, although the general English sentiment model from Stanford classified (7) the statement as *neutral*. The code *positive* was chosen because the attribute in

9 All syntactic relations' names used in this thesis follow the taxonomy of Universal Dependencies v2. For more detailed information, see http://universaldependencies.org/u/dep/.

the expression *"finding information"* is *"fast"*, which positively influences a consumer regarding the purpose of the selected service.

Besides the presented codes, the expressed sentiment and addressed aspects were compared (8) with an overall quality rating on the five-point scale on the level of the whole quality review of a service consumer.

Following sub-chapters describe findings regarding proposed research questions of this study. The first chapter explains from which statements a review is build and where the most important information can be found. The second chapter describes how reviewers formulate service aspects and how expressed aspects can be classified in general. Third chapter analyses expressed sentiment and differences between general English sentiment model and tailored sentiment analysis. The last chapter refers to the main issue of how service quality is expressed in full text and what is the difference between structured and unstructured quality feedback.

6.3.1 Composition of a quality review

Open-coding of quality reviews on a statement level resulted in twelve different statement types. The identified statement types are presented in table 6.1. *Quality*, *explanation, experience and emotion statements* were very common among reviews. Statement analysis helped to understand that consumers form a review using statements with different roles.

Two or more *quality statements* were found in each analysed quality review. Quality statements answer the question, what an aspect is like. The example of a statement, *"The layout is confusing for new users"*, clearly describes the quality of the *layout* aspect by the word *confusing* which is a quality judgement.

Quality statements were often accompanied by *explanation statements* in which reviewers presented supporting arguments for their quality judgements. For example, the clause *"I've been studying for 5 years, and I still do not really know it"*, explains the previously expressed quality statement *"it's unnecessarily complicated"*.

Quality comparison is a special type of quality statement. It does not explicitly express quality judgement, but expresses a result of comparison between services or service parts. The sentence, *"I saw a better school system"*, is a good example, even though it is not evident what the name of the better system is.

Emotion expressions appear to have the same language patterns as quality statements. The difference is that the emotion expression statements do not answer the question of what an aspect is like – *slow*, *fast*, *confusing*, *unstable*, but rather of how consumers feel towards the aspect – *disappointed*, *satisfied*, *happy*.

Statement type	Answered question	Example
Quality statement	What is a service aspect like?	*"it's unnecessarily complicated"*
Emotion expression	How does respondent feel about a service aspect?	*"That was incredibly disappointing"*
Explanation	Why does respondent perceive quality in this way?	*"I've been studying for 5 years, and I still do not really know it" explains "it's unnecessarily complicated"*
Experience	How does respondent use a service or similar services?	*"I only use the 'My studies' section, sometimes the 'Public information' portal and the 'Document server'"*
Suggestion	What should change and how?	*"it should be made more user friendly"*
Preference	What does respondent prefer?	*"The most important information is in 'Moje studium'"*
Quality comparison	What is better and how?	*"I saw a better school system"*
Structure statement	How does respondent perceive a service structure?	*"There is also eLearning"*
Observation	What is a reality related to a service like?	*"so many students log in from mobile devices"*
Summary	What are the most significant aspects like?	*"In the end, required information is available"*
Benefit statement	What are the benefits of using a service?	*"it can basically even help you with your studies"*
Cost statement	What are the costs of using a service?	*"You will need approximately 2 weeks to navigate"*

Table 6.1: Statement types

Most of the reviews contained also *experience statements.* This kind of statement does not contain any information about quality. Consumers only refer to their own experience with a service, to how they use it or how they use similar services. An example clause looks as follows, *"I only use the 'My studies' section, sometimes the 'Public information' portal and the 'Document server'"*, which refers to use patterns of a service consumer and generally to an experience with the reviewed service.

Consumers express a particular type of experience in *structure statement.* It contains pieces of a service structure in the way a consumer sees it. For example, the sentence *"there is also eLearning"* claims that the part of service mentioned in a previous sentence contains another part called *eLearning*. It is possible that a consumer wrote about the part because he wants to express the perceived importance of that part.

There were also statements that explicitly articulate importance of certain parts of a service – *preference statements*. They answer the question, what does respondent prefer. *"The most important information is in 'Moje studium'"* can be used as an example. Although preference statement looks similar to a quality comparison, they differ in purpose.

Besides the quality judgements, experience and preference, customers express results of their own observation of reality that surrounds a certain service – *observation statements*. For example, in the sentence *"so many students log in from mobile devices"*, consumer observes behaviour of other consumers – how they use a service.

Benefit and *cost statements* were rare, but interesting for service diagnosis. Benefit statement describes subjective benefits seen by consumers. For example, the sentence *"it can basically even help you with your studies"* expresses a possible effect of using a service. On the other hand, cost statements describe, what costs a using of service causes. In the example sentence *"You will need approximately 2 weeks to navigate"*, the cost is time needed to use a service properly.

Suggestion statements hold valuable information. Although it was sporadic, consumers shared suggestions about how a better service looks like from their point of view. This kind of statement answers the question about what should change and how. For example, the sentence *"it should be made more user friendly"* asks for a better user-friendly service experience.

Summary statement is also worth mentioning. Although the summary statement is not freestanding, the respondent concludes the most important statements: *"In the end, required informations are available"*.

From service quality point of view, the crucial part of the quality analysis is to identify how a consumer perceives (sentiment) a certain part of a service (aspect). Regarding that, the most valuable information can be found in quality statements.

Language analysis of quality statements showed that in most cases both aspects and sentiments are expressed in two dependent words. These kinds of word dependencies were coded as a *core* of the statement. Word dependencies coded as a core in this research are following, ordered by frequency, even though the frequency is not significant in such a small sample:

JJ-nsubj → NN, JJ-nsubj → PRP, JJ-ccomp → VB, JJ-nsubj → NNS, NN-nsubj → NN,
VBG-nsubj → NN, VBZ-nsubj → NN, JJ-csubj → VBG, JJ-dep → VB, JJ-nsubj → WDT,
JJ-xcomp → VB, NN-nmod → NNS, NS-amod → JJ, VBD-nmod → NN, VB-dobj → PRP,
VBN-csubjpass → VBG, VBN-nsubjpass → NN, VBN-nsubjpass → PRP, VB-nsubj → NN,
VB-dobj → NN

The core word dependency term corresponds with *feature-opinion pair* from (Zhuang, Jing, Zhu 2006), or *aspect-sentiment pair* according to (Liu 2015).

Further analysis showed which part of *core* dependency is more likely a service aspect and which is an expressed sentiment. Differences are described in chapters 6.3.2 and 6.3.4.

Apart from *core dependencies*, *aspect* and *sentiment dependencies* were also likely to be present in the data sample. Aspect dependencies extend aspect part of *core* dependency if the aspect is expressed in more than one word. Sentiment dependency for multi-word sentiment expressions plays the same role. In keeping with opinion mining theory (Liu 2015), *intensity* and *orientation modifiers* were present as special types of dependencies that also extend sentiment part of core dependency.

Other dependencies that were rarely present were coded as *actor, condition* and *ability.* All these dependencies are hard to interpret in relation to service quality. *Actor* is a kind of dependency that tells who caused an activity. It could be a consumer, but also service personnel or a service artefact.

Condition is an interesting dependency because it sets boundaries in which quality judgement is made. For example, the condition *"on mobile"* refers to a platform in quality judgement *"it's clumsy on mobile"*. S*ubjectivity modifier* dependency is also worth noting mentioning. It is a formulation of words that strongly modifies subjectivity, *"I think"* for example.

Type	**Count**	**Type**	**Count**
core	35	ability	4
aspect	25	actor	3
sentiment	7	condition	2
sentiment intensity modifier	12	subjectivity modifier	9
sentiment orientation modifier	4		

Table 6.2: Coded types of word dependencies

6.3.2 Aspects

Chapter 3 resulted in the model of service (figure 3.1), where components and attributes were the building blocks. It is possible to address these blocks using aspects from the review text. Four types of aspects were identified in the research sample. These types are defined by usage of three different parts of speech: noun, adjective and verb. Table 6.3 describes these aspect types.

Aspect expression	**Typical part of speech**	**Description**
Component	Noun	A service or a service part
Activity	Verb	An activity that a service performs or an activity that consumer performs while using a service
Attribute	Adjective or Adverb	An attribute of service or of a service component separable from sentimental expression
Implicit	Pronoun	Common knowledge or expressed in previous sentences

Table 6.3: Aspect types

Although the service model (figure 3.1) consisted of components and attributes, more aspect kinds were found in the sample of reviews. Components and attributes correspond with the respective concept from the model, whereas *activities* can be seen as a special type of component.

Component aspect expression

Component aspect expressions are usually expressed in one or two words. One word expression is always based on any form of a noun. Nouns are accompanied by another noun, as compound or nominal modifier, or by adjective as adjectival modifier.

Parsed expressions coded as component aspect were: *NN, NNS, NN-compound→NN, NN-nmod→NN, NN-amod→JJ, NN-nmod→NNS, NNS-amod→JJ*.

Component expression dependencies with adjectives are questionable and must always be communicated with service experts that know the focal service. For example, in expression "*basic navigation"*, *basic* is an attribute of the same navigation or is it a completely different aspect?

Activity aspect expression

A constituent of *activity aspect* expression is usually a verb word that also refers to a component aspect expressed by a noun or a pronoun. Activity aspects carry latent information about at least two aspects in one expression. For example, *"find schedule"* points to the activity *finding*, but also to the component *schedule*, whereas *"find marks"* refers to the same activity with a different component – *marks*. Other examples are *"change theme"*, *"add plugin"* or *"choose design"*. A second aspect may be in a position of verb subject or object. The second latent aspect is not always explicitly linked to a concrete aspect, what is illustrated in expressions such as *"do something"* or "*it works"*.

Parsed expressions coded as activity aspect were: *VBZ, VB, VBG, VB-dobj→NN, VB-dobj→NNS, VBD-dobj→PRP, VB-dobj→NN-nmod→NN, VBG-dobj→NN, VB-nmod→NN*.

Implicit aspect expression

Aspects are implicitly expressed by pronouns. They point at certain explicit aspects from previous clauses or refer to common knowledge of consumers. It could help to express general statements towards a whole service, such as *"I like it"*. Besides implicitly expressed pronouns, general statements are made by using words, such as *"everything" or "overall"*.

Parsed expressions coded as implicit aspect were: *"it"/PRP, "them"/PRP, "which"/WDT.*

The study also tested how the general English model from Stanford can match a reference to an explicit aspect of a previous clause. It is possible because an output of the model can mark nouns with the same

reference code as the pronoun from following clauses and sentences. However, there were only few meaningful matches, but there is still a potential for modelling a more specialised parser with use of a service-related language.

Attribute aspect expression

Liu (Liu 2015) sees implicit aspects not only as pronouns that substitute nouns but also as an implicitly stated nouns in adjectival or other multi-word expressions. For example, *expensive* refers to *price.* In this study, similar findings were identified. In order to distinguish them from previously defined implicit aspects expressed by pronouns, a new category of *attribute aspects* was formed.

Attribute aspect, in context of this thesis, implies what an object is like. For example, *water* could be *warm* or *cold*, *still* or *sparkling.* Attributes are strongly connected with expressed sentiment but are context dependent. Consumers of a thermal bath may see *warm water* as positive, whereas the same consumers may see *warm water* as negative in case of beverage.

Parsed expressions coded as attribute aspect were: *JJ, VBN, NN, RB, JJ-nmod:npmod→NN, JJ-amod→JJ, JJ-nsub:xsubj→VB.*

Attribute aspect expressions were typically used in one word as adjectives. Adverbs and nouns were also coded as attribute aspects. Nouns coded as attribute aspects refer to the same inference as mentioned – *expensive* becomes *price*, but they are already expressed in a noun form by reviewers. A line between attribute and component is thin, because it is context specific and subjective, especially for abstract service parts such as information systems.

Besides aspect types identification, following observations were done:

Clauses were coded with more aspects because the aspects appear in multiple dimensions. For example, the phrase *"hard to find"* in the statement *"sometimes its hard to find information what I need"* refers to *Navigation* aspect. It also seems to be part of *Ease of use* aspect, because the sentence contains the word *"hard to"*, which is a constraint in using a system. Last part of the sentence refers to an information need of a consumer and can be seen as *Information support* aspect. Another example from the same review *"It's nice that I can change theme of InSIS"* refers to *Customisation, Appearance* and *Themes*.

Aspects are often expressed together with sentiment in one expression. For example, in the statement *"system is chaotic"*, *system* is one clear opinion target and *chaotic* behaves as sentiment word, but it is also an expression of *User interface* aspect because it refers to an orderliness of a system to a user.

Opinion mining techniques presented in chapter three have limited options in the means of aspect extraction. It is necessary to ensure that aspects can be extracted as multiple words and various parts of

speech. For example, reducing words to nouns only for capturing all possible aspects is not well-founded. It is also important to ensure that the selected method can extract more than one aspect from a single clause and moreover, that a single word can contain both sentiment and service aspect. There is also the issue of aspect grouping, but harnessing of implicit aspects seems to be the most problematic part.

6.3.3 Model of service aspects

Without specific knowledge of a service background, such as service documentation, the model of service aspects was built from aspects expressed in reviews. The modelling was done according to the same conditions consumers have when they write a quality review; it was based on the act of using a service, the voice of customer and a certain level of IT knowledge and knowledge of reality in which the service is consumed. The model presented in figure 6.2 is subjective and cannot be generalised to e-service, nor this particular service. The model illustrates, how a consumer's mental model of service composition may be formalised.

The model uses a simplified ER notation. It contains entities and relationships. There is one central entity – *service*. All relationships are oriented in the direction to the central point. The direction implies how the quality of all components is composed and how it is in accordance with the service model of component hierarchy and attributes presented in chapter 3.2.

Squares represent aspects expressed in reviews and coded as aspects. The aspects are at the lowest possible level of generalisation. Most of the aspects were left in the same form as in the reviews; attribute aspects that were converted into the noun form and generalised are the only exception. For example, *responsive* was modelled as *responsiveness*, but also *clumsy* was modelled as *responsiveness* because the reviewer referred to the ability of a website to accommodate to a mobile screen. **All the attribute aspect expressions were modelled in an *attribute of* relationship.**

All of the activity aspect expressions were modelled as sub-services of two types: *informational services* and *functional services.* These two generalisations were not explicitly stated in reviews, but correspond to general statements about the ability to *work* or *do something* for functional services and to *find information* for informational services. The two types of sub-service are in line with two other types of service from the theoretical division of IT services: information, application, infrastructure and support (Voříšek, Basl 2008). Only the object aspect expressions were modelled in all relations.

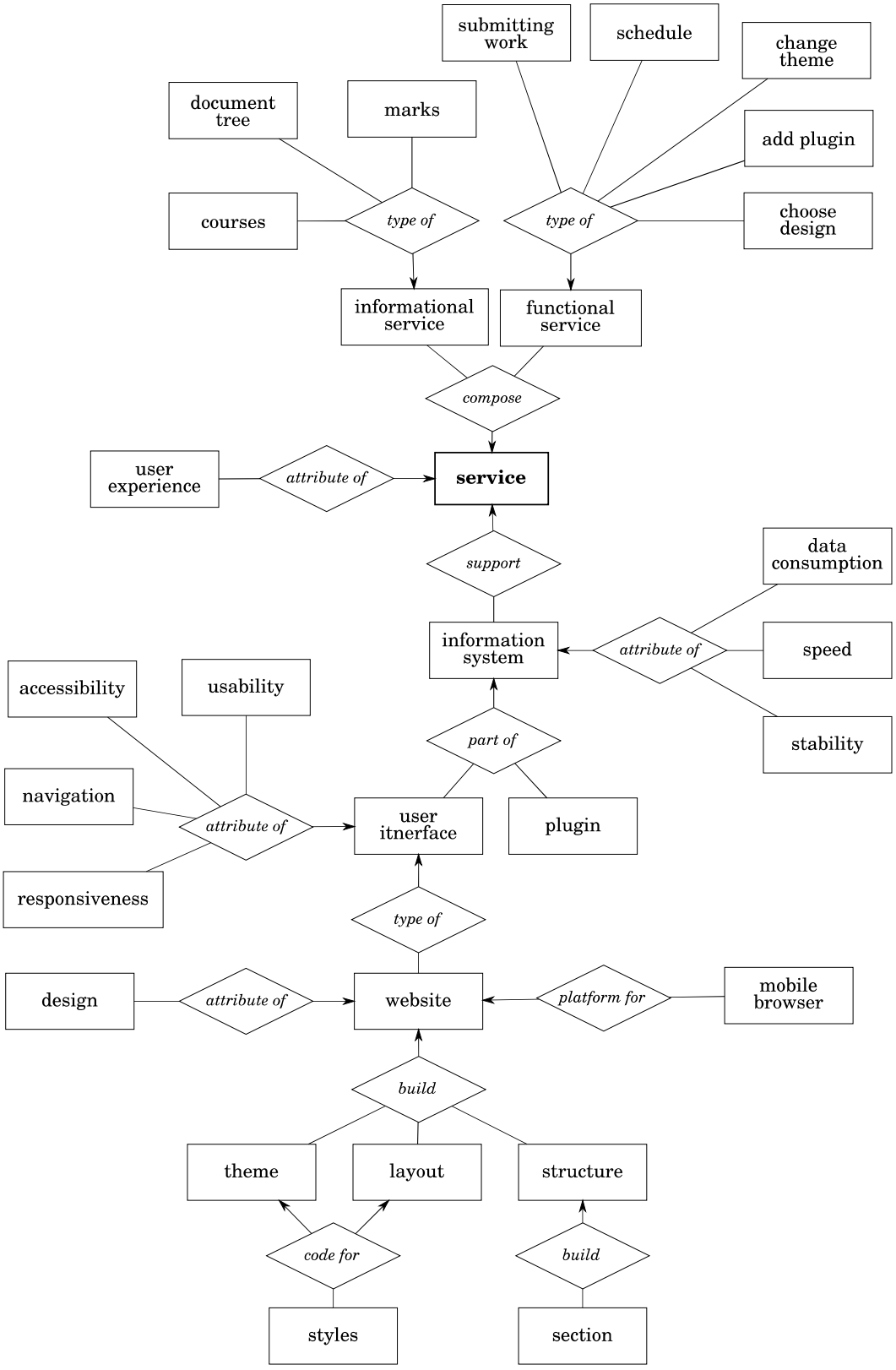

Figure 6.2: Model of a service aspects (author)

6.3.4 Sentiment

All aspects presented in the model Figure 6.2 were connected in the text with a certain sentiment orientation and intensity. Sentiment orientation and intensity were coded on a scale of five ordinal categories widely used in the literature (Kang, Park 2014; Ribeiro, Araújo, Gonçalves, André Gonçalves, Benevenuto 2016; Taboada, Brooke, Tofiloski, Voll, Stede 2011): *very negative, negative, neutral, positive and very positive*. The scale joints sentiment polarity (*negative, neutral, positive*) with sentiment intensity (*weakly, strongly*). For example, the statement *"accessibility of system is not a problem"* was coded as *positive*, whereas the sentence *"that was incredibly disappointing"* as *very negative*.

For comparison statement, the codes *worst* and *better* were used instead of *negative* and *positive*. The neutral category such as *same* or *similar* could have also been used, but was not supported by the occurrence in the data. An example of negative sentiment orientation in comparison statement is *"I saw a better school system"*, which means that the consumer perceives school system worse than another.

The coded sentiment was compared to results of the sentiment prediction of the Stanford NLP library with pre-trained general English model (Manning, Bauer, Finkel, Bethard, Surdeanu, McClosky 2014). The codes vary in 43 percent of cases. The differences can be explained by (a) a coder bias, (b) the service specific language, or (c) a classifier error.

The sentiment classifier was trialled on general English during the review process, so the probability of error is low. Nonetheless, the language of service reviewers varies from the general English, because the reviewers were not native speakers and the language is service specific, and model does not necessarily contain all used words or their shifted meaning. A coder bias is possible because the coding is subjective and it was not cross-validated with other coders. Table 6.4 illustrates differences in sentiment codes in detail.

	Stanford general English classification		
Coded sentiment	**negative**	**neutral**	**positive**
negative	18	4	1
neutral	8	4	0
positive	6	4	11
very negative	1	0	0
worse	0	1	0

Table 6.4: Sentiment classification difference

As language analysis from chapter 6.3.1 showed, besides the aspect, **sentiment appears as the other part of a core of a quality statement, but can be expressed in more words**. Adjectives were the dominant part

of speech for words coded as sentiment words. Nonetheless, adverbs, nouns and verbs were also present. Following list contains all expressed parts of speech with affiliated dependency.

> *JJ/nsubj, JJ/ccomp, JJ/xcomp, JJ/csubj, JJ/dep, JJ/nsub, MD/advcl, NN/nmod, NN/nsubj, RB/advmod, RB/neg (attribute negation), VBG/nsubj, VBN/nsubjpass, VBN/csubjpass*

As discussed in the previous chapters, **sentiment is often carried together with aspect by a single word**. It applies particularly to attribute aspect expressions. For example, the word *fast* describing an information system is the aspect *speed* but also carries a positive sentiment.

Sentiment intensity and orientation are modified by dependent adverbs. *Advmod* was the syntactic relation for an intensity modification. The orientation was modified using *neg* relation. The orientation and intensity modifications are the other reason **for the use of multi-word attributes in sentiment analysis.**

The sentiment expressed by reviewers is strongly dependent on a specific context. The finding supports the presented opinion that lexicon-based sentiment analysers are less effective. For example, in the statement *"buttons are small"*, the sentiment was coded as negative. It is hard to guess what the reviewer expected to be the right *buttons* size. If the statement was followed by an explanatory statement *"it is hard to hit them"*, the sentiment of *small* would be clear. An adverb used as sentiment intensifier might be another possible clue. In the statement *"buttons are too small"*, it is obvious that the reviewer expected bigger buttons. For some aspects, readers can understand sentiment, regardless of the fact whether the word is accompanied by other words or not. For example, "*small screen"* would be probably perceived negatively, whereas "*small price"* might be seen positively.

The sentiment expressed in emotion statements is easier to analyse than the one in quality statements because it is less context dependent. Sentiment words coded in emotion expressions were: *satisfied, disappointed, like.* In contrast, sentiment words in quality statements were, for example: *complicated, unstable, small.*

6.3.5 Relation between review rating and review body

This chapter builds on the analysis of gathered reviews. The analysis enabled to see the review as a set of statements of a certain type, a set of aspects of a certain type and the corresponding aspect sentiment. The goal of this chapter is to explain the research questions Q1 and Q4 more in depth: How can consumers' feedback imply service quality? What is the interplay between structured consumers' feedback and unstructured consumers' feedback like? Additional sub-question of Q4 was explored: Can text be used for quality diagnosis only?

Reviewers tend to use the review rating as an overall quality metric. It is supported by the observation of sentimental expressions in summary sentences such as *"Overall, I would have rated it as a weak*

average" together with rating of two points from five, *"To summarize, I am perfectly satisfied with the services the university information system can provide us"* with full five stars rating or *"it is fine"* with three stars out of five. Unfortunately, not every review contains such overall quality judgements.

After the analysis, **new questions arose**: What does reviewers motivate to write a summary statement? How can reviewers be stimulated to make such statement? Unfortunately, these questions cannot be answered without an additional survey. Especially, the first question may bring to light to the issue of identifying reviews written with less effort; the reviews where overall rating tends to correspond less with actual perceived quality, the reviews that may distort overall quality statistics (Zhang, Yu, Li, Lin 2016).

According to the literature review discussion in chapter 2, high overall rating means that the service quality exceeded expectation or met the expectation of high-quality service. It is an interesting assumption, but it cannot be validated easily. In the analysed reviews, **different strategies for scale rating were observed**. One reviewer rated the service quality with five points and used the expression *"perfectly satisfied"* in the summary statement, whereas other reviewer used the expression *"I like"* with the same five-point rating. The possible explanations of the second case are: (a) The expression mirrors actual meeting of expectations of high-quality service. (b) The reviewer tends to use moderate language. (c) The reviewer tends to rate quality higher. However, these hypotheses cannot be confirmed in this research.

According to the observed compliance of quality rating and evaluation of summary statement, the question *Q4* can be answered: **Unstructured feedback complies with structured feedback in the particular case of an expressed summary statement, however different strategies of writing and rating may shift relationship in both directions.**

The possible compliance also supports the additional research sub-question, **a text can be used for quality diagnosis only, but it has a significant limitation**. The desired formulation can be found in a summary statement, but the diagnosis of such statement has the same issue as quality rating on a scale. **The quality of particular aspects is hidden, and thus, service cannot be improved; it is even possible that quality will decrease after an improvement due to lack of understanding consumers' point of view.**

Evaluation of aspects weights and quality influence relationships

The research question *Q4* was partially explained regarding the interplay between summary statements and quality rating. However, the interplay between expressed non-summary statements and overall rating remains unexplored. Statement types, expressed aspects and sentiment were considered for each review to evaluate the interplay.

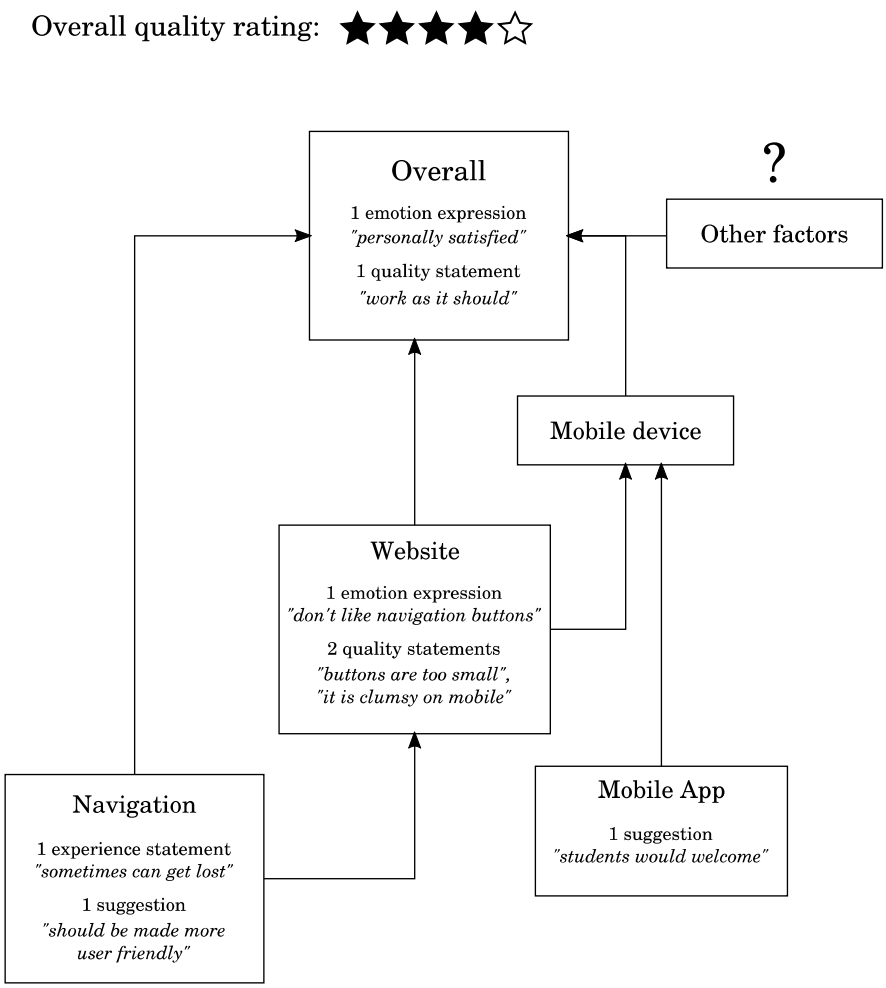

Figure 6.3: Example review content (author)

Figure 6.3 illustrates a possible structure of the relationship between individual aspect groups. The figure is based on a review sample. The aspects mentioned in the review were generalised into more abstract groups. A possible influence of perceived quality is modelled between groups using arrows. The figure was created for illustration although it is obvious that these kinds of relationships cannot be certainly captured.

The example review was rated with four rating points from five possible. The review contained two summary statements: one emotion expression *"personally satisfied"* and one quality statement *"work as it should"*, which clearly refer to the meeting of expectations. The emotion statement supports the fact that the rating is higher than average.

What does not correspond is the relationship of overall rating or statement to aspect groups. All mentioned aspects are linked to a negative sentiment. From this example, another partial answer to the research question *Q4* is clear, **overall quality judgement does not always result only from mentioned aspects**.

The perceived weight of the aspects mentioned in the example must be on a low level, and other factors with positive sentiment must be considered.

In cases where explicit weights are not expressed in a review, like in case of the sample review, the weights cannot be satisfactorily set. The only way is an approximative setting, but this setting is limited to one situation only. The situation was described in (Song, Lee, Yoon, Park 2015) and is based on the following assumption. **If only two aspects are expressed, one with negative sentiment and the second with positive sentiment, and the rating is high, the relative importance of the positive aspect is higher than the relative importance of the negative aspect** – the positive aspect overweighted the negative one. However, this situation happens rarely, because reviewers usually mention more than two aspects. If more aspects appear in a review, it is right to assume that only a group of positive or negative aspects overweights the other group.

Another review from the analysed sample illustrates the findings from the literature. The respondent wrote mainly about bad *user experience* caused by *navigation* problems. On the other hand, respondent stated that the *information support* is on the expected level. No other service aspects were mentioned. The limitation of the review to these two groups of aspects was supported by the summary statement that mentioned the both, navigation and information support. As per the overall quality rating of *two points*, the *navigation* aspect was perceived probably as more important than the *information support*.

However, the assumption of aspects weighting has another limitation. It was observed that **the overall rating was influenced by other unexpressed aspects.** The question is how strong that influence might be. According to the literature review, consumers mention only the aspects that did not meet or exceed their expectations (Zhang, Yu, Li, Lin 2016), which is in contradiction to the example review where non-present aspects influenced the overall quality completely.

6.4 Chapter summary and discussion

The purpose of this study was to explore consumers' point of view from a different perspective. The previous studies examined quality on the aggregated level where all opinions of consumers are put together. This kind of aggregation is necessary because it enables service managers to see service quality in the bigger picture. However, there is a problem linked with the nature of service – service is perishable and depends on many circumstances. Moreover, each consumer perceives quality and quality dimension in different ways. Without the understanding of how quality is captured in a review, it is not possible to aggregate the overall quality of service. This issue is addressed directly by the research question *Q1* and is supported by examination of the relationship between quality rating and full-text review body, *Q4*.

In order to answer these research questions, it was necessary to collect and analyse data first. The data was collected from students who are also consumers of the university information system service. An online survey with the typical structure of an online review served as the method for their gathering.

Because more perspectives needed to be considered, more than one open coding process was undertaken. The analyses considered seven particular coding steps: (1) sentence coding, (2) language pattern coding, (3) parsed dependencies coding, (4) aspect coding, (5) sentiment pattern coding, (6) sentiment coding, (7) comparison with an overall quality rating.

To enable the coding, each review needed to be divided into single clauses and classified by a parser into word dependencies and parts-of-speech. The sentiment of clauses was also classified for comparison purpose.

Regarding this particular qualitative study, five research questions were set. The first raised question was, *what type of statements do consumers use to build a review.* **Open coding of statements resulted in twelve statement types.** The statement with the highest potential for quality diagnosis was the *quality statement.* The *emotion expression statement* was similar. It does not mention specific service attributes, but expresses emotions towards service aspects. Another type that relates to quality was the *quality comparison*. Close to the quality comparison, but only stating that one aspect is more important than another, was the *preference statement.* All quality statements were supported by the *explanation statements*. Reviewers also described their *experience* with the service or *observation* of reality that surrounds it. They also made *suggestions* which represents a great potential for service improvement. Also, the *benefits* and *costs* of using the service were stated. The reviews were also *summarised*, which offers an important view on consumer' perception of final quality judgements.

Regarding the weights settings, the *preference* and *structure statements* and *summary* are useful. However, no review in the sample contained enough information to set the meaningful aspect weighs.

The second question, *what kind of language patterns do reviewers use*, was **answered by parsing sentences with the Stanford English parser, that outputted word dependencies, and open coding of each dependency**. The dependencies were coded according to their role in expressing of quality. Only quality and emotion statements were analysed because they contain information about service aspects and quality. It was found that aspects are often expressed with a sentiment in one dependency. The dependency was, regarding service quality, called the *core dependency.* If an aspect or sentiment was expressed in the multiple words, the dependency was coded as *aspect* or *sentiment dependency.* Sentiment was modified, in accordance with the opinion mining literature, using two kinds of dependencies – *sentiment intensity modifier, sentiment orientation modifier.* Other dependencies were *ability, actor, condition*, and *subjectivity modifier.*

The third question, *how do reviewers address service aspects*, was **answered by coding expressed aspects and aspect expression patterns**. It was observed that aspects could be seen, in keeping with the opinion mining literature, as *component* or attribute. It was interesting that aspects were expressed as an *activity.* Activity aspects were mapped mostly to sub-services. The last observation was that aspects were often *implicit.* Implicit aspects were, in contrast to Liu's point of view, expressed only by pronouns. The language patterns that appeared in aspect expressions were identified.

It was surprising that clauses were usually coded with more than one aspect because aspects appeared in multiple dimensions and were also often expressed together with sentiment in one expression.

The fourth question, *how do reviewers express sentiment*, was **answered by coding expressed sentiment and sentiment expression patterns**. As it was already mentioned, the sentiment often appeared together with aspects in one expression. It was also found that the sentiment expressed by reviewers is strongly dependent on a specific context. The sentiment was modified by adverbial modifiers in intensity and negation modifiers in orientation. The language patterns that appeared in sentiment expressions were identified.

The last question, *how does overall review rating relate to review body*, was answered by evaluation of the relationship between quality rating and count, sentiment and categories of aspects. Special attention was paid to quality expressed in summary statements. Based on the small sample of reviews, it was found that there is no clear relationship between quality rating and content of review body. Only summary statements were similar to quality rating, but there were small contradictions as well. These contradictions can be explained by different reviewer strategies.

The main finding is that the aspect sentiment alone cannot explain overall quality rating. According to the **observations made in this study, it is clear that aspects have different weights, which are hard to extract from a text**. Moreover, **the overall quality is clearly influenced by non-expressed aspects**.

Individual service quality trees

The service quality literature proposed the aspect hierarchy tree (Hu, Liu 2004; Liu 2012, 2015), but this concept should be modelled on the individual level. The findings of this study call for a different mechanism for quality aggregation and diagnosis.

What does an optimal service quality diagnosis look like? On account of perceived service quality as an individual concept, it is necessary to diagnose service quality on the same level – an individual's level. Each review should be transformed into a quality model – more specifically into a tree of quality attributes. All individual's quality trees should be placed in a database to enable their querying.

A quality tree reflects how a consumer understands a service; it is a mental model of service regarding its structure, quality and emotions, a result of implicit quality expectation confirmation. A tree is composed of weighted relations between service quality attributes and corresponding emotions including sentiment and its orientation and intensity.

Qualitative research in this chapter found that the quality tree cannot be created based on a single consumer review because consumers mention only a few quality attributes. Full quality structure, influence relationships and weights remain hidden.

Service quality research should focus on finding a way to capture such quality trees. It could also be used for perceived quality prediction in case of planning a change in production quality. Nevertheless, certain constraints still exist. One of the main issues is that no consumer review uncovers a complete quality tree. There are always non-expressed attributes and a latent perceived service quality composition.

There are at least two possible ways how to complete a quality tree: (1) to ask additional questions to consumers, (2) to predict non-expressed quality attributes from consumers behaviour and existing data.

Chapter 7: Discussion

7.1 Introduction

The goal of the thesis was to find how can consumers' feedback imply service quality. This chapter addresses the proposed research questions and discusses how the literature reviews, the case studies and the qualitative study presented in this thesis answered these questions. For illustration, figure 7.1 shows the examined relationships introduced in the first chapter.

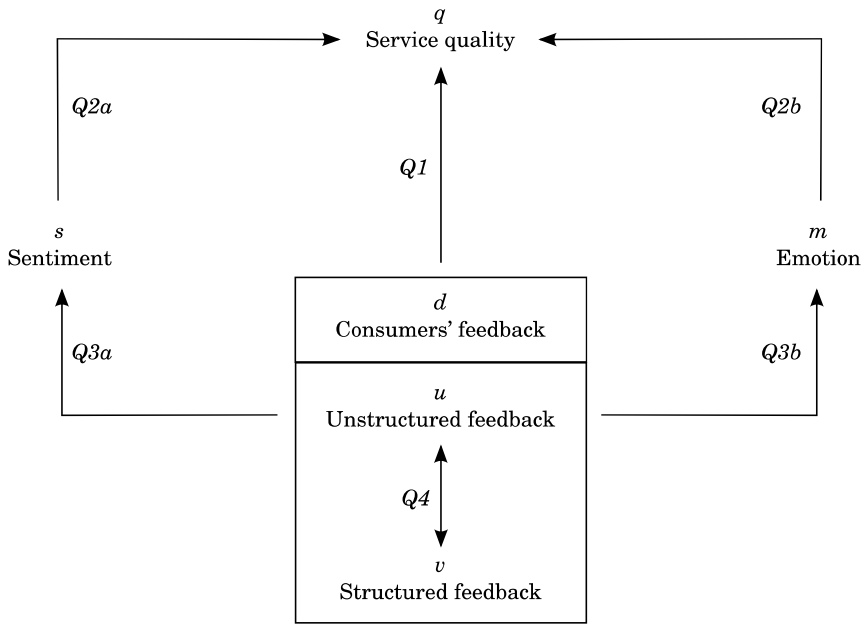

Figure 7.1: Illustration of the problem definition (author)

Q1 How can consumers' feedback *d* imply service quality *q*?

Q2a How can consumers' feedback sentiment *s* imply service quality *q*?

Q2b How can consumers' feedback emotionality *m* imply service quality *q*?

Q3a How can be consumers' feedback sentiment *s* extracted from *d*?

Q3b How can be consumers' feedback emotionality *m* extracted from *d*?

Q4 What is the interplay between structured consumers' feedback *c* and unstructured consumers' feedback *x*?

The research question *Q1* was answered using the research questions Q2 and *Q3*. The both questions have two branches *a* and *b*. The branch *a* refers to sentiment classification and *b* refers to emotion classification.

7.2 Q2a: How can consumers' feedback sentiment imply service quality?

None of the reviewed studies from chapter 3.6 explains the relationship of consumers' feedback to service quality in a sufficient way. Three of the studies tried to map user content to SERVQUAL quality dimensions (Palese, Piccoli 2016; Song, Lee, Yoon, Park 2015; Duan, Cao, Yu, Levy 2013). **The SERVQUAL dimensions were validated in many studies, but they do not fit all contexts**, especially not the e-service area, where, among others, the authors of SERVQUAL identified a new dimensionality. The dimensionality would be even different in every other service branche such as in e-retailing or video-streaming services. In contrast to the previously mentioned studies, (Ashton, Evangelopoulos, Prybutok 2015) used aspect groups that appear naturally in the feedback, but did not frame the study with quality research and they limited the dimensions to complaints categories only. (James, Calderon, Cook 2017) found aspect groups and mapped them to general service dimension of technical and system quality.

From the reviewed studies, as from the service quality literature, a controversy over expectation measuring implies. (Song, Lee, Yoon, Park 2015) captured expectation separately from perceived service performance. According to their work, the expectations are expressed by an overall frequency of aspect mentions. **This approach should not be accepted because of these three reasons**: (a) Expectation is an individual concept and cannot be calculated across all consumers. (b) Frequency of an aspect mention relates only to aspect importance and not to expected quality. (c) Expressed attitude towards service quality is already a result of quality expectation and perception comparison.

The review in chapter 2 showed that service quality is strongly connected with an individual and it is influenced by various factors. It is measured using service quality dimensions. Quality does not have only multiple dimensions, but also multiple levels. And thus, **quality can be expressed in a hierarchical model of quality attributes**, where quality dimensions are on the top level.

Service quality model is not a stable construct, and thus it needs to be tailored for different contexts. Especially, it is not recommended to use models of traditional service quality for e-services. Service consumers perceive a different importance of quality dimensions. The quality dimension is an abstract construct that groups quality attributes of the similar type. General model of service quality is illustrated in the figure 2.1.

Service can also be modelled of service components and their attributes in a hierarchical way. The service component could be another service or an artefact. The service attributes and components are generally called service aspects. The model of a service using opinion mining terminology was constructed in chapter 3.2. The joint model of service quality is illustrated in the figure 7.2. **The joint model poses the background of the research question *Q2***. A sub-question should be formulated as "*What is the character of the relationship between service aspect and quality attribute?"*

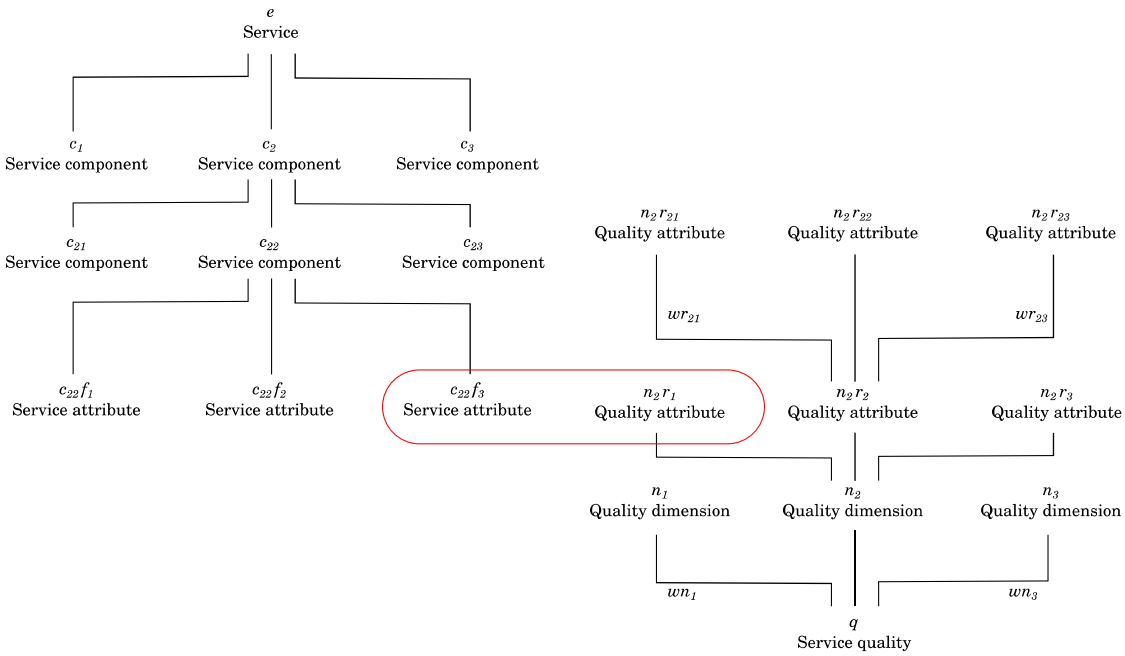

Figure 7.2: Joint model of service quality (author)

Based on the literature review from chapter 2, the most frequent tool for service quality diagnosis is a survey. **Quality surveys are used for measuring levels of agreement with quality evaluation statements.** For example, (Li, Tan, Xie 2002) proposed "Online ordering is simple" as one of the statements from the *competence* quality dimension. It implies from the example that a statement already contains a sentiment with a certain orientation and intensity. Respondents can moderate the intensity of the expressed sentiment or turn its polarity by choosing a different level of agreement with the statement. The formulation of these statements has to be carefully considered. **The main challenge is to reflect particular quality dimensions with expressed service aspects and to evaluate the sentiment value resulting from the statement formulation and the chosen level.** Apart from statements, Likert-type scales are often used for overall service quality evaluation or for evaluation of a particular service dimension. Instead of a whole statement, a simple label such as *speed*, *location* or *support* is provided as a guideline to quality reviewers.

Unstructured consumer feedback **offers more information than structured quality surveys**. The research in chapter 6 found that reviews contain *quality statements*, *comparisons*, *emotion expressions*, *explanations*, *consumers' experience*, *observations*, *preference*, *suggestions*, *structure statements*, *cost and benefit statements* and *summaries*. However, the quality diagnosis based on unstructured feedback is much more challenging.

The literature review from chapter 3 and the case study from chapter 4 proved that it is possible to extract service aspects and consumers' sentimental attitude from feedback. **The task is the same as in the case of close-ended survey statements: How to link service aspects with certain quality dimensions?** However, in this case, it is different because the text is freely formulated by consumers.

There is no arguing that it is possible to extract service aspects from consumers' feedback. It was proved by the case study in chapter 4 and by the reviewed studies from chapter 3.6. In the case study, the aspect was extracted using the topic analysis. **The analysis resulted in two service quality attributes, two core sub-services and two IT sub-services**. Although, according to the broadest e-service quality model (Yang, Fang 2004), the only service quality dimensions that clearly correspond with the model were the *access* and service *support* topics that can be seen as Yang's *communication.* The case study used the same method for quality dimension exploration as Yang. The only difference is that, in case of this study, the data sample was limited to one particular e-service. Therefore, the results cannot be generalised to all e-banking services, but **for this focal service, aspect groups relate directly to quality dimensions**.

Another challenge is that, unlike quality surveys, **the unstructured feedback cannot guarantee the coverage of all dimensions in one particular review**. It is happening due to the fact that **quality is not measured directly but is only self-reported and therefore the quality evaluation relies heavily on memory**. The events that are connected with more intensive emotions are easier to remember for an individual. The statement is supported by the review in chapter 2.5.3 and also by the neuroscience literature (Dolcos, LaBar, Cabeza 2004).

From the service quality diagnosis point of view, **only important or highly affective events are incorporated into service quality review**. An intensive sentiment towards a service may be caused by exceeding or not meeting the expectations of a consumer. The sentimental opinion is then expressed in the text and connected, according to syntactic rules, with a service aspect.

The assumption of important or affective aspect's presence does not mean that quality can be diagnosed only from expressed aspects. The research in chapter 6 found that the **perceived quality, measured by overall rating score, can be strongly influenced by non-expressed aspects**.

Even with all quality attributes expressed in text, there will still be an issue of attributes and dimension weights – the power of influence on overall quality. The service quality literature review described a few attempts of how quality dimensions should be weighted, but they were applied to pre-defined close-ended surveys, not to unstructured quality reviews. The research in chapter 6 resulted in an idea of quality trees.

On account of perceived service quality as an individual concept, it is necessary to diagnose service quality on the same level – an individual's level. Each review should be transformed into a quality model – more specifically to a tree of quality attributes. All individuals' quality trees should be placed in a database where they can be available for a complex quality querying.

Although there are issues with evaluation of quality expressed in a review full text, it has been proved in the literature review, and it is supported by the study in chapter 6, that the aspect sentiment implies service quality. It is only necessary to group aspects in the right quality dimensions and to weight them regarding consumers' individual perception. **Based on these findings, it is possible to answer positively the research question *Q2a* that the sentiment expressed in consumers' feedback implies service quality.**

7.3 Q2b: How can consumers' feedback emotionality imply service quality?

Unlike Q2a, the relationship between feedback emotionality and service quality was simplified. The examination of sentiment in the previous question was conducted regarding the service and quality model structure. Because the relationship between consumers' feedback emotionality and service quality has not been examined in the literature before, this thesis brings an exploratory analysis of the relationship. The case study in chapter 5 presents a statistical model based on binary logistic method.

The model explains the relationship with a high significance. **Service quality expressed as an agreement with the statement that refers to a high level of service quality can be explained by the intensity of modal emotions**. The modal emotion that strongly positively implies service quality is joy. Contrarily, sadness and disgust imply service quality strongly negatively. The findings are supported by the relationship between feedback emotionality and service aspects.

Based on these findings, it is possible to answer positively the research question *Q2b* that the emotionality of consumers' feedback implies service quality.

The relation of the emotionality of particular quality attributes to service quality poses the same problem as the relation of the sentiment of particular quality attributes to service quality discussed in the previous chapter. However, once attributes are weighted and linked with quality dimensions, extracted emotions offer deeper insight into consumers' quality judgements and attitudes towards service.

7.4 Q3a: How can be consumers' feedback sentiment extracted from unstructured feedback?

The part 3.3 answers the research question *Q3a* by reviewing opinion mining techniques for consumers' opinion extraction and classification of sentiment and emotions. The analysis of consumers' feedback can

be undertaken on different levels. The highest level is a whole message, review or email. The lower level is a sentence or a single service aspect. The lower levels provide higher fidelity of service image perceived by a consumer but are more difficult to extract. Service components or features then have to be matched with sentiment target aspects from consumer's feedback. The sentiment is usually classified according to the intensity and orientation of words derived from the lexicon that has a certain relationship to the sentiment target aspect or according to supervised model predictions.

This thesis presents, in chapter 4, a method of extraction on the document level. The review represents a document and is connected with a concrete group of service aspects. The body of the review is the subject of sentiment prediction. The sentiment is predicted using the Naïve Bayes learner trained on the Cornell movie-review database. The approach is supported by the reviewed papers from chapter 3. The same approach was used by (Duan, Cao, Yu, Levy 2013), but on the sentence level. (Song, Lee, Yoon, Park 2015) used dictionary approach on the word/phrase level. And (James, Calderon, Cook 2017) used dictionary approach on the document level.

Based on that, it is possible to claim that there are various ways of how to extract sentiment from consumers' unstructured feedback. The best approach relies on the level of analysis and data preparation.

7.5 Q3b: How can be consumers' feedback emotionality extracted from unstructured feedback?

The case study examined the possibilities of emotionality extraction. The study compared the lexicon approach in which the emotional vector based on associations with Plutchik's eight modal emotions was constructed. The same emotion model was used for supervised machine learning. The study showed that the best results were performed by learners for *surprise*, *fear and anticipation* (99.6% - 97.7%). Worse results were returned by learners for *anger, disgust* and *sadness* (92% - 88%). The worst results were returned by *trust* and *joy* (78% - 54%). The sufficient performance was shown, in this task, by Fuzzy Rule Learner, Gradient Boosted Trees Learner, Random Forest and PNN Learner (DAA), SVM RBF, RProp MLP Learner, Decision Tree Learner, Naive Bayes Learner. The learning showed that the results of emotion learning do not rely on stop-words and that uni-gram features perform slightly better than uni-gram and bi-gram features together and also better than bi-gram features. The lexicon approach showed the significantly worse score.

Based on the case study from chapter 5, it is possible to claim that emotionality can be successfully extracted from consumers' feedback using supervised machine learning. The predicted emotionality was seen as an intensity of the eight modal emotions.

7.6 Q4: What is the interplay between structured and unstructured consumers' feedback?

The qualitative study from chapter 6 examined the interplay more in-depth. It was found that reviewers tend to use review rating as an overall quality metric.

According to the literature review discussion in chapter 2, the high overall rating means that a service quality exceeded expectation or met the expectation of high-quality service. It is an interesting assumption, but it cannot be validated easily. In the analysed reviews, **different strategies for scale rating were observed**. For example, a reviewer rated service quality with five points and used the expression *"perfectly satisfied"* in the summary statement, whereas another reviewer used the expression *"I like"* with the same five-point rating. The possible explanations of the second case are: (a) The expression mirrors the actual meeting of expectations of high-quality service. (b) The reviewer tends to use moderate language. (c) The reviewer tends to rate quality higher. However, these hypotheses were not possible to confirm in this research and thus they are proposed for a future research.

According to the observed compliance of quality rating and evaluation of summary statement, the question *Q4* was answered: **Unstructured feedback can comply with structured feedback in the particular case of expressed summary statement, however different strategies of writing and rating may shift the relationship in both directions.**

For other aspects that were not mentioned in the summary statement, it is hard to confirm the nature of their relationship to quality rating. Although the literature review in chapter 2 claimed statistical significance, it was clear from the observed samples that the relation can be influenced by other factors and that the weights of aspects are hard to extract. On the other hand, in particular cases, the overall quality rating can help to set the aspect weights.

Chapter 8: Conclusion

This thesis focused on the service quality diagnosis from the consumer's point of view. It examined the service that a service provider delivers over information technologies to a large number of consumers, also known as the e-service. The thesis reflects the move towards the digital economy in which it is no longer possible to discuss service quality face to face, but instead, a lot of digital content is generated by service consumers.

The opinion mining is a research discipline that analyses a natural language in textual form and results in information about expressed opinions. This thesis explored sentiment classification, emotion classification and aspect mining methods from this area.

The goal of the thesis was to find *how consumers' feedback can imply service quality* by answering four research questions. The object of the research, in general, was the e-service quality. The subject of the research was the online consumer feedback.

The thesis achieved the goal by answering four research questions. *Q1*, *how can consumers' feedback imply service quality*, is the core question that is explained with use of three sub-questions - *Q2a*, *how can consumers' feedback sentiment imply service quality, Q2b*, *how can consumers' feedback emotionality imply service quality, Q3a, how can be consumers' feedback sentiment extracted from unstructured feedback, Q3b, how can be consumers' feedback emotionality extracted from unstructured feedback,* and *Q4, what is the interplay between structured consumers' feedback and unstructured consumers' feedback.*

Several models for measuring service quality exist, they differ in quality conceptualisation and dimensionality. Researchers from different fields proposed different quality dimensions; for the focal service, e-service quality dimensions were also defined and discussed in the case study in chapter 4. Service quality models and dimensions are described in the second chapter.

The third chapter explained the opinion mining methods, including the aspect extraction technique comparison, (*Q3*) and presented the first critical literature review of the papers that use opinion mining for service quality diagnosis (*Q2* and *Q3*).

In addition to the mentioned literature reviews, this work presented two case studies and one qualitative study. The first case study from chapter 4 examined e-banking feedback using sentiment and topic analysis (*Q2a* and *Q3a*). The second case study from chapter 5 explored emotionality of consumers' feedback (*Q3b*) and its relation to service quality (*Q2b*). The qualitative study from chapter 6 brought the light on *Q4* and *Q1*.

Based on chapter 7 where research questions were discussed, it is possible to say that the goal of the thesis was fulfilled.

Bibliography

ALANEZI, Mohammed Ateeq, KAMIL, Ahmed and BASRI, Shuib, 2010. A proposed instrument dimensions for measuring e-government service quality. *International Journal of U- & E-Service, Science & Technology*. December 2010. Vol. 3, no. 4, p. 1–17.

AMAZON.COM INC., 2017. About Customer Ratings. [online]. 2017. [Accessed 1 February 2017]. Available from: https://www.amazon.com/gp/help/customer/display.html?nodeId=200791020

ANTONS, David, KLEER, Robin and SALGE, Torsten Oliver, 2015. Mapping the Topic Landscape of JPIM , 1984-2013: In Search of Hidden Structures and Development Trajectories. *Journal of Product Innovation Management*. 2015. P. n/a-n/a. DOI 10.1111/jpim.12300.

ASHTON, Triss, EVANGELOPOULOS, Nicholas and PRYBUTOK, Victor R., 2015. Quantitative quality control from qualitative data: control charts with latent semantic analysis. *Quality & Quantity*. 20 May 2015. Vol. 49, no. 3, p. 1081–1099. DOI 10.1007/s11135-014-0036-5.

BIRD, Steven, KLEIN, Ewan and LOPER, Edward, 2009. *Natural Language Processing with Python*. 1 edition. O'Reilly Media. ISBN 9780596516499.

BITNER, Mary Jo, ZEITHAML, Valarie A. and GREMLER, Dwayne D, 2010. Technology's Impact on the Gaps Model of Service Quality. In: *Handbook of Service Science*. Boston, MA : Springer US : Springer. p. 197–218. Service Science: Research and Innovations in the Service Economy 1865-4932 TA -. ISBN 978-1-4419-1627-3.

BLEI, David M. and LAFFERTY, John D., 2007. A correlated topic model of Science. *The Annals of Applied Statistics*. 2007. Vol. 1, no. 1, p. 17–35. DOI 10.1214/07-AOAS136.

BLEI, David M, NG, Andrew Y and JORDAN, Michael I, 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003. Vol. 3, no. Jan, p. 993–1022.

BOULDING, William, KALRA, Ajay, STAELIN, Richard and ZEITHAML, Valarie A., 1993. A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions. *Journal of Marketing Research*. 1993. Vol. 30, no. 1, p. 7–27. DOI 10.2307/3172510.

BRADLEY, M M, GREENWALD, M K, PETRY, M C and LANG, P J, 1992. Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, memory, and cognition*. 1992. Vol. 18, no. 2, p. 379–390. DOI 10.1037/0278-7393.18.2.379.

BRADY, Michael K. and CRONIN JR, J Joseph, 2001. Some New Thoughts on Conceptualizing Perceived Service Quality: A Hierarchical Approach. *Journal of Marketing*. 2001. Vol. 26, no. 3, p. 34–49.

BROWN, Stephen W. and SWARTZ, Teresa A., 1989. A Gap Analysis of Professional Service Quality. *Journal of Marketing*. 1989. Vol. 54, no. 2, p. 92–98.

BROWN, Susan a, VENKATESH, Viswanath and GOYAL, Sandeep, 2014. Expectation Confirmation in Information Systems Research: a Test of Six Competing Models. *MIS Quarterly*. 2014. Vol. 38, no. 3, p. 729-A9.

BRUCKNER, Tomáš, BUCHALCEVOVÁ, Alena, CHLAPE, Dušan, ŘEPA, Václav, STANOVSKÁ, Iva and VOŘÍŠEK, Jiří, 2012. *Tvorba informačních systémů*. ISBN 978-80-247-7903-4.

BURGET, Radim, KARÁSEK, Jan and SMÉKAL, Zdeněk, 2011. Recognition of emotions in Czech newspaper headlines. *Radioengineering*. 2011. Vol. 20, no. 1, p. 39–47.

CADOTTE, Ernest R., WOODRUFF, Robert B. and JENKINS, Roger L., 1987. Expectations and norms in models of consumer satisfaction. *Journal of Marketing Research*. 1987. Vol. 24, no. August, p. 305–14.

CALINSKI, Tadeusz and HARABASZ, Joachim, 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*. 1974. Vol. 3, no. 1, p. 1–27.

CARUANA, Albert and PITT, Leyland, 1997. INTQUAL - an internal measure of service quality and the link between service quality and business performance. *European Journal of Marketing*. 1997. Vol. 31, no. 8, p. 604–616.

CASELLA, George and GEORGE, Edward I., 1992. Explaining the Gibbs Sampler. *The American Statistician*. 1992. Vol. 46, no. 3, p. 167–174.

CHAUDHURI, Arjun, 2006. *Emotion and reason in consumer behavior*. Routledge. ISBN 9781136406898.

CHEN, Danqi and MANNING, Christopher D, 2014. A Fast and Accurate Dependency Parser using Neural Networks. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [online]. 2014. p. 740–750. Available from: https://cs.stanford.edu/~danqi/papers/emnlp2014.pdf

CHOI, WoongChul and YOO, DaeHun, 2009. Software assurance towards better IT service. *Journal of Service Science*. 2009. Vol. 1, no. 1, p. 31–56. DOI 10.1007/s12927-009-0003-1.

CHOUDHURY, Koushiki, 2014. Service quality and word of mouth: a study of the banking sector. *International Journal of Bank Marketing*. 2014. Vol. 32, no. 7, p. 612–627. DOI 10.1108/IJBM-12-2012-0122.

CHOWDHURY, Jhinuk, REARDON, James and SRIVASTAVA, Rajesh, 1998. *Alternative modes of measuring store image: An empirical assessment of structured versus unstructured measures*. 1998.

CRONIN JR, J Joseph and TAYLOR, Steven A, 1992. Measuring Service Quality: A Reexamination and Extension. *Journal of Marketing*. July 1992. Vol. 56, no. 3, p. 55. DOI 10.2307/1252296.

CRONIN JR, J Joseph and TAYLOR, Steven A, 1994. SERVPERF versus SERVQUAL: Reconciling Performance-Based and Perceptions-Minus-Expectations Measurement of Service Quality. *JOURNAL OF MARKETING*. 1994. Vol. 58, no. 1, p. 125–131. DOI 10.2307/1252256.

DABHOLKAR, P. a., THORPE, D. I. and RENTZ, J. O., 1995. A Measure of Service Quality for Retail Stores: Scale Development and Validation. *Journal of the Academy of Marketing Science*. 1995. Vol. 24, no. 1, p. 3–16. DOI 10.1177/009207039602400101.

DAVIS, R D and BAGOZZI, R P, 1989. User acceptance of computer technology: a comparison of two theoretical models. *Management Science*. 1989. Vol. 35, no. 8, p. 22,982.

DOLCOS, Florin, LABAR, Kevin S. and CABEZA, Roberto, 2004. Interaction between the Amygdala and the Medial Temporal Lobe Memory System Predicts Better Memory for Emotional Events. *Neuron*. 2004. Vol. 42, no. June, p. 855–863.

DUAN, Wenjing, CAO, Qing, YU, Yang and LEVY, Stuart, 2013. Mining Online User-Generated Content: Using Sentiment Analysis Technique to Study Hotel Service Quality. In: *2013 46th Hawaii International Conference on System Sciences*. 2013. p. 3119–3128. ISBN 978-1-4673-5933-7.

EL-BAYOUMI, Janice G, 2012. Evaluating IT service quality using SERVQUAL. In: *Proceedings of the ACM SIGUCCS 40th annual conference on Special interest group on university and college computing services - SIGUCCS '12*. New York, New York, USA: ACM Press. 2012. p. 15. SIGUCCS '12. ISBN 9781450314947.

ERKAN, Gunes and OZGUR, Arzucan, 2007. Semi-Supervised Classification for Extracting Protein Interaction Sentences using Dependency Parsing. In: *EMNLP-CoNLL*. 2007. p. 228–237.

EVERITT, B. S., LANDAU, S. and LEESE, M., 2001. *Cluster Analysis*. 4th. London: Arnold Publishers.

FASSNACHT, M. and KOESE, Ibrahim, 2006. Quality of Electronic Services: Conceptualizing and Testing a Hierarchical Model. *Journal of Service Research*. 2006. Vol. 9, no. 1, p. 19–37. DOI 10.1177/1094670506289531.

FINLAYSON, Mark Alan, 2014. Java Libraries for Accessing the PrincetonWordnet: Comparison and Evaluation. In: *Proceedings of the 7th International Global WordNet Conference (GWC 2014)*. 2014. p. 78–85.

GAO, Guodong (Gordon), GREENWOOD, Brad N, AGARWAL, Ritu and MCCULLOUGH, Jeffrey S, 2015. Vocal Minority and Silent Majority: How Do Online Ratings Reflect Population Perceptions of Quality. *MIS Quarterly*. 2015. Vol. 39, no. 3, p. 565–590. DOI 10.2139/ssrn.2629837.

GAO, Guodong, GU, Bin and LIN, Mingfeng, 2006. The Dynamics of Online Consumer Reviews. In: *Proceedings of the International Conference on Web Information Systems Engineering (WISE'06)*. 2006. p. 1–6.

GARVIN, David a., 1984. *What Does "Product Quality" Really Mean?* 1984. ISBN 0019-848X.

GEFEN, David, 2002. Customer loyalty in e-commerce. *Journal of the association for information systems*. 2002. Vol. 3, no. 1, p. 2.

GOOGLE INC., 2017. Google Play Music – Android Apps on Google Play. [online]. 2017. [Accessed 1 February 2017]. Available from: https://play.google.com/store/apps/details?id=com.google.android.music

GORMLEY, C and TONG, Z, 2015. *Elasticsearch: The Definitive Guide*. O'Reilly Media. ISBN 9781449358501.

GOTLIEB, Jerry B, GREWAL, Dhruv and BROWN, Stephen W, 1994. Consumer Satisfaction and Perceived Quality: Complementary or Divergent Constructs? *Journal of Applied Psychology*. 1994. Vol. 79, no. 6, p. 875–885.

GRIFFITHS, Thomas L and STEYVERS, Mark, 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*. 2004. Vol. 101, no. suppl. 1, p. 5228–5235. DOI 10.1073/pnas.0307752101.

GRÖNROOS, Christian, 1984. A Service Quality Model and Its Marketing Implications. *European Journal of Marketing*. 1984. Vol. 18, no. 4, p. 36–44. DOI 10.1108/EUM0000000004784.

GRÖNROOS, Christian, 2000. *Service management and marketing: a customer relationship management approach*. John Wiley & Sons Incorporated.

GRONROOS, Christian and VOIMA, P, 2013. Critical service logic: making sense of value creation and

co-creation. *Journal of the Academy of Marketing Science*. 2013. Vol. 41, no. 2, p. 133–150. DOI 10.1007/s11747-012-0308-3.

GRÜN, Bettina and HORNIK, Kurt, 2009. topicmodels: An R Package for Fitting Topic Models. *Journal Of Statistical Software*. 2009. Vol. 40, no. 13.

HAN, Sang-lin and BAEK, Seung, 2004. Antecedents and Consequences of Service Quality in Online Banking : An Application of the. *Advances in consumer Research*. 2004. Vol. Volume 31, p. 208–214.

HASSAN, Ahmed, 2010. What ' s with the Attitude ? Identifying Sentences with Attitude in Online Discussions. *Computational Linguistics*. 2010. No. October, p. 1245–1255.

HOFMANN, Thomas, 2001. Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*. 2001. Vol. 42, no. 1–2, p. 177–196. DOI 10.1023/A:1007617005950.

HU, Minqing and LIU, Bing, 2004. Mining and summarizing customer reviews. In: *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*. 2004. p. 168–177. ISBN 1581138889.

HU, Nan, PAVLOU, Paul A. and ZHANG, Jennifer, 2014. Can Online Word-of-Mouth Communication Reveal True Product Quality? *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)*. 2014. No. January 2017, p. 175–203. DOI 10.1145/1134707.1134743.

HU, Nan, PAVLOU, Paul A and ZHANG, Jennifer, 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. In: *Proceedings of the 7th ACM conference on Electronic commerce*. ACM. 2006. p. 324–330. ISBN 1595932364.

HU, Nan, ZHANG, Jie and PAVLOU, Paul a., 2009. Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*. 2009. Vol. 52, no. 10, p. 144–147. DOI 10.1145/1562764.1562800.

JAMES, Tabitha L., CALDERON, Eduardo D. Villacis and COOK, Deborah F., 2017. Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Systems with Applications*. 2017. Vol. 71, p. 479–492. DOI 10.1016/j.eswa.2016.11.004.

JIA, Ronnie, REICH, Blaize Horner and PEARSON, J Michael, 2008. IT service climate: An extension to IT service quality research. *Journal of the Association of Information Systems*. 2008. Vol. 9, no. 5, p. 294–320.

JIANG, J J, KLEIN, G and CARR, C L, 2002. Measuring information system service quality: Servqual from the other side. *MIS QUARTERLY*. 2002. Vol. 26, no. 2, p. 145–166. DOI 10.2307/4132324.

JIANG, Ling, JUN, Minjoon and YANG, Zhilin, 2016. Customer-perceived value and loyalty: how do key service quality dimensions matter in the context of B2C e-commerce? *Service Business*. 2016. Vol. 10, no. 2, p. 301–317. DOI 10.1007/s11628-015-0269-y.

JINGJUN, Xu, BENBASAT, Izak and CENFETELLI, Ronald T, 2013. Integrating Service Quality with System and Information Quality: An Empirical Test in the E-Service Context. *Mis Quarterly*. 2013. Vol. 37, no. 3, p. 777–794.

JOSHI, Mahesh, DAS, Dipanjan, GIMPEL, Kevin and SMITH, a. Noah, 2010. Movie Reviews and Revenues: An Experiment in Text Regression. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010. p. 293–296. ISBN 1-932432-65-5.

KAHNEMAN, Daniel and TVERSKY, Amos, 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*. 1979. Vol. 47, no. 2, p. 263–292.

KANG, Daekook and PARK, Yongtae, 2014. Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications*. 2014. Vol. 41, p. 1041–1050. DOI 10.1016/j.eswa.2013.07.101.

KEERTHI, S. S., SHEVADE, S. K., BHATTACHARYYA, C. and MURTHY, K. R. K., 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*. 2001. Vol. 13, no. 3, p. 637–649. DOI 10.1162/089976601300014493.

KETTINGER, W J and LEE, C C, 1994. Perceived Service Quality and User Satisfaction with the Information Services Function. *DECISION SCIENCES*. 1994. Vol. 25, no. 5–6, p. 737–766. DOI 10.1111/j.1540-5915.1994.tb01868.x.

KETTINGER, W J and LEE, C C, 1997. Pragmatic perspectives on the measurement of information systems service quality. *MIS QUARTERLY*. 1997. Vol. 21, no. 2, p. 223–240. DOI 10.2307/249421.

KOH, Noi Sian, HU, Nan and CLEMONS, Eric K., 2010. Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*. 2010. Vol. 9, no. 5, p. 374–385. DOI 10.1016/j.elerap.2010.04.001.

KOZAK, Marcin, 2012. "A Dendrite Method for Cluster Analysis" by Caliński and Harabasz: A Classical Work that is Far Too Often Incorrectly Cited. *Communications in Statistics - Theory and Methods*. 15 June 2012. Vol. 41, no. 12, p. 2279–2280. DOI 10.1080/03610926.2011.560741.

KRITIKOS, Kyriakos, PERNICI, Barbara, PLEBANI, Pierluigi, CAPPIELLO, Cinzia, COMUZZI, Marco, BENRERNOU, Salima, BRANDIC, Ivona, ESZ, Attila Kert, PARKIN, Michael and CARRO, Manuel, 2013. A survey on service quality description. *ACM Computing Surveys*. 2013. Vol. 46, no. 1, p. 1–58. DOI 10.1145/2522968.2522969.

LADHARI, Riadh, 2008. Alternative measures of service quality: a review. *Managing Service Quality*. 2008. Vol. 18, no. 1, p. 65–86. DOI 10.1108/09604520810842849.

LADHARI, Riadh, 2010. Developing e-service quality scales: A literature review. *Journal of Retailing and Consumer Services*. 2010. Vol. 17, no. 6, p. 464–477. DOI 10.1016/j.jretconser.2010.06.003.

LEE, Sangno, BAKER, Jeff, SONG, Jaeki and WETHERBE, James C., 2010. An Empirical Comparison of Four Text Mining Methods. *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences (HICSS'10)*. 2010. P. 1–10. DOI 10.1109/HICSS.2010.48.

LEONG, Lai Ying, HEW, Teck Soon, LEE, Voon Hsien and OOI, Keng Boon, 2015. An SEM-artificial-neural-network analysis of the relationships between SERVPERF, customer satisfaction and loyalty among low-cost and full-service airline. *Expert Systems with Applications*. 2015. Vol. 42, no. 19. DOI 10.1016/j.eswa.2015.04.043.

LEPMETS, Marion, CATER-STEEL, Aileen, GACENGA, Francis and RAS, Eric, 2012. Extending the IT service quality measurement framework through a systematic literature review. *Journal of Service Science Research*. June 2012. Vol. 4, no. 1, p. 7–47. DOI http://dx.doi.org/10.1007/s12927-012-0001-6.

LI, Hongxiu and SUOMI, Reima, 2009. A Proposed Scale for Measuring E-service Quality. *International Journal of u- and e-Service, Science and Technology*. 2009. Vol. 2, no. 1, p. 1–10.

LI, Y N, TAN, K C and XIE, M, 2002. Measuring web-based service quality. *Total Quality Management*. August 2002. Vol. 13, no. 5, p. 685–700. DOI 10.1080/0954412022000002072.

LIM, Ji Yeon, YOON, Jae Yoel, KIM, Lee Joon and KIM, Ung Mo, 2012. Information Extraction of Review Using LIWC. *International Journal of Future Computer and Communication*. 2012. Vol. 1, no. 2, p. 91–94.

LINDER-PELZ, Susie, 1982. Toward a theory of patient satisfaction. *Social science & medicine*. 1982. Vol. 16, no. 5, p. 577–582.

LIU, Bing, 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*. 2012. Vol. 5, no. 1, p. 1–167.

LIU, Bing, 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. First edit. New York, NY, USA: Cambridge University Press. ISBN 978-1-316-29832-9.

LIU, Qian, LIU, Bing, ZHANG, Yuanlin, KIM, Doo Soon and GAO, Zhiqiang, 2016. Improving Opinion Aspect Extraction Using Semantic Similarity and Aspect Associations. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*. 2016. P. 2986–2992.

LO, Shuchuan, 2008. Web service quality control based on text mining using support vector machine. *Expert Systems with Applications*. 2008. Vol. 34, no. 1, p. 603–610. DOI 10.1016/j.eswa.2006.09.026.

LÓPEZ, Andrea, DETZ, Alissa, RATANAWONGSA, Neda and SARKAR, Urmimala, 2012. What patients say about their doctors online: A qualitative content analysis. *Journal of General Internal Medicine*. 2012. Vol. 27, no. 6, p. 685–692. DOI 10.1007/s11606-011-1958-4.

LU, Bin, OTT, Myle, CARDIE, Claire and TSOU, Benjamin K., 2011. Multi-aspect sentiment analysis with topic models. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2011. p. 81–88. ISBN 9780769544090.

LU, Yue, MEI, Qiaozhu and ZHAI, Cheng Xiang, 2011. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval*. 2011. Vol. 14, no. 2, p. 178–203. DOI 10.1007/s10791-010-9141-9.

LUKYANENKO, Roman, PARSONS, Jeffrey and WIERSMA, Yolanda F., 2014. The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content. *Information Systems Research*. 2014. Vol. 25, no. 4, p. 669–689. DOI 10.1287/isre.2014.0537.

LUSCH, Robert F and VARGO, Stephen L, 2006. The Service-Dominant Logic of Marketing: Dialog. *Debate, and Directions, ME Sharpe, Armonk, NY*. 2006. Vol. 10.

MA, Jianbin, XUE, Bing and ZHANG, Mengjie, 2016. A Profile-Based Authorship Attribution Approach to Forensic Identification in Chinese Online Messages. In: *Intelligence and Security Informatics*. Springer International Publishing. p. 33–52. ISBN 978-3-319-31863-9.

MALHOTRA, Naresh K, 1982. Information Load and Consumer Decision Making. *Journal of Consumer Research*. 1 March 1982. Vol. 8, no. 4, p. 419–430.

MANNING, Christopher D, BAUER, John, FINKEL, Jenny, BETHARD, Steven J, SURDEANU, Mihai and MCCLOSKY, David, 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* [online]. 2014. p. 55–60. ISBN 9781941643006. Available from: http://aclweb.org/anthology/P14-5010

MAURI, Aurelio G., MINAZZI, Roberta and MUCCIO, Simonetta, 2013. A Review of Literature on the Gaps Model on Service Quality: A 3-Decades Period: 1985–2013. *International Business Research*. 2013. Vol. 6, no. 12, p. 134–144. DOI 10.5539/ibr.v6n12p134.

MCCALLUM, Andres and NIGAM, Kamal, 1998. A Comparison of Event Models for Naive Bayes Text Classification. In: *AAAI/ICML-98 Workshop on Learning for Text Categorization*. 1998. p. 41–48.

MCDONALD, Malcolm, FROW, Pennie and PAYNE, Adrian, 2011. *Marketing Plans for Services : A Complete Guide* [online]. Hoboken, UNKNOWN: Wiley. ISBN 9781119951865. Available from: http://ebookcentral.proquest.com/lib/vsep/detail.action?docID=698012

MELE, Cristina and POLESE, Franesco, 2011. Key Dimensions of Service Systems in Value-Creating Networks. In: *Service Science: Research and Innovation in the Service Economy*. p. 37–59. ISBN 978-1-4419-8269-8.

MEUTER, Matthew L, OSTROM, Amy L, ROUNDTREE, Robert I, BITNER, Mary Jo and ENCOUNTERS, Service, 2000. Self-Service Technologies: Understanding Customer Satisfaction with Technoiogy-Based Service Encounters. *JOURNAL OF MARKETING*. 2000. Vol. 64, no. 3, p. 50–64. DOI 10.1509/jmkg.64.3.50.18024.

MIKOLOV, Tomas, CORRADO, Greg, CHEN, Kai and DEAN, Jeffrey, 2013. Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of the International Conference on Learning Representations (ICLR 2013)*. 2013. p. 1–12. ISBN 1532-4435.

MILLER, George A., 1995. WordNet: A Lexical Database for English. *Communications of the ACM*. 1995. Vol. 38, no. 11, p. 39–41.

MILLIGAN, G. W. and COOPER, M. C., 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. 1985. Vol. 50, no. 2, p. 159–179.

MOHAMMAD, Saif M. and TURNEY, Peter D., 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*. 2013. Vol. 29, no. 3, p. 436–465. DOI 10.1111/j.1467-8640.2012.00460.x.

MUDAMBI, Susan M and SCHUFF, David, 2010. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*. 2010. Vol. 34, no. 1, p. 185–200. DOI Article.

NEWMAN, David and WELLING, Max, 2009. Distributed Algorithms for Topic Models. *Journal of Machine Learning Research*. 2009. Vol. 10, p. 1801–1828.

OXFORD DICTIONARIES, 2018. *Definition of service in English by Oxford Dictionaries* [online]. 2018. Available from: https://en.oxforddictionaries.com/definition/civic

PALESE, Biagio and PICCOLI, Gabriele, 2016. Online Reviews as a Measure of Service Quality. In: *2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium, Dublin 2016*. 2016.

PANG, B. and LEE, L., 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* [online]. 2004. p. 271. Available from: http://portal.acm.org/citation.cfm?id=1218990

PANG, Bo, LEE, Lillian and VAITHYANATHAN, Shivakumar, 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. [online]. 2002.

DOI http://dx.doi.org/10.3115/1118693.1118704. Available from: http://arxiv.org/abs/cs/0205070

PARASURAMAN, A., ZEITHAML, Valarie A. and MALHOTRA, Arvind, 2005. E-S-QUAL A Multiple-Item Scale for Assessing Electronic Service Quality. *Journal of Service Research*. 2005. Vol. 7, no. X, p. 1–21. DOI 10.1177/1094670504271156.

PARASURAMAN, A, L, Berry L and ZEITHAML, Valarie A., 1991. Refinement and Reassessment of the SERVQUAL Scale. *JOURNAL OF RETAILING*. 1991. Vol. 67, no. 4, p. 420–450.

PARASURAMAN, A, ZEITHAML, Valarie A. and BERRY, Leonard L, 1988. SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality. *JOURNAL OF RETAILING*. 1988. Vol. 64, no. 1, p. 12–40.

PENNEBAKER, James W, MEHL, Matthias R and NIEDERHOFFER, Kate G, 2003. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology*. 2003. Vol. 54, p. 547–77. DOI 10.1146/annurev.psych.54.101601.145041.

PITT, L F, WATSON, R T and KAVAN, C B, 1997. Measuring information systems service quality: Concerns for a complete canvas. *MIS QUARTERLY*. 1997. Vol. 21, no. 2, p. 209–221. DOI 10.2307/249420.

PITT, Leyland F., WATSON, R T and KAVAN, C B, 1995. SERVICE QUALITY - A MEASURE OF INFORMATION-SYSTEMS EFFECTIVENESS. *MIS QUARTERLY*. 1995. Vol. 19, no. 2, p. 173–187. DOI 10.2307/249687.

PLATT, John C., 1998. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*. 1998. ISBN 0262194163 9780262194167.

PLUTCHIK, Robert, 1980. A general psycho-evolutionary theory of emotion. *Theories of Emotions*. 1980. Vol. 1, p. 3–31.

PLUTCHIK, Robert, 2001. The nature of emotions. *American Scientist*. 2001. Vol. 89, no. 4, p. 344–350.

POSNER, Jonathan, RUSSELL, James A and PETERSON, Bradley S, 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*. 2005. Vol. 17, p. 715–734. DOI 10.10170S0954579405050340.

QIU, Guang, LIU, Bing, BU, Jiajun and CHEN, Chun, 2009. Expanding domain sentiment lexicon through double propagation. *IJCAI International Joint Conference on Artificial Intelligence*. 2009. P. 1199–1204.

QIU, Guang, LIU, Bing, BU, Jiajun and CHEN, Chun, 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics* [online]. 2011. Vol. 37, no. 1, p. 9–27. DOI 10.1162/coli_a_00034. Available from: http://www.mitpressjournals.org/doi/10.1162/coli_a_00034

QU, Zhe, ZHANG, Han and LI, Haizheng, 2008. Determinants of online merchant rating: Content analysis of consumer comments about Yahoo merchants. *Decision Support Systems*. 2008. Vol. 46, no. 1, p. 440–449. DOI 10.1016/j.dss.2008.08.004.

QUIRK, Chris, MENEZES, Arul and CHERRY, Colin, 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In: *43rd Annual Meeting of the Association for Computational Linguistics*. 2005. p. 271–279. ISBN 1932432515.

REDCHUK, Andrés, 2010. *Service Quality Measurement: a New Methodology*. Rey Juan Carlos University.

RIBEIRO, Filipe N., ARAÚJO, Matheus, GONÇALVES, Pollyanna, ANDRÉ GONÇALVES, Marcos and BENEVENUTO, Fabrício, 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* [online]. 2016. Vol. 5, no. 1, p. 1–30. DOI 10.1140/epjds/s13688-016-0085-1. Available from: http://dx.doi.org/10.1140/epjds/s13688-016-0085-1

ROBINSON, Regan, GOH, Tiong Thye and ZHANG, Rui, 2012. Textual factors in online product reviews: A foundation for a more influential approach to opinion mining. *Electronic Commerce Research*. 2012. Vol. 12, no. 3, p. 301–330. DOI 10.1007/s10660-012-9095-7.

RUIJIN, Zhang and YUNCHANG, Zhang, 2010. Service Quality, Customer Satisfaction and Customer Loyalty of Mobile Communication Industry in China. . 2010. Vol. 20, no. 3, p. 269–277.

SALAS-ZARATE, Maria del Pilar, LOPEZ-LOPEZ, Estanislao, VALENCIA-GARCIA, Rafael, AUSSENAC-GILLES, Nathalie, ALMELA, Angela and ALOR-HERNANDEZ, Giner, 2014. A study on LIWC categories for opinion mining in Spanish reviews. *Journal of Information Science*. 2014. Vol. 40, no. 6, p. 749–760. DOI 10.1177/0165551514547842.

SANTOS, Jessica, 2003. E-service quality: a model of virtual service quality dimensions. *Managing Service Quality: An International Journal*. 1 June 2003. Vol. 13, no. 3, p. 233–246. DOI 10.1108/09604520310476490.

SCHEMBRI, S. and SANDBERG, J., 2011. The experiential meaning of service quality. *Marketing Theory*. 2011. Vol. 11, p. 165–186. DOI 10.1177/1470593111403221.

SCHERER, Klaus R, 2005. What are emotion? And how can they be measured? *Social Science Information Sur Les Sciences Sociales*. 2005. Vol. 44, no. 4, p. 695–729. DOI 10.1177/0539018405058216.

SEBASTIANI, Fabrizio, 2002. Machine learning in automated text categorization. *ACM Computing Surveys*. 2002. Vol. 34, no. 1, p. 1–47. DOI 10.1145/505282.505283.

SETH, Nitin, DESHMUKH, S G and VRAT, Prem, 2005. Service quality models: a review. *International Journal of Quality & Reliability Management*. 1 December 2005. Vol. 22, no. 9, p. 913–949. DOI 10.1108/02656710510625211.

SONG, Bomi, LEE, Changyong, YOON, Byungun and PARK, Yongtae, 2015. Diagnosing service quality using customer reviews: an index approach based on sentiment and gap analyses. *Service Business*. 2015. DOI 10.1007/s11628-015-0290-1.

STENT, Amanda, CHOI, Jinho D, ST, West, ST, West and YORK, New, 2015. It Depends : Dependency Parser Comparison Using A Web-based Evaluation Tool. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015. p. 387–396. ISBN 9781941643723.

STRAUSS, A L and CORBIN, J M, 1990. *Basics of qualitative research: grounded theory procedures and techniques*. Sage Publications. ISBN 9780803932500.

TABOADA, Maite, BROOKE, Julian, TOFILOSKI, Milan, VOLL, Kimberly and STEDE, Manfred, 2011. Lexicon-Based Methods for Sentiment Analysis. *Association for Computational Linguistics*. 2011. Vol. 37, no. 2.

TAUSCZIK, Y.R. and PENNEBAKER, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*. 2010. Vol. 29, no. 1, p. 24–54. DOI 10.1177/0261927X09351676.

TAYLOR, Sharon, IQBAL, Majid and NIEVES, Michael, 2007. *ITIL Service strategy*. Stationery Office. IT infrastructure library. ISBN 9780113310456.

TAYLOR, Sharon, LLOYD, Vernon and RUDD, Colin, 2007. *ITIL Service design*. Stationery Office. IT infrastructure library. ISBN 9780113310470.

TSANG, Alex S L and PRENDERGAST, Gerard, 2009. Is a "star" worth a thousand words? *European Journal of Marketing*. 2009. Vol. 43, no. 11/12, p. 1269–1280. DOI 10.1108/0309060109898.

TSANG, N.K.F., LAI, M.T.H. and LAW, R., 2010. Measuring E-service quality for online travel agencies. *Journal of Travel and Tourism Marketing*. 2010. Vol. 27, no. 3, p. 306–323. DOI 10.1080/10548401003744743.

VAN BON, Jan, JONG, Arjen de, KOLTHOF, Axel, PIEPER, Mike, ROZEMEIJER, Eric, TJASSING, Ruby, VAN DER VEEN, Annelies and VERHEIJEN, Tieneke, 2007. *IT service management: an introduction*. First edit. Van Haren Publishing, Zaltbommel. ISBN 978-90-8753-051-8.

VANDYKE, T P, KAPPELMAN, L A and PRYBUTOK, V R, 1997. Measuring information systems service quality: Concerns on the use of the SERVQUAL questionnaire. *MIS QUARTERLY*. 1997. Vol. 21, no. 2, p. 195–208. DOI 10.2307/249419.

VANPARIA, Bhavesh and TSOUKATOS, Evangelos, 2013. COMPARISON OF SERVQUAL, SERVPERF, BSQ AND BANKQUAL SCALE IN BANKING SECTOR. *Confronting Contemporary Business Challenges Through Management Innovation*. 2013. P. 2405–2430.

VENCOVSKY, Filip, BRUCKNER, Tomas and SPERKOVA, Lucie, 2016. Customer Feedback Analysis: Case of E-banking Service. In: *10th European Conference on Information Systems Management: ECISM 2016*. 2016. p. 404.

VENCOVSKY, Filip, LUCAS, Benjamin, MAHR, Dominik and LEMMINK, Jos G. A. M., 2017. Comparison of Text Mining Techniques for Service Aspect Extraction. In: *Proceedings of the 4th European Conference on Social Media ECSM 2017*. Vilnius, Lithuania: Academic Conferences and Publishing International Limited. 2017. p. 297–307. ISBN 978-1-911218-47-0.

VOŘÍŠEK, J and BASL, J, 2008. *Principy a modely řízení podnikové informatiky*. Oeconomica. ISBN 9788024514406.

VOŘÍŠEK, Jiří and POUR, Jan, 2012. *Management podnikové informatiky*. Praha: Professional Publishing. ISBN 9788074311024.

WALTERS, J. Hart, 1961. Structured or Unstructured Techniques? *Journal of Marketing*. 1961. Vol. 24, no. 4, p. 58.

WANG, Yi, BAI, Hongjie, STANTON, Matt, CHEN, Wen Yen and CHANG, Edward Y., 2009. PLDA: Parallel latent dirichlet allocation for large-scale applications. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2009. p. 301–314. ISBN 3642021573.

WIEBE, Janyce, WILSON, Theresa and CARDIE, Claire, 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*. 2005. Vol. 39, no. 2–3, p. 165–210. DOI 10.1007/s10579-005-7880-9.

WIKIMEDIA COMMONS, 2011. Plutchik's wheel of emotions. [online]. 2011. Available from:

https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg

WOLFINBARGER, Mary and GILLY, Mary C., 2003. eTailQ: Dimensionalizing, measuring and predicting etail quality. *Journal of Retailing*. 2003. Vol. 79, no. 3, p. 183–198. DOI 10.1016/S0022-4359(03)00034-4.

XIA, Lan and BECHWATI, Nada Nasr, 2010. Word of Mouse: The role of cognitive personalization in online consumer reviews. *Journal of Interactive Advertising*. 2010. Vol. 9, no. 1, p. 3–13. DOI 10.1080/15252019.2008.10722143.

YANG, Zhilin and FANG, Xiang, 2004. Online service quality dimensions and their relationships with satisfaction: A content analysis of customer reviews of securities brokerage services. *International Journal of Service Industry Management*. 2004. Vol. 15, no. 3, p. 302–326. DOI 10.1108/09564230410540953.

YANG, Zhilin, JUN, Minjoon and PETERSON, Robin T., 2004. Measuring customer perceived online service quality: Scale development and managerial implications. *International Journal of Operations & Production Management*. 2004. Vol. 24, no. 11, p. 1149–1174. DOI 10.1108/01443570410563278.

YARIMOGLU, Emel Kursunluoglu, 2014. A Review on Dimensions of Service Quality Models. *Journal of Marketing Management*. 2014. Vol. 2, no. 2, p. 79–93.

YARIMOGLU, Emel Kursunluoglu, 2015. A Review of Service and E-Service Quality Measurements: Previous Literature and Extension. *Journal of Economic and Social Studies*. 2015. Vol. 5, no. 1, p. 169–201.

ZEITHAML, Valarie A., PARASURAMAN, A, MALHOTRA, A and CENTRAL, Proquest, 2002. Service quality delivery through Web sites: A critical review of extant knowledge. *JOURNAL OF THE ACADEMY OF MARKETING SCIENCE*. 2002. Vol. 30, no. 4, p. 362–375. DOI 10.1177/009207002236911.

ZHANG, X.a, YU, Y.b, LI, H.c and LIN, Z.d, 2016. Sentimental interplay between structured and unstructured user-generated contents. *Online Information Review*. 2016. Vol. 40, no. 1, p. 119–145. DOI 10.1108/OIR-04-2015-0101.

ZHANG, Xianfeng, LUO, Jifeng and LI, Qi, 2012. Do different reputation systems provide consistent signals of seller quality: A canonical correlation investigation of Chinese C2C marketplaces. *Electronic Markets*. 2012. Vol. 22, no. 3, p. 155–168. DOI 10.1007/s12525-012-0092-4.

ZHU, YanChun and ZHANG, Wei, 2012. Can average score of online reviews reveal product's real quality? *Proceedings - 2012 International Conference on Management of e-Commerce and e-Government, ICMeCG 2012*. 2012. P. 30–33. DOI 10.1109/ICMeCG.2012.94.

ZHUANG, Li, JING, Feng and ZHU, Xiao-Yan, 2006. Movie review mining and summarization. *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06* [online]. 2006. P. 43. DOI 10.1145/1183614.1183625. Available from: http://portal.acm.org/citation.cfm?doid=1183614.1183625

Glossary

Emotion	Emotion is a kind of homeostatic process in which behaviour mediates progress toward equilibrium (Plutchik 2001).
Feeling	Feeling is a subjective cognitive representation, reflecting a unique experience of mental and bodily changes in the context of being confronted with a particular event (Scherer 2005).
Online review	Online reviews are a common form of socialized data, representing spontaneously shared opinions by customers on review platforms (Mudambi, Schuff 2010).
Opinion mining	The opinion mining is a field of study that analyses people's opinions, sentiment, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text (Liu 2015).
Satisfaction	Satisfaction is defined as primarily and affective response to a specific consumptive experience (Linder-Pelz 1982) or an emotional response (Cadotte, Woodruff, Jenkins 1987).
Sentiment words	Natural aspects as they are words in a language for expressing positive or negative sentiments. (Liu 2015)
Service aspect	A service aspect is a component or attribute of a service that customers see as crucial when assessing the service. Modified from (Song, Lee, Yoon, Park 2015) to fit name conventions from (Liu 2015). See chapter 3.2.
Text classification	Text classification (or text categorization) is to label the natural language texts with thematic categories from a pre- defined set (Sebastiani 2002).
Topic model	Topic model is probabilistic latent variable model of documents that exploit the correlations among the words and latent semantic themes (Blei, Lafferty 2007).

List of figures

List of tables

Appendices

Study	**(Lo 2008)**	**(Duan, Cao, Yu, Levy 2013)**	**(Ashton, Evangelopoulos, Prybutok 2015)**	**(Song, Lee, Yoon, Park 2015)**	**(Palese, Piccoli 2016)**	**(James, Calderon, Cook 2017)**
Object	Web service quality	Hotel service quality	Online retail service quality	Mobile navigation service quality	Online retail service quality	Healthcare service quality
Subject	Messages for a web manager on a discussion forum	Online reviews	Open-ended question from survey after the service cancellation	Online reviews	Online reviews	Online reviews
Quality measurement	Complain rate / p-chart	Service performance	Quality attributes / p-chart	Expectation confirmation / P (weighted aspect sentiment polarity) - E (total relative aspect frequency)	Service performance	Quality attributes
Dimensionality/ Classes	Technical problems, Complaints, Other	SERVQUAL (tangibles, reliability, assurance, responsiveness, empathy)	Labelled clusters (The service, Wait time too long for product to arrive (mail), Delays in shipping and receiving, Product availability waiting time is too long, The customer wants a more extensive selection, Received damaged products)	Service aspect hierarchy identified by domain experts	SERVQUAL (tangibles, reliability, assurance, responsiveness, empathy)	Systems, interpersonal, and technical
Level	Document / one class	Sentence / one class	Document / multiple classes	Word	Sentence (aspect), Document (sentiment)	Document
Methods	Supervised learning / SVM classification, Keyword extraction / TF-IDF	Supervised learning / Naïve Bayes classification for sentiment and SERVPERF dimensions	Unsupervised learning / LSA classification	Aspect extraction based on dictionary and part of speech, sentiment classification based on dictionary	Multi-aspect sentiment analysis Topic analysis / Gibbs sampling and Latent Dirichlet Allocation with seed words	Topic analysis / Latent Dirichlet Allocation aspect mining, Dictionary base sentiment classification

Appendix I: Literature review of opinion mining approaches in service quality field summary (chapter 3.6, p. 37)