FAKULTÄT FÜR !NFORMATIK

Faculty of Informatics

Opinion Mining

Sentiment Analysis for Cultural Institutions

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Jörg Dubsky
Matrikelnummer 0725059

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Wolfdieter Merkl

Wien, 30.09.2012

(Unterschrift Jörg Dubsky) (Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Opinion Mining

Sentiment Analysis for Cultural Institutions

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Jörg Dubsky
Registration Number 0725059

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Wolfdieter Merkl

Vienna, 30.09.2012

(Signature of Author) (Signature of Advisor)

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.ac.at

Erklärung zur Verfassung der Arbeit

Jörg Dubsky
Seckendorfstraße 4/2/19, 1140 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Jörg Dubsky)

Acknowledgements

This work would not have been possible without the support from my advisor Prof. Dr. Wolfdieter Merkl under whose guidance I chose this topic. He was abundantly helpful and has assisted me in numerous ways.

A special thank goes to my colleague Christina Stödtner. The discussions and coffee breaks I had with her were invaluable.

I am grateful to all my friends, most importantly to Christian, Markus and Daniel for being more than friends over years.

My final words go to my family. I want to thank my father Dietmar, my sister Heike and my brother-in-law Jan for their unconditional support and love in whatever I pursue. Finally, I am so proud of my niece Amelie which always brings a smile to my face.

Abstract

The increasing distribution of social networks in the Web and the associated active participation of people results to a strong increase of so called "user-generated content". Internet users publish and share personal information such as age, education, employment, relationship status, interests, backgrounds and photos on various platforms. Another part is the writing of personal experiences and opinions about various topics. Obviously, this is also true for cultural institutions. Museums and their exhibitions are evaluated and reviewed from many points of view. This information represents an important aspect of the human decision process. At the present time many people use the Internet to obtain quickly reviews about products, movies, music and leisure activities. Other opinions can therefore strongly influence the decision to purchase, or referring to cultural institutions, the choice to visit a museum. However, to obtain the most complete and objective opinion of several people, a number of recessions is necessary. However, the search and reading is an exhausting and time-consuming task.

To simplify this process, so-called sentiment analysis systems (also known as opinion mining) are used. It will automatically scan documents on subjective opinions and filter these out. This can be accomplished at three levels of detail: The entire document, individual sentences within a document or individual aspects within sentences can be extracted and evaluated as positive or negative. Hence, such systems promise not only benefits to consumers, but also profit and non-profit organizations can benefit from the large advantages.

This master's thesis developed a "Feature-Based Sentiment Analysis" model for cultural institutions. German-language text documents from a variety of social networks and online communities are analyzed. Initially, discussed aspects and properties of museums (e.g., exhibitions, entrance fee, museum shop, restaurant, etc.) are extracted. Afterwards, the corresponding subjective opinions are determined (positive, neutral, negative). The designed system uses algorithms from the field of machine learning and semantic rules to automate this process. The developed prototype visualize the extracted topics with the help of diagrams and allows the summarization of a single cultural institution and the comparison of two cultural institutions by their opinionated topics. The evaluation results in a precision of 0.6170 and recall of 0.6744. In conclusion, improvements in the topic extraction module are needed to enhance the whole system.

Kurzfassung

Die stetig wachsende Verbreitung von sozialen Netzwerken im Internet und die damit einhergehende aktive Beteiligung der Menschen, führte zu einem starken Anstieg des sogenannten "user-generated content". Internetnutzer veröffentlichen und teilen persönliche Informationen wie Alter, Ausbildung, Arbeit, Beziehungsstatus, Interessen, Herkunft und Fotos auf verschiedensten Plattformen. Ein weiterer Bestandteil ist das Verfassen von persönlichen Erfahrungen und Meinung über verschiedenste Themen. Dies gilt natürlich auch für kulturelle Institutionen. Museen und deren Ausstellungen werden aus vielerlei Gesichtspunkten beurteilt und rezensiert. Diese Informationen stellen einen wichtigen Aspekt bei dem menschlichen Entscheidungsprozess dar. In der heutigen Zeit nutzen viele Menschen das Internet um schnellst möglichst Erfahrungsberichte über Produkte, Filme, Musik und Freizeitaktivitäten zu erhalten. Die Meinung anderer kann folglich die eigene Kaufentscheidung oder, bezugnehmend auf kulturelle Institutionen, die Wahl des zu besuchenden Museums stark beeinflussen. Um jedoch eine möglichst vollständige und objektive Meinung mehrere Menschen zu erhalten, ist eine Vielzahl von Rezessionen notwendig. Jedoch ist die Suche und das Lesen eine anstrengende und zeitintensive Aufgabe.

Um diesen Prozess für den Menschen zu vereinfachen, werden sogenannten Sentiment Analysis Systeme (auch bezeichnet als Opinion Mining) verwendet. Dabei wird in Textdokumenten automatisiert nach subjektiven Meinungen gesucht und gefiltert. Dies kann auf drei Detaillierungsgraden vollzogen werden: Das gesamte Dokument, einzelne Sätze innerhalb eines Dokumentes oder einzelne Aspekte innerhalb von Sätzen können extrahiert und als positiv oder negativ bewertet werden. Solche Systeme bieten folglich nicht nur den Verbrauchern, sondern auch profit und non-profit Organisationen große Vorteile.

Diese Master's Thesis entwickelt ein "Feature-Based Sentiment Analysis" Model für kulturelle Institutionen. Dabei werden deutschsprachige Textdokumente aus verschiedensten sozialen Netzwerken und Online-Communitys analysiert. Dabei werden zunächst besprochene Aspekte und Eigenschaften von Museen (z.B. Ausstellungen, Eintrittspreis, Museumsshop, etc.) extrahiert. Anschließend werden die dazugehörigen, subjektiven Meinungen (positiv, neutral, negativ) bestimmt. Das konzipierte System verwendet Algorithmen aus dem Bereich des maschinellen Lernens und ein semantisches Regelwerk um diesen Prozess automatisiert zu bearbeiten. Der entwickelte Prototyp visualisiert die extrahierten Themen mittels Diagrammen und ermöglicht somit die Zusammenfassung einzelner kultureller Institutionen sowie den Vergleich zweier Museen anhand der bewerteten Aspekte. Die Evaluierung des Gesamtsystems ergab einen Precision-Wert von 0,6170 und einen Recall-Wert von 0,6744.

Contents

CHAPTER 1

Introduction

1.1 Motivation

Due to the increasing number of Internet users worldwide, the user-generated content growing to an uncountable amount of data. At the end of 2011, more than 32%[1] of the world population use Internet for various reasons. The success of smart phones and mobile Web have contributed that people are online any time and everywhere. In addition, thanks to the success of social networking services, Internet has become a unique platform for sharing and expression what matters to them by spreading their activities, reviews, photos and videos. Certainly such content includes personal opinions and remarks about everything and everyone, published on social communities, forums and blogs. The number of visitors is strikingly high and is growing daily. With 901 million monthly active users at the end of March 2012[2], almost 40% of all Internet users participate on Facebook. On the microblogging platform Twitter, 140 million active users[3] express them-self in 140 characters. But also customer review websites and forums are heavily used. For example, thousands users publish product reviews on Amazon or their local business reviews on Qype or TripAdvisor. Therefore, this previously mentioned social media platforms are used as data source for opinionated user comments.

Certainly not only private people use these services. The recent study by Bundesverband Digitale Wirtschaft e.V. (BVDW) revealed that 72% of 188 interviewed German enterprises participate on social media platforms [BVDW, 2011]. These resources are used for public relations (PR), customer retentions, product launches, advertising campaigns, market research, information retrieval and opinion mining about products. But also cultural institutions in the German-speaking area joined social networks and use their benefits. For example, the Kunsthis-

[1]cf. http://www.internetworldstats.com/stats.htm (accessed on August 30^{th}, 2012)

[2]cf. http://newsroom.fb.com/content/default.aspx?NewsAreaId=22 (accessed on August 30^{th}, 2012)

[3]cf. http://blog.twitter.com/2012/03/twitter-turns-six.html (accessed on August 30^{th}, 2012)

torische Museum Vienna has currently about 8.000 Facebook fans[4], 300 Twitter follower[5] and hundreds of comments on local review sites. They use this media platforms to spread information, stay in contact with visitors, practise social commerce and get an overview about the customer behaviour.

However, less than 50% of all surveyed companies measure these network activities or practise monitoring. But nearly 40% are interested in a social media analysis. A structural analysis can provide information about the demographic structures as well as quality and quantity relations between users. Enterprises are also highly interested to extract user's opinion about their brand, products, their features and competitors. Surely, also cultural institutions are interested to find out what people think about their services. A sentiment analysis can extract topics which are interesting for visitors and what they think and feel about that. The process automation and summarization on a social media analysis Web platform is a very interesting scientific and personal challenge.

1.2 Problem Definition

The Web contains a wealth of opinions about brands, products, people and other issues. Certainly users express their opinion about cultural institutions in social networks and local review sites. The use of online available data would be more cost efficient in contrast to traditional approaches like questionnaires and interviews of customers. However the extraction of the user's sentiment results in many problems. The challenge is to filter and relate this pile of information and use it for opinion mining. For realization, both semantic and structural analysis is necessary.

This master thesis will focus on the semantic analysis of user generated content about cultural institutions in German speaking area. Gathering and filtering of data provided by social media platforms is necessary for this analysis. The derived research question is the design of a model which is able to classify topics and discover the sentiment of users about these subjects. Other primary problems are the common speech of Internet users which includes mistakes in writing and syntax. Due to the evaluation the design of a data warehouse which combines content from social networking services and local review websites represents a further challenge. To obtain a good data source selection, the following five different service provider are selected: Facebook, Foursquare, TripAdvisor, Twitter and Qype.

1.3 Aim of the Work

The master thesis will be composed of a theoretical and a practical part.

The objective of the theoretical part is to find a new mathematical model based on recent scientific findings. In a first step the model should be able to analyze and categorize user comments. This includes discovering and weighting of keywords. Afterwards, the model maps the personal opinion to the related subject. The aim is to discover the topics of user comments and classify them according to their sentiment.

[4]cf. `http://www.facebook.com/KHMWien` (accessed on August 30th, 2012)
[5]cf. `https://twitter.com/#!/KHM_Wien` (accessed on August 30th, 2012)

The practical part will be a joint Web platform which is able to execute the mathematical model and visualize the semantic interrelations. Generally speaking, the platform provides information and benchmarks about cultural institutions for a specific group of people. The targeted users are employees of museums and interested persons in general. Users search for a specific museum or exhibition and the application starts the topic classification of written comments and categorizes positive, neutral and negative statements. The primary objective is the mapping of personal opinions to discovered topics and benchmark user sentiment. Therefore, museums are able to find out which assessment criteria need improvements and which not. In addition to the sentiment analysis, the system answers questions about the quality and quantity of relations between users. This feature will be realized by my colleague Christina Stödtner.

1.4 Methodological Approach

In my master's thesis the approach follows three different methods which can be classified to the phases analysis, design and evaluation.

The analysis phase allows a better understanding of the current situation and characterizes the problem more detailed. First, the selection of social media platforms is essential to get qualified and quantified information. Secondly, the consolidation of the user generated content into a data warehouse is precondition for further research. Based on the data collection the theory respectively the hypothesis will be derived and developed.

The design phase deals with the finding of the mathematical model and the draft of a prototype. The model is important to get consistent data about cultural institutions and a valid comparison of museums. To evaluate the mathematical model, functional and structural testing is required. The purpose of the tests is to validate the model and to verify or refuse the stated hypothesis. This will be realized by manual testing of the classification and the semantic opinion mining. Mostly predefined data collections are used as training, testing and evaluation sets. Since there is no collection available which combines the content from Facebook, Foursquare, TripAdvisor, Twitter and Qype, the evaluation will be executed by a self developed data warehouse.

1.5 Structure of the Thesis

The following master's thesis consists of five main chapters and is divided into two main parts: Chapter 2 and 3 should provide theoretical basics and chapter 4 and 5 describes the practical part of this work.

Chapter 2 - Introduction
This chapter gives an introduction to opinion mining and sentiment analysis. First, an explanation of the term sentiment analysis as well as commonly used terminology and components in this research field are given. Furthermore, the reader can find reasons for the demand and application of sentiment analysis in the area of cultural institutions. Finally, problems and challenges which are currently known and unsolved will be presented.

Chapter 3 - State of the Art
The third chapter discuss the previous work by various researchers which are basis for this thesis. Different and highly promising approaches are dedicated to the classification by the help of sentiment phrases, text classification methods or score functions. After reading this chapter, the reader should be able to understand existing algorithms and to have an overview about the current state of the art.

Chapter 4 - Implementation
In this chapter the realisation of the novel approach for a sentiment analysis in the area of cultural institutions is presented. The implemented sentiment analysis model with their applied technology and resources are provided to the reader. The sub modules which consists of the data pre-processing module, the opinion lexicon module, the topic extraction module, the sentiment extraction module and the summarization are described in detail. Apart from this, the final approach with their sentiment classification result is given.

Chapter 5 - Evaluation
The evaluation chapter describes theoretical diverse evaluation methods which are well-established in this research field. Afterwards, this methods are applied on the novel approach and their results are presented.

Chapter 6 - Summary and Future Work
Finally, the last chapter summarizes this work and presents possible further research tasks and unrealised features.

CHAPTER 2

Basics of Sentiment Analysis

So far the term *sentiment analysis* was often mentioned in this thesis. The following chapter declares the exact meaning and basic components of sentiment analysis. Furthermore, the demand for sentiment analysis in the area of cultural institutions and possible applications gives reasons for the importance of this research field. In conclusion, known and unsolved problems and challenges are introduced.

2.1 Sentiment Analysis

Sentiment analysis, also known as opinion mining, deals with the extraction of opinions in textual content and the determination whether they express a positive, negative or neutral sentiment. It is an interdisciplinary scientific discipline which combines computational linguistics and information retrieval. Computational linguistics deals with the rule-based or statistical modelling of natural language. Information retrieval is about searching, recovering and interpreting information out large amounts of data.

One of the main reasons for the lack of study on opinions is that there was a very little written opinionated text before the emerge of the Internet. Before the Web, if an individual needs to make a decision, he or she asks for opinions from families and friends. However, the Internet has dramatically changed the way that people express their views and opinions. Nowadays, people express their views on almost anything in social networks, Internet forums, discussion groups and blogs, which are collectively called the *user-generated content* [Liu, 2010].

In general, there are two main types of textual information on the Web: *Facts* and *opinions*. A fact is a true objective expression about an entity in the world. In contrast, an opinion is a subjective expression that describes people's sentiment and feeling on something [Liu, 2010]. Therefore, sentiment analysis deals with the processing of opinionated terms. The term *subjective* is often falsely used as a synonym for *sentiment*. Subjectivity is the linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs and speculations (private states). Therefore, a private state is not open to objective observation or verification. Thus, a subjective analysis classifies content into objective or subjective. In contrast, sentiment analysis identifies the sentiment that a person may hold towards an entity. Compared to subjectivity analysis sentiment analysis is a finer grain analysis compared to subjectivity analysis and is usually divided into the categories positive, negative and neutral (see Table 2.1).

Subjective analysis	**Sentiment analysis**
Subjective	Positive Negative
Objective	Neutral

Table 2.1: Classification categories of subjective analysis and sentiment analysis [Banea, 2008]

Some researches count early projects on beliefs as forerunners of the area of sentiment analysis (e.g., [Carbonell, 1981]). Others state that the year 2001 marks the beginning of widespread awareness of the opportunities that sentiment analysis raise. Subsequently, in very recent years there have been published hundreds of papers published on the subject by numerous researchers [Pang and Lee, 2008]. In [Bhuiyan et al., 2009], Buhiyan, Xu and Jøsang have classified the unambiguous research opportunities in this research field. Figure 2.1 illustrates the classification of the research field of sentiment analysis.

In reference to [Liu, 2006], the authors define *sentiment classification* and *feature-based opinion mining* as main tasks of evaluative texts.

Sentiment classification considers the assignment of text as being positive, negative or neutral. Some researchers practice sentiment classification at *document-level*, where others practice it on *sentence-level*. However, no details are discovered about what people like or dislike. In the area of cultural institutions, and especially museums, internet users generate content about exhibitions, their highlights, prices and other characteristics in several social media web platforms. Therefore, sentiment classification on museum reviews would result in the categorization of positive or negative documents, but would not consider aspects that have been commented on in a review. Methods are based on identification of opinion words and phrases by *dictionary based* and *corpus-based approaches*. Dictionary based approaches use synonyms and antonyms to determine word sentiments (positive, negative or neutral) on a set of manual seed opinion words. In contrast, corpus-based approaches find co-occurrence patterns of words to detect the sentiment of words, phrases and documents.

As mentioned before, a positive or negative evaluated text does not mean that the reviewer prefers or despises everything of the object. To obtain such detailed aspects, feature-based opinion mining deals with discovering aspects and topics in a review that people like or dislike. Thus, this analysis is on sentence-level. For example, the Qype comment (see Figure 2.2) evaluates different aspects about the Vienna museum “Schloss Belvedere”. The review differentiates assessment criteria into the categories exhibition in general, the number of visitors, photo-allowance and pricing. Intuitively, the exhibition itself can be ranked rather positive. However, the number of visitors, the ban on photography as well as the entry price are negative. The automated extraction of these categories and the sentiment classification is part of feature-based opinion mining.

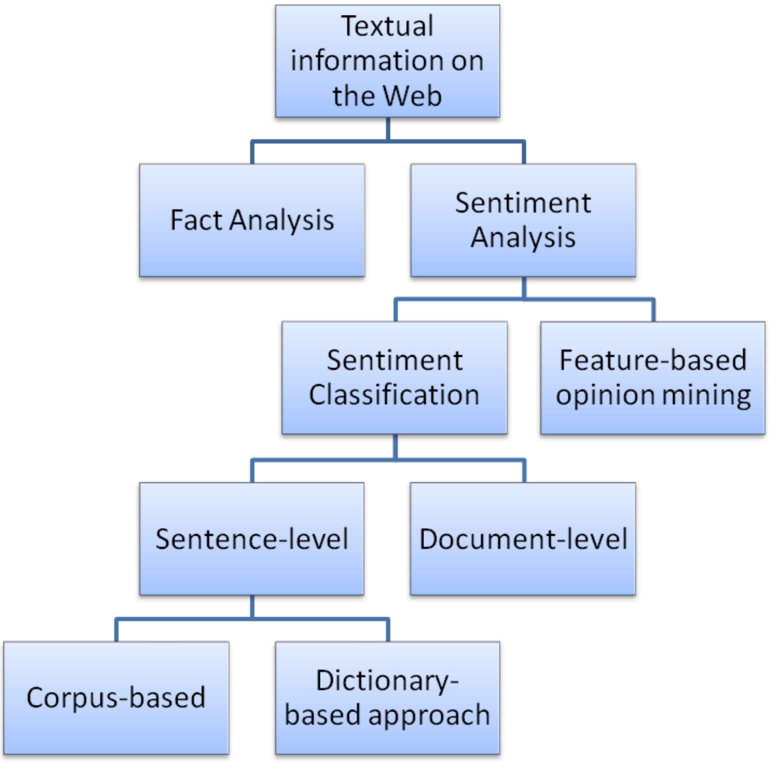


Figure 2.1: Classification of sentiment analysis research [Bhuiyan et al., 2009]

This master's thesis applies feature-based opinion mining to analyze user-generated content in the domain of cultural institutions. However, to operate research on this scientific field, knowledge about sentiment classification is essential and fundamental. Therefore, also sentiment classification is discussed in detail in the state of the art section in Chapter 3.

2.2 Components of Sentiment Analysis

The following section defines terminology and basic concepts used in the scientific area of sentiment analysis. With reference to the scientific publication "Sentiment Analysis and Subjectivity" by Bing Liu [Liu, 2010] and the help of examples, terms and concepts are introduced and illustrated.

- **Opinion holder** The person that expresses the opinion is called opinion holder h. In the review (see Figure 2.2), "ki_schorsch" is the author and therefore the opinion holder.

- **Document** The document d consists of a sequence of sentences $S = s_1, s_2, ..., s_n$. In reply to this master's thesis, a document can be a review on local review websites, comments on social network platforms, a forum post or a blog that evaluates a set of objects. The review, written by "ki_schorsch", shown in Figure 2.2 presents a document.

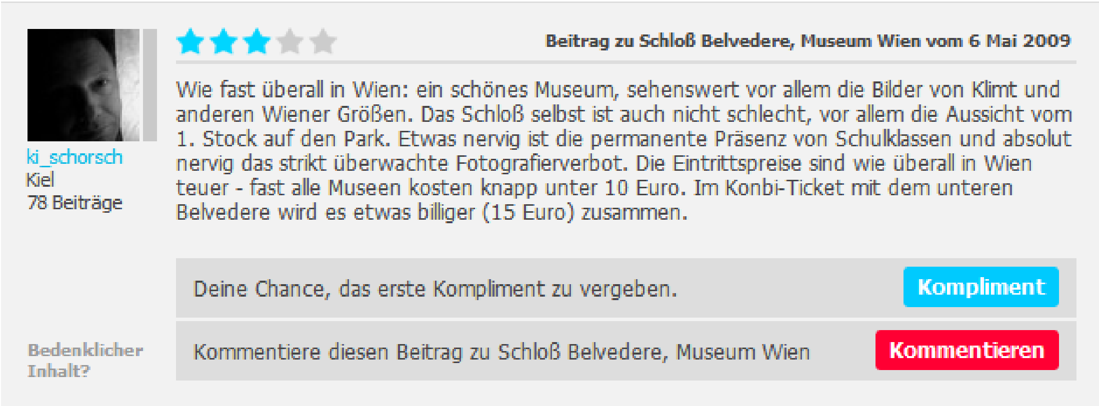


Figure 2.2: Example of a user review of Schloss Belvedere obtained from Qype.com

- **Object** An object O is an entity which can be a brand, product, person, topic, event and many more. An object is associated by components (and sub-components) c and a set of attributes a. In the domain of cultural institutions, an object can be for instance a museum or an exhibition. Sub-components of a museum are their exhibition rooms, wardrobe, shop, toilet, restaurant or coffee bar. Attributes are the location, size, entry fee or opening hours.
  Xiaowen Ding, Bing Liu and Philip S. Yu [Ding et al., 2008] define an object as follows:

  > An object O is an entity which can be a product, person, event, organization, or topic. It is associated with a pair, $O : (T, A)$, where T is a hierarchy or taxonomy of components (or parts), sub-components, and so on, and A is a set of attributes of O. Each component has its own set of subcomponents and attributes.

- **Topic** A topic t represents an attribute of the object. In other domains, e.g. in the product domain, the term *feature* or *aspect* is used quite common used. However, in the domain of cultural institutions, the term feature may not sound natural and therefore uses the term *topic*. The “Eintrittspreis” (entry fee) or the “Fotografiererlaubnis” (photo-allowance) are topics which are reviewed by “ki_schorsch”.

- **Explicit topic** If a topic t or any of its synonyms appears in a sentence s, t is called an explicit topic in s.

  Example 1: “Eintrittspreis” (admission fee) in the following sentence is an explicit topic: “Die Eintrittspreise sind wie überall in Wien hoch” (the admission fee is high as everywhere in Vienna)

- **Implicit topic** If neither t nor any of its synonyms appear in s but t is implied, then t is called an implicit topic in s.

Example 2: The "Besucheranzahl" (number of visitors) is an implicit topic in the following sentence as it does not appear in the sentence but it is implied: "Etwas nervig ist die permanente Präsenz von Schulklassen" (a bit annoying is the permanent presence of school classes)

- **Opinion** In contrast to a fact which represents objective statements about an object, an opinion on a topic t is a positive, negative or neutral emotion on t from an opinion holder.

  Example 3: "Ein schönes Museum" (a beautiful museum) illustrates a positive opinion about the museum.

  By combination of the prior described concepts, Bing Liu [Liu, 2010] defines an opinion as a quintuple $(e_j, a_{jk}, s_{oijkl}, h_i, t_l)$,where

  e_j is a target entity,

  a_{jk} is a topic or aspect of the entity e_j,

  s_{oijkl} is the sentiment value of the opinion from opinion holder h_i on topic a_{jk} of entity e_j at time t_l. s_{oijkl} is usally positive, negative or neutral,

  h_i is an opinion holder,

  t_l is the time when the opinion is expressed.

- **Semantic orientation (polarity)** The semantic orientation (polarity) of an opinion on a topic t states whether the opinion is positive, negative or neutral.

- **Opinion passage on a feature** An opinion passage on a topic t of an object o is a group of sentences in a document d that expresses a positive, negative or neutral opinion on t.

  Example 4: The following two sentences reflect the opinion about the entry fee. "Die Eintrittspreise sind wie überall in Wien teuer - fast alle Museen kosten knapp unter 10 Euro. Im Kombi-Ticket mit dem unteren Belvedere wird es etwas billiger (15 Euro) zusammen." (The ticket prices are expensive as everywhere in Vienna - almost all museums cost just below 10 Euro. With the Kombi ticket which includes the Lower Belvedere, it is slightly cheaper (15 Euro).)

- **Comparative opinion** A comparative opinion expresses a relation of similarities or differences between two or more objects by the opinion holder based on some features of the objects.

  Example 5: The following opinionated sentence compares the entry fee of two cultural institutions. "10 Euro für das Technische Museum sind für mich okay, aber nicht 15 Euro für den Zoo." (10 Euros for the Technical Museum are okay for me, but not 15 Euros for the Zoo.)

2.3 The Demand for Sentiment Analysis in the Area of Cultural Institutions

Pang and Lee introduced their widely cited paper "Opinion Mining and Sentiment Analysis" [Pang and Lee, 2008] with the following sentence: *"'What other people think' has always been an important piece of information for most of us during the decision-making process"*. Of course, this statement holds also in the domain of cultural institutions. Long before the Internet was established, many people asked friends for recommendations about a museum, leisure activity, church, school or work. However, the Web allows that people find out numerous opinions and experiences of others by reading their reviews and comments in social networks, review sites, forums and blogs in a quick and uncomplicated way. With the evolution of Web 1.0 in the mid 1990s to Web 2.0 since 2004, not the technology per se changed: Rather the way of communication and interaction shifted dramatically. In contrast to classical websites where users are limited to passive viewing of content, Web 2.0 sites allows users to collaborate with each other and to create content. Before the explosion of Web 2.0, museum websites informed statically about opening hours, admission price, exhibitions, guides and other services[1].

But nowadays, cultural institutions present the participation on several platforms on their websites and invite visitors for active discussions and sharing their opinions. They spread information and discuss on Facebook, Google+, Tumblr, Twitter, and use the photo sharing services Flickr and Instagram, furthermore they make use of video platforms like YouTube and Vimeo and podcasts provide audio files. For example, the "museum moderner kunst stiftung ludwig wien" (mumok) links directly to the social media platforms Facebook[2], Twitter[3], Youtube[4] and their iTunes Podcasts[5]. Interestingly, nearly 5%[6] of the annual visitors "like" the mumok Facebook page. The training dataset contains 13 cultural institutions which are discussed in detail in Section 5.2.1. Figure 2.3 shows the percentage of these museums (n = 13) which are opinionated discussed on each social network profile. By manual selection, an institution was tagged as opinionated discussed on a social network when at least 10 user posts contain a subjective opinion. In addition, the user comments must have been written in a period of 6 months. Obviously it can be assumed that on local review sites (TripAdvisor and Qype) almost all museums are evaluated. The user-generated content consists of comments and recommendations which contains personal opinions. However, the flood of information and data generated by the users can no longer be processed manually. Furthermore, cultural institutions may not have the resources to train a staff member or appoint a social media analyst. To get a fast and complete overview, an automatic sentiment analysis will be a very important tool to extract user opinions and trending

[1]For example, the "museum moderner kunst stiftung ludwig wien" (mumok) in 2001, `http://web.archive.org/web/20010715040810/http://www.mumok.at/dpages/willkommen.htm`, (accessed on August 30^{th}, 2012)

[2]`http://www.facebook.com/MUMOK`

[3]`https://twitter.com/#!/MUMOKWien`

[4]`http://www.youtube.com/user/mumokvienna`

[5]`itunes.apple.com/de/podcast/mumok-podcast/id268538422`

[6]Calculated by the number of Facebook likes (`http://www.facebook.com/MUMOK`) divided by the average number of visitors per year of the annual visitors (`http://www.wienkultur.info/page.php?id=98`) (accessed on August 30^{th}, 2012)

topics and is therefore highly demanded.

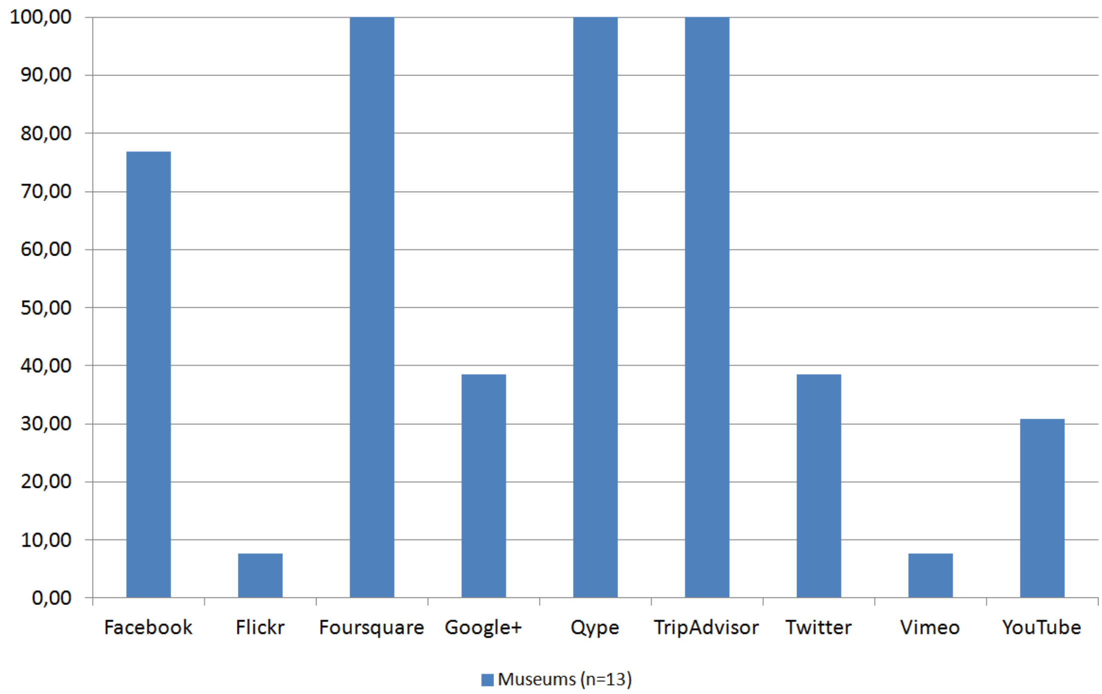


Figure 2.3: Percentage of museums which are opinionated discussed on their own social network profiles (as of August 25, 2012)

2.4 Applications

In the introduction (see Chapter 1), an application of sentiment analysis is used as a motivating example: The social media analysis for cultural institutions. But also consumers and visitors can use numerous possibilities of opinion mining. The following section seeks to enumerate applications for cultural institutions as well as for their visitors.

2.4.1 Applications for Cultural Institutions

Nowadays, information about customers is highly valuable and is crucial for the success of businesses. Also cultural institutions should use the power of a sentiment analysis for increasing their number of visitors, improve their services and exhibitions, and to understand and identify the target audience.

Since institutions spend time and huge sums for shaping their image and promoting their exhibitions, such text analysis can be used as monitoring and marketing tool. Institutions can measure the success of marketing campaigns, promotions and advertisings by analysis of the user-generated content. The extraction of quantity and quality of user comments about museums, their exhibitions and events allows a reflection of marketing activities and their realization. Furthermore, opinion mining allows answering questions like “Has the number of visitors decreased in the last years?”. The subjective judgements about entry fee, location, quantity and quality of the exhibition and staff are indicators for resolving such questions. The use of online available data might be a cost efficient instrument in conjunction with traditional approaches like questionnaires and interviews of customers. In general, the detected sentiments can be applied for investigation of the customer satisfaction and suggested improvements. Furthermore, a competitor analysis could assess strengths and weaknesses of current and potential competitors.

In addition, cultural institutions can benefit through better advertisement placement. For search engines and social networking services that display ads, it is useful to bring up ads when relevant positive sentiments are detected or hide ads when relevant negative statements are discovered [Pang and Lee, 2008]. For example, when an user writes positive about a museum, he or she should be informed about current exhibitions and other news. In contrast, in the case of negative commenting, he or she might possibly get no ads about the discussed museum. However, more sophisticated systems can detect topics which are negative rated and afterwards selectively post ads to change users opinion. When a reviewer writes negatively about the entry fee, the system can pop up ads which promote price discount or free entrance days.

A further field of application is the review spam detection. Because many people read some reviews before visiting a museum, the user-generated content is an important decision criterion. If the discussion about an institution is mainly negative in many comments, the potential visitor will most likely choose another one. In contrast, if many people express their positive sentiment about an exhibition, the chance that a reader visits it rises. Hence, positive opinions can result in significant financial profits. Intuitively, it involves the danger of manipulated reviews. Bing Liu [Liu, 2006] classifies two main objectives for writing spam reviews: The promotion of own objects by writing of unearned positive reviews (hype spam) and the damage of competitors by writing unfair negative reviews (defaming spam) [Liu, 2006]. Therefore, applications which detect spammers and their misleading comments are more than needed.

2.4.2 Applications for Customers and Visitors

As mentioned before, businesses can benefit enormously by opinion mining and their applications. However, also customers and, in the area of cultural institutions, visitors can use this technology for information gathering.

Possible applications are recommendation systems which don't suggest cultural institutions that gain a lot of negative feedback [Pang and Lee, 2008]. Besides of the kind of museum or the location, recommendations on local review sites or social networks can include subjective user opinions. It would be a useful feature which protects users from bad surprises. For instance, when visitors at which museums they are (e.g., by check-in on Foursquare or Facebook), social networks recommend other museums from users which are visited the same and other museums. However, the recommendation of mostly negative rated museums, based on objective and subjective reasons is not satisfactory.

Opinion-aggregation and comparison websites can be used for summarization of an uncountable mass of reviews and user comments. Customers can easily get an overview about different aspects of cultural institutions by entering a specific name or feature. By submission of a museum, the system discovers the topics of user comments and classifies them according to their sentiment. But also filtering of sentiment text by a particular aspect (e.g., the museum shop) can be part of a comparison website and eases the finding of positive evaluated institutions by own selection criterion.

2.5 Problems and Challenges

2.5.1 Object Identification

Without the knowledge about the object on which an opinion is expressed, the opinion is not very helpful. E.g. in the sentence “Das Museumsquartier ist schön und bietet viel zu sehen.” (The Museumsquartier is beautiful and offers much to see) the object “Museumsquartier” is discovered. This task is similar to the research field of *named entity recognition* (NER). NER is the task of finding and classifying elements in unstructured text into predefined categories (usually including the three types $PERSON$, $ORGANIZATION$ and $LOCATION$). The categories are called Named Entities (NE) and can be synonymously used as proper names. The classification can be done by hand-crafted rules (e.g. define an NE as $PERSON$ when Mr. or Mrs. occurs in the subset) or machine learning methods (e.g. the Maximum Entropy Classifier or the Hidden Markov Model Classifier) [Scheurer, 2012]. The following Figure 2.4 illustrates an exemplary in- and output of a NER system:

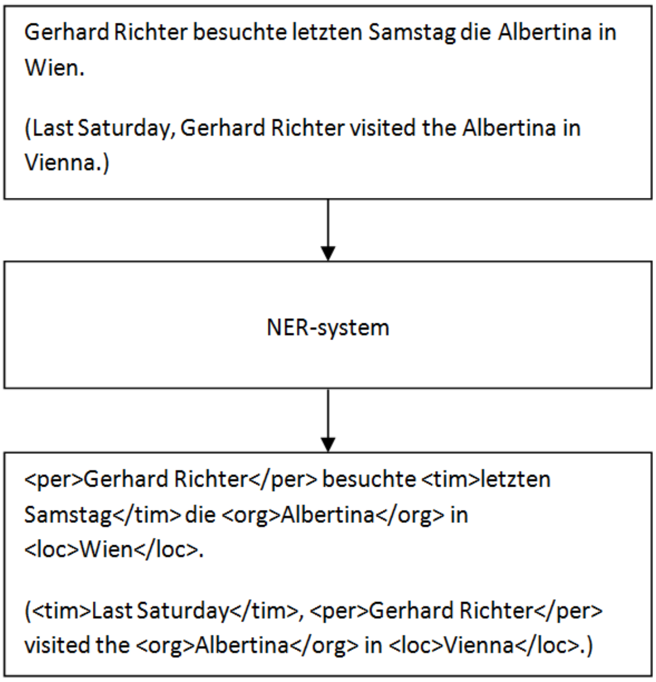


Figure 2.4: Exemplary in- and output of a NER system

The most commonly used evaluation measures for NER systems are $recall$ (fraction of retrieved instances that are relevant), $precision$ (fraction of relevant instances that are retrieved) and $F - measure$ (combines precision and recall) which are described in detail in Chapter 5.1. [Florian et al., 2003] presents a framework which combines four classifiers. Robust linear classifier, maximum entropy, transformation-based learning and a hidden Markov model classifier are combined under different conditions. When no training resources are used (e.g., gazetteers) the system achieves a F-measure of 91.63 on English data. By the integration of names, locations and person gazetteers and more general data, the F-measure error was further reduced by a factor of 15 to 21%. The system achieves also satisfactory evaluation results by classification of German datasets with a precision of 84.60%, recall of 61.93% and a F-measure of 71.51.

However, there is a difference: In opinion mining applications, users want to compare competing objects. For example, in the review sentence "Die Anreise zum Museum mit den Wiener Linien ist am günstigsten und schnellsten." (the journey to the museum with the Wiener Linien is the cheapest and fastest.), the transport service "Wiener Linien" is not a competing object in the domain of cultural institutions. Obviously is a competing object almost invariably relevant. However, it can not be stated in general that a non-competing object is irrelevant. For instance, the non-competing object "Wiener Linien" can be a relevant opinionated object about an cultural institution. In conclusion, it can concluded that a competing object is relevant, but it can not concluded that a non-competing object is irrelevant. Therefore, one of the issues which must be solved is the separation of relevant and irrelevant objects. [Pang and Lee, 2010]

2.5.2 Topic Extraction and Synonym Grouping

A further problem deals with the extraction of topics and the finding of their synonyms. The opinion about the topic "Eintrittspreis" (admission fee) can be expressed in many ways: "Der Preis ist viel zu hoch" (the price is too high), "Der Eintritt ist vollkommen gerechtfertigt" (admission is fully justified) or "Die Karten kosten zu viel" (the tickets cost too much) are just three of dozen phrases. The automate synonym grouping is one of the major challenges and is not solved satisfactory and domain independent up today. Current researches mainly detect nouns with a good recall, but a low precision. Therefore, mostly all discovered nouns are topics, however not all relevant topics are retrieved. [Pang and Lee, 2010]

2.5.3 Opinion Orientation Classification

This problem can be sub-divided into two tasks: The subjectivity and the sentiment classification. The subjectivity classification determines whether there is an opinion on a feature. Therefore, the process deals with the identification of objective and subjective sentences. If subjective sentences are discovered, the sentiment classification algorithm categorizes either into positive, negative or neutral. However, even in the same domain, same words indicate different sentiments. The word "lange" (long) indicates in the sentences "Die Öffnungszeiten sind sehr lange" (the opening hours are very long) a positive opinion. However, in the sentence "Die Menschenschlange an der Kasse war sehr lang" (People queue at the cash was very long), "lang" (long) indicates a negative opinion. [Pang and Lee, 2010] Existing approaches are based on supervised and unsupervised methods as well as with the help of lexical resources which are discussed in the following chapter.

2.5.4 Language Correction and Interpretation

Even by solution of the above mentioned problems and the errorless implementation, the system would never detect and analyze all opinions correctly. The complex and misapplied German grammar, the poor spelling and punctuation, the lack of capitals and the use of abbreviations would falsify the result. Therefore a well working spell checker is needed to increase the number of correct classified opinions. Furthermore, it is not unusual for Internet users to write sarcasm and ambiguity sentences which also leads to a wrong sentiment classification. With the current state of technology and the help of natural language processing, it is not possible to overcome all of these problems.

CHAPTER 3

State Of The Art

The following chapter offers an overview about key research topics of opinion mining and their related work. The sections present the state of the art about the most studied topic sentiment classification, the generation of opinion lexicons and finally the feature-based sentiment analysis.

3.1 Sentiment Classification

Sentiment classification can be carried out in two different ways: The classification on document and on sentence-level. Document-level sentiment classification considers the whole user review or comment as positive, negative or neutral. Of course, each individual sentence implies a opinion or not. By sentence-level sentiment classification, the opinionated sentences can also be allocated to a semantic orientation (positive, negative or neutral). [Liu, 2010]

3.1.1 Document-Level Sentiment Classification

Given a set of evaluative texts D, the sentiment classifier determines each document $d \in D$ as a positive, negative or neutral opinion. Most existing solutions are based on supervised and unsupervised learning methods which are introduced below. But before, a remark about the application area: Researchers assume that a review is written from a single opinion holder which expresses his or her opinions on single objects. This is valid for customer reviews of specific institutions, however this assumption is not valid for all types of resources. [Liu, 2010] In Web forums and blogs, authors express opinions on multiple institutions in a single posting. The classification of such content on document-level leads to unsatisfactory results and should be processed on sentence-level or with the help of the feature-based approach.

Classification Based on Supervised Learning

Supervised Learning is a classic machine learning task for topic-based text classification. It is appropriated by a small number of classes, the classes are already known and some training data is available. Supervised classification can be applied for predicting the class to which an object is likely to belong. [Gupta, 2006] Therefore, also sentiment classification can be formulated as a supervised learning problem. The classes are presented by the sentiment labels positive, negative and neutral and the training data are the user-generated content extracted from Web 2.0 platforms. Since many comments are already rated by “stars”, the testing data is also available. However, there are some differences between topic-based text and sentiment classification which are formulated by Bing Liu [Liu, 2010]:

- In topic-based classification, topic related words are important, but
- in sentiment classification, not surprisingly, sentiment words that indicate a opinion are important (e.g., great, bad, excellent, etc.).

In 2002, Pang et al. [Pang et al., 2002] analyzed the three machine learning methods naïve Bayes, maximum entropy classification and support vector machines (SVM) for classifying movie reviews as positive or negative. However, the researches conclude that the machine learning methods do not perform as well on sentiment as on traditional topic-based classification which is argued by the more challenging problems (see Section 2.5). Anyhow, it has been shown that n-grams as classification features performed well with either naïve Bayes or SVM. An n-gram is an n-character slice from a given sequence of text. Depending on the size of an n-gram, it is referred to as unigram (n = 1), bigram (n = 2) or trigram (n = 3). Larger sizes are referred to by the value of n (e.g., four-gram, five-gram, etc.). For instance, the trigrams that can be generated from “sentiment analysis” are “sen”, “ent”, “nti”, “tim”, “ime”, “men”, “ent”, “nt_”, “t_a”, “_an” and so forth. In [Cavnar and Trenkle, 1994], the researchers use n-gram-models for language detection and subject classification. After counting all n-grams, the most frequent are added to a category profile (e.g., language). Following, the profile of unknown text (document profile) is then compared to every category listed. By calculating the distance (sometimes called Out-Of-Place measure) between the document profile and all category profiles, the text is assigned to the category with the minimum distance (see Figure 3.1).

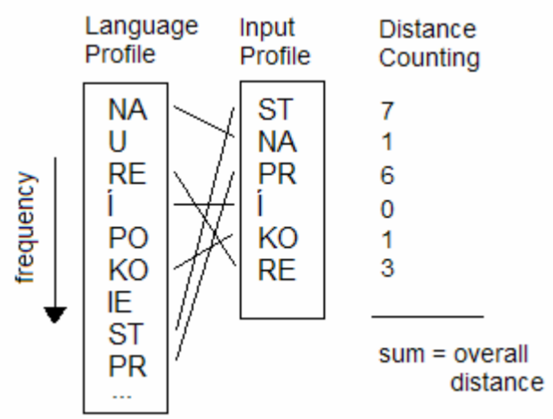


Figure 3.1: Comparison between unknown text and one of languages [Ölvecký, 2005]

Other scientists used, apart from n-grams, more features for the training of sentiment classifier. A feature commonly used in classic tropic-based text classification is the *term frequency* and *term presence*. Term frequency measures of how often a term is found in a document. In contrast, term presence only reflects, with the help of binary-valued vectors, if a term occurs (1) or not (0). Pang et al. [Pang et al., 2002] obtained a better performance and higher accuracies of naïve Bayes and SVMs by using presence. Various researches (e.g., [Hatzivassiloglou and McKeown, 1997] or [Hatzivassiloglou and Wiebe, 2000]) find out that adjectives are important indicators of opinions. Therefore, the part of speech (POS) identification of each word is a useful feature. However, not only adjectives express opinions: Also nouns and verbs can be strong indicators and should not be ignored. Pang et al. [Pang et al., 2002] using only adjectives and disregard other part of speeches, which results in a worse performance. The classification performs much better with the same number of most frequent word-level unigrams.

The feature of discovering words that are commonly used to express positive or negative sentiments is called *opinion words* (also known as *polar words*, *opinion-bearing words* and *sentiment words*). Such positive opinion words are for example “good”, “awesome” or “beautiful”. “Terrible”, “bad” or “wack” are negative examples. Not only individual words, also *opinion phrases* can express sentiments. For example, the proverb “all that glitters is not gold” implies a negative opinion on an object. Both individual words and opinion phrases are instrumental to sentiment analysis. Other researches have dealt with *syntactic dependency*. Otero [Otero, 2008] assume that syntactic dependencies play a central role in the field of semantic processing and interpretation. A syntactic dependency will be defined as a binary operation that takes as arguments the meaning of the two related words (the *head* and the *dependent*), and gives as result a more elaborate arrangement of their meaning. Syntactic structure is described in terms of whether a particular noun is the subject or agent of the verb. The subject has the grammatical function in a sentence of relating its phrase by means of the verb to any other elements present in the sentence. In contrast, a grammatical agent is the cause or initiator of an event. Figure 3.2 illustrates the subject dependency between “read” (the head) and “John” (the dependent).

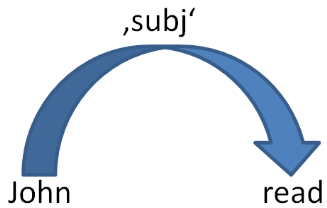


Figure 3.2: Subject dependency between “read” (the head) and “John” (the dependent) [Otero, 2008]

Words dependency based features allow a semantic interpretation of syntactic dependencies and are mainly generated with the help of dependency trees. For instance, Figure 3.3 illustrates the dependency tree of the sentence “John read the short book”. The tree contains the following syntactic functions: adj (adjective), det (determiner), obj (object) and subj (subject).

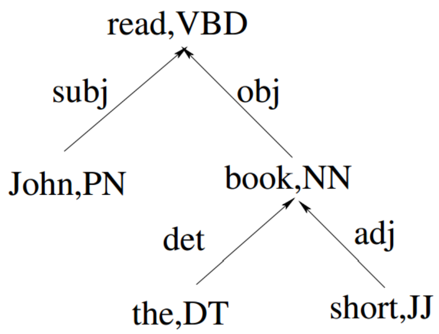


Figure 3.3: Dependency tree [Wilson et al., 2004]

The handling with *negation words* is also an important concern in sentiment analysis. Their appearances often change the opinion orientation: For example, the sentence "I don't like this museum" is negative. However, in the sentence "Not only the shop but also the restaurant is commendable", the word "not" does not change the opinion. Das and Chen [Das et al., 2001] improve the accuracy results by attaching the term "NOT" to words occurring close to such negation words (no or don't). The term "like" in the sentence "I don't like this museum" is converted to "like-NOT".

Apart from the classification as positive or negative opinions, other researches predicted rating scores of user-generated content. Pang and Lee [Pang and Lee, 2005] discuss an algorithm which determines an author's evaluation with respect to a multi-point scale (e.g., one to five stars). Finally, based on the fact that sentiment classification is highly sensitive to the domain, a classifier trained using texts from one domain often performs poorly on other domains. Liu [Liu, 2010] found the reason in the different usage of words and language constructs for expressing opinions in different domains.

As already mentioned in the preface, sentiment classification on document-level can be processed with the help of the feature-based approach. [Abbasi et al., 2008] proposed a system design for classification of the Web forum opinions in multiple languages which focuses on document-level classification on sentiment only (see Figure 3.4).

The design has two major steps: extracting an initial set of features and performing feature selection. The feature extraction module incorporated syntactic, lexical and structural features in the sentiment classification attribute set. These features are more generic and applicable across languages. However, semantic features were not used since these attributes are context dependent. Such features are topic and language specific. For instance, the set of positive polarity words describing a good music album may not be applicable to discussions about racism. The initial feature set consists of the following 14 different feature categories:

1. **POS n-grams**: Frequency of part-of-speech tags (e.g., NN for nouns)

2. **Word roots**: Frequency of roots (e.g., act, bio)

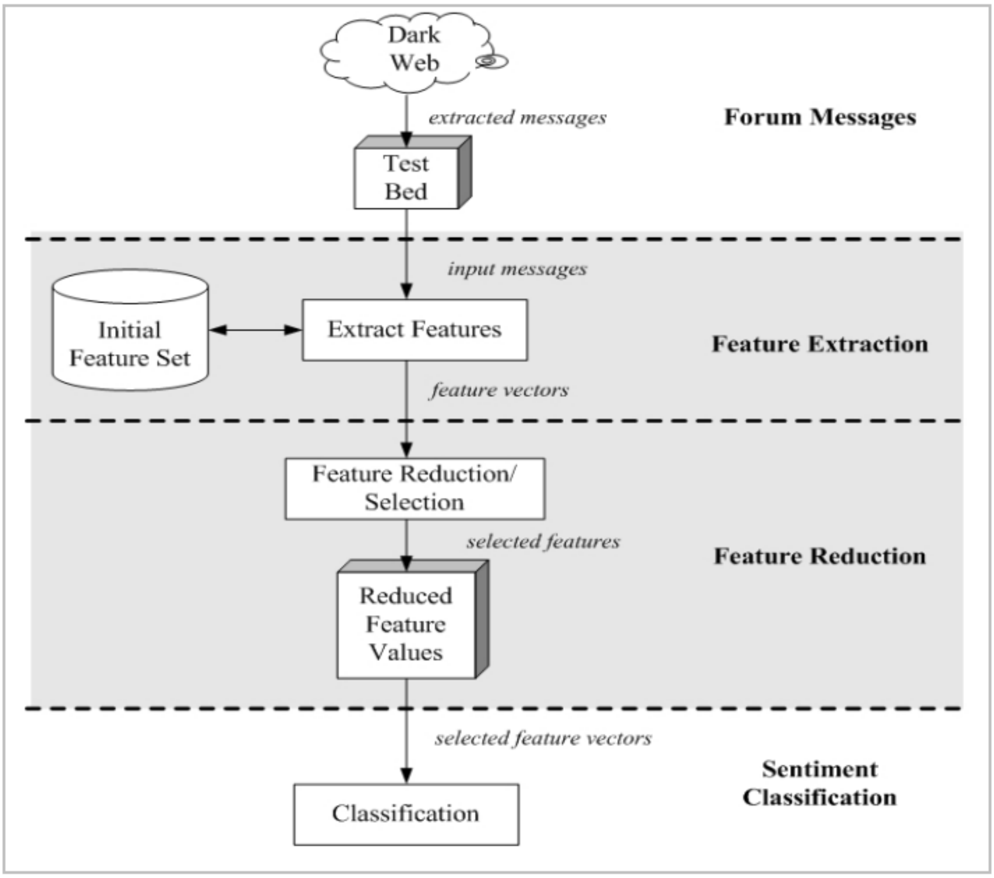


Figure 3.4: Design of the sentiment classification system [Abbasi et al., 2008]

3. **Word n-grams**: Word n-grams (e.g., senior editor, editor in chief)
4. **Punctuation**: Occurrence of punctuation marks (e.g., ! ; : , . ?)
5. **Letter n-grams**: Frequency of letters (e.g., a, b, c)
6. **Character n-grams**: Character n-grams (e.g., abo, out, ut, ab)
7. **Word lexical**: Total words, % character per word
8. **Character lexical**: Total character, % character per message
9. **Word length**: Frequency distribution of 1-20 letter words
10. **Vocabulary richness**: Richness (e.g., hapax legomena, Yule's K)
11. **Special character**: Occurrence of special character (e.g., #$%&*+)
12. **Digit n-grams**: Frequency of digits (e.g., 100, 17, 5)

13. **Structural**: Has greeting, has url, requoted content, etc.

14. **Function words**: Frequency of function words (e.g., of, for, to)

The feature selection module uses the Entropy Weighted Genetic Algorithm (EWGA) which uses the information gain (IG) heuristic to weight the various sentiment attributes. For the IG, the authors used the Shannon [Shannon, 2001] entropy measure, which is formulated mathematically as follow:

$$IG(C, A) = H(C) - H(C|A) \tag{3.1}$$

where, $IG(C, A)$ is the information gain for feature A,

$H(C) = -\sum_{i=1}^{n} p(C = i) \log(p(C = i))$ is the entropy across sentiment classes C,

$H(C|A) = -\sum_{i=1}^{n} p(C = i|A) \log(p(C = i|A))$ is the specific feature conditional entropy and

n is the total number of sentiment classes.

If the number of positive and negative sentiment messages is equal, the entropy across sentiment classes C $H(C)$ is 1. In addition, the information gain for each attribute A will vary along the range of 0 to 1 with higher values indicating greater information gain. All features with an information gain greater than 0.0025 (i.e., IG(C,A) > 0.0025) are selected. The experiment produces a very good result on the benchmark movie dataset of English and Arabic written reviews.

Classification Based on Unsupervised Learning

In contrast to supervised learning, unsupervised learning tries to find hidden structures in which the examples given to the learner are unlabelled. The learner are based on opinion words and phrases. Turney [Turney, 2002] describes an algorithm which are based on fixed syntactic phrases by application of the following three steps:

Step 1: Phrases extraction
As mentioned previously, phrases containing adjectives or adverbs are good indicators of opinions. The algorithm extracts two sequent words, where one word is an adjective or adverb and the other is a content word. Therefore, a combination of words are extracted if their POS tags conform to any of the patterns in Table 3.1. According to the Penn Treebank Project[1], words are tagged with the following abbreviations:

[1]cf. `http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html` (accessed on August 30^{th}, 2012)

NN (Noun, singular)	RBS (Adverb, superlative)
NNS (Noun, plural)	VB (Verb, base form)
JJ (Adjective)	VBD (Verb, past tense)
RB (Adverb)	VBN (Verb, past participle)
RBR (Adverb, comparative)	VBG (Verb, gerund or present participle)

	First word	Second word	Third word (Not Extracted)
1	JJ	NN or NNS	anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Table 3.1: Patterns of POS tags for extracting two-word phrases

Example: According to the pattern in row 1 (see Table 3.1), in the sentence "The museum has disgusting facilities", "disgusting facilities" will be extracted.

Step 2: Estimation of the extracted phrases

The second step measures the orientation of the extracted phrases using the pointwise mutal information (pmi):

$$pmi(term_1, term_2) = \log(\frac{p(term_1, term_2)}{p(term_1)p(term_2)}) \tag{3.2}$$

where $p(term_1, term_2)$ is the co-occurrence probability of $term_1$ and $term_2$, which simply express the probability that the terms occur together. $p(term_1)p(term_2)$ gives the probability that the terms co-occur if they are statistically independent. Therefore, the ratio between these two values is a measure of the degree of statistical dependence of one of the words when we observe the other. [Liu, 2010]

The opinion orientation (oo) of a phrase is measured based on its association with a positive term (e.g., beautiful) and its association with a negative term (e.g., poor):

$$oo(phrase) = pmi(phrase, “beautiful'') - pmi(phrase, “poor'') \tag{3.3}$$

The probabilities are calculated by request queries to a given document set and collecting the number of $hits$. For each document search, a search engine tool gives usually the number of relevant documents to the query. By searching the two terms together and separately, we can estimate the probabilities in Equation 3.2. Let $NEAR$ a function which constrains the search to documents that contain the words within ten words of one another. Thus, Equation 3.3 can be rewritten as:

$$oo(phrase) = \log(\frac{hits(phraseNEAR“beautiful'')hits(“poor'')}{hits(phraseNEAR“poor'')hits(“beautiful'')}) \tag{3.4}$$

Step 3: Recommendation
The algorithm computes the oo of all phrases and afterwards average of a given review. If the average is positive, the review is classified as positive, negative or neutral.

3.1.2 Sentence-Level Sentiment Classification

As the name sentence-level sentiment classification suggest, it does not analyze the sentiment of the whole document, but of each sentence. Like the document-level classification, the sentence-level sentiment classification does not extract object features or topics that have been commented on [Bhuiyan et al., 2009]. Liu [Liu, 2010] differ in two sub-tasks for sentence-level sentiment classification:

- **Subjectivity classification**, which determines whether a sentence s is subjective or objective and
- **sentence-level sentiment classification**, which determine whether a subjective sentence express a positive or negative opinion.

Since both problems are classification problems, traditional supervised learning methods are applicable and have been widely studied. Wiebe et al. performed subjectivity classification using the naïve Bayes classifier [Wiebe and Mihalcea, 2006] and support vector machines [Wilson et al., 2004]. Wilson et al. [Wilson, 2005] present an approach of phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. The approach enables the automatic identification of contextual polarity for a large subset of sentiment expressions, where achieving results are significantly better than baseline which is provided by the classification of words and their prior polarity (positive, negative, both, neutral). The presented model is divided into two sub processes: For the first step, the researchers concentrate on whether words and phrases are neutral or polar in context (neutral-polar classification which uses 28 features). For the second step, they take all words and phrases marked as polar in step one and focus on identifying their contextual polarity (polarity classification which uses 10 features). A detailed explanation of all features can be found in [Wilson, 2005].

Sentence-level sentiment classification assume that a sentence expresses a single opinion from a single opinion holder. However, this assumption is only valid for simple sentences with one opinion (e.g., “The restaurant of the museum is awesome.”). But, many sentences express more than one opinion. For example, the sentence “The restaurant of the museum is awesome but the facilities are disgusting” includes a positive and a negative sentiment. In [Wilson et al., 2004], the researches discover that not only single sentences may contain multiple opinions, but also subjective and factual clauses. Also the categorization by strength into the levels neutral, low, medium and high is discussed in [Wilson et al., 2004]. The classification is realized by a combination of PREV-strength and SYNTAX-strength features. Previously established types of clues (PREV) are words and phrases (clues) which include entries of manually developed resources by other researchers of polar adjectives and verbs, subjective nouns as well as extraction patterns. The SYNTAX clues consist of the following five classes from each word w in every dependency tree:

- **root**(w, t): word w with POS tag t is the root of a dependency tree (i.e., the main verb of the sentence).
- **leaf**(w, t): word w with POS tag t is a leaf in a dependency tree (i.e., it has no modifiers).
- **node**(w, t): word w with POS tag t.
- **bilex**(w, t, r, wc, tc): word w with POS tag t is modified by word w_c with POS tag t_c, and the grammatical relationship between them is r.
- **allkids**$(w, t, r_1, w_1, t_1, ..., r_n, w_n, t_n)$: word w with POS tag t has n children. Each child word w_i has POS tag t_i and modifies w with grammatical relationship r_i, where $1 \leq i \leq n$.

To adapt the PREV and SYNTAX clues to strength classification, the method uses a training data set of opinion annotations [Wilson and Wiebe, 2003], where each expression is characterized by a number of attributes: Who is expressing the opinion, who or what is the target of the opinion, the type of attitude expressed by the opinion, and the subjective strength (neutral, low, medium, and high). To filter the clues and organize them into new sets based on strength these annotations are used. For each clue c and strength s, the probability of c being in an annotation of strength s is calculated by $(strength(c)) = s$. For $s = neutral$, this is the probability of c being in a neutral-strength annotation. If $P(strength(c)) = s \geq T$, for some threshold T, put c in the set for strength s. In the experiments of Wiebe, Wilson and Hwa, they set $T = (P(strength(word)) = s) + 0.25$ *or* 0.95 if $(P(strength(word)) = s) + 0.25 \geq 1$.

The authors find out that expressions marked as *high* often contain words that are very infrequent. Interestingly, only 33% of low-strength annotations are negative, compared to 78% of high-strength annotations. The researches deduce that the strength of subjective expressions will be informative for discovering polarity.

3.2 Opinion Lexicon Generation

Since opinion words and phrases are essential in many sentiment classification tasks, the following chapter discusses how these words and clauses are generated. Altogether, they are summarized into an *opinion lexicon*. Opinion words can be divided into *base* and *comparative type*. Examples of base type words are *good*, *wonderful*, *fantastic* as well as *bad*, *terrible* and *poor*. In contrast, terms of the comparative type are applied by superlative and, as we can already guess, comparative opinions (e.g., *better*, *best* and *worse*). However, the sentence “Museum A is better than museum B”do not express an opinion that any of the two museums is good or bad. Rather it says that museum A is better compared to museum B. Therefore, words of the comparative type cannot be used in the same way as words of the base type. Liu [Liu, 2010] groups comparative relations into the following four main groups:

- **Non-equal gradable comparisons**: Relations of the type *greater* or *less than* which implies an order.
- **Equative comparisons**: Relations of the type *equal to* which are stated that two objects and their features are equal.
- **Superlative comparisons**: Relations of the type *greater* or *less than all others* which rank one object over all others.
- **Non-gradable comparisons**: Relations that compare features of two or more objects, but do not grade them.

Interestingly, such sentences or phrases usually have a keyword or key phrase. Keywords are indicative words or phrases for comparisons (e.g., better, prefer, exceed, outperform, on the other hand, etc). [Jindal and Liu, 2006] found manually a list of 30 words by going through a subset of comparative sentences. Afterwards, they used WordNet (see the following Section 3.2) to find their synonyms, which results in a final list of 69 words. In addition, they included 9 phrases such as *on the other hand*, *as far as*, etc. to the list of keywords. The researchers state that words with POS tags of JJR (Adjective, comparative), RBR (Adverb, comparative), JJS (Adjective, superlative) and RBS (Adverb, superlative) are also good indicators. Therefore, the set of keywords K is defined as:

$$\begin{aligned} K = \{&JJR,\ RBR,\ JJS,\ RBS\} \bigcup \\ &\{words\ such\ as\ favour,\ prefer,\ win,\ beat,\ but,\ etc.\} \bigcup \\ &\{phrases\ such\ as\ number\ one,\ up\ against,\ etc.\} \end{aligned} \tag{3.5}$$

Jindal and Liu [Jindal and Liu, 2006] found out that 83 keywords and key phrases identify 98% of the comparative sentences with a precision of 32%.

However, to collect a complete list of opinion words, the following three approaches are commonly used: Manual approach, dictionary-based approach and corpus-based approach. Intuitively, the manual word detection is very time- and therefore cost-consuming and just with the help of the two other automated processes attainable. *Dictionary based approaches* starts with the selection of a small set of opinion words and their sentiment. Afterwards, by the help of online dictionaries (e.g. *WordNet* for English words or *GermaNet* German words) synonyms and antonyms are searched and added to the seed list of opinion words. This process stops when no more new words are found. Surely, manual revision is needed to add, correct or remove words. However, the approach has a deficit: Sentiment words which are domain dependent can be not discovered. For example, a *long* battery lifetime by technical devices is positive, however *long* queues in museums are negative. The *corpus-based approach* can overcome this limitation. The key concept, called *sentiment consistency*, was proposed by Hazivassiloglou and McKeown [Hatzivassiloglou and McKeown, 1997] in 1997. Likewise the dictionary based approach, the algorithm starts with a seed of opinion words to find others. Following, the words and a set of linguistic constraints are used to identify additional opinion words and their sentiment. For example, the AND constraint says that conjoined adjectives usually have the same sentiment (e.g., “The museum is beautiful *and* extraordinary”). Other constraints are OR, BUT, EITHER-OR and NEITHER-NOR. Another approach which extracts domain specific opinion words was introduced by Qiu et al. [Qiu et al., 2009]. The four researchers show that object features are almost associated with opinion words. The method, called *double propagation*, use extracted opinions and features to identify new opinion words and new features. Also, this algorithm is iterative and ends when no more terms can be found. In summary, the corpus-based approach is not as effective as the dictionary-based approach. The preparation of a huge corpus which covers all words is almost impossible. Furthermore, even a complete lexicon cannot guarantee an accuracy of 100%. The sentence “I search for a good family museum” contains the opinion word *good*, but express no sentiment. Therefore, additional methods and rules are needed.

Part Of Speech

Part of speech (POS) tagging deals with the determination of the correct parts of speech for a sequence of words. This natural language processing technique classifies the part-of-speech of a word to a linguistic category. The *Stuttgart-Tübingen Tagset (STTS)* is a set of 54 tags for marking German words with part-of-speech labels. The tag-set consists of 48 normal POS tags (nouns, verbs, adjectives, etc.) and 6 tag categories for foreign words, compounds, non-words and punctuation marks. The tag table used in this master’s thesis was developed by the University of Stuttgart and the University of Tübingen in 1995. The used POS classification with description and examples are declared in Table A.2 in the annex.

For detecting phrases which represent a user’s sentiment, patterns of tags are needed. A pattern includes a sequence of POS tags which identify a sentiment statement. The *sentiment extraction module* (see Chapter 4.2.4) is dedicated to the research and finding of POS patterns. But before, existing lexical resources which include a predefined list of words and their POS tag are discussed.

Numerous English lexical databases exist which classify nouns, verbs, adjectives and adverbs. For example, *WordNet*[2] is one of the largest free available dataset which includes currently 155,287 unique strings, 117,659 synsets[3] and more than 200,000 word-sense pairs. Based on this lexical database, Esuli and Sebastiani [Esuli and Sebastiani, 2006] designed the extension *SentiWordNet*[4] which enables the mapping of synsets to numerical scores $Obj(s)$, $Pos(s)$ and $Neg(s)$. These measures describing how objective, positive and negative the terms contained in the synset are. However, there exist only few German databases with qualitative and quantitative content. The *GermaNet*[5], based on the same WordNet technology, as well as *Sentiment-*

[2]WordNet `http://wordnet.princeton.edu` (accessed on August 30^{th}, 2012)

[3]Synonyms (words that denote the same concept and are interchangeable in many contexts) are grouped into unordered synsets

[4]SentiWordNet `http://sentiwordnet.isti.cnr.it/` (accessed on August 30^{th}, 2012)

[5]GermaNet `http://www.sfs.uni-tuebingen.de/lsd/` (accessed on August 30^{th}, 2012)

Wortschatz[6] (or SentiWS) will be presented and discussed in the next section. To ensure a proper analysis, minor modifications were needed due the domain of Cultural Institutions, which will be also part of this section.

GermaNet

GermaNet is a lexical-semantic database for German word senses which uses conceptual ontological information with lexical semantics, within and across word classes [Hamp and Feldweg, 1997]. The database relates German nouns, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets. Furthermore, semantic relations between these synsets are defined. Currently, GermaNet's version 7.0 contains 74,612 synsets and 99,523 lexical units. It has been developed at the research group for General and Computational Linguistics Division of Computational Linguistics of the Linguistics Department, University of Tübingen since 1997[7]. As mentioned before, the researchers use the Princeton WordNet technology for the database format, interface and database compilation. For application of the online ontology, a license is needed and is therefore not been applied in the model. However, the relations and principles of synsets are fundamental and useful for the further own research and implementation. In [Hamp and Feldweg, 1997], two basic types of relations are distinguished:

- **Lexical relations** which hold between different lexical realizations of concepts.
- **Conceptual relations** which hold between different concepts in all their particular realizations.

Synonymy and **antonymy** are lexical relations which hold for all word classes. An example for synonymy are the nouns "Karotte" and "Möhre" (carrot). An example for antonymy are "männlich" (male) and "weiblich" (female).

Pertainmy relates denominal adjectives with their nominal base ("geblümt" (flowered) with "Blume" (flower)), deverbal nominalizations with their verbal base ("Entdeckung" (discovery) with "entdecken" (to discover)) and deadjectival nominalizations with their respective adjectival base ("Müdigkeit" (tiredness) with "müde" (tired)).

The **hyponymy** relation, also called "is-a-relation", holds for words whose semantic field is included within that of another word, its *hypernym*. For instance, the word "Fußball" (football) is the hypernym of the word "Ball" (ball) and vice versa, the word "Ball" (ball) is a hyponymy of "Volleyball" (volleyball).

Meronymy, or "has-a-relation", holds only for nouns and is a part-whole relation. An example for meronymy is "Stern" (star) standing in the meronymy relation to "Sternbild" (constellation).

The **subevent** relation models temporal inclusions of verbs by the characteristic that the first event is always a subevent of the second. For example "schnarchen" (snoring) and "schlafen" (sleeping).

[6] SentimentWortschatz `http://asv.informatik.uni-leipzig.de/download/sentiws.html` (accessed on August 30^{th}, 2012)

[7] GermaNet `http://www.sfs.uni-tuebingen.de/lsd/index.shtml` (accessed on August 30^{th}, 2012)

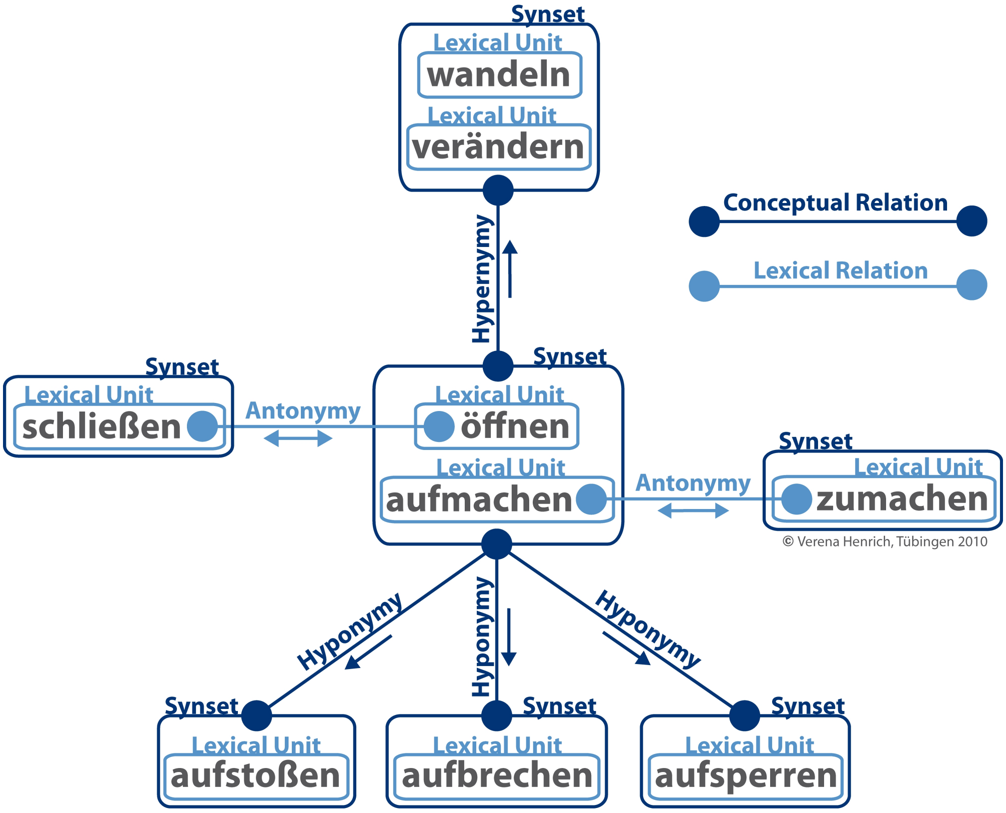


Figure 3.5: GermaNet example with multiple relations

The Figure 3.5[8] above illustrates an example including multiple relations and concepts. Besides these relations, GermaNet marks (by a “?”) and defines artificial concepts to avoid unmotivated co-hyponyms and systematic structuring of the data. The authors define **lexical gabs** which means that they can be expected to be expressed in other languages. For example, in Figure 3.6, a *nobleman* is a co-hyponymy to the three other hyponymy of human. However, the first three are related to a certain education and *nobleman* to a state of person from birth on. The additional artificial concept *?educated human* is modelled in Figure 3.7.

SentimentWortschatz

SentimentWortschatz (SentiWS) is a public available German-language resource for sentiment analysis, published by Robert Remus, Uwe Quasthoff and Gerhard Heyer [Remus et al., 2010]. The current version of SentiWS (v1.8c) contains 1,650 positive and 1,818 negative words. The resource consists of a positive and a negative text file, which implies the word, the positive and

[8]cf. `http://www.sfs.uni-tuebingen.de/lsd/documents/illustrations/relations/gn_relations_example` (accessed on August 30th, 2012)

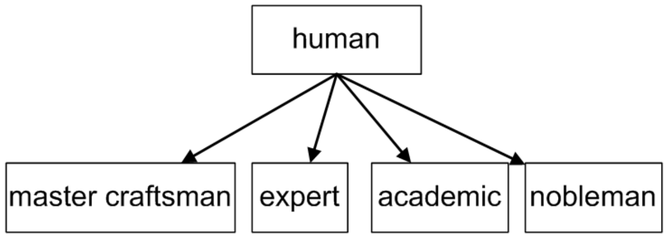


Figure 3.6: Lexical gaps without artificial concepts

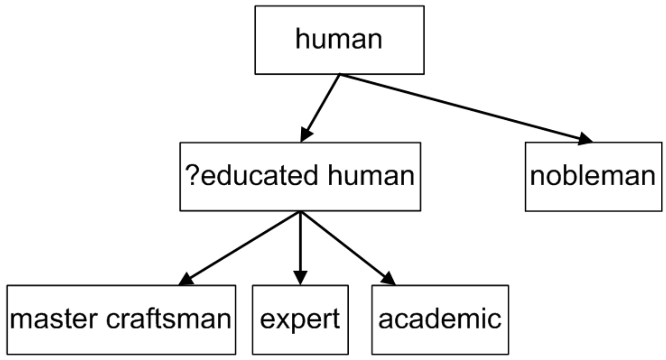


Figure 3.7: Lexical gaps with artificial concepts

negative polarity weighting within the interval of -1 and +1, their part of speech (POS) tag and, if available, their inflections. The dictionary is structured as follows:

$$< Word > | < POStag >< Polarityweight >< Infl1 >, ..., < Inflk >$$

As mentioned above, the polarity weight is a value between -1 and +1. In [Remus and Heyer, 2010], the authors give an overview about the computation and the applied approaches. The weight is calculated by utilizing a method, called Pointwise Mutual Information (PMI), which was first suggested by [Church and Hanks, 1990]. This approach was used to determine the semantic orientation and its strength of adjectives by [Turney, 2002]. In reference to [Remus and Heyer, 2010], the semantic orientation SO of a given word w is calculated from the strength of its association A with a manually-selected set of positive seed words P minus the strength of its association with a set of negative seed words N.

$$SO - A(w) = \sum_{p \in P} A(w, p) - \sum_{n \in N} A(w, n) \tag{3.6}$$

The word w is classified as having a positive semantic orientation when $SO - A(w)$ is positive. In contrast, when $SO - A(w)$ is negative the word w have a negative semantic orientation. The absolute value of $SO - A(w)$ can be considered the strength of its semantic orientation. The researchers used the following German seed sets P and N:

$$P = \left\{ \begin{array}{c} \text{gut, schön, richtig, glücklich,} \\ \text{erstklassig, positiv, großartig,} \\ \text{ausgezeichnet, lieb, exzellent,} \\ \text{phantastisch} \end{array} \right\} \tag{3.7}$$

$$N = \left\{ \begin{array}{c} \text{schlecht, unschön, falsch,} \\ \text{unglücklich, zweitklassig,} \\ \text{negativ, scheiße, minderwertig,} \\ \text{böse, armselig, mies} \end{array} \right\} \tag{3.8}$$

In reference to Section 3.1.1, the semantic associations $A(w;p)$ and $A(w;n)$ are then calculated using the PMI. In Equation 3.9, the PMI between two words w_1 and w_2 is defined as:

$$PMI(w_1, w_2) = \log(\frac{P(w_1, w_2)}{P(w_1) * P(w_2)}) \tag{3.9}$$

The probabilities that a word occurs or the words w_1 and w_2 co-occur were estimated using frequencies in a German-language corpus which consists approximately 100 Million sentences.

In case that $SO - A(w)$ of a word that was entered as being positive is negative (or vice versa). However, if $SO - A(w)$ of a term is null, is either removed, put in the opposite class or its weight is set to the minimum weight of its class. Finally, all weights are scaled to the interval of [-1; +1] and rounded to 4 decimal places. Interestingly, the distribution of the absolute weights follow a Zipf's law distribution, which means that a small number word forms have high weights, some word forms have medium weights and a large amount of word forms have little or very little weights [Remus and Heyer, 2010] (see Figure 3.8).

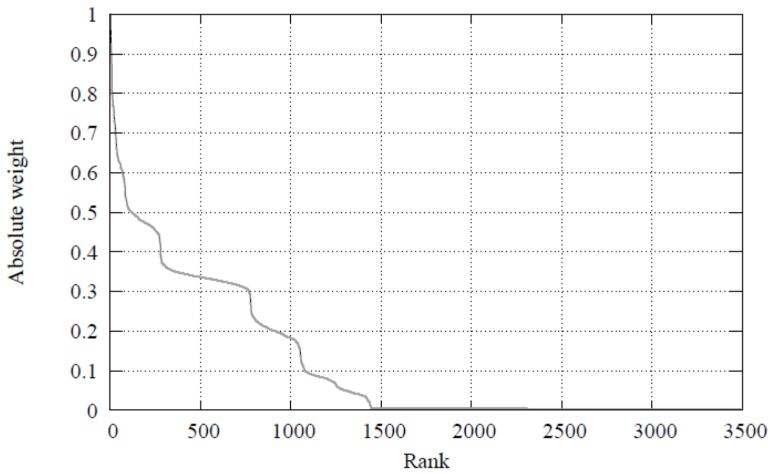


Figure 3.8: The distribution of the absolute weights [Remus and Heyer, 2010]

3.3 Feature-Based Sentiment Analysis

As mentioned before, sentiment analysis at document- or sentence-level classifies the whole document or sentence as being positive or negative. However, also a positive reviewed document does not imply that the author has positive opinions on all topics or features. Of course, that is also valid for negative opinionated documents: Not all aspects must be negative evaluated. In general user reviews and comments include positive and negative aspects. To discover such details, the opinion quintuple $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$ (see Section 2.2) is needed. Furthermore, all synonyms W_{jk} and feature indicators I_{jk} must be identified. Also the opinion holder, the object name and the time when every post is submitted is useful in some applications. The extraction of these three tasks a collectively named as *Named Entity Recognition* (NER) and well studied [Liu, 2010]. Sarawagi [Sarawagi, 2008] comprehensively summarizes algorithms which are using statistical methods for entity extraction. Besides of the NER, Liu divided feature-based sentiment analysis into the following two key mining tasks:

- **Feature extraction**: Identification of object features or aspects (e.g., in the sentence "The museum shop is too expensive", the "shop" is the aspect) and
- **Opinion orientation identification**: The determining whether the opinions on the features are positive, negative or neutral (e.g., in the above sentence, the "shop" is negative opinionated).

These two tasks are challenging and need further subtasks for problem solving. Therefore, the next two chapters describe identification algorithms on free formatted reviews for features and opinions.

3.3.1 Feature extraction

Research so far has shown that unsupervised learning methods are efficient for finding features that are nouns and noun phrases. For instance, Hu and Liu [Hu and Liu, 2004] presents a method consisting of two steps which requires a large number of reviews.

Step 1: Frequent nouns
The first step finds frequent nouns and noun phrases with the help of POS tagger. Their occurrence are counted and only terms with the highest frequency are possible features. This strategy is argued with the fact, that users which frequently talked about an aspect, they are usually important and expressed as nouns.

Step 2: Finding infrequent features
However, also not often mentioned features can be express an opinion. For detection of such infrequent features, opinions words are used: The same opinion word can be used to describe different object features. For instance, the "admission fee" is found as a frequent feature. In the sentence "The admission fee is expensive", "expensive" is a negative opinion word. By knowing this word, the feature "shop" can be extracted in the sentence "The museum shop is expensive". The *double propagation*(see Section 3.2) can be also used for feature extraction.

The idea is similar to step 2 above and starts with a set of opinion words. However, it applies dependence relations of opinion words and features which are described by dependency grammar. The iterative algorithm results in a set of syntactic rules for extracting opinion words and features. But also approaches of topic modelling and clustering are used for feature extraction in user-generated content. For instance, Titov and McDonald [Titov and Mcdonald, 2008] propose a joint model (called Multi-Aspect Sentiment (MAS) model) of text and aspect ratings for extracting text to be displayed in sentiment summaries. The model achieves high accuracy, is general and can be used for segmentation in other applications as well as domains. A clustering based method is proposed by Su et al. [Su et al., 2008] to identify explicit and implicit features. The experimental results based on real Chinese web reviews demonstrate that the method outperforms the state-of-art algorithms.

However, after feature extraction, the grouping of synonyms and the mapping of explicit to implicit features are two additional challenges. Because of the importance in the research field of sentiment analysis, both tasks and their related works are described below.

The grouping of synonyms is essential for feature analysis. For instance, when users write about the admission price they do not only use this term. Authors use synonyms like "admission charge", "entry fee", "entrance fee" or simply "admission". Thereby thesaurus dictionaries (WordNet or GermaNet) can help to identify and group similar words. Carenini et al. [Giuseppe Carenini and Zwart, 2005] proposed a method based on similarity metrics by including user- and domain-specific prior knowledge of the evaluated entity. The researchers turned the task of feature extraction into one of term similarity mapping features into a user-defined taxonomy of the entity's features with a promising accuracy.

Contrary to the assertion by Liu that mostly nouns reference to an opinion or topic, other researchers find out that adjectives are important indicators for opinions. [Hatzivassiloglou and McKeown, 1997] identified and validated constraints from conjunctions on the positive or negative polarity of conjoined adjectives. Mostly, adjectives which are conjoined with "und" (and) are usually of the same orientation. For instance, the conjoined terms "gesetzmäßig und gerecht" (legitimate and fair) imply the same polarity. The situation is reversed for but, which usually connects two adjectives of different orientations. The presented model uses this information and identifies adjectives in the following four stages:

1. All conjunctions of adjectives are extracted from the document corpus.

2. A log-linear regression model combines information from different conjunctions to determine if each two conjoined adjectives are of same or different orientation.

3. A clustering algorithm separates the adjectives into two subsets of different orientation. It places as many words of same orientation as possible into the same subset.

4. The average frequencies in each group are compared and the group with the higher frequency is labelled as positive.

This approach produces a correct classification of attributes between 78% and 92%, depending on the data test set.

As mentioned before, feature extraction indicate many adjectives and adverbs as feature indicators. However, even individual adjectives and adverbs can imply an object feature or topic. For example, the adjective “expensive” in the sentence “The museum is too expensive” can be mapped to the implicit topic “admission fee”. But, such words are domain dependent and their meaning differs often. Liu [Liu, 2010] illustrate this problem as follows: The word “heavy” usually describes the weight of an object, however in the sentence “The traffic is heavy”, “heavy” does not describe the weight of the traffic. He proposed the manual preparation of a mapping list during training data annotation. However, not many other researchers has been working on this matter and therefore an efficient approach is missing.

3.3.2 Opinion Orientation Identification

This section discusses the identification of opinions expressed on features of an object. Surely, the introduced methods on sentence-level (see Chapter 3.1.2) can be applied. Apart from this approaches, Ding et al. [Ding et al., 2008] proposes a lexicon-based approach which is divided into the following 4 steps:

Step 1: Identifying opinion words and phrases
Each positive word (e.g., “great”) or phrase is assigned the opinion score +1, each negative word (e.g., “bad”) is assigned the opinion score -1 and each neutral word (e.g., “okay”) is assigned the score 0.

Step 2: Handling negations
Negation words are used to revise the opinion score. For instance, the phrase “not great” turns the score of “great” into -1.

Step 3: But-clauses
In English and German sentences, “but”, “however” or “except” compare two features. Therefore, the sentiment before such a word and after are opposite to each other and change the opinion score. For instance, the score of the sentence “The facilities are not great[-1], but the admission fee is okay[0]” is after applying this step changed to “The facilities are not great[-1], but the admission fee is okay[+1]”. In the training dataset which consists of 438 user comments from Facebook, Foursquare, TripAdvisor, Twitter and Qype, only 7.3% of the comments contain words for comparison. However, such words are not used primarily to compare two features. For example, in the sentence “Very nice museum, unique collection, but also a nice cafe”, the term “but” is used as a conjunction. In the German-written dataset, only 6.25% of the comments which contains the term “aber” are used for comparison. In relation to the whole dataset, 0.45% of reviewers use such words to compare two features. Therefore, the fixed changing of opinion scores is not satisfactory and would falsify the result.

Step 4: Aggregation opinions
To determine the final opinion score on each object feature $f_i \in f_1, ..., f_m$ in the sentence s, the

following opinion aggregation function is applied:

$$score(f_i, s) = \sum_{op_j \in s} \frac{op_j * so}{d(op_j, f_i)}, \tag{3.10}$$

where op_j is an opinion word or phrase, $d(op_j, f_i)$ is the distance between feature f_i and opinion word op_j in s and $op_j * so$ is the opinion score of op_j. If the final score is positive, then the opinion on feature f_i is positive. Otherwise, if the score is negative, the feature is negative. A well known shortcoming is that opinion words alone cannot discover all sentiment expressions. This can be overcome by additional rules which are composed of an expression on the left and the implied opinion on the right, formally written as:

$$expression \longrightarrow opinion$$

Liu introduced [Liu, 2010] some basic rules, which are listed below:

Rule
A negative opinion word implied a negative opinion,   *a positive opinion word implied a positive opinion.*   1. Neg ⟶ Negative   2. Pos ⟶ Positive
Negated negative opinion words inverted to a positive opinion,   *negated positive opinion words inverted to a negative opinion.*   3. Negation Neg ⟶ Positive   4. Negation Pos ⟶ Negative
Positive opinion if the object feature are in the   *desired value range, otherwise negative.*   5. Desired value range ⟶ Positive   6. Below or above the desired value range ⟶ Negative
Decreasing or increasing the quantities associated with   *some opinionated terms change the sentiment of opinions.*   7. Decreased Neg ⟶ Positive   8. Decreased Pos ⟶ Negative   9. Increased Neg ⟶ Negative   10. Increased Pos ⟶ Positive

Table 3.2: Basic rules

In the first rule, Neg is a negative opinion word or phrase and Pos is a positive opinion word or phrase. Therefore, the derived sentiment is negative related on an arbitrary topic. Visa versa, the same applies for a positive opinion word or phrase Pos. Rule 3 and rule 4 represents the effect of negation words. Negations are for instance “not” or “no” which invert the sentiment of the positive or negative opinion word or phrase. In some domains, a topic or feature may have a desired value range (rule 5 and rule 6). If it is above or below the normal range it is negative, otherwise the opinion is regarded as positive. For example, “the entrance fee is 2 (or 20) Euro”. Therefore, whether the entrance fee is rated as positive or negative is highly dependent on the domain. Rules 7 to 10 processes the case that the decreasing or increasing of quantities conjunct with some opinionated topics change the orientations of the opinions. [Liu, 2010] illustrates these rules with the following example out of the medical domain: “This drug reduced my pain significantly”. In this sentence, “pain” is a negative opinion word. The reduction of “pain” indicates a desirable effect of the drug. Therefore, the decreased pain implies a positive opinion on the drug.

With the help of conceptual rules, opinions can be discovered easily. However, the creation of complex and combined rules is a difficult task as well as domain dependent. The finding of such rules in the area of cultural institution is part of Chapter 4.

3.3.3 Summary Generation

After discussing the two key mining tasks, one important step is missing: The generation of the final feature-based review summary. [Hu and Liu, 2004] proposed an opinion summarization model which categorize products by their opinion polarity (see Figure 3.9).

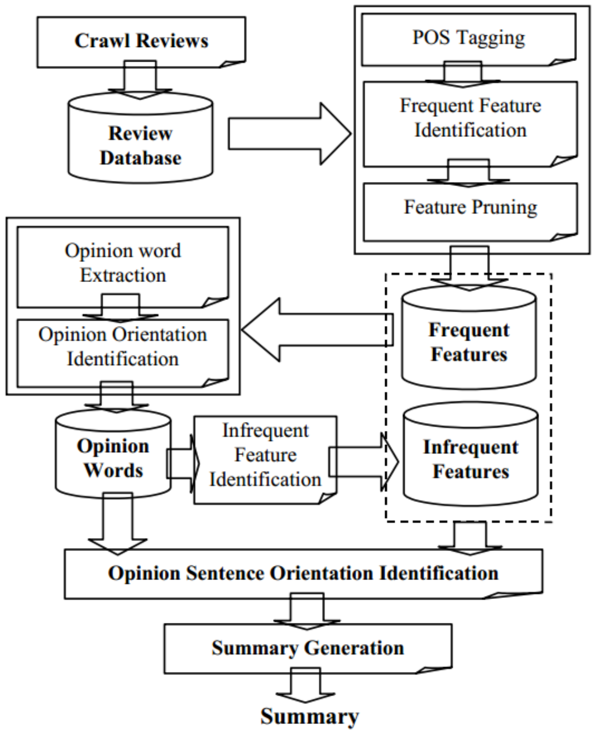


Figure 3.9: Feature-based opinion summarization [Hu and Liu, 2004]

Their most representative work in this area of study follows three main steps:

1. Mining product features that have been commented on by customers (see Section 3.3.1),
2. identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative (see Section 3.3.2), and
3. summarizing the results.

Initially, the system first downloads or crawls all the reviews and put them in the review database. Afterwards, the system finds those frequent features that many people have expressed their opinions on with the help of frequent feature identification and feature pruning. Then, the opinion words are extracted using the resulting frequent features and orientations of the opinion words which are identified with the WordNet lexical database (see Section 3.2). Using the extracted opinion words, the system then finds those infrequent features. In a next step, the orientation of each opinion sentence is identified. All of these steps are described in the sections before. In the last step, a final summary is produced by the *Summary Generation* module which follows these sub-steps [Hu and Liu, 2004]:

- For each identified feature or topic, the related opinion sentences are put into positive and negative categories according to the opinion sentences orientations.
- Afterwards, a measure is computed which shows how many reviews give positive or negative opinions to the feature.
- All features are ranked according to the frequency of their appearances in the reviews. Feature phrases appear before single word features as phrases normally are more interesting to users.

Hu and Liu suggest in their paper a variant for illustrating a feature summary. The following shows an example summary for the feature "shop" of a museum.

Feature: shop
Positive: 6

- The museum shop is highly recommended.
- After visiting the museum, you should visit the well-stocked museum shop.
- The museum shop offers many nice and inexpensive gifts.
- ...

Negative: 2

- The exhibition was great, but the museum shop was overpriced.
- The museum shop is expensive and confusing.

In 2005, Liu et al. illustrated an opinion summarization of bar graph style which is also categorized by product features [Liu, 2005]. In their paper, the authors proposed an analysis system (called Opinion Observer) with a visual component to compare consumer opinions of different products. In contrast to the feature summary of Hu and Liu in 2004, users can clearly see the strengths and weaknesses of each product. Figure 3.10 illustrates the idea by comparing customer opinions of two museums along different topic dimensions, i.e., admission fee, exhibition, staff, shop and child friendliness.

Figure 3.10: Visual comparison of consumer opinions on two museums, in reference to [Liu, 2005].

CHAPTER 4

Implementation

This chapter presents a detailed description of the practical part of this work. The first chapter provides an overview about the approach in general. In Section 4.2, the feature-based sentiment analysis model with their four modules is explained. Finally, the system output and their graphical conversion are presented.

4.1 Approach

With the help of the theoretical background from Chapter 3, a feature-based sentiment analysis system has been developed which is specified for user-generated content in the domain of cultural institutions, written in German language. The primary objective is the automatically extraction of topics and personal opinions in unstructured text documents in order to map the discovered opinions to the associated topic. The system is implemented as a PHP web-application and consists of five modules: The data pre-processing module, the opinion lexicon module, the topic extraction module, the sentiment extraction module and the summarization module (see Figure 4.1).

The data pre-processing module extracts and cleans the user-generated content from the Web and stores the data into the *Review Database*. The opinion lexicon module analyzes the user-generated content to fill the SentiWS database with domain specific words. Nouns, verbs and adjectives which imply a topic are added to the opinion lexicon. Afterwards, the topic extraction feature find opinionated topics with the help of unsupervised learning methods, especially POS pattern. Ensuing, the sentiment extraction module identifies the opinion orientation (positive, negative or neutral) of the discovered topics by multiplying the polarity weight of the topic and the opinion word. Finally, the summarization module groups similar topics with the help of a thesaurus and manually detected synonyms in the domain of cultural institutions.

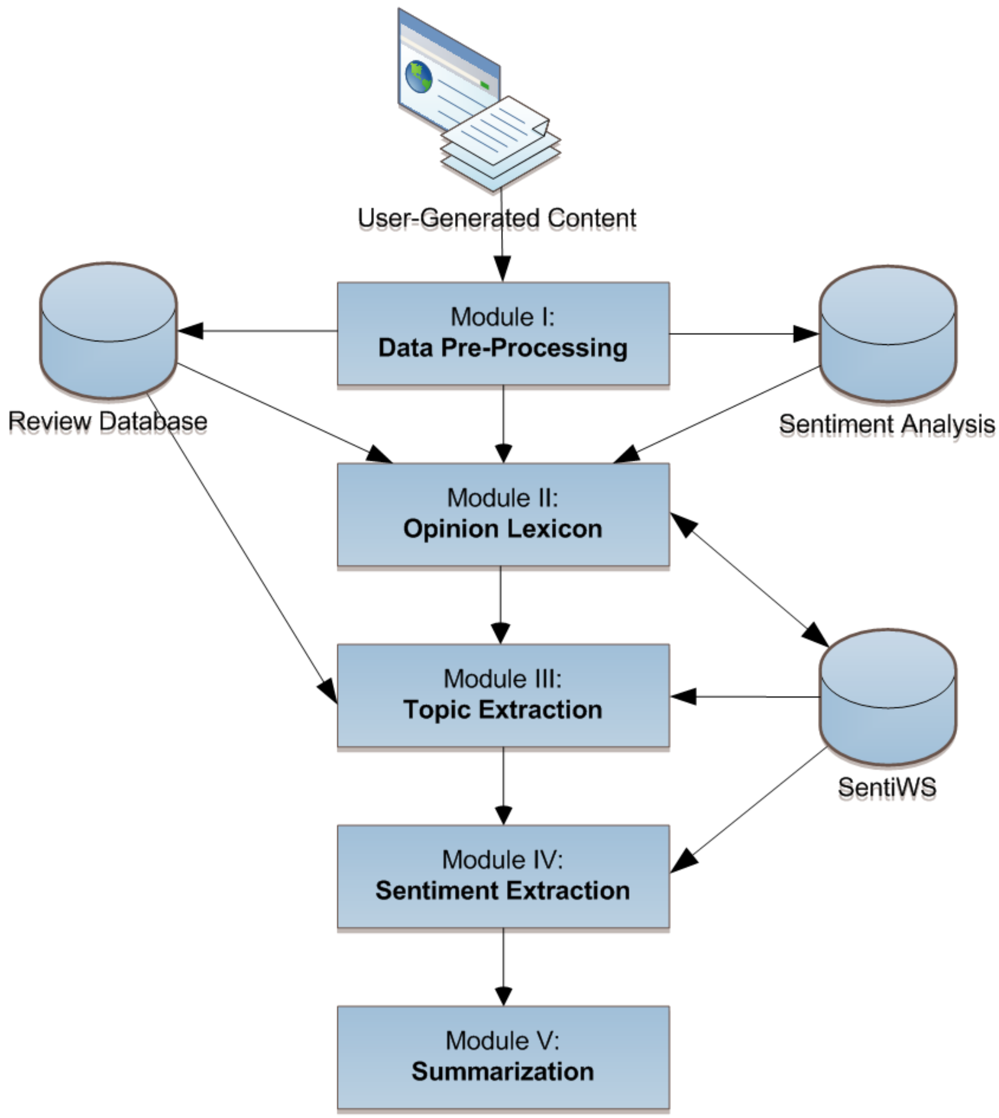


Figure 4.1: Architecture of the opinion mining system

4.2 Model

4.2.1 Data Pre-Processing

Data pre-processing is an important step in the data mining process. Based on the fact that Internet users also publish irrelevant, redundant, noisy and unreliable information, the cleaning of the extracted data is required. Therefore, before the implementation can commence, the *removing of stop words*, the *stemming*, the *part of speech (POS) tagging* and the *term frequency* computation of each word has been carried out. The next sections specify these four sub processes. The results of these data pre-processing are used to complete the opinion lexicon and extraction of topics and sentiments.

For further research, the complete pre-processed training data set is provided as a CSV and SQL file and contains the id, data source, institution, review, review without stop words, stemmed review and POS tagged review. In addition, the *Sentiment Analysis* database which contains the computed term frequency measures for each word in the training data set is provided.[1]

Stop Word Removal

Intuitively, not every word in a sentence is required for sentiment analysis. For instance, the Foursquare tip "*Mit einer Jahreskarte an der Info lassen sich die Schlangen an den Kassen umgehen.*" (With an annual ticket at the information, queues can be avoided at the tills.) contains conjunctions, articles, adverbs, pronouns, etc. (*function words*) which have little lexical meaning or have ambiguous meaning. These words are unessential for topic and sentiment extraction and can be filtered at first. Not a single definite list of stop word is existing, however several proposals by researchers. This master's thesis combines the German stop word list from Dr. Martin Porter[2] and Ranks.nl[3]. Altogether, 265 words are collated which are summarized in Table A.1. The stop word removal algorithm is applied on all reviews and remove these words. Additionally, all punctuation marks and special characters are replaced. Afterwards, the edited reviews are stored separately in the database. This step improves and speeds up further analytical tasks. The sentence of the introductory example is changed to "*Mit Jahreskarte Info lassen Schlangen Kassen umgehen*" which implies nonetheless all meaningful words.

Stemming

Stemming is the process for reducing inflected or derived words to their stem. A German stemmer should reduce the words "beeindruckendes", "beeindruckender", "beeindruckendste" or "beeindruckendster" to the word "beeindruckend" (awesome). This step is useful for many information retrieval assessments. In the research field of sentiment analysis, topics and sentiments can be combined. In 1968, Julie Beth Lovins published the first stemmer[4]. However, the most widely used stemmer algorithm was published in 1980 by Dr. Martin Porter. In 2001, he builds the "Snowball" framework[5] for writing stemming algorithms for several languages. Since this thesis analyze reviews written in German language, the Porter stemmer was programmed

[1] Training set for cultural institutions (CSV and SQL file): `www.dub-in-sky.de/Masterthesis/Trainingset.zip`

[2] available at `http://snowball.tartarus.org/algorithms/german/stop.txt` (accessed on August 30^{th}, 2012)

[3] available at `http://www.ranks.nl/stopwords/german.html` (accessed on August 30^{th}, 2012)

[4] The Lovins stemming algorithm, available at `http://snowball.tartarus.org/algorithms/lovins/stemmer.html` (accessed on August 30^{th}, 2012)

[5] Snowball: A language for stemming algorithms, available at `http://snowball.tartarus.org/texts/introduction.html` (accessed on August 30^{th}, 2012)

for PHP by following these steps[6]:

Notes:
The following letters are *vowels*: a, e, i, o, u, y, ä, ö, ü
R1 is the region after the first non-vowel following a vowel, or is the null region at the end of the word if there is no such non-vowel.
R2 is the region after the first non-vowel following a vowel in R1, or is the null region at the end of the word if there is no such non-vowel.

- **Step 1:**
  (a) Replace *ß* by *ss*, and put *U* and *Y* between vowels into upper case.
  (b) R1 and R2 are first set up in the standard way (see the note on R1 and R2), but then R1 is adjusted so that the region before it contains at least 3 letters.

- **Step 2:**
  Search for the longest among the following suffixes,

  (a) *em* | *ern* | *er*
  (b) *e* | *en* | *es*
  (c) *s* (preceded by a valid s-ending)

  and delete if in R1. If an ending of group (b) is deleted, and the ending is preceded by *niss*, delete the final *s*.

  For example: äckern -> äck, ackers -> acker, armes -> arm, bedürfnissen -> bedürfnis

- **Step 3:**
  Search for the longest among the following suffixes,

  (a) *en er est*
  (b) *st* (preceded by a valid st-ending, itself preceded by at least 3 letters)

  and delete if in R1.

  For example: derbsten -> derbst by step 1, and derbst -> derb by step 2, since b is a valid st-ending, and is preceded by just 3 letters)

- **Step 4:**
  Search for the longest among the following suffixes, and perform the action indicated.
  end | *ung*
  delete if in R2

[6] The steps of the German stemming algorithm: `http://snowball.tartarus.org/algorithms/german/stemmer.html` (accessed on August 30^{th}, 2012)

if preceded by *ig*, delete if in R2 and not preceded by *e*
ig | ik | isch
delete if in R2 and not preceded by *e*
lich | *heit*
delete if in R2
if preceded by *er* or *en*, delete if in R1
keit
delete if in R2
if preceded by *lich* or *ig*, delete if in R2

- **Step 5:**
  Turn *U* and *Y* back into lower case, and remove the umlaut accent from a, o and u.

To the best of my knowledge, there is no German stemming implementation for PHP available. For this reason, the function was programmed in reference to the rules of Dr. Martin Porter (see above) and the JavaScript implementation of Joder Illi[7]. The complete $stemm(\$word)$ function can be downloaded on my personal web server `http://www.dub-in-sky.de/Masterthesis/GermanPorterStemmerPHP.zip`.

POS Tagging

With the help of the SentiWS database (see Section 3.2), each review is POS tagged as follows: Does SentiWS contains the word itself or is the word an inflection, the database returns the POS tag, sentiment and polarity. Afterwards, the information is attached behind the word, formally structured as

$$[posTag|sentiment|polarityWeight]$$

For instance, the user comments about the Museumsquartier Vienna "*In der Kantine ist der Nudelauflauf zu empfehlen!!!*" is after stop word removal and POS tagging stored as "*Kantine[NN|neutral|0.004] Nudelauflauf empfehlen[VVINF|positive|0.004]*". The word *Kantine* is identified as a neutral noun and *empfehlen* as a positive infinitive. However, *Nudelauflauf* is not an element of the SentiWS database. When the tagging process is finished, each tagged review is separately stored into the review database.

The limited number of classified words results in a poor analysis outcome. By applying SentiWS on unprepared reviews, only 6% of words are correctly POS tagged (see Table 4.1).

There are two causes for this result: On the one hand, the occurrence of non-German reviews and on the other hand, the appearance of nouns which are not classified. For this reason, language detection and lexical modifications are needed to raise the total percentage of tagged words.

[7] German Porter Stemmer in JavaScript: `https://gist.github.com/942312` (accessed on August 30^{th}, 2012)

Data source	Qype	Foursquare	Twitter	TripAdvisor	Total
Number of Reviews	116	217	168	129	630
Number of words	11975	3139	2240	9704	27058
Number of POS tagged words	739	116	70	698	1623
% of tagged words	**6,17%**	**3,70%**	**3,13%**	**7,19%**	**6,00%**

Table 4.1: POS tagged words

Language Detection

Since this master's thesis focuses on the analysis of German written user comments, the data extraction process should only select German reviews, tweets and tips. One might think that this task should be slightly realized by the usage of APIs. However, some social media platforms or local review websites provide no public API or an API parameter for language restricting.
Foursquare offers no API method to retrieve only German user reviews. In contrast, the Twitter API queries foreign-language Tweets which are tagged with a "de" language code. However, the manual language evaluation of the predefined data set finds out, that more than 7% are wrongly classified. The verification of the data extraction process discovered two main problems, which can be summarized on the one hand by

- the **missing language classification methods** and on the other hand by
- the **wrong language classification** of existing detection methods

by some APIs. Both problems could be solved by manual evaluation and language identification. However, this task is very time consuming and expensive. Furthermore, the model works only with predefined data sets. A language detection method would allow the analysis of dynamic and actual data sources, without the need of human preprocessing.

The automatic language detection of a given text is a complex task. Several algorithms and scientific papers circulate in the World Wide Web for solving this problem. This master's thesis evaluates an approach by calculating recall, precision and F-measure.

Founded in 1999 by Stig S. Bakken, the community driven **PHP Extension and Application Repository (PEAR)** project provides a structured library of open-source code for PHP users. PEAR code is written in PHP and segmented in packages, which expresses a separate project with its own development team. For language detection, the $Text\_LanguageDetect$ package is provided for free download and allows the identification of 52 human languages from text. The implemented detection method is based on the *n-gram-based text categorization* approach by William B. Cavnar and John M. Trenkle [Cavnar and Trenkle, 1994]. The method returns a list of all possible languages which are detected by correlating ranked 3-gram frequencies to a table of 3-gram frequencies of known languages.

For evaluation, discovering of true positive, false positive, true negative and false negative is necessary. However, it is not required that the method detects for each comment the correct language, but rather, if the text is written in German or not. For example, it is not necessary that the Foursquare tip

"Un museu natural como cualquier otro. Si puedes ahorrartelo, hazlo."

is classified as a Spanish written comment, but of course as a non-German review.

The following Tables (4.2 and 4.3) presents the result of the classification by applying the $detect(String)$ method of more than 200 Foursquare and more than 150 Twitter entries.

		Text is written in German	
		True	**False**
Classified as	**Positive**	76	0
German text	**Negative**	5	136

Table 4.2: Language classification of Foursquare entries

		Text is written in German	
		True	**False**
Classified as	**Positive**	131	0
German text	**Negative**	3	34

Table 4.3: Language classification of Twitter entries

In reference to Chapter 5.1 precision, recall and F-measure are calculated (see Table 4.4). Especially the ideal precision value and the high recall argue for the n-gram-based text categorization method and their application.

Feature	Precision	Recall	F-measure
Language detection	1	0,95	0,98

Table 4.4: Evaluation of language classification (PEAR) for Foursquare and Twitter

By applying the language detection method, the percentage of POS tagged words was increased by 0.41%. Surely this not a crucial increment, however a required component to avoid subsequent analysis errors.

Data source	Qype	Foursquare	Twitter	TripAdvisor	Total
Number of Reviews	116	76	133	129	454
Number of words	11975	1328	1877	9704	24884
Number of POS tagged words	739	89	70	698	1596
% of tagged words	**6,17%**	**6,70%**	**3,73%**	**7,19%**	**6,41%**

Table 4.5: POS tagged words after language detection

Term Frequency

For further analyzing tasks, certain measures are needed which count and weight terms to find out how important a word is to a document in relation to others. The term frequency tf_{ij} measures the occurrence frequency of term i in document j. However, the length of a document is not taken into account. The inverse document frequency (idf) is a normalization factor for the characteristics of term distribution in the whole document collection. Therefore, the $tf * idf$ measure reflects how important a word is to a document in a collection or corpus. It combines multiplicatively tf and idf components and can be mathematically formulated as:

$$tf * idf(t_i, d_j, N) = tf(t_i, d_j) * log\frac{N}{df_i} \tag{4.1}$$

, where t_i is the term i,
d_j is the document j,
N is the total number of documents in the corpus and
df_i is the number of documents where the term i appears.

For instance, the term "Museum" (museum) totally occurs in German written reviews 215 times (N = 215). In a randomly selected Qype review (134 words), the term occurs two times. Following the previously defined formulas, the term frequency for "Museum" is therefore:

$$tf = (2/134) = 0.015$$

In the training set of 630 documents, *Museum* occurs in 215 of these. Then, the idf is calculated as:

$$idf = log(630/215) = 2.93$$

Therefore, the $tf * idf$ of the term "Museum" is:

$$tf * idf = 0.015 * 2.93 = 0.044$$

In conclusion, the $term\ frequency$, $inverse\ document\ frequency$ and the combined $term\ frequency - inverse\ document\ frequency$ is computed for each word in the data set. By using real-time data from social media platforms, new extracted content is analyzed with the help of these measures. All values are stored separately into the *Sentiment Analysis Database* and contains the document id, the word, and the corresponding tf, idf and $tf * idf$ measure.

4.2.2 Opinion Lexicon Module

The objective of the opinion lexicon module is the POS tagging, the computation of the polarity weight and the discovering of the sentiment (positive, negative or neutral) of all words which imply a topic or sentiment. Basis for this process is the SentiWS database which was introduced in Section 3.2. The resource includes positive and negative adjectives, adverbs, nouns and verbs. However, only 6% of all words in the training set are tagged. To improve the percentage of tagged words, an algorithm is applied which finds new words iteratively by already known words and their term frequency.

Feature I: Single opinion and topic word detection using $tf * idf$

An intuitive approach is the analysis of the word frequency. The idea behind this method is that words which are frequently occurring are possible topics. Otherwise, words which are infrequently used are may be domain specific words and can also imply a topic. The objective is to find a value range which implies both, infrequent and frequent topics, and exclude non-opinion words. For this reason, the module uses the $tf * idf$ measures, which states how important a word is to a document within a whole document corpus. The numbers are the result of the data pre-processing module. In a first step, all words are listed by their occurrence in the document corpus (term frequency). However, it was not possible to derive a generic pattern for topic and opinion word detection. Therefore, a list of opinion related words which are not already stored in the SentiWS database was manual selected. On the basis of this list and the associated $tf*idf$ measures, the range was observed with the following approximation method: The $tf*idf$ measure of all words were compared with the $tf * idf$ measure of the manual detected words. The objective is to find a boundary which includes as many as possible topic and opinion words and exclude as many useless words as possible which imply no topic or opinion. A comparison of Figure 4.2 and Figure 4.3 shows, that almost all words are in the range from about 2 to 6. Therefore, in this range it can not be derived whether a word implies a topic or opinion, or none of them. However, the following characteristic can be determined: More topic and opinion words are in the range about 6.1. The evaluation (see Section 5.3.1) showed that nearly 10% of topic and opinion words are above a $tf * idf$ value of 6.1. Furthermore, the word set contains only 0.41% of non-topic or non-opinion words. For this reason, words with a $tf * idf$ measure greater than 6.1 are added to the SentiWS database.

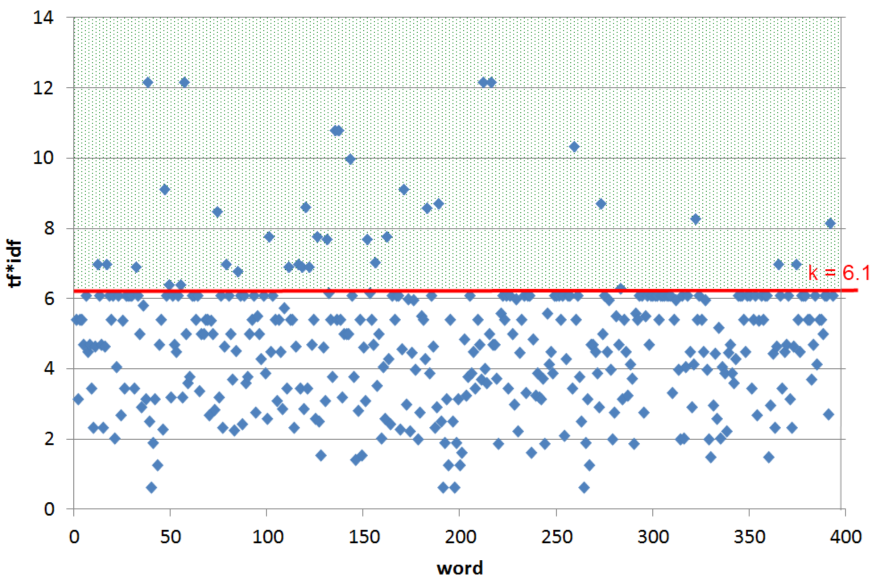


Figure 4.2: td*idf value range of words which imply a topic or opinion

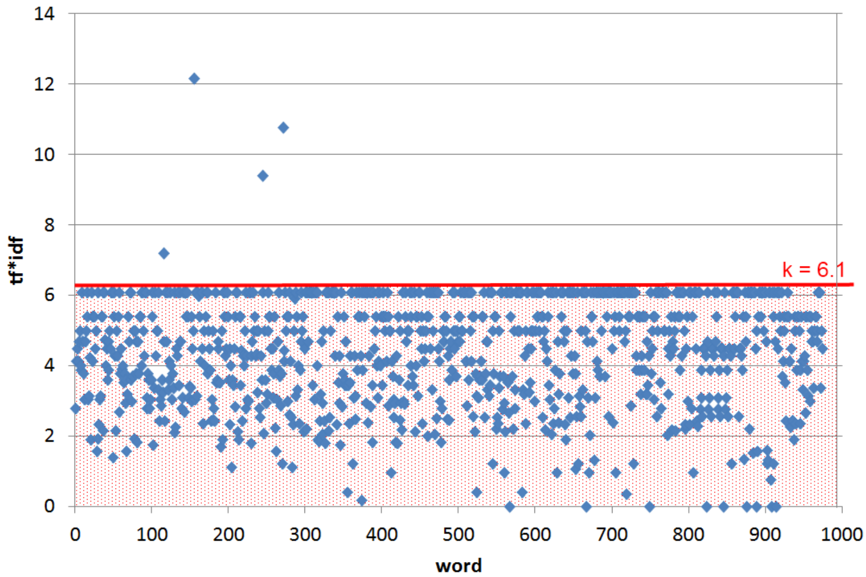


Figure 4.3: td*idf value range of words which imply non-topic and non-opinion word

To improve the result, the algorithm uses this range to expand the opinion lexicon by new and unknown real-time social media data. Afterwards, the polarity weight is calculated in reference to Section 3.2. In contrast to the described method by [Remus et al., 2010], the set of positive or negative seed words are not manually selected. Rather the set consists of all already known positive or negative weighted opinion words in the SentiWS database. Afterwards the semantic association $A(w;p)$ and $A(w;n)$ is computed using the PMI (see Section 3.2). The PMI between two words w_1 and w_2 is defined by:

$$PMI(w_1, w_2) = \log(\frac{P(w_1, w_2)}{P(w_1) * P(w_2)}) \tag{4.2}$$

For instance, word w_1 is "Wechselausstellung" (temporary exhibition) and w_2 is "phantastische" (amazing), than $P(Wechselausstellung)$ is the probability that "Wechselausstellung" occurs and $P(Wechselausstellung \,\&\, phantastische)$ is the probability that both words co-occur. The PMI is calculated with all positive words in the set (e.g., "schöne" (beautiful), "exzellente" (excellent), etc.). The $A(w;p)$ is the sum of all these values. The same applies for the negative words and results in the $A(w;n)$ measure. Afterwards, the semantic orientation $SO-A(w)$ is computed by the subtraction of $A(w;p)$ and $A(w;n)$:

$$SO - A(w) = \sum_{p \in P} A(w,p) - \sum_{n \in N} A(w,n) \tag{4.3}$$

All weights are scaled to the interval of [-1; 1] and rounded to 4 decimal places with +1.0 being absolutely positive and -1.0 being absolutely negative [Remus et al., 2010]. Finally, the word itself, the polarity weight and the sentiment (positive, negative or neutral) is added to the SentiWS database.

So far, only topics were identified which consists of one single word. However, topics and proper names of exhibitions or institutions can for sure consists of two or more words. For instance, the name of the exhibition "Monet bis Picasso - Die Sammlung Batliner" (Monet to Picasso - The Batliner Collection) consists of five words. For this reason, the opinion lexicon module uses n-grams to discover such topics. Therefore word-level n-grams (for n = 2,3,...,n_{max}) are extracted and the number of times they appeared are counted. In contrast to the discovering of single words, the term frequency of word pairs are not summed-up over the whole data set and all cultural institutions. Based on the fact that names of exhibitions depends on the cultural institution, the term frequency in documents about a respective museum expresses more about the content of word pairs. For example, a name of a specific exhibition from the *Albertina* in Vienna will not occur in user-generated content about the *Deutsches Museum* in Munich.

After reviewing the n-grams, three types of word pairs can be identified:

1. Word pairs which imply a topic (e.g., "graphischen Sammlung"),
2. word pairs which include a opinion word and a topic (e.g., "schöne Wechselausstellung") and
3. word pairs which imply no opinion word and no topic (e.g., "und die").

Feature II: Composited topic word detection using double propagation

For processing case 1, the most frequent bi-grams (n = 2) are extracted. However, the number of occurrences says per se nothing whether the word implies a topic and/or an opinion word. It was also not possible to find value ranges or other patterns which discover such words. Therefore, also useless word pairs (e.g., "und die" [and the], "ist in" [is in], "mit dem" [with the]), etc.) are outputted by the n-gram function. To detect topics which consist of more than one word (e.g., "permanenten Ausstellung" [permanent exhibition]), "wechselnde Ausstellungen" [current exhibition], "graphischen Sammlung" [graphic collection], etc.), the module uses the double propagation method (see Section 3.2). The iterative algorithm uses already extracted opinions and topics to identify new opinion words and new topics and ends when no more terms can be found. If a word is an already collected noun within the word pair (e.g., "Ausstellung" [exhibition]) and the word before is a known verb or adjective with a neutral polarity weight (e.g., "permanent(e)" [permanent], "wechseln(de)" [changing]), the word pair is added as a neutral topic to the SentiWS database. Furthermore, word pairs are identified as composited topics which are consisting of two words with uppercased initial letters (e.g., "Handschriften Laubers" or "Bibliotheca Albertina"). However, this rule holds in the case that one of the words is not yet covered as a positive or negative adjective or verb. With this restriction, word pairs at the beginning of the sentence (e.g., "Schöne Wechselausstellung" [beautiful temporary exhibition]) are not extracted as a composited topic. After processing all discovered word pairs in the bi-gram, these steps are repeated for the n-grams until no further topics can be found (see Algorithm 1. Thereby, the parameter $nMax$ defines the maximum number of n-grams which should be generated. Ordinarily, the algorithm starts with bi-grams (n = 2).

Algorithm 1: DetectCompositedTopics(D, n, nMax)

```
input: A set of documents D of a cultural institution
input: The n-gram parameter n
input: Maximum number of n-grams nMax
tag all words which are already known;
NGramList ← NGrams(D);
for i ← 1 to NGramList.size() do
    if word(n).POStag == NN then
        if word(n − 1).POStag == ADJ or ADV or V then
            if word(n − 1).polarity == neutral then
                SentiWS.AddTopic(NGramList[i]);
    if word(n − 1) has a capital initial letter AND word(n) has a capital initial letter
    then
        if word(n − 1).POStag AND word(n).POStag == null then
            SentiWS.AddTopic(NGramList[i]);
if nMax != n then
    DetectCompositedTopics(D, n + 1, nMax);
```

Feature III: Single topic and opinion word detection using double propagation

As mentioned in Section 3.1.1, *opinion words* are adjectives, adverbs or verbs before or after a noun which imply a positive or negative sentiment (i.e., “schön” (beautiful), “ungenügend” (unsatisfactory), etc.). Mostly, a topic is enveloped from an opinion word and vice versa (i.e., “ein schönes Museum”, “die Ausstellungsräume sind schlecht”). The algorithm uses this characteristic to find out not yet recognized topics and opinion words.

To detect word pairs which include an opinion word and a topic (case 2), the algorithm is divided into two procedures: The first procedure indicates topics and domain specific terms with the help of already known opinion words. The second procedure detects unrecognised opinion words when a recorded topic before or after occurs. The algorithm will be illustrated by using the following exemplary word pairs: “schöne Wechselausstellung” (beautiful temporary exhibition), “extravagante Wechselausstellung” (extravagant temporary exhibition) and “extravagante Ausstellungsstücke” (extravagant exhibits) (see Figure 4.4).

At first, the $POSTag(D)$ method tags all words in the document list which are included by default in the SentiWS database. In reference to Section 4.2.1, the information is attached behind each word as $[posTag|sentiment|polarityWeight]$. For each word in the document set, the algorithm checks if the word is tagged and therefore already included in the database. Otherwise, if the word is not tagged and has a capital initial letter, the sentence structure will be reviewed. Is the word before or after a sentiment adjective, adverb or verb, the $Add(word)$ procedure stores the word as a new topic into the database (i.e., the word “Wechselausstellung”, because of the adjective “schön”). In the case that at least one word has been added, the

Step	Word 1	Word 2
1	schöne adjective, positive, included in the SentiWS	Wechselausstellung unknown word
2	schöne	Wechselausstellung added to the SentiWS as noun, neutral, based on the adjective
3	extravagante unknown word	Wechselausstellung noun, neutral, included in the SentiWS
4	extravagante added to the SentiWS as adjective, positive, based on the adjective	Wechselausstellung
5	extravagante adjective, positive, included in the SentiWS	Ausstellungsstücke unknown word
6	extravagante	Ausstellungsstücke added to the SentiWS as noun, neutral, based on the adjective

Figure 4.4: Example of topic and opinion word detection

$DetectOpinionWords(D)$ (see Algorithm 3) procedure is called to detect opinion words. In exactly the same way as before, the $DetectOpinionWords(D)$ procedure tags all words which are included by default in the SentiWS database and new added by the $DetectTopics(D)$ (see Algorithm 2 procedure. However, the procedure retrieves nouns to determine opinion words around them. If it is an opinion word before or after a noun, and the word is not yet recorded, the word is added to the database (i.e., the adjective “extravagante”, based on the noun “Wechselausstellung”). If new words are added, the algorithm rolls back to the $DetectTopics(D)$. Therefore, the noun “Ausstellungsstücke” can be added on the basis of the adjective “extravagant”. The iterative algorithm terminates when no more topics or opinion words can be found. However, the algorithm achieves only good results under the assumption that nouns are written with a capital initial letter.

4.2.3 Topic Extraction

As the name suggests, the topic extraction module deals with the discovering of discussed topics. As mentioned in Section 4.1, topic extraction is based on unsupervised learning methods. With the help of phrase structures the module extracts topics. However, topics which are not sentiment related, but objective statements (facts) are ignored. In contrast to the opinion lexicon module, only topics are extracted which are related to an opinion word. For instance, the sentence “Das Museum hat einen Shop und ein Restaurant” (The museum has a shop and a restaurant) discusses the topics “Shop” and “Restaurant”, but reflects no sentiment on these terms. In contrast, the terms will be extracted in the sentence “Das Museum hat einen schönen Shop und ein günstiges Restaurant” (The museum has a nice shop and a reasonably priced restaurant) based on the adjectives “schön” and “günstig”.

For discovering phrase structures, topics from 438 reviews from Facebook, Foursquare, TripAdvisor, Twitter and Qype are manually filtered out. By processing the POS tagged reviews

Algorithm 2: DetectTopics(D)

input: A set of documents D

$DTagged \leftarrow$ `SentiWS.POSTag(`D`);`
$wordsAdded$ = false;
for $i \leftarrow 1$ **to** $DTagged.size()$ **do**
  if $word[i].POStag$ == *null* **then**
    if $word[i]$ *has a capital initial letter* **then**
      if $word[i-1].POStag$ *OR* $word[i+1].POStag$ == $ADJorADVorV$ **then**
        $opinionWord = word[i-1]$ OR $word[i+1]$;
        if $opinionWord.polarity$ *!= neutral* **then**
          $SentiWS.Add(word[i])$;
          $wordsAdded$ = true;

if $wordsAdded$ == *true* **then**
  `DetectOpinionWords(`D`);`

Algorithm 3: DetectOpinionWords(D)

input: A set of documents D

tag all words which are already known;
$DTagged \leftarrow$ `SentiWS.POSTag(`D`);`
$wordsAdded$ = false;
for $i \leftarrow 1$ **to** $DTagged.size()$ **do**
  if $word[i].POStag$ == *NN* **then**
    if $word[i-1].POStag$ *OR* $word[i+1].POStag$ = *null* **then**
      $opinionWord = word[i-1]$ OR $word[i+1]$;
      $SentiWS.Add(word[i])$;
      $wordsAdded$ = true;

if $wordsAdded$ == *true* **then**
  `DetectTopics(`D`);`

from the training set, the phrase structures close to a topic (n words before and n words after a topic) are analyzed. By variation the parameter n on a scale from 1 to 5, the POS tags of the n neighbours before and after a topic are printed out. The objective is, to find patterns of phrase structures which indicate sentiment related topics. For this reason, the algorithm returns a list of POS tagged words before and after each manual selected topic of n-neighbours. However, the analysis resulted in the conclusion that in German written reviews, the distance between topic word and opinion words is arbitrary and therefore no pattern can be deductive reasoned. In the worst case, the opinion word is at the beginning and the topic word at the end of a sentence (e.g., "**Schön** ist auch die hinauf führende **Freitreppe**"). Based on this awareness, the module discovers sentiment related topics with the help of another characteristic: The number of topic and opinion words occurring in a sentence. Depending on the number of words, the algorithm processes each case in a different approach. For instance, if two topic words and one opinion word occurs in a sentence, the algorithm discover which topic word is the sentiment related one. The following four characteristics can be determined and differentiated:

- **Case 1:** The sentence includes a topic word ($n_{tw} = 1$) and an opinion word ($n_{ow} = 1$) (e.g., "Ein **schönes Museum**")
- **Case 2:** The sentence includes more than one topic word ($n_{tw} > 1$) and an opinion word ($n_{ow} = 1$) (e.g., "Das **Restaurant** in dem *Museum* ist sehr **schön**")
- **Case 3:** The sentence includes a topic word ($n_{tw} = 1$) and more than one opinion word ($n_{ow} > 1$) (e.g., "Das **Museum** ist **günstig** aber **überfüllt**")
- **Case 4:** The sentence includes more than one topic word ($n_{tw} > 1$) and more than one opinion word ($n_{ow} > 1$) (e.g., "Das Museum hat einen **schönen Shop** und ein **günstiges Restaurant**")

The first and most simple case, which assumes that exactly one topic word and exactly one opinion word occurs, is implemented as following: The algorithm disjoints each review on sentence-level. Therefore, the user-generated content is divided in case of a full stop ("."). Afterwards, with the help of the SentiWS database is checked whether a word or word pair in the sentence is a topic. If this is the case, the detected opinion word is analyzed. In the training data set, opinion words are adjectives, adverbs or verbs. Has the opinion word a positive or negative polarity weight and is therefore positive or negative opinionated, the topic word is extracted as a sentiment discussed topic. In this case, the distance of the topic and opinion word is irrelevant, since the opinion word don't have to follow directly the topic word or vice versa. The derived rule can be formulated as:

Case	Rule	Topic Word tw_n	Opinion Word ow_n	Constraints
1	1	n = 1 (NN or ADJ)	n = 1 (ADJ or ADV or V)	polarity weight of ow_1 = positive or negative

Table 4.6: Case 1, Rule 1 (one topic word and one opinion word)

However, the German language is much more complex. More than one topic word can appear in a sentence while just one opinion word occurs. To process the second case, the model analyzes the whole structure of a sentence. The challenge is to find the right topic word which is related to the opinion word.

For instance, with the help of the completive SentWS database, the words "Museum" and "Sitzgelegenheiten" are identified as topics in the sentence "In dem Museum sind die zu wenigen

Sitzgelegenheiten zu bemängeln". Based on the German grammar, the sentence can also be written as "Die wenigen Sitzgelegenheiten im Museum sind zu bemängeln". Therefore, the position of the opinionated topic ("Sitzgelegenheiten") and the opinion word ("bemängeln") provides no information for the topic extraction process. The related topic word is not always closer to the opinion word than the unrelated topic word ("Museum"). After reviewing the training dataset, sentences with one opinion word and more than one topic word have one characteristic: The topic words can be classified hierarchically in an "is-part-of-relation" or "has-a-relation" (see Section 3.2). Almost always, the word at the lowest hierarchical level is the related topic word. Referring to the example at the beginning, "Sitzgelegenheiten" is a part of a "Museum". Therefore, the noun "Sitzgelegenheiten" is the sentiment related topic. For this reason, a relation tree with domain specific topic words is created and implemented. All topics occurring in sentences which follow this characteristic are printed out and manually mapped to a relation tree. The following Figure 4.5 shows an exemplary part of the tree.

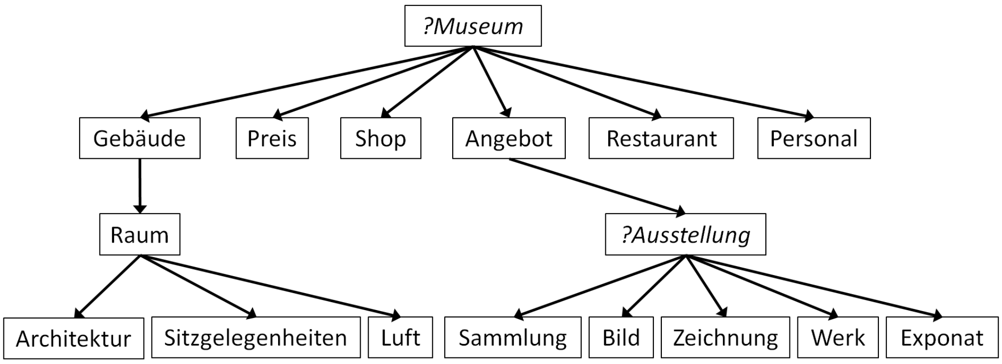


Figure 4.5: Dependence tree

In reference to Section 3.2, artificial concepts are used and marked by a "?" to avoid repeatedly mentions. For example, the concept "?Ausstellung" can imply the word "Ausstellung" itself, but also specific names of exhibitions, or the terms "Wechselausstellung", "Sonderausstellung", "Dauerausstellung", etc. The rule can be formulated as (Table 4.7):

Case	Rule	Topic Word tw_n	Opinion Word ow_n	Constraints
2	2	n > 1 (NN or ADJ)	n = 1 (ADJ or ADV or V)	Polarity weight of ow_1 = positive or negative; Topic word on the lowest level of the hierarchy

Table 4.7: Case 2, Rule 2 (more than one topic words and one opinion word)

However, this rule can be revoked in the case of enumerations. For instance, the user opinion "Das Museum, die Architektur, das Restaurant und der Shop haben mir sehr gut gefallen" evaluates each topic as positive. Consequently, rule 3 extract each topic in a sentence as sentiment related topic word if they are separated by commas (Table 4.8):

Case	Rule	Topic Word tw_n	Opinion Word ow_n	Constraints
2	3	n > 1 (NN or ADJ)	n = 1 (ADJ or ADV or V)	polarity weight of ow_1 = positive or negative; Topic words are separated by commas

Table 4.8: Case 2, Rule 3 (enumeration of more than one topic words and one opinion word)

In conclusion, rule 3 does not use the relation tree to exclude topics. The reversion of rule 2 leads to the extraction of each word in the enumeration as a topic.

Case 3 fires when a sentence includes exactly one topic word and more than one opinion words. This is the case when users express their sentiment about a topic with more than one opinion word. In the German language, such adjectives, adverbs and verbs can be conjunct with[8]:

> und, sowohl als auch, weder noch, und auch nicht, und noch weniger, einerseits, anderseits, nämlich, oder, entweder oder, sonst, aber, jedoch, sondern, nur, außer, weil, da, denn, daher, deshalb, obwohl, trotzdem, selbst wenn, damit, so dass, soweit, sofern, vorausgesetzt, angenommen, bis, seit, seitdem, sobald, bevor, während, als, solange, nachdem, wenn, als, jedesmal, wenn, jetzt, da, wenn, falls, wenn es so ist, es sei denn, wenn es nicht so wäre, denn, wenn, ob, ob oder ob nicht, ob oder ob nicht, als ob, wie, so ... wie, (größer) als, so viel (...) wie, nicht mehr als, verglichen mit

To illustrate this point, in the sentence “Es gibt ein viel zu kleines und teures Café” are two negative adjectives which are connected with the conjunction „und“. Both opinion words refer to the topic word “Café”. For sure, this rule also holds when opinion words are enumerated and/or connected with conjunctions (e.g., “Die Sonderausstellung ist absolut sehenwert, groß, aufregend aber leider teuer”). Therefore, rule 4 can be expressed as follow (Table 4.9):

Case	Rule	Topic Word tw_n	Opinion Word ow_n	Constraints
3	4	n = 1 (NN or ADJ)	n > 1 (ADJ or ADV or V)	Polarity weight $ow_1, .., ow_n$ = positive or negative; Opinion words are separated by commas or/and conjunctions

Table 4.9: Case 3, Rule 4 (one topic words and more than one opinion words)

[8] cf. `http://www.braesicke.de/conaisc.htm` (accessed on August 30^{th}, 2012)

Case 4 implies that more than one topic word and more than one opinion word occur in a sentence. Therefore, the objective is to map the right opinion words to their related topic. Obviously, a topic can be described by one (Case 1) or more than one (Case 3) opinion words. In the phrase, “Ansonsten zu bemängeln: wenige Sitzgelegenheiten und ein zu kleines und teures Café”, the topic “Sitzgelegenheiten” is opinionated by one adjective (“wenige”) and the topic “Café” by two adjectives (“kleines” and “teures”). Furthermore, not all occurring topics must be sentimental discussed (Case 2). Therefore, case 4 can combine all previous treated cases. For processing, the sentence must be divided into subsets.

In a first step, the sentence is analyzed for commas. The sentence is separated by each comma which was not detected within an enumeration in rule 3 or rule 4. If there is no comma in the sentence, the algorithm skips the first step. In the next step, the distance between each topic word and opinion word is calculated in each subset or sentence. Thereby, enumeration and conjunction are handled as one single opinion word (see Figure 4.6).

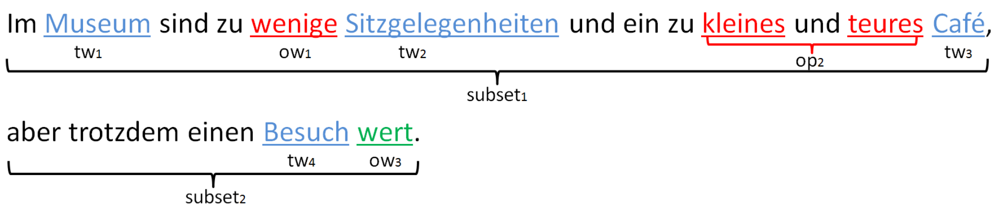


Figure 4.6: Sentence with more than one topic word and more than one opinion word

Topics are extracted as sentiment related which are closest to opinion words. In the first subset, the distance between each opinion word and each topic word is measured. The topic word with the minimum distance is marked as a sentiment topic. Table 4.10 represents the distance calculation of ow_1 to tw_1, tw_2 and tw_3. Since, the fourth topic word tw_4 is in the second subset, the word is not listed in the table.

Opinion Word ow_n	**Topic Word tw_n**	**Distance between ow_n and tw_n**
ow_1	tw_1	3
$\boldsymbol{ow_1}$	$\boldsymbol{tw_2}$	**1**
ow_1	tw_3	8

Table 4.10: Distance calculation (step 1)

Because of the minimum distance between ow_1 and tw_1, the word “Sitzgelegenheiten” is marked as a sentiment related topic. The word is removed from the list and then the algorithm calculates the distance between ow_2 to tw_1 and tw_3. As shown by Table 4.11, tw_3 is also identified as a sentiment related topic. Since there are no more opinion words left in $subset_1$, the algorithm terminates and the word “Museum” is not extracted as a sentiment discussed topic. In $subset_2$, exactly one topic word and one opinion word occurs. Therefore, rule 1 of the first case is applied and the word “Besuch” is extracted. The generalized rule is formulated in Table 4.12.

Opinion Word ow_n	Topic Word tw_n	Distance between ow_n and tw_n
ow_2	tw_1	8
$\boldsymbol{ow_2}$	$\boldsymbol{tw_3}$	**1**

Table 4.11: Distance calculation (step 2)

Case	Rule	Topic Word tw_n	Opinion Word ow_n	Constraints
4	5	n > 1 (NN or ADJ)	n > 1 (ADJ or ADV or V)	Polarity weight ow_n = positive or negative; Topic words tw_n with a minimum distance to an opinion word ow_n

Table 4.12: Case 4, Rule 5 (more than one topic words and more than one opinion words)

Algorithm 4: TopicExtraction(D)

input: A document D
$Document \leftarrow$ `SentiWS.POSTag(`D`);`
$Sentences \leftarrow$ `Document.split(.);`
for $i \leftarrow 1$ **to** $Sentences.size()$ **do**
  tw = getNumberOfTopicWords($Sentence[i]$);
  ow = getNumberOfOpinionWords($Sentence[i]$);
  if tw == *1 AND* ow == *1* **then**
    `Rule1(`tw, ow`);`
  if tw > *1 AND* ow == *1* **then**
    `CheckEnumeration` *returns* $true$ *or* $false$ *if the topic words are an enumeration or not*
    if `CheckEnumeration(`$tw_1, .., tw_n$`) ==` $false$ **then**
      `Rule2(`$tw_1, .., tw_n, ow$`);`
    if `CheckEnumeration(`$tw_1, .., tw_n$`) ==` $true$ **then**
      `Rule3(`$tw_1, .., tw_n, ow$`);`
  if tw == *1 AND* ow > *1* **then**
    `Rule4(`$tw_1, ow_1, .., ow_n$`);`

```
for i ← 1 to Sentences.size() do
    if tw > 1 AND ow > 1 then
        Sentence.split returns subsets of a sentence, separeted by commas (not
        enumerations) Subsets ← Sentence.split(,);
        for j ← 1 to Subsets.size() do
            tw_subset = getNumberOfTopicWords(Subset[i]);
            ow_subset = getNumberOfOpinionWords(Subset[i]);
            if tw_subset == 1 AND ow_subset1 == 1 then
                Rule1(tw, ow);
            if tw_subset > 1 AND ow_subset1 == 1 then
                Rule2(tw_1, .., tw_n, ow);
                Rule3(tw_1, .., tw_n, ow);
            if tw_subset == 1 AND ow_subset1 > 1 then
                Rule4(tw_1, ow_1, .., ow_n);
            if tw_subset > 1 AND ow_subset1 > 1 then
                Rule5(tw_1, .., tw_n, ow_1, .., ow_n);

function Rule1(tw, ow) {
if ow.polarity != neutral then
    ExtractTopic(tw);
}
function Rule2(tw_1, .., tw_n, ow) {
CheckHierarchy returns the topic word with the lowest hierarchy depth
identifiedTopicWord = CheckHierarchy(tw_1, .., tw_n);
if ow.polarity != neutral then
    ExtractTopic(identifiedTopicWord);
}
function Rule3(tw_1, .., tw_n, ow) {
if ow.polarity != neutral then
    ExtractTopic(tw_1, .., tw_n);
}
function Rule4(tw_1, ow_1, .., ow_n) {
if CheckEnumeration(ow_1, .., ow_n) == true then
    if opinionWord.polarity != neutral then
        ExtractTopic(tw_1);
}
function Rule5(tw_1, .., tw_n, ow_1, .., ow_n) {
calculateDistance(ow_subset1, .., ow_subsetN);
ExtractTopic(tw_1, .., tw_n);
}
```

4.2.4 Sentiment Extraction

The fourth module deals with the detection of the sentiment and opinion orientation from the extracted topics. In reference to the topic extraction module in the previous section, sentiments are discovered for each different case. However, there are other factors which must be taken into consideration. There are rules established for each of the four cases which handle opinion phrases and negations words (Step 1 and Step 2). These steps are needed to determine the precise polarity weight of each opinion word or phrase. The calculation of topic's sentiment (positive, neutral or negative) depends on the respective case (Step 3).

Step 1: Opinion phrases
In a first step, the analyzed sentences from the third module are passed to the sentiment extraction module. For each detected topic, the associated opinion word is inspected. Intuitively, opinions are expressed not only with a single word. After reviewing the training dataset, the evaluation revealed that an opinion phrase consists nearly always of two words. The following four patterns are identified:

POS tag ow_1	POS tag ow_2	Example opinion phrase
Adjective (ADJX)	Verb (VVINF)	gut gefallen
Adjective (ADJX)	Noun (NN)	hohe Qualität
Adjective (ADJX)	Personal Pronoun (PRF)	eignet sich
Adjective (ADJX)	Adjective (ADJX)	schön große

Table 4.13: POS patterns of opinion phrases

The polarity weight of the opinion phrase is calculated by the average of the polarity weight of both opinion words:

$$polarityWeight_{phrase} = \frac{(polarityWeight_{ow1} + polarityWeight_{ow2})}{2} \tag{4.4}$$

Step 2: Negation words
Negation words have been ignored so far. However, such terms revise the opinion score. For example, "das Museum ist nicht lehrreich" (the musem is not informative), the term "nicht" turns the polarity weight of the opinion word "lehrreich". There are several ways to negate opinion words which are listed below[9]:

- kein (no, not any, not a, none, no one, nobody)
- nicht (not)
- gar nicht, überhaupt nicht (not at all)
- nicht mehr (no more, no longer, any more)

[9] In reference to `http://class.georgiasouthern.edu/german/grammar/gr-neg.htm` (accessed on August 30^{th}, 2012)

- nie mehr (not ever [again])
- nie (never)
- noch nicht (not yet)
- noch nie (not ever)
- nichts (nothing, not...anything)
- niemand (nobody, not ... anybody)

Therefore, the algorithm searches for these terms in front of an opinion word or phrase. When this occurs, the polarity weight of the opinion word is inverted. For instance, “schönes” has a positive polarity weight of 0.0081, which is calculated by default in the SentiWS resource by the PMI (see Section 3.2). In the subset “kein schönes Museum”, the score reverses to a negative polarity weight of -0.0081 and therefore to a negative opinionated topic.

Step 3: Sentiment extraction
In case 1 (see Table 4.6) exactly one opinion word/phrase and exactly one topic word occurs in a sentence. Therefore to determine the sentiment of a discussed topic, the polarity weight of the opinion word is multiplied by the polarity weight of the topic word. If the result is a positive value, the topic is interpreted as positive opinionated. Otherwise, if the result is a negative value, the topic is interpreted as negative opinionated. If the opinion word and the topic word have a neutral polarity weight, the topic word is determined as neutral reviewed.

In the second case (see Table 4.7 and Table 4.8), more than one topic word and exactly one opinion word occurs within a sentence. This case is divided into two sub cases: In the first, the opinion refers to just one topic word, but more than one topic word appear. In this connection the sentiment is defined by multiplication as in Case 1 and only one topic is highlighted as positive, neutral or negative. In the second sub case, the opinion word refers to any topic word within an enumeration. Therefore, each topic is marked as sentiment discussed term depending on the polarity weight of the opinion word and the topic word itself. The formula for Case 1 and Case 2 is as follows:

$$sentiment(tw_n) = polarityWeight_{twn} * polarityWeight_{ow1} \tag{4.5}$$

Case 3 (see Table 4.9), exactly one topic word and more than one opinion word were written in a sentence. The opinion words are conjoined by enumerations or conjunctions. Intuitively, not every word has to be positive or negative. Instead, some words can be positive, whereas others can be negative or neutral. The following sentence is to illustrate this point: “Das Museum ist sehenswert, billig aber klein”. For this reason, any opinion word polarity is involved for the sentiment estimation. The polarity weights are summed up and divided by the number of opinion words. As in the previous cases, the sentiment is derived from the positive or negative result. The formula can be written as:

$$sentiment(tw_1) = polarityWeight_{tw1} * \sum_{i=1}^{n} \frac{(polarityWeight_{ow1} + ... + polarityWeight_{own})}{n} \tag{4.6}$$

As mentioned in the previous section, case 4 can be a combination of prior cases. The occurrence of more than one opinion word and more than one topic word can be operated in sentences and subsets as follows: When exactly one opinion word or phrase evaluates one topic word, the sentiment is determined according to Equation 4.5. If n opinion words describe one topic word, Equation 4.6 is applied. Vice versa, when an opinion word judge about more than n topics, again Equation 4.5 is used for calculating the users sentiment. In addition, to include the distance between each opinion word/phrase and topic word $d(ow_i, tw_n)$, the opinion score for each topic t in a sentence s is determined. For instance, in the sentence "Die Sonderausstellung ist sehenswert", the distance $d(sehenswert, Sonderausstellung)$ is 2. In the sentence "Die Sonderausstellung, die wir letztes Wochenende mit der ganzen Familie erfahren durften, ist sehenswert", the distance $d(sehenswert, Sonderausstellung)$ is 12. The aggregation function (see Section 3.3.2) was as follows adjusted:

$$score(tw_n, s) = \sum_{ow_i \in s} \frac{polarityWeight(ow_i)}{d(ow_i, tw_n)} \tag{4.7}$$

where ow_i is an opinion word or phrase, $d(ow_i, tw_i)$ is the distance between topic tw_i and opinion word ow_i in s and $polarityWeight(ow_i)$ is the polarity weight of ow_i. If the score is positive, then opinion on topic t is positive. Otherwise, the sentiment to topic t is negative.

4.2.5 Summarization

The objective of the fifth module is on the one hand the summarization of a single cultural institution, and of the other hand the comparison of two cultural institutions by their opinionated topics. The feature-based sentiment classification is summarized using a website. The "Social Media Analysis for Cultural Institutions" platform is written in the server-side scripting language PHP. Apart from the sentiment analysis, a structural analysis will be implemented by my colleague Christina Stödtner which answers questions about the quality and quantity of relations between users. The combination of both scientific analyses (sentiment and structural) in the area of web science allows a complete and meaningful benchmark tool for cultural institutions. For the single summarization, the process is similar to conventional search or analysis websites: The user enters or select a cultural institution and after clicking on the "analyze"-button, the results will be visualised appropriately. The extracted topics and sentiments are represented by pie charts. By clicking on a sentiment, all user comments with their filtered words which imply a topic or sentiment are displayed. For this reasons, the web service uses and combines different graphic technologies which are described below.

- **Pure Javascript Tabs** For separation of the two analysis categories, Matt Walker[10] implemented the tabs with the help of HTML5 and CSS3. It allows an easy adaptability and a well looking content separation.

- **Google Chart Tools** The extracted topics and sentiments of a cultural institution should be represented graphically with the help of pie charts. Thereby a topic is reflected by a pie chart which chart elements are the sentiments positive, negative and neutral. The Google Chart Tools[11] provides a package which implements exactly these requirements. The easy configuration and data interface allows a ideal data visualization.

- **Cascading Style Sheets (CSS)** Text, buttons, tables and other elements are individual customized with CSS. Some HTML elements (e.g. the popup window and the "analyze"-button[12]) uses already existing presentation semantics.

The following screen design (see Figure 4.7) intended to provide an impression of the graphical interface and the summarization of the user's sentiment.

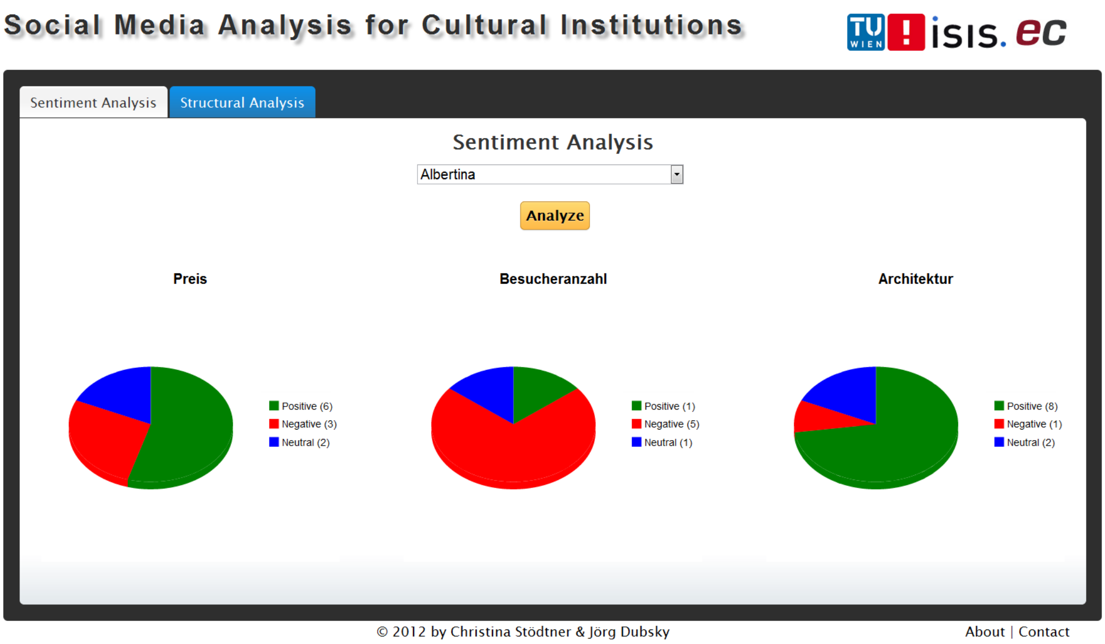


Figure 4.7: Screen design of the "Social Media Analysis for Cultural Institutions" web platform

[10] Pure Javascript Tabs are documented at `http://www.my-html-codes.com/javascript-tabs-html-5-css3` (accessed on August 30^{th}, 2012)

[11] Google Pie Charts are documented at `https://google-developers.appspot.com/chart/interactive/docs/gallery/piechart` (accessed on August 30^{th}, 2012)

[12] Popup Modal Window is documented at `http://87studios.net/webtuts/build-a-popup-modal-window-using-the-jquery-reveal-plugin/` (accessed on August 30^{th}, 2012)

By clicking on an extracted topic, the related user reviews are displayed in a pop-up window. All topic and opinion words are highlighted depending on the computed sentiment (positve = green, neutral = blue, negative = red) (see Figure 4.8).

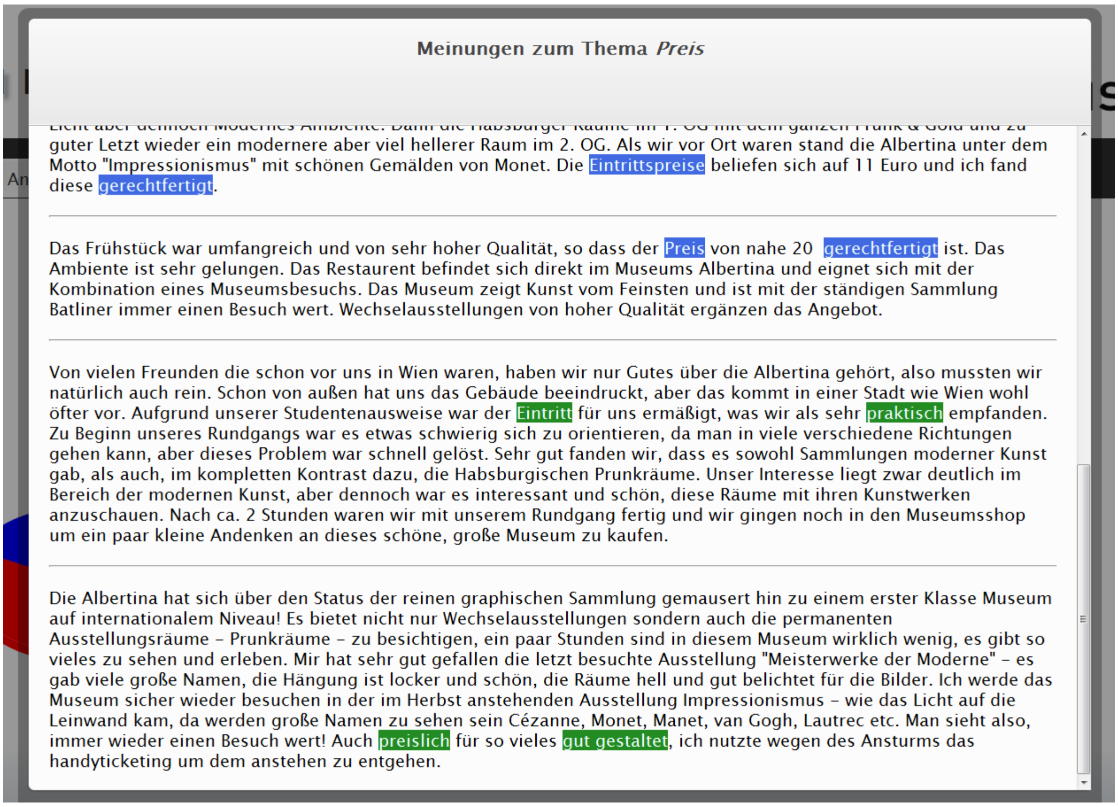


Figure 4.8: Pop-up window with the highlighted user-generated content

In reference to Section 3.3.3, the summary of topics was adjusted and is carried out by the following sub-steps:

1. Each identified topic is put into positive, neutral or negative categories according to the opinion orientation.
2. Every word is returned to its root word with the help of the Porter stemmer algorithm (see Section 4.2.1)
3. Words with the same root word are grouped (e.g., “Preise”, “preislich” and “Preises” are grouped to the topic “Preis”)
4. A topic and any of its synonyms are also summarized with the help of the dependence tree (e.g., “Eintritt”, “Kartenpreis” and “Eintrittspreis” are grouped to the topic “Preis”)
5. Afterwards, a measure is computed which shows how many reviews give positive, neutral or negative opinions to the topic.

6. All pie charts which including the opinionated topics are ranked according to the frequency of their appearances in the reviews.

For the comparison of two museums, the user selects two museums and after clicking on the "analyze"-button, the results will be visualised in a column chart. In contrast to the single summarization, only topics which are discussed in both museums are displayed. This restriction allows a clear and quick overview of two comparing institutions. Obviously, the summary steps above are performed for each museum before they are compared. Afterwards, on the x-axis the topics are mapped. The y-axis shows the total number of each positive or negative rated topic. Figure 4.9 illustrates the graphical implementation of the comparison module.

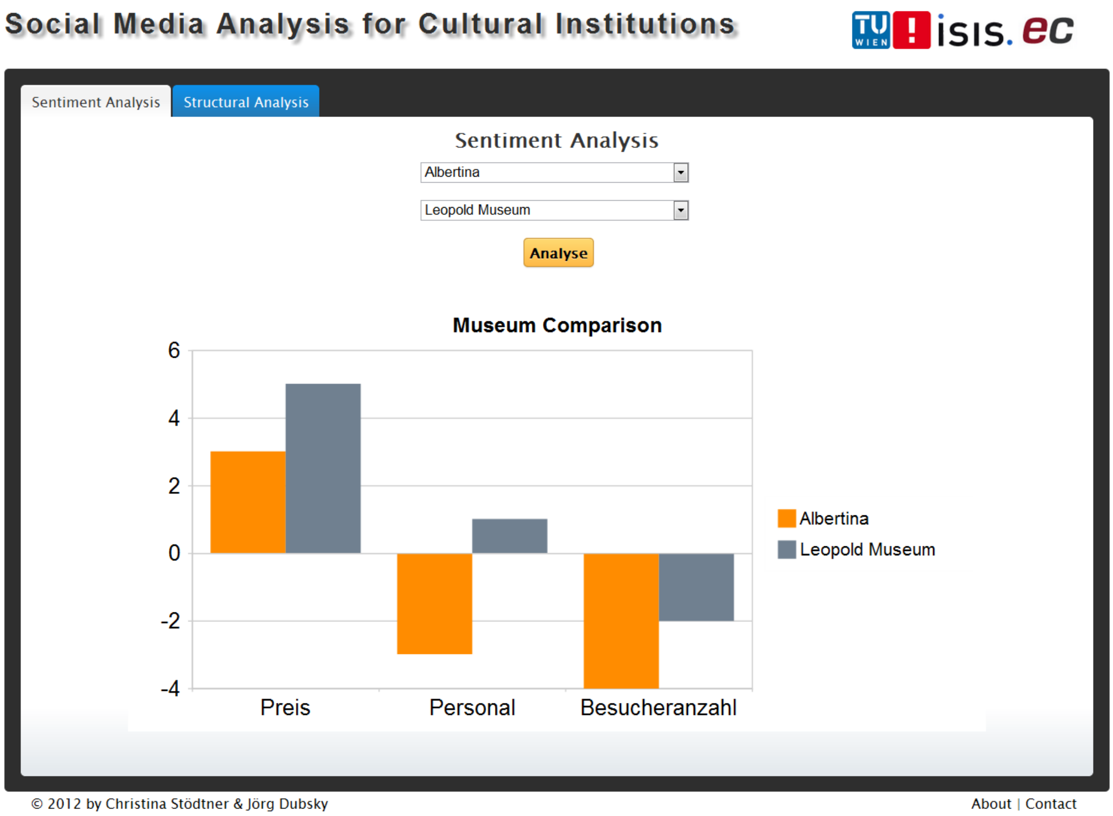


Figure 4.9: Screenshot of the comparison of two cultural institutions

CHAPTER 5

Evaluation

5.1 Evaluation in Sentiment Analysis

In other scientific fields, the evaluation of research results with analytical methods is common used. This method proves whether the system is correct and complete by measuring the efficiency and checking the formal specification of the problem. However, sentiment classification is not evaluable and solvable as a formal logic problem by deductions. Therefore, experimental evaluation of a classifier is typically conducted experimentally which states the ability to take the right classification decision by measuring its effectiveness. [Tang et al., 2009] The effectiveness can be plotted by *precision and recall* as well as *correlation coefficient and relative error* which are formally described in the next sections.

5.1.1 Precision and recall

In the traditional research of information retrieval, *precision* is defined as the fraction of retrieved instances that are relevant and *recall* as the fraction of relevant instances that are retrieved. In other words, precision and recall are calculated with the following formulas:

$$precision = \frac{Number\ of\ relevant\ instances\ retrieved}{Total\ number\ of\ instances\ retrieved} \tag{5.1}$$

$$recall = \frac{Number\ of\ relevant\ instances\ retrieved}{Total\ number\ of\ relevant\ documents} \tag{5.2}$$

However, there is a trade-off between precision and recall: A high precision returns relevant instances but misses many other useful instances. In contrast, a high recall returns most relevant documents but includes lots of junk. Therefore, an effective classifier needs both: High precision and recall value (see Figure 5.1).

For classification tasks, the terms true positives, true negatives, false positives and false negatives compare the results of the classifier under test with external judgements. The terms positive

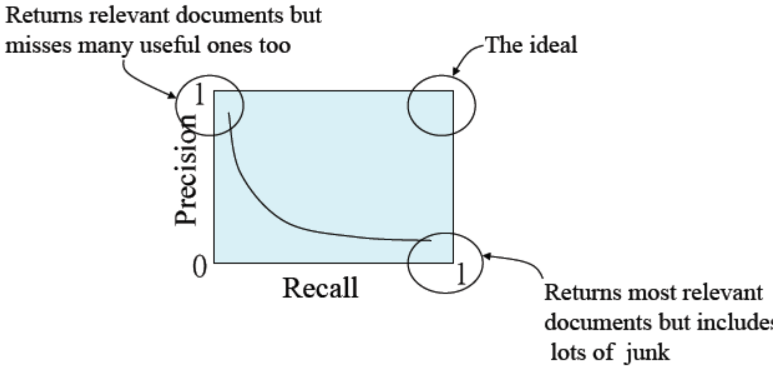


Figure 5.1: Trade-off between precision and recall [Merkl, 2011]

and negative refer to the classifiers' observation and the terms true and false refer to whether that prediction corresponds to the external judgement. This is illustrated by the following Table 5.1:

		Condition	
		True	**False**
Test	**Positive**	True positive (TP)	False positive (FP)
outcome	**Negative**	False negative (FN)	True negative (TN)

Table 5.1: Classification between TP, TN, FP and FN

In relation to sentiment classification, precision is the number of correct cases in the system output based on the manual judgement divided by the number of all cases that the system assigns either a positive or negative sentiment. In contrast, recall is the ratio of correct cases that the system assigned compared to the base of all cases where a human analyst associated either positive or negative sentiments manually. [Tang et al., 2009] Generally and formally expressed as:

$$Precision = \frac{\#\ of\ correct\ cases\ based\ on\ the\ manual\ judgement}{\#\ of\ all\ cases\ that\ the\ system\ assigned\ either\ as\ positive\ or\ negative} \tag{5.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.4}$$

$$Recall = \frac{\#\ of\ correct\ cases\ based\ on\ the\ manual\ judgement}{\#\ of\ all\ cases\ that\ the\ human\ assigned\ either\ a\ positive\ or\ negative} \tag{5.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.6}$$

The traditional *F-measure* combines precision and recall and is also known as the *F1 measure*, because both values are evenly weighted.

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.7}$$

However, in the domain of sentiment classification of user-generated content it is often acceptable to calculate *accuracy* instead of recall. Tang et al. [Tang et al., 2009] argue that by the fact that customers are not always interested in having all possible reviews, but just a few positive and a few negative. Therefore, sometimes accuracy is more important than recall and is defined as:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.8}$$

5.1.2 Correlation coefficient and relative error

The *correlation coefficient* is a standard measure of the degree to which two variables are linearly related. [Mishne and de Rijke, 2006] For instance, the measure can detect the fluctuation of a sentiment which are predicted by the model. A correlation coefficient of 1 means that there is a perfect linear relation between the prediction and the actual values. However, a coefficient of 0 means that the prediction is completely unrelated to the actual values. [Tang et al., 2009]

$$CorrCoefficient = \frac{S_{PA}}{S_P * S_A}, where \tag{5.9}$$

$$\begin{aligned} S_{PA} &= \frac{\sum_i (p_i - \bar{p}) * (a_i - \bar{a})}{n-1}, \\ S_P &= \frac{\sum_i (p_i - \bar{p})^2}{n-1}, \\ S_A &= \frac{\sum_i (a_i - \bar{a})^2}{n-1}, \text{ where} \end{aligned}$$

p_i is the estimated value for document i
a_i is the actual value for instance i
$\bar{p}$, $\bar{a}$ is the average of p respectively a and
n is the total number of instances.

The *relative error* represent the mean difference between the actual and the estimated values and is defined by

$$RelError = \frac{\sum_i (|p_i - a_i|)}{\sum_i (|a_i - \bar{a}|)} \tag{5.10}$$

5.1.3 Benchmark

In [Tang et al., 2009], the authors consider that experiments are comparable only if the following three conditions are:

1. On exactly the same collection (i.e., same documents and same categories)
2. With the same "split" between training set and test set
3. With the same evaluation measure, and the same parameter values

Therefore, a lack of just one of these three conditions make the comparison of the experimental results almost impossible. For this reason, researchers evaluate their classifier with standardized data sets. For instance, the *Amazon Product Review Data*[1] collection contains more that 5.8 million reviews and is applied in many papers. The data set was proposed in [Jindal and Liu, 2008] and implies information about reviewers, review texts, ratings and product info. Another, frequently used dataset is used to benchmark sentiment classification on movie reviews. The *Internet Movie Database*[2] consists of 42,000 reviews in HTML form. However, to the best of my knowledge, there exists no collection which is generally applied as training and testing set in the domain of cultural institutions. Therefore, an own composed data set will be discussed in the next section.

5.2 Dataset

For feature extraction and sentiment classification the Web Service applies a predefined training dataset of 438 entries. The document corpus consists of user generated content from Facebook, Foursquare, TripAdvisor, Twitter and Qype that was extracted during April 2012 to June 2012. The data collection is based on an arbitrary selection of cultural institutions in the German speaking area which are listed below:

1. **Albertina (Vienna, Austria)** The Albertina is a museum which exhibits approximately 65,000 drawings, modern graphic works, photographs and architectural drawings.[3] The museum uses Facebook, Twitter, publishes videos via the video-sharing website Vimeo and gets good rating from local review sites.
2. **Deutsches Museum (Munich, Germany)** The Deutsches Museum is the world's largest museum of technology and science with about 28,000 exhibited objects from 50 fields of science and technology.[4] They participate on Facebook and they are highly recommended by reviewers. However, the museum does not attend on Twitter and other social media platforms.

[1] available at `http://liu.cs.uic.edu/download/data/` (accessed on August 30^{th}, 2012)
[2] available at `http://www.imdb.com/reviews/index.html` (accessed on August 30^{th}, 2012)
[3] cf. `http://www.albertina.at/die_sammlung` (accessed on August 30^{th}, 2012)
[4] cf. `http://www.deutsches-museum.de/information/` (accessed on August 30^{th}, 2012)

3. **Grünes Gewölbe (Dresden, Germany)** The Grünes Gewölbe is a historic museum that contains the largest collection of treasures in Europe.[5] There exists no own Facebook Page or Twitter Account. But there are many user reviews written on Foursquare and other local review platforms.

4. **Haus der Musik (Vienna, Austria)** Opened in 2000, the Haus der Musik is the first museum of sound and music in the world.[6] They use the social networks Facebook, Twitter and YouTube. Surely, many reviews are written on Foursquare, Tripadvisor, Qype and other portals.

5. **Kunsthistorisches Museum (Vienna, Austria)** The Kunsthistorisches Museum is an art museum and had a yearly average of about 600,000 visitors.[7] With the help of Facebook, Twitter, Google+ and YouTube, the museum stay in contact with their visitors.

6. **Leopold Museum (Vienna, Austria)** The Leopold Museum is home of one of the largest collections of modern Austrian art and known for the Egon Schiele and Gustav Klimt collection.[8] Also this cultural institution uses Facebook to spread information. In addition, since 2009 a YouTube channel is actively used.

7. **Technisches Museum Wien (Vienna, Austria)** Opened in 1918, the Technisches Museum Wien exhibit the history of technique in several categories.[9] Apart from numerous reviews on rating websites, users state their comments and opinions on the Facebook Page.

8. **MUMOK (Vienna, Austria)** The Museum Moderner Kunst Stiftung Ludwig Wien (MUMOK) has a collection of 7,000 modern and contemporary art works by Andy Warhol, Pablo Picasso, Gerhard Richter, Roy Lichtenstein and many more.[10] Also this museum uses the two biggest social network platforms Facebook and Twitter. In addition, they manage a Flickr group and a YouTube Channel.

9. **Museum für angewandte Kunst (Vienna, Austria)** The Museum für angewandte Kunst (MAK) provides a collection of applied and contemporary art.[11] Besides the attendance on Facebook, the museum publishes films on YouTube.

10. **Museumsquartier (Vienna, Austria)** The Museumsquartier (MQ) is one of the ten largest cultural areas in the world. The MQ is home of large art museums like the Leopold Museum and the MUMOK. Furthermore, the MQ contains the Tanzquartier, an international

---

[5] cf. `http://www.skd.museum/de/museen-institutionen/residenzschloss/gruenes-gewoelbe/index.html` (accessed on August 30^{th}, 2012)

[6] cf. `http://www.hausdermusik.at/das-klangmuseum/16.htm` (accessed on August 30^{th}, 2012)

[7] cf. `http://www.khm.at/das-museum/` (accessed on August 30^{th}, 2012)

[8] cf. `http://www.leopoldmuseum.org/de/sammlung-leopold/hauptwerke` (accessed on August 30^{th}, 2012)

[9] cf. `http://www.technischesmuseum.at/geschichte` (accessed on August 30^{th}, 2012)

[10] cf. `http://www.mumok.at/sammlung/die-sammlung/` (accessed on August 30^{th}, 2012)

[11] cf. `http://www.mak.at/das_mak/standorte/expositur/mak_wien_1` (accessed on August 30^{th}, 2012)

centre for dance, the Architekturzentrum Wien, the Kunsthalle Wien and other institutions.[12] Therefore, it combines numerous cultural institutions and uses social media platforms like Facebook and Twitter. Additionally, they link on their website to Foursquare and Flickr.

11. **Naturhistorisches Museum (Vienna, Austria)** With approximately 30 million objects, the Naturhistorisches Museum Wien (NHM) is one of the most important natural history museums of the world.[13] Surprisingly, the museum conducts no own Facebook Page or Twitter account. However, based on the high number of visitors, many comments on Foursquare, Qype and Tripadvisor exist.

12. **Schloss Belvedere (Vienna, Austria)** The Belvedere is a historical building which houses the Belvedere museum. The Upper Belvedere houses the collection of Austrian art including the world's largest Gustav Klimt collection.[14] The institution participate on Facebook, Twitter, Youtube and Flickr.

13. **Technorama (Winterthur, Switzerland)** The Technorama is the biggest science centre in Switzerland which allows visitors to understand natural and scientific phenomena.[15] In contrast to many museum, the Technorama has currently no own Facebook Page and a Twitter account is missing. However, qualitative and quantitative reviews can be found on Foursquare, Tripadvisor and Qype.

Surely, not all platforms on which cultural institutions participate are relevant for a sentiment analysis. The extracted values and their notation depend on the data source. However, every web platform provides an application programming interface (API) which allows an easy data access without the adoption of self-written wrappers. The following chapter gives an overview about the applied data sources, their APIs and summarizes which values are extracted.

5.2.1 Data sources

The World Wide Web provides a huge amount of objective and subjective content. For sentiment analysis, content that reflects opinions are relevant and required. Therefore, the selection of data sources was focused on Web 2.0 platforms, where users are predestinated to share their personal information and opinions. This master's thesis focuses on social networks and local review web services, that include:

- the most used social network site **Facebook**,
- the microblogging web service **Twitter**,
- the location-based social networking website **Foursquare**,

[12] cf. `http://www.mqw.at/de/das+mq/ueber+das+mq/` (accessed on August 30^{th}, 2012)
[13] cf. `http://www.nhm-wien.ac.at/museum` (accessed on August 30^{th}, 2012)
[14] cf. `http://www.belvedere.at/de/schloss-und-museum` (accessed on August 30^{th}, 2012)
[15] cf. `http://www.technorama.ch/ueber-uns/das-technorama/` (accessed on August 30^{th}, 2012)

- the local review website **Qype** and
- the travel website **TripAdvisor**.

Of course, also web blogs spread the individual opinion of an author. However, a blog entry demonstrates only one opinion and does not provide an opinion-mashup of many people. In addition, discovering relevant and recent blogs is a difficult ambiguous task. Therefore, the dataset does not include content from web blogs.

The selection of data sources will be more closely examined below. In addition, queries and resources which are used to extract the user-generated content are considered in detail. For these queries, IDs of the cultural institutions are needed and was manually found out. The next section discusses also the received JSON[16] object which is a data-interchange format, completely language independent and easy to parse and generate for machines and humans.

Facebook

Facebook, founded in 2004, is the most used social network, which enables users to stay in contact with friends and discovers what's going on in the world. Private people, politicians, celebrities, profit and non-profit organizations can create private profiles or public pages for sharing what's on his or her mind. Therefore, also cultural institutions benefits of the provided features of Facebook. Museums can easily disseminate information and see what other people say about its institution. This huge amount of data allows to find out users mood about different topics. In addition, almost all of the cultural institutions in the training dataset have a Facebook page and publish frequently content (see Section 2.3). For this reason, Facebook is a very respective source which reflects and mashup the opinion of many people. However, not all provided information is relevant for a sentiment analysis. The analysis database collect only user comments and recommendations from the respective Facebook page. Demographic data, related interests and other personal information of the reviewer imply per se no opinion and are therefore not gathered.
The Facebook Graph API [17] presents the social graph with all objects (people, photos, events and pages) and the connections between them (friend relationships, shared content and photo tags). Every object in the social graph has a unique ID which allows the access of properties by requesting. For instance, the Albertina Museum has object ID 207782682950 and therefore the object can be fetched at `https://graph.facebook.com/207782682950`. However, to get additional information about a user and their comments, an *access token* for authorized requests is needed. The Page object supports real-time updates including statuses, posts and conversations which can be requested as follow:

```
get status updates
https://graph.facebook.com/{OBJECTID}/statuses?access_token={ACCESS_TOKEN}
get posts and comments
https://graph.facebook.com/{OBJECTID}/posts?access_token={ACCESS_TOKEN}
```

[16]Official JSON website: `http://www.json.org` (accessed on August 30^{th}, 2012)

[17]Facebook Graph API, available at `http://developers.facebook.com/docs/reference/api/` (accessed on August 30^{th}, 2012)

Foursquare

Launched in March 2009, Foursquare quickly grows up to a community of 20 million people worldwide. The social service helps to keep up with friends, discover what's nearby, save money and unlock deals. In addition, users can *check in* where they are: In Cafés, restaurants, universities, hotels and of course cultural institutions. By writing a *tip*, people can suggest things to other people on Foursquare. This comments can also include personal opinions which are useful for sentiment analysis.
Also Foursquare provides a public API[18] which enables the discovering of check ins, their history, where friends are, tips and recommendations. For data extraction, an unique *Client ID* and *Client Secret* are needed. For sentiment analysis, only tips are relevant and extracted. Similar to Facebook, a ID of the venue respectively the cultural institution (e.g. 4b058894f964a520c4ce22e3 of the Albertina Museum) is required to request tips. With the help of other optional parameters the results can be sorted by friends, recent and popular or limited to a number up to 500. A tip request of a venue is structured as follow:

```
get tips of a venue
https://api.foursquare.com/v2/venues/{VENUE_ID}/tips?sort=recent&limit=500&
    client_id={CLIENT_ID}&client_secret={CLIENT_SECRET}
```

The JSON response includes the text of the tips, the authors, the creation dates and if available, other personal information.

TripAdvisor

TripAdvisor, founded in 2000, is a travel website which aggregates travel information, user reviews and opinions of travel-related content. With more than 50 million unique monthly visitors and over 60 million reviews and opinions, TripAdvisor is the world's largest travel community. Besides hotels and restaurants, also cultural institutions and attractions are discussed and rated in numerous reviews. Therefore, the platform is a predestinated source for opinion mining. However, the API content is for licensed partners only and is not publicly accessible. For the training set, more than 200 reviews are manually filtered out and stored into the database.

Twitter

Twitter is a microblogging service that enables users the sending and reading of text-based posts ("Tweets") of up to 140 characters. The social network was launched in 2006 and gained worldwide popularity rapidly. In 2012, Twitter counts over 140 million active users which generates over 340 million tweets daily and retrieves over 1.6 billion search queries per day. Thereby, people use the hashtag symbol # before relevant keywords or phrases in their Tweet to categorize those. Therefore, Tweets which are hashtagged with the name of the cultural museum reflects and mashup opinions. However, accounts and Tweets from cultural institutions are not very useful for sentiment analysis. But, the active participation is helpful to start conversations and

[18] Foursquare API, available at `https://developer.foursquare.com/` (accessed on August 30th, 2012)

to spread information.
The Tweets are extracted with the help of the public REST API[19]. In contrast to other APIs, the simple search method needs no authentication and is defined as followed:

```
get Tweets of a search term
http://search.twitter.com/search.json?q={TERM}
```

The JSON response includes appropriate Tweets, their authors, creation dates, language and geoloction[20], if available.

Qype

Qype, founded in Hamburg in 2006, is Europe's largest user-generated local review site with 22 million visits per month. The 2 million reviewers comment in the categories eat and drinking, nightlife, shopping, services, hotels, events and arts & entertainment. Therefore, also Qype represents a good source of opinions for cultural institutions and is a selected source of the training data set. In contrast to their competitor TripAdvisor, Qype provides a public and well-documented API[21]. For the usage, a API consumer key is needed for accessing the API resources. As in all REST APIs all data is modelled with resources which have an unique URL. To get all reviews for a specific place or cultural institution, the particular place ID is needed. The query is semantically structured as followed:

```
get reviews of certain place
http://api.qype.com/v1/places/{PLACE_ID}/reviews/de.json?consumer_key={
    CPONSUMER_KEY}
```

The response contains the content text, author, language, rating and creation date.

5.2.2 Data Extraction

As already mentioned, the user-generated content is extracted from social media platforms and local review sites by usage of API queries. This section aims to provide a brief overview about the implementation of the request and response handling. Exemplary, a JSON request of a Twitter search will be presented in Listing 5.1. JSON, or JavaScript Object Notation, is a text-based open standard which is developed for human-readable data interchange[22]. It was specified by Douglas Crockford and is described in RFC 4627[23]. JSON is composed of five forms which are displayed in Figure 5.2[24]: Objects, arrays, values, strings and numbers. An *object* is an

[19]Twitter API available at `https://dev.twitter.com/docs/apietwas` (accessed on August 30th, 2012)

[20]Geolocation is the identification of the real-world geographic location of a mobile phone or an Internet-connected computer

[21]Qype API available at urlhttp://www.qype.co.uk/developers/api (accessed on August 30th, 2012)

[22]cf. `http://www.json.org/` (accessed on August 30th, 2012)

[23]The application/json Media Type for JavaScript Object Notation (JSON) `http://tools.ietf.org/html/rfc4627` (accessed on August 30th, 2012)

[24]The elements of JSON are taken from `http://www.json.org/` (accessed on August 30th, 2012)

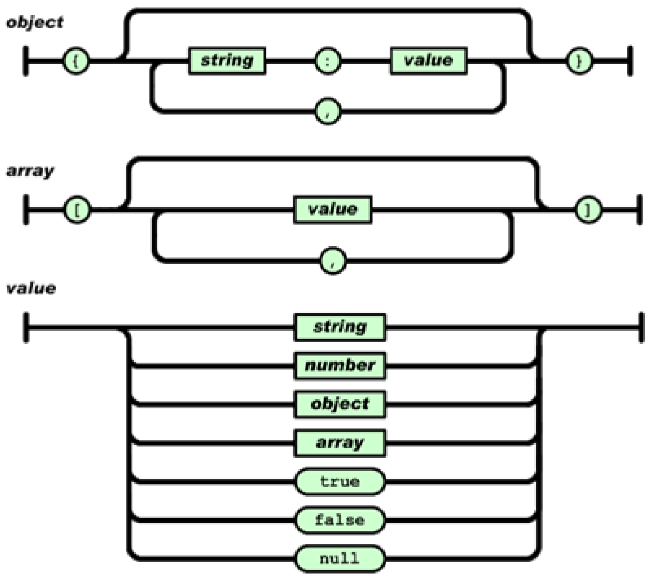
unordered set of name or value pairs. An object begins with a *left brace* ({) and ends with *right brace* (}). Each name is followed by a *colon* (:) and the name/value pairs are separated by *commas* (,). An *array* is an ordered collection of values. An array begins with *left bracket* ([) and ends with *right bracket* (]). The values are separated by *commas* (,). A *value* can be a string in double quotes ("string"), a number, a boolean value (true, false), an object or an array. A *string* is a sequence of zero or more Unicode characters. A character is represented as a single character string. A *number* is a positive or negative integer, decimal or complex number.

First, the $jQuery.getJSON(url)$ function (line 1) sends the string to the server with the request. Secondly, the server processes the query and response the result in a JSON format. The response includes the content and attributes which are stored in the $data$ variable (line 2). Finally, each Tweet is saved into the database ($saveTweet()$ function, line 3) as an object which contains the name of the target cultural institution, the data source, the unique id, the opinion holder, the time when the opinion is expressed, language and of course the content itself.

```
1 $.getJSON('http://search.twitter.com/search.json?q=Kunsthistorisches%Museum&
     rpp=100&callback=?&lang=de', function(data) {
2   $.each(data.results, function(index,item){
3     saveTweet('Twitter', 'Kunsthistorisches Museum', item.id, created_at,
         from_user, item.iso_language_code, item.text);
4   });
5 });
```

Listing 5.1: JSON request of an Twitter search

The user-generated content from Facebook, Foursquare, TripAdvisor and Qype is extracted in the same way.



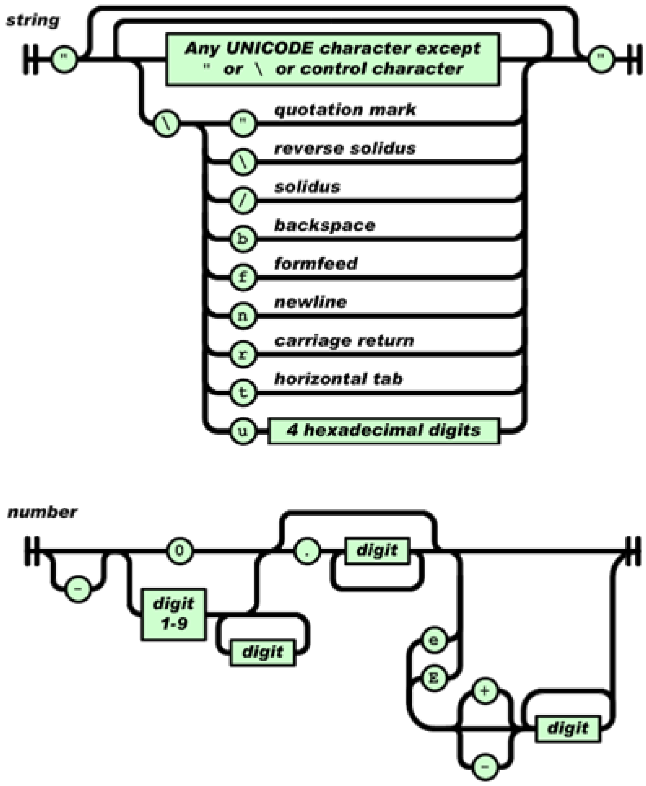


Figure 5.2: Elements of JSON

5.3 Evaluation Results

The following section discusses the evaluation results of the implemented opinion mining system. In a first step, the opinion lexicon module, the topic extraction module and the sentiment extraction module are evaluated separately. In a second step, the opinion mining system for cultural institutions which combines every module as total system is analyzed.

For each module, the precision, the recall and the F-measure is calculated using the training dataset. Precision shows the probability that a retrieved word, topic or sentiment (depending on the module) is relevant. Recall is the probability that a randomly selected relevant word, topic or sentiment is retrieved in the extraction. The F-measure combines precision and recall by evenly weighting of both values. Afterwards, the percentages of wrongly and correctly identified terms are mapped into a diagram based on the absolute outcome. A final discussion analyzes the results and gives indications about possible weaknesses and suggestions for improvement.

5.3.1 Evaluation of the Opinion Lexicon Module

The $td * idf$-feature, introduced in Section 4.2.2, is used to discover single topic and opinion words. The objective was to find a $td * idf$ boundary k which separates topic and opinion words from useless words. Thereby, the percentage of identified topics and opinion words which lie in the range should be as high as possible. Vice versa, the number of words which imply no topic or opinion should be as low as possible in this range. The following Table 5.2 shows the approximation for the parameter k. In the value range of $k = 6.1\ and\ k \geq \infty$, the resulting bag of words contains 9.92% of topic and opinion words, and 0.41% of words which does not. The remaining 89.67% of words are in a range between $k = 0\ and\ k < 6.1$ and consists of topic and opinion as well as of non-topic and non-opinion words.

boundary k	**topic and opinion words in the range (%)**	**non-topic and non-opinion words in the range (%)**
5,9	31,04	26,46
6	29,52	26,36
6,1	**9,92**	**0,41**
6,2	9,41	0,41
6,3	9,16	0,41
6,4	8,65	0,41

Table 5.2: k-boundary evaluation

With the help of these extracted words and the default SentiWS words, feature II and feature III are evaluated. Feature II discovers topics which consist of two or more words. The third Feature extract topic and opinion words iteratively by already known words. The following Table 5.3 presents the result of the precision, recall, F-measure and accuracy calculation of all applied features in the opinion lexicon module.

Approach	**Precision**	**Recall**	**F-measure**
Feature I	0.907	0.0992	0.0894
Feature II	0.9206	0.9508	0.4677
Feature III	0.8145	0.6418	0.359
Feature I + II + III	0.824	0.8052	0.4072

Table 5.3: Evaluation results of the opinion lexicon module

Feature I achieves an excellent precision value due to the restricting k-boundary. Therefore, the probability is high that an extracted word implies a topic or opinion. However, fewer than 10% of the manual detected lexicon words are discovered by this feature which argues the very low recall and F-measure. Therefore, a $tf * idf$ boundary is not such a good classifier for discovering all lexical and non-lexical words in the context of opinion mining. Nevertheless, words that are extracted are correctly classified. The second feature deals with the discovering of topic words which consists of more than two words. The applied algorithm uses n-grams and achieves good results in precision and recall. However, word pairs which include adverbs are not discovered. For instance, the terms “Stillleben von Cézanne”, “Herzogs Albert von Sachsen-Teschen” or “Bild von Aussen” are ignored by reason of the occurring adverb “von”. To improve the result, a further rule might be implemented to enclose such word pairs.

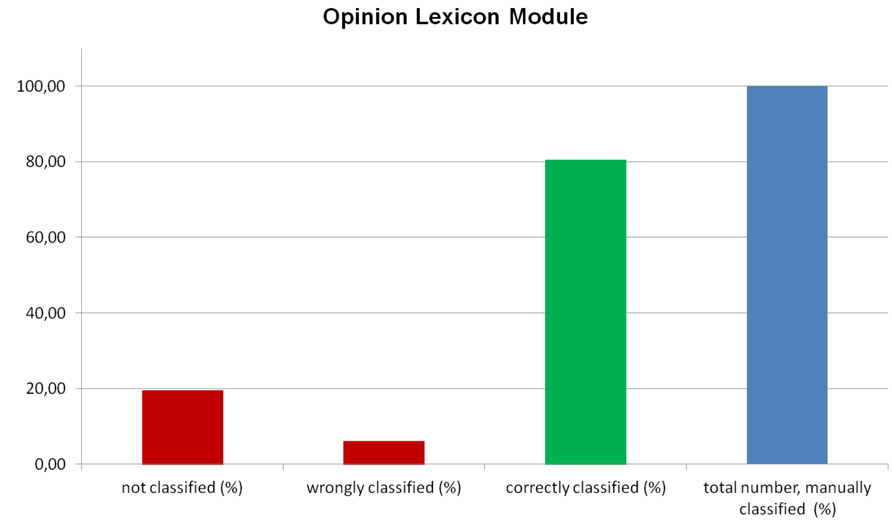


Figure 5.3: Not classified, wrongly classified and correctly classified terms in the opinion lexicon module

Feature III detects single topic and opinion words by using the double propagation method. The iterative algorithm achieves that more than 80% of the retrieved words are correctly classified. To raise the precision and recall, the initial set of domain depending words needs to be extended. Therefore, the quality of Feature III is highly dependent on the initial word set of Feature I. In the use of all three features together, a satisfactory precision and recall value can be presented: Both values are located above 80%. In summary, the opinion lexicon module contains extracted words and word pairs which are with a percentage of 80.52% (n = 843) correctly identified. 25.66% (n = 269) of words are incorrectly identified, which contains correct words which are not extracted (not classified = 19.48%), and incorrect words which are extracted (wrongly classified = 6.18%) (see Figure 5.3).

5.3.2 Evaluation of the Topic Extraction Module

This module deals with the evaluation of the topic extraction module. The objective of the module is the extraction of topics for which an opinion is expressed in the user-generated content. For this reason, the algorithm differentiates sentences between four cases (see Section 4.2.3). The evaluation shows the following results:

The computed precision shows that nearly 63% of the extracted topics are relevant. In summary, 71.43% (n = 204) of topics are correctly identified. However, 5.86% (n = 17) of the terms are misclassified and even 28.57% (n = 82) of the topics are not identified (see Figure 5.4). It

Approach	Precision	Recall	F-measure
Case 1 + 2 + 3 + 4	0.6316	0.7143	0.3352

Table 5.4: Evaluation results of the topic extraction module

can be derived that the implemented rules are good base, but leave room for further improvement. Especially, the distance calculation between opinion word/phrase and topic word are not satisfying. The German language is much more complex than the minimum distance unambiguously determine the correct topic word. In addition, the topic extraction strongly depends on the completeness of the opinion lexicon module. In addition, the module does not consider implicit and ironic discussed topics. For example, the topic "Sicherheitsbediensteten" is not extracted in the sentence "Was macht ein Museum aus: Inspiration, interessante Eindrücke, einmalige und herausragende Exponate, Wohlbefinden, Service. Service? Das kann aktuell bei den Sicherheitsbediensteten schon mal ins Auge gehen.". Furthermore, the applied algorithm detects no topics which are opinionated intersentential. However, about the relatively high recall value can be stated that relatively few non-topic words are extracted.

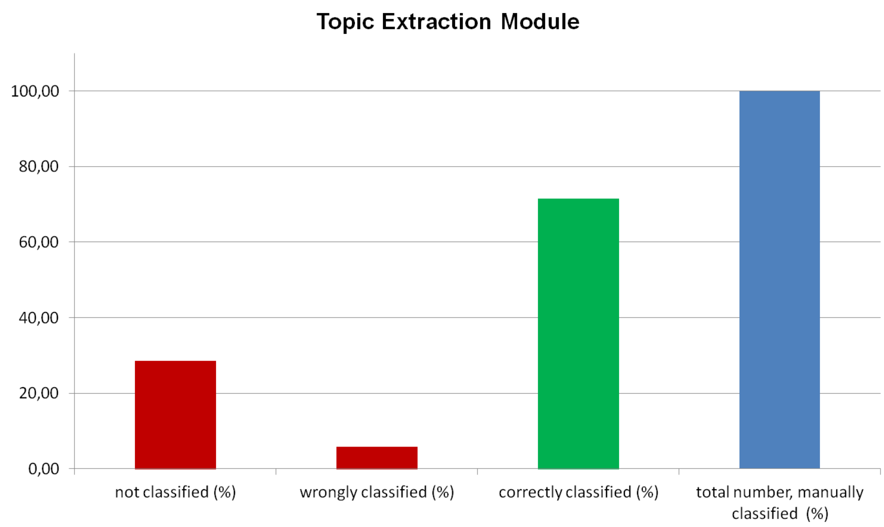


Figure 5.4: Not classified, wrongly classified and correctly classified terms in the topic extraction module

5.3.3 Evaluation of the Sentiment Extraction Module

The sentiment extraction module computed the polarity weight, and therefore the sentiment of each detected topic from the previous module. The algorithm follows three steps: Sentiment computation of opinion phrases, treatment of negation words and the sentiment extraction. The following table presents the evaluation results of the positive and negative extracted topics.

Approach	Precision	Recall	F-measure
Step 1 + 2 + 3 (Sentiment positive)	0.9706	0.9296	0.4748
Step 1 + 2 + 3 (Sentiment negative)	0.8333	0.7692	0.4072

Table 5.5: Evaluation results of the sentiment extraction module

Especially the extraction of positive sentiments work very well as can be seen in the high precision and recall value. On the one hand, there are not many words incorrect classified as positive (recall = 0.9296), and on the other hand, there are almost all positive sentiments extracted (precision = 0.9706). The number of computed negative sentiments is also satisfactory, but nevertheless lower by almost 14% in contrast to positive sentiments.

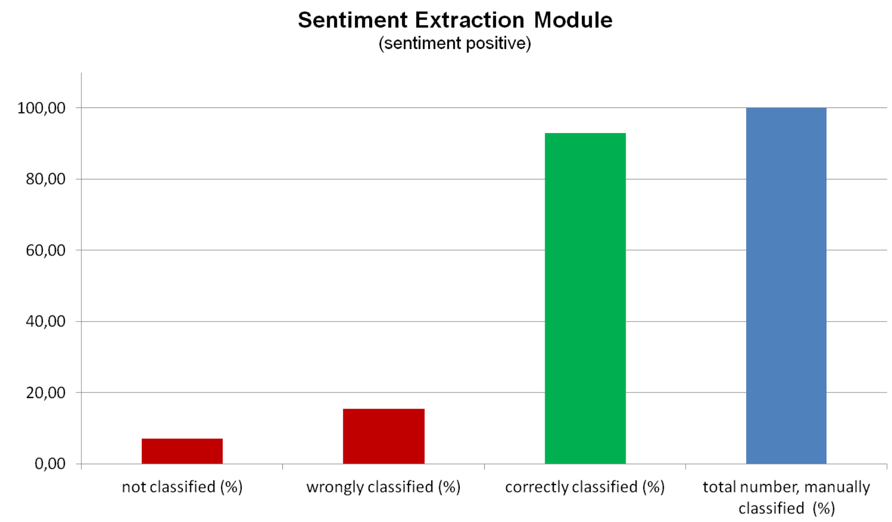


Figure 5.5: Not classified, wrongly classified and correctly classified terms in the sentiment extraction module

However, the recall value drops below 77%. Therefore, a large number of negative sentiments are wrongly identified (in summary 25.89%). Interestingly, there are 23.08% of negative sentiments not extracted, in contrast to comparatively few 7.04% of positive sentiments. In relation to the summarization chart of the positive sentiments (see Figure 5.5), it can be deducted that about 20% (n = 28) of the positive as well as more than 20% (n = 20) of the negative opinionated topics are incorrect classified (see Figure 5.6).

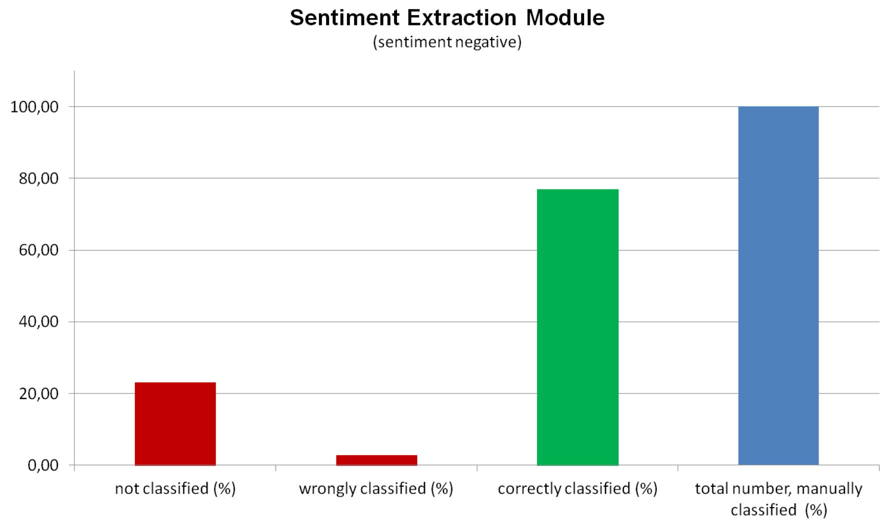


Figure 5.6: Not classified, wrongly classified and correctly classified terms in the sentiment extraction module

5.3.4 Evaluation of the Opinion Mining System

Finally, this section discusses the evaluation of the whole opinion mining system which combines the previous modules. Generally speaking, a system is only as good as the sum of its parts. Therefore, a very well functioning sentiment computation module is useless when the topic extraction module detects no topics. The other way around, a good topic extraction is useless when the related sentiments are incorrectly classified. However, the developed opinion mining system for cultural institutions achieves a precision of 0.6170 and recall of 0.6744. Because visitors are not always interested in having all possible reviews, but just a few positive and a few negative, the accuracy measure is calculated. In the developed system, the balance between retrieved positive and negative is not yet optimal and needs further improvements.

Approach	Precision	Recall	F-measure	Accuracy
Opinion Lexicon + Topic Extraction + Sentiment Extraction	0.6170	0.6744	0.3222	0.1259

Table 5.6: Evaluation results of the opinion mining system

In summary, 69.05% (n = 232) of topics are correctly classified. The summarization diagram (see Figure 5.7) of the opinion mining system that weaknesses lies in the wrong number of incorrect classified sentiment topics. Nearly 40% (n = 151) are incorrect classified. This measure includes topics which are not classified (33.34%, n = 112) and terms which are falsely

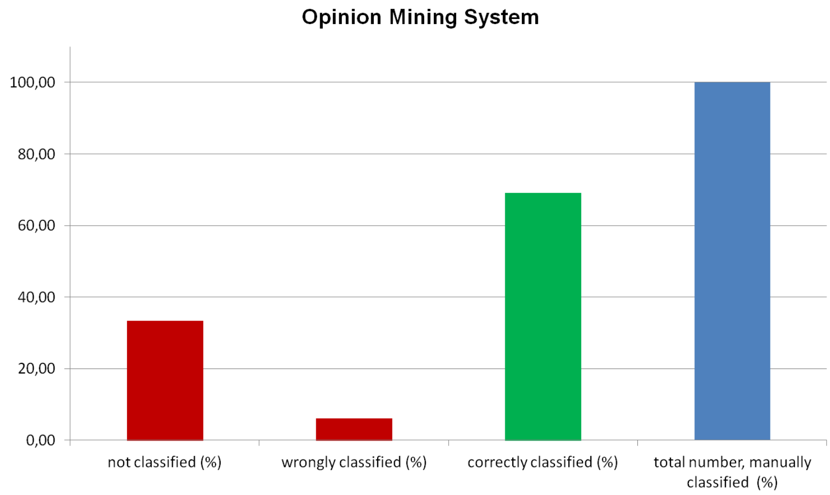


Figure 5.7: Not classified, wrongly classified and correctly classified terms in the opinion mining system

classified as topics (6.02%, n = 39). The summarization diagram (see Figure 5.7) of the opinion mining system that weaknesses lies in the wrong number of incorrect classified sentiment topics. Nearly 40% are incorrect classified. Especially the double propagation algorithm for single topic and opinion word detection in feature III of the opinion lexicon module (see Section 4.2.2) needs improvements. The iterative algorithm searches new words by already collected words and terminates when no more topics or opinion words can be found. Figure 5.8 illustrates a snipped example of terms which are identified as a single topic or opinion word in a network diagram. Mistakenly identified terms are written in red. It was started with the term "schön" for discovering topic words which are occurring before or after. After this step completes, all new discovered topics are used to detect new opinion words. For instance, the term "Räume" is used to detect the opinion word "prachtvollen". However, the algorithm achieves only good results under the assumption that nouns are written with a capital initial letter. In this figure, 29.54% of the words are wrongly identified. However, also such words can lead to further correct identifications. For example, the term "modernen" implies no opinion per se, but identifies 6 new domain specific topic words. Therefore, it is difficult to detect patterns for dropping wrong findings. Nevertheless, to reduce this outcome, improvements in the topic extraction module are needed. In addition, an implicit topic detection feature, a language correction feature and the analysis over several sentences are possible approaches.

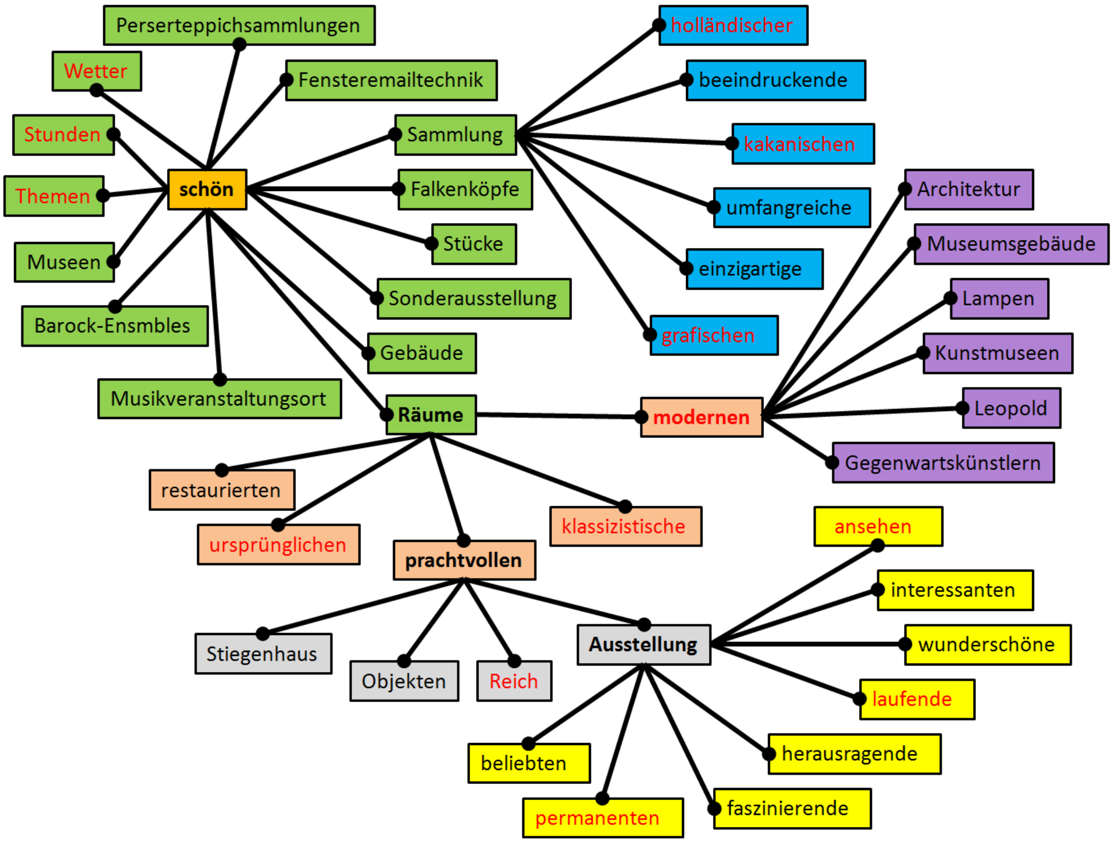


Figure 5.8: Network diagram of the single topic and opinion word detection using double propagation

CHAPTER 6

Summary and Future Work

6.1 Summary

In this Master's thesis an approach for sentiment analysis in the domain of cultural institutions is presented. Sentiment analysis, or opinion mining, is an interdisciplinary scientific discipline which combines computational linguistics and information retrieval. In contrast to traditional Natural Language Processing (NLP) which deals with the application of computational models to text data (e.g., automatic translation between languages, dialogue systems which allow a human to interact with a machine using natural language, etc.), sentiment analysis deals with the extraction of opinions in text documents. Discovered opinions are classified whether they express a positive, neutral or negative sentiment. With this relatively new scientific discipline are many problems accompanied. The object identification, topic extraction, synonym grouping, sentiment classification and language correction have not yet been solved satisfactorily.

Previous research can be classified into the scientific areas of sentiment classification, opinion lexicons and the feature-based sentiment analysis. Sentiment classification can be carried out on document- and sentence-level. Document-level sentiment classification considers the whole user review as positive, negative or neutral. However, each individual sentence can imply an opinion or not. By sentence-level sentiment classification, the opinionated sentences can also be allocated to a semantic orientation (positive, negative or neutral). Since opinion words and phrases are essential in sentiment classification tasks. Such words and clauses are filtered out and summarized into an opinion lexicon. As mentioned before, sentiment analysis at document- or sentence-level classifies the whole document or sentence as being positive or negative. However, also a positive reviewed document or sentence does not imply that the author has positive opinions on all discussed topics. Of course, that is also valid for negative opinionated documents and sentences: Not all aspects must be negative evaluated. For this reason, feature-based sentiment classification discovers discussed topics on an object and determines their related sentiment.

This work presents a prototype of a feature-based sentiment analysis system for cultural institutions. The approach aims at automatic extraction of topics and personal opinions in unstructured text documents. Those documents are extracted from social networking services and contain reviews as well as user-comments. The system is implemented as a PHP web-application and consists of five modules: The data pre-processing module, the opinion lexicon module, the topic extraction module, the sentiment extraction module and the summarization module. The data pre-processing module prepares the data by the removing of stop words, the stemming, the part of speech (POS) tagging and the computation of term frequencies of each word. The objective of the opinion lexicon module is the computation of the polarity weight, POS tagging and the discovering of the sentiment (positive, negative or neutral) of words which imply a topic or sentiment. For this reason, the module applies three features for single topic word, composited topic word and opinion word detection. The topic extraction module deals with the discovering of sentiment related topics. Based on the number of occurring topics (n_{tw}) and opinions (n_{ow}) in a sentence, the module differentiated between four cases: In case 1, the sentence includes a topic word and an opinion word (e.g., "Ein **schönes Museum**"). In case 2, the sentence includes more than one topic word and an opinion word (e.g., "Das **Restaurant** in dem *Museum* ist sehr **schön**"). In case 3, the sentence includes a topic word and more than one opinion word (e.g., "Das **Museum** ist **günstig** aber **überfüllt**"). Finally, in case 4, the sentence includes more than one topic word and more than one opinion word (e.g., "Das Museum hat einen **schönen Shop** und ein **günstiges Restaurant**"). These differentiations allow the assignment of a topic word to the correct opinion word with the help of verifications based on enumerations and conjunctions. N-grams, dependency trees, double propagation and distance calculation are applied features. The sentiment extraction module deals with the detection of the sentiment and opinion orientation from extracted topics. With the help of POS pattern of opinion phrases and the treatment of negation words, the sentiment is computed using score functions. The purpose of the last module is on the one hand the summarization of a single cultural institution, and of the other hand the comparison of two cultural institutions by their opinionated topics by diagrams.

Each module is assessed by precision, recall and F-measure. The evaluated dataset contains about 400 reviews from five social networking and local review platforms Facebook, Foursquare, Twitter, TripAdvisor and Qype. The opinion lexicon module contains extracted words and word pairs which are with a percentage of 80.52% correctly identified. 25.66% of words are incorrectly identified, which contains correct words which are not extracted and incorrect words which are extracted. The computed precision of the topic extraction module shows that nearly 63% of the extracted topics are relevant. It can be derived that the implemented features are a good base, but need undoubtedly further improvements. In particular, the number of incorrectly identified words (about 30%) must be reduced. However, the sentiment extraction module delivers very good results. About 90% of the computed sentiments are correctly identified which advocates the applied methods. But, based on the fact that a system is only as good as the sum of its parts, the opinion mining system achieved more than 60% correct classified opinions, but nearly 40% are incorrect classified. To reduce this outcome further work and research is indispensable.

6.2 Future Work

1. **Automatic topic detection and POS tagging**
   This master's thesis extract topics by counting term frequency. However, manual word detection is necessary for getting topics which are highly related to the domain of cultural institutions. Therefore, the process of topic detection and discovering of their POS tags should be implemented automatically. The POS tagging could be easily realized by using a commercial language detection API[1]. This modification is condition for creating a domain independent sentiment analysis framework.

2. **Comparative sentence detection**
   Another useful model feature would be the comparison of different cultural institutions. For example, the following sentence compares two museums: "Es werden viel mehr Werke in Museum A ausgestellt als in Museum B". The identification of such sentences and the topic allocation as positive for museum A and negative for museum B would complete the model. The manual evaluation of the dataset results that such sentences does not predominate. In other domains however, it makes more sense to include comparative sentences. For instance, in technical product reviews it is more common that internet users compare features.[2] Consequently a comparative sentence detection module would allow a more representative sentiment classification result.

3. **Opinion spam detection**
   Another possible extension deals with the detection of opinion spam. As already known, internet users spread their opinion about cultural institutions, brands, products and persons. On the one hand, this informations are used by people to find opinions of existing users before deciding to go into a museum or not. On the other hand, museums can identify problems and find out information about their competitors. The manipulating and affectation of user reviews is therefore highly critical and must be filtered out. [Jindal and Liu, 2008] has classify three different types of opinion spam:

   - Untruthful opinions, which mislead readers by intentional misdirection of reviews,
   - reviews on brands only, which do not comment about features but only about the brand or institution as such, and
   - non-reviews, which containing no opinions by postings of advertisements, questions, answers or random texts.[3]

   Of course, the adoption of such feature makes the classification result of users sentiment even more valuable.

[1] For example Google Tranlator API `https://developers.google.com/translate/` (accessed on August 30^{th}, 2012)

[2] cf. [Liu, 2006, Page 412]

[3] cf. `http://www.cs.uic.edu/~liub/FBS/opinion-spam-WSDM-08.pdf` (accessed on August 30^{th}, 2012)

4. **Data source extension**
   To comprise the opinion of several people, the extension of further data sources is required. During this time, numerous social network platforms, review websites, blogs and forums exists and sprout from scratch. To get a consistent overview, the finding and inclusion of relevant for future work and analysis improvements. Certainly not all web services provide open APIs for data access and therefore the implementation of web crawlers[4] is a further challenge.

5. **Implict topic detection**
   Based on the fact that many users express their opinion not directly, a further module is needed to extract also implicitly described topics and their sentiment. However, this problem is up today not satisfactorily solved. Nevertheless, those feature would increase the number of correct identified opinions enormously.

6. **Language correction**
   The complex and misapplied German grammar, the poor spelling and punctuation, the lack of capitals and the use of abbreviations would falsify the result. Therefore a well working spell checker is needed to increase the number of correct classified opinions. With the help of a language correction module the misspelled term “Restaurent” can be transformed to “Restaurant”. In addition, it is not unusual for Internet users to write sarcasm and ambiguity sentences which also leads to a wrong sentiment classification.

[4] Web crawlers, also known as spiders or robots, are computer programs that automatically browse the World Wide Web and subsequently download websites and extract their relevant content. [Liu, 2006, Page 273]

Bibliography

[Abbasi et al., 2008] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information and System*, 26(3):12:1–12:34.

[Banea, 2008] Banea, C. (2008). Subjectivity and sentiment analysis. Slides by Carmen Banea based on presentations by Jan Wiebe (University of Pittsburg) and Bing Liu (University of Illinois).

[Bhuiyan et al., 2009] Bhuiyan, T., Xu, Y., and Jøsang, A. (2009). State-of-the-art review on opinion mining from online customers ' feedback. In *Proceedings of the 9th Asia-Pacific Complex Systems Conference*, pages 385–390.

[BVDW, 2011] BVDW (2011). Bundesverband Digitale Wirtschaft e.v. *Einsatz von Social Media in Unternehmen*.

[Carbonell, 1981] Carbonell, J. (1981). *Subjective understanding, computer models of belief systems*. Computer science: Artificial intelligence. UMI Research Press.

[Cavnar and Trenkle, 1994] Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

[Church and Hanks, 1990] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. volume 16, pages 22–29, Cambridge, MA, USA. MIT Press.

[Das et al., 2001] Das, S. R., Chen, M. Y., Agarwal, T. V., Brooks, C., shee Chan, Y., Gibson, D., Leinweber, D., Martinez-jerez, A., Raghubir, P., Rajagopalan, S., Ranade, A., Rubinstein, M., and Tufano, P. (2001). Yahoo! for amazon: Sentiment extraction from small talk on the web. In *8th Asia Pacific Finance Association Annual Conference*.

[Ding et al., 2008] Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, Web Search and Data Mining 2008 (WSDM'08), pages 231–240, New York, NY, USA. ACM.

[Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422.

[Florian et al., 2003] Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings Of Conference on Computational Natural Language Learning (Conll-2003)*, pages 168–171.

[Giuseppe Carenini and Zwart, 2005] Giuseppe Carenini, R. N. and Zwart, E. (2005). Extracting knowledge from evaluative text. In *In Proceedings of the 3rd International Conference on Knowledge Capture*, pages 11–18. ACM Press.

[Gupta, 2006] Gupta, G. (2006). *Introduction To Data Mining With Case Studies*. Prentice-Hall Of India Pvt. Limited.

[Hamp and Feldweg, 1997] Hamp, B. and Feldweg, H. (1997). Germanet - a lexical-semantic net for german. In *Proceedings of Association for Computational Linguistics (ACL) workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

[Hatzivassiloglou and McKeown, 1997] Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, European Chapter of the Association for Computational Linguistics (EACL) '97, pages 174–181. Association for Computational Linguistics.

[Hatzivassiloglou and Wiebe, 2000] Hatzivassiloglou, V. and Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, Conference on Computational Linguistics 2000 (COLING'00), pages 299–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 68—–177.

[Jindal and Liu, 2006] Jindal, N. and Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) conference on Research and development in information retrieval*, SIGIR '06, pages 244–251, New York, NY, USA. ACM.

[Jindal and Liu, 2008] Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, Web Search and Data Mining 2008 (WSDM'08), pages 219–230. ACM.

[Liu, 2005] Liu, B. (2005). Opinion observer: Analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM Press.

[Liu, 2006] Liu, B. (2006). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc.

[Liu, 2010] Liu, B. (2010). Sentiment analysis and subjectivity. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

[Ölvecký, 2005] Ölvecký, T. (2005). N-Gram based Statistics Aimed at Language Identification. In Bieliková, M., editor, *Proceedings of IIT.SRC 2005: Student Research Conference in Informatics and Information Technologies, Bratislava*, pages 1–7. Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

[Merkl, 2011] Merkl, W. (2011). Course material "Information Search on the Internet", Vienna University of Technology.

[Mishne and de Rijke, 2006] Mishne, G. and de Rijke, M. (2006). Capturing global mood levels using blog posts. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 145–152.

[Otero, 2008] Otero, P. G. (2008). The meaning of syntactic dependencies. In *Linguistik online (www.linguistik-online.de/35_08/gamallo.pdf)*. Universität Bern, Philosophisch-historische Fakultät.

[Pang and Lee, 2005] Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Association for Computational Linguistics (ACL'05)*, pages 115–124.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

[Pang and Lee, 2010] Pang, B. and Lee, L. (2010). Sentiment analysis: A multi-faceted problem. *Science*, 25(1):76–80.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings Of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

[Qiu et al., 2009] Qiu, G., Liu, B., Bu, J., and Chen, C. (2009). Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st international jont conference on Artifical intelligence*, International Joint Conferences on Artificial Intelligence 2009 (IJCAI'09), pages 1199–1204.

[Remus et al., 2010] Remus, R., Quasthoff, U., and Heyer, G. (2010). SentiWS - a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10).*

[Remus and Heyer, 2010] Remus, U. Q. R. and Heyer, G. (2010). SentiWS - a publicly available german-language resource for sentiment analysis. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Sarawagi, 2008] Sarawagi, S. (2008). *Information Extraction.* Foundations Trends in Databases, Now Publishers Inc., Indian Institute of Technology, CSE, Mumbai, India.

[Scheurer, 2012] Scheurer, D. (2012). *Named Entity Recognition - Techniques and Evaluation*. GRIN Verlag.

[Schiller et al., 1999] Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, Universität Stuttgart, Universität Tübingen.

[Shannon, 2001] Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

[Su et al., 2008] Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., Swen, B., and Su, Z. (2008). Hidden sentiment association in chinese web opinion mining. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 959–968, New York, NY, USA. ACM.

[Tang et al., 2009] Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Ssystems With Applications*, 36(7):10760–10773.

[Titov and Mcdonald, 2008] Titov, I. and Mcdonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings Association for Computational Linguistics: Human Language Technologies (ACL'08: HLT)*, pages 308–316.

[Turney, 2002] Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics (ACL'02), pages 417–424, Stroudsburg, PA, USA.

[Wiebe and Mihalcea, 2006] Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Proceedings Association for Computational Linguistics (ACL'06)*, pages 1065–1072.

[Wilson, 2005] Wilson, T. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354.

[Wilson and Wiebe, 2003] Wilson, T. and Wiebe, J. (2003). Annotating opinions in the world press. In *Proceedings of the Fourth SIGdial Workshop on Discourse and Dialogue*, pages 13–22.

[Wilson et al., 2004] Wilson, T., Wiebe, J., and Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*, pages 761–769.

APPENDIX A

Resources

A.1 Stop Words

The following table lists all 265 stop words which are meaningless for topic and sentiment detection, collated from Dr. Martin Porter[1] and Ranks.nl[2]. The list can be downloaded as SQL and CSV file from `www.dub-in-sky.de/Masterthesis/Stopwords.zip`.

Table A.1: Stop words collection

Stopwords							
aber	ander	auf	darum	deines	deshalb	dieser	einer
aber	andere	aus	das	dem	desselben	dieses	eines
alle	anderem	bei	dass	demselben	dessen	dir	einig
allem	anderen	bin	daß	den	dich	doch	einige
allen	anderer	bis	dasselbe	denn	die	dort	einigem
aller	anderes	bist	dazu	denselben	dies	du	einigen
alles	anderm	da	dein	der	diese	durch	einiger
als	andern	dadurch	deine	derer	dieselbe	ein	einiges
also	anderer	daher	deinem	derselbe	dieselben	eine	einmal
am	anders	damit	deinen	derselben	diesem	einem	er
an	auch	dann	deiner	des	diesen	einen	es

[1]available at `http://snowball.tartarus.org/algorithms/german/stop.txt` (accessed on August 30^{th}, 2012)

[2]available at `http://www.ranks.nl/stopwords/german.html` (accessed on August 30^{th}, 2012)

Stopwords							
einer	haben	ja	man	nein	solchen	vom	weshalb
eines	hat	jede	manche	nicht	solcher	von	wie
einig	hatte	jedem	manchem	nichts	solches	vor	wieder
einige	hatten	jeden	manchen	noch	soll	während	wieso
einigem	hattest	jeder	mancher	nun	sollen	wann	will
einigen	hattet	jedes	manches	nur	sollst	war	wir
einiger	hier	jene	mein	ob	sollt	waren	wird
einiges	hin	jenem	meine	oder	sollte	warst	wirst
einmal	hinter	jenen	meinem	ohne	sondern	warum	wo
er	ich	jener	meinen	sehr	sonst	was	woher
es	ihm	jenes	meiner	seid	soweit	weg	wohin
etwas	ihn	jetzt	meines	sein	sowie	weil	wollen
euch	ihnen	kann	mich	seine	über	weiter	wollte
euer	ihr	kannst	mir	seinem	um	weitere	würde
eure	ihre	kein	mit	seinen	und	welche	würden
eurem	ihrem	keine	muß	seiner	uns	welchem	zu
euren	ihren	keinem	muss	seines	unse	welchen	zum
eurer	ihrer	keinen	müssen	selbst	unsem	welcher	zur
eures	ihres	keiner	musst	sich	unsen	welches	zwar
für	im	keines	mußt	sie	unser	wenn	zwischen
gegen	in	können	müsst	sind	unsere	wer	
gewesen	indem	könnt	musste	so	unses	werde	
hab	ins	könnte	nach	solche	unter	werden	
habe	ist	machen	nachdem	solchem	viel	werdet	

A.2 Stuttgart-Tübigen Tagset (STTS)

The Stuttgart-Tübingen Tagset (STTS) [Schiller et al., 1999] is a set of 54 POS-tags for marking German words with part of speech labels. The tag-set consists of 48 normal POS-tags (nouns, verbs, adjectives, etc.) and 6 tag categories for foreign words, compounds, non-words and punctuation marks. The following Table describes each tag with a description and an example.

Table A.2: Stuttgart-Tübingen Tagset (STTS)

POS	Description	Examples
ADJA	attributives Adjektiv	[das] große [Haus]
ADJD	adverbiales oder prädikatives Adjektiv	[er fährt] schnell, [er ist] schnell
ADV	Adverb	schon, bald, doch
APPR	Präposition; Zirkumposition links	in [der Stadt], ohne [mich]
APPRART	Präposition mit Artikel	im [Haus], zur [Sache]
APPO	Postposition	[ihm] zufolge, [der Sache] wegen
APZR	Zirkumposition rechts	[von jetzt] an
ART	bestimmter oder unbestimmter Artikel	der, die, das, ein, eine
CARD	Kardinalzahl	zwei [Männer], [im Jahre] 1994
FM	Fremdsprachliches Material	[Er hat das mit “] A big fish [” übersetzt]
ITJ	Interjektion	mhm, ach, tja
KOUI	unterordnende Konjunktion mit “zu” und Infinitiv	um [zu leben], anstatt [zu fragen]
KOUS	unterordnende Konjunktion mit Satz	weil, daß, damit, wenn, ob
KON	nebenordnende Konjunktion	und, oder, aber
KOKOM	Vergleichskonjunktion	als, wie
NN	normales Nomen	Tisch, Herr, [das] Reisen
NE	Eigennamen	Hans, Hamburg, HSV
PDS	substituierendes Demonstrativpronomen	dieser, jener
PDAT	attribuierendes Demonstrativpronomen	jener [Mensch]

POS	Description	Examples
PIS	substituierendes Indefinitpronomen	keiner, viele, man, niemand
PIAT	attribuierendes Indefinit-pronomen ohne Determiner	kein [Mensch], irgendein [Glas]
PIDAT	attribuierendes Indefinit-pronomen mit Determiner	[ein] wenig [Wasser], [die] beiden [Brüder]
PPER	irreflexives Personalpronomen	ich, er, ihm, mich, dir
PPOSS	substituierendes Possessivpronomen	meins, deiner
PPOSAT	attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]
PRELS	substituierendes Relativpronomen	[der Hund ,] der
PRELAT	attribuierendes Relativpronomen	[der Mann ,] dessen [Hund]
PRF	reflexives Personalpronomen	sich, einander, dich, mir
PWS	substituierendes Interrogativpronomen	wer, was
PWAT	attribuierendes Interrogativpronomen	welche [Farbe], wessen [Hut]
PWAV	adverbiales Interrogativ- oder Relativpronomen	warum, wo, wann, worüber, wobei
PAV	Pronominaladverb	dafür, dabei, deswegen, trotzdem
PTKZU	“zu” vor Infinitiv	zu [gehen]
PTKNEG	Negationspartikel	nicht
PTKVZ	abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad
PTKANT	Antwortpartikel	ja, nein, danke, bitte
PTKA	Partikel bei Adjektiv oder Adverb	am [schönsten], zu [schnell]
TRUNC	Kompositions-Erstglied	An- [und Abreise]

POS	Description	Examples
VVFIN	finites Verb, voll	[du] gehst, [wir] kommen [an]
VVIMP	Imperativ, voll	komm [!]
VVINF	Infinitiv, voll	gehen, ankommen
VVIZU	Infinitiv mit “zu”, voll	anzukommen, loszulassen, loszureißen
VVPP	Partizip Perfekt, voll	gegangen, angekommen
VAFIN	finites Verb, aux	[du] bist, [wir] werden
VAIMP	Imperativ, aux	sei [ruhig !]
VAINF	Infinitiv, aux	werden, sein
VAPP	Partizip Perfekt, aux	gewesen
VMFIN	finites Verb, modal	dürfen
VMINF	Infinitiv, modal	wollen
VMPP	Partizip Perfekt, modal	gekonnt, [er hat gehen] können
XY	Nichtwort, Sonderzeichen enthaltend	3:7, H2O, D2XW3
\$,	Komma	,
\$.	Satzbeendende Interpunktion	. ? ! ; :
\$(	sonstige Satzzeichen; satzintern	- [,]()

A.3 Document Corpus

For further research, the complete pre-processed training dataset is provided as a CSV and SQL file (`www.dub-in-sky.de/Masterthesis/Trainingset.zip`). The database table contains the id, data source, institution, review, review without stop words, stemmed review and POS tags of 438 reviews from Facebook, Foursquare, TripAdvisor, Twitter and Qype. The following screenshots of user comments showing weaknesses of the feature-based sentiment analysis system which should give an inspiration for possible extensions for further research. Extracted topics are highlighted in green (positive), blue (neutral) or red (negative). In addition, topics which can not be extracted from the system are highlighted in yellow.

The following user review about the Leopold Museum which is obtained from the local review website Qype. Almost all sentiment discussed topics are identified. However, the topic "Bilder von Egon Schile" is not classified as positive. Composited topic words which contains a preposition ("von") are not yet discovered.

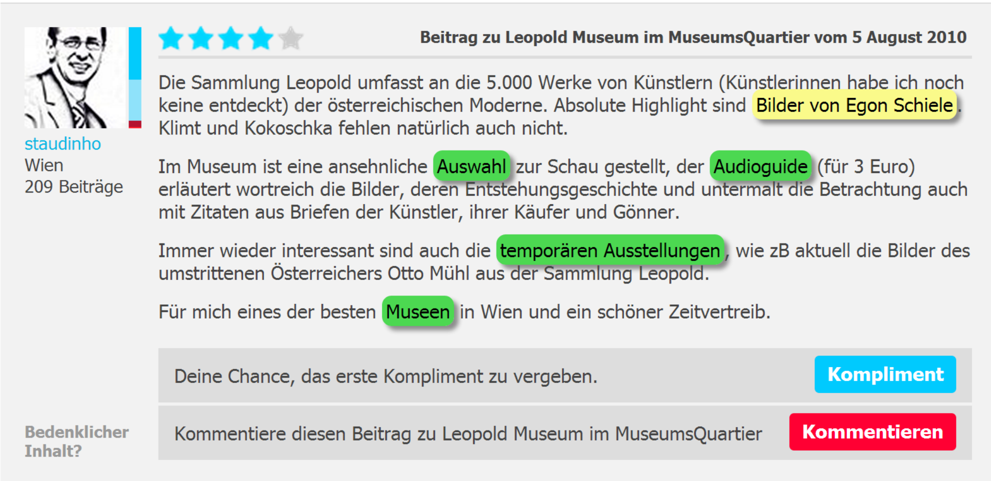


Figure A.1: User review about the Leopold Museum (obtained from Qype)

In the following Foursquare "tip" about the Museumsquartier, the opinionated discussed topic "WLAN" is not classified as positive. The system is not able to analyze reviews which are written in mixed and colloquial languages.

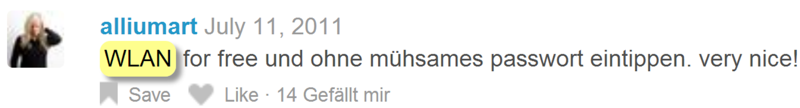


Figure A.2: User review about the Museumsquartier (obtained from Foursquare)

Nearly all topics in the user review about the Albertina are correctly identified and classified as positive or neutral. However, the developed system is not able to interpret and identify misspelled words (“Restaurent”). A language correction module would overcome this weakness.

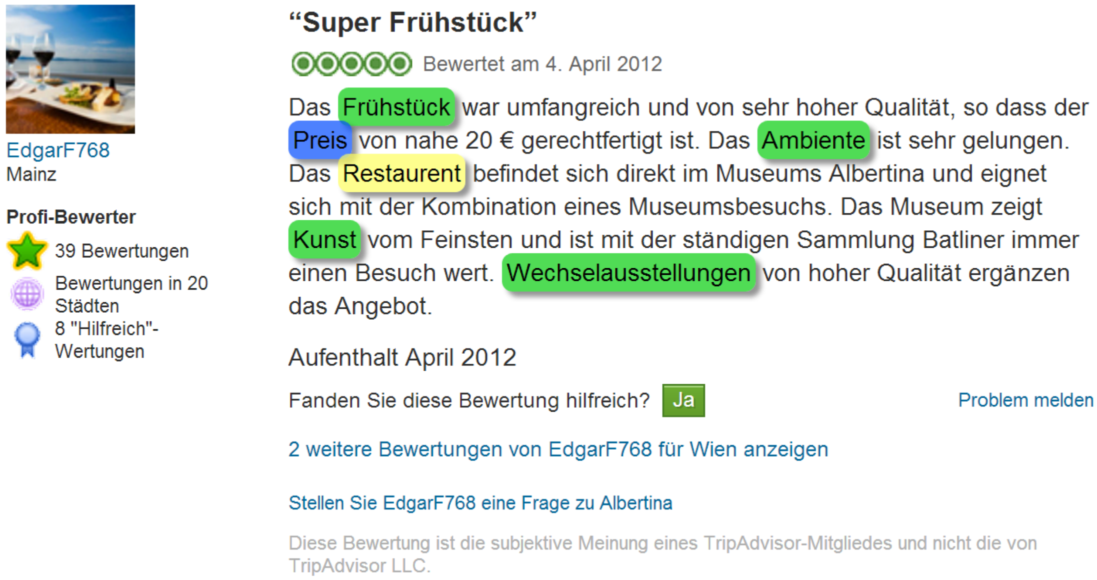


Figure A.3: User review about the Albertina (obtained from TripAdvisor)

The following user statement reviews the Museumsquartier on the social network Foursquare. The domain specific term “Hallenburger” is not extracted, because the opinion word “geilen” was not identified. In addition, the topic word “Preis” is not identified as neutral discussed aspect. The system is not able to interpret opinion phrases like “auch noch zu verkraften”.

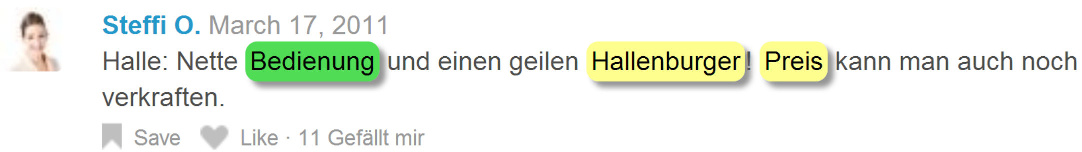


Figure A.4: User review about the Museumsquartier (obtained from Foursquare)

This Facebook comment reviews the Museum Moderner Kunst Stiftung Ludwig Wien (MUMOK). The term "Ausstellung" can be correctly identified as a negative discussed aspect. However, the topic "aufseher/wärter" are not detected, because the opinion words "störend und ungut" are in the following sentence. The system is not able to analyze topics over several sentences. In addition, the misspelling (disregard of the capitalization) complicates the identification process.

Figure A.5: User review about the MUMOK (obtained from Facebook)